Predicting the Need for Urgent Instructor Intervention in MOOC Environments

ALRAJHI, LAILA,MOHAMMED

Predicting the Need for Urgent Instructor Intervention in MOOC Environments

Laila Mohammed Alrajhi

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

Department of Computer Science

Durham University in the United Kingdom

2023

DEDICATION

To the sake of Allah, all praise is due to him.

To my dear parents, Mohammed and Lolwa Alrajhi

To my patient husband, Mazen Aljumai

To my lovely children

To my affectionate brothers and sisters

To my supportive supervisors, Prof. Alexandra I. Cristea

Predicting the Need for Urgent Instructor Intervention in MOOC Environments

Laila Mohammed Alrajhi

Submitted for the Degree of Doctor of Philosophy

2023

ABSTRACT

In recent years, massive open online courses (MOOCs) have become universal knowledge resources and arguably one of the most exciting innovations in e-learning environments. MOOC platforms comprise numerous courses covering a wide range of subjects and domains. Thousands of learners around the world enrol on these online platforms to satisfy their learning needs (mostly) free of charge. However, the retention rates of MOOC courses (i.e., those who successfully complete a course of study) are low (around 10% on average); dropout rates tend to be very high (around 90%). The principal channel via which MOOC learners can communicate their difficulties with the learning content and ask for assistance from instructors is by posting in a dedicated MOOC forum. Importantly, in the case of learners who are suffering from burnout or stress, some of these posts require urgent intervention.

Given the above, urgent instructor intervention regarding learner requests for assistance via posts made on MOOC forums has become an important topic for research among researchers. Timely intervention by MOOC instructors may mitigate dropout issues and make the difference between a learner dropping out or staying on a course. However, due to the typically extremely high learner-to-instructor ratio in MOOCs and the often-huge numbers of posts on forums, while truly urgent posts are rare, managing them can be very challenging — if not sometimes impossible. Instructors can find it challenging to monitor all existing posts and identify which posts require immediate intervention to help learners, encourage retention, and reduce the current high dropout rates.

The main objective of this research project, therefore, was thus to mine and analyse learners' MOOC posts as a fundamental step towards understanding their need for instructor intervention. To achieve this, the researcher proposed and built comprehensive classification models to predict the need for instructor intervention. The ultimate goal is to help instructors by guiding them to posts, topics, and learners that require immediate interventions.

Given the above research aim the researcher conducted different experiments to fill the gap in literature based on different platform datasets (the FutureLearn platform and the Stanford MOOCPosts dataset [1]) in terms of the former, three MOOC corpora were prepared: two of them gold-standard MOOC corpora to identify urgent posts, annotated by selected experts in the field; the third is a corpus detailing learner dropout. Based in these datasets, different architectures and classification models based on traditional machine learning, and deep learning approaches were proposed.

In this thesis, the task of determining the need for instructor intervention was tackled from three perspectives: (i) identifying relevant posts, (ii) identifying relevant topics, and (iii) identifying relevant learners. Posts written by learners were classified into two categories: (i) (*urgent*) intervention and (ii) (*non-urgent*) intervention. Also, learners were classified into: (i) requiring instructor intervention (*at risk of dropout*) and (ii) no need for instructor intervention (*completer*).

In identifying posts, two experiments were used to contribute to this field. The first is a novel classifier based on a deep learning model that integrates novel MOOC post dimensions such as numerical data in addition to textual data; this represents a novel contribution to the literature as all available models at the time of writing were based on text-only. The results demonstrate that the combined, multidimensional features model proposed in this project is more effective than the text-only model. The second contribution relates to creating various simple and hybrid deep learning models by applying plug & play techniques with different types of inputs (word-based or word-character-based) and different ways of representing target input words as vector representations of a particular word. According to the experimental findings, employing Bidirectional Encoder Representations from Transformers (BERT) for word embedding rather than word2vec as the former is more effective at the intervention task than the latter across all models. Interestingly, adding word-character inputs with BERT does not improve performance as it does for word2vec. Additionally, on the task of identifying topics, this is the first time in the literature that specific language terms to identify the need for urgent intervention in MOOCs were obtained. This was achieved by analysing learner MOOC posts using latent Dirichlet allocation (LDA) and offers a visualisation tool for instructors or learners that may assist them and improve instructor intervention. In addition, this thesis

[1] https://datastage.stanford.edu/StanfordMoocPosts/

contributes to the literature by creating mechanisms for identifying MOOC learners who may need instructor intervention in a new context, i.e., by using their historical online forum posts as a multi-input approach for other deep learning architectures and Transformer models. The findings demonstrate that using the Transformer model is more effective at identifying MOOC learners who require instructor intervention.

Next, the thesis sought to expand its methodology to identify posts that relate to learner behaviour, which is also a novel contribution, by proposing a novel priority model to identify the urgency of intervention building based on learner histories. This model can classify learners into three groups: low risk, mid risk, and high risk. The results show that the completion rates of high-risk learners are very low, which confirms the importance of this model. Next, as MOOC data in terms of urgent posts tend to be highly unbalanced, the thesis contributes by examining various data balancing methods to spot situations in which MOOC posts urgently require instructor assistance. This included developing learner and instructor models to assist instructors to respond to urgent MOOCs posts. The results show that models with undersampling can predict the most urgent cases; 3x augmentation + undersampling usually attains the best performance. Finally, for the first time, this thesis contributes to the literature by applying text classification explainability (eXplainable Artificial Intelligence (XAI)) to an instructor intervention model, demonstrating how using a reliable predictor in combination with XAI and colour-coded visualisation could be utilised to assist instructors in deciding when posts require urgent intervention, as well as supporting annotators to create high-quality, gold-standard datasets to determine posts cases where urgent intervention is required.

DECLARATION

The work and experiments in this thesis are based on research carried out within the Artificial Intelligence and Human Systems Group (AIHS) at the Department of Computer Science at Durham University, UK. No part of this thesis has been submitted elsewhere for any other qualification or degree and it is all the author's work unless referenced to the contrary in the below.

List of Publications

Accepted papers included in this thesis:

- **Chapter 4:** Alrajhi, L., Alharbi, K., & Cristea, A. I. (2020). A multidimensional deep learner model of urgent instructor intervention need in MOOC forum posts. In *Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Athens, Greece, June 8–12, 2020, Proceedings 16* (pp. 226-236). Springer International Publishing.
- **Chapter 4:** Alrajhi, L., & Cristea, A. I. (2023, May). Plug & Play with Deep Neural Networks: Classifying Posts that Need Urgent Intervention in MOOCs. In *International Conference on Intelligent Tutoring Systems* (pp. 651-666). Cham: Springer Nature Switzerland.
- **Chapter 5:** Alrajhi, L., Alharbi, K., Cristea, A. I., & Pereira, F. D. (2022, November). Extracting the Language of the Need for Urgent Intervention in MOOCs by Analysing Text Posts. In *International Conference on Web-Based Learning* (pp. 161-173). Cham: Springer International Publishing.
- **Chapter 6:** Alrajhi, L., Alamri, A., & Cristea, A. I. (2022, June). Intervention Prediction in MOOCs Based on Learners' Comments: A Temporal Multi-input Approach Using Deep Learning and Transformer Models. In *Intelligent Tutoring Systems: 18th International Conference, ITS 2022, Bucharest, Romania, June 29–July 1, 2022, Proceedings* (pp. 227-237). Cham: Springer International Publishing. (**Nominated for the best paper award**).

- **Chapter 7:** Alrajhi, L., Alamri, A., Pereira, F. D., & Cristea, A. I. (2021). Urgency analysis of learners' comments: an automated intervention priority model for MOOC. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17* (pp. 148-160). Springer International Publishing.
- **Chapter 8:** Alrajhi, L., Alamri, A., Pereira, F. D., Cristea, A. I., & Oliveira, E. H. (2023). Solving the imbalanced data issue: automatic urgency detection for instructor assistance in MOOC discussion forums. *User Modeling and User-Adapted Interaction*, 1-56.
- **Chapter 9:** Alrajhi, L., Pereira, F. D., Cristea, A. I., & Aljohani, T. (2022, July). A Good Classifier is Not Enough: A XAI Approach for Urgent Instructor-Intervention Models in MOOCs. 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* (pp. 424-427). Cham: Springer International Publishing.
- **Chapter 9:** Alrajhi, L., Pereira, F. D., Cristea, A. I. & Alamri, A., (2023). Serendipitous Gains of Explaining a Classifier - Artificial versus Human Performance and Annotator Support in an Urgent Instructor-Intervention Model for MOOCs. In *Proceedings of the Workshop on Human Factors in Hypertext.*

Accepted papers not included in this thesis:

- Alharbi, K., Alrajhi, L., Cristea, A. I., Bittencourt, I. I., Isotani, S., & James, A. (2020). Data-Driven analysis of engagement in gamified learning environments: A methodology for real-time measurement of MOOCs. In *Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Athens, Greece, June 8–12, 2020, Proceedings 16* (pp. 142-151). Springer International Publishing.
- Yu, J., Alrajhi, L., Harit, A., Sun, Z., Cristea, A. I., & Shi, L. (2021). Exploring bayesian deep learning for urgent instructor intervention need in mooc forums. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17* (pp. 78-90). Springer International Publishing.

- 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).
- Aljohani, T., Cristea, A. I., & Alrajhi, L. (2022, July). Bi-directional Mechanism for Recursion Algorithms: A Case Study on Gender Identification in MOOCs. 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* (pp. 396-399). Cham: Springer International Publishing.

Forthcoming Publications

- **Chapter 2:** Alrajhi, L., Alshehri, M., & Cristea, A. I. (2023). Systematic Literature Review of Identifying Instructor Intervention Need in MOOC Discussion Forums. ACM Computing Surveys. (under review)

ACKNOWLEDGEMENTS

I would like to thank Almighty Allah for helping and granting me the strength, determination, patience, and guidance to complete this doctoral thesis. Nothing can be accomplished without Allah's blessings. Alhamdulillah.

For me, words cannot deliver and describe my feelings, thanks, and appreciation to the lovely people around me.

First, I would like to express my deep gratitude and thanks to my PhD supervisor Professor Alexandra I. Cristea for her invaluable guidance, encouragement, advice, and support. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries. Her insightful feedback has been instrumental in shaping the direction of this thesis.

Next, I would like to express my deepest gratitude and thanks my lovely parents, Mohammed and Lolwa, for their prayers, love, unlimited support throughout my life, and encouragement. Thank you for always being there. None of this would have been possible without their motivation. No words can express my love for you.

My deepest thanks go to my small family, my wonderful husband, Mazen, for his understanding, help, generous support, and continuous encouragement during the tough times. Without his patience, I would not have been able to complete this work. Thank you for being my best friend. Also, to my lovely kids, Faisal, Abdulaziz, Nawaf, and Mohammed you are my happiness in life. Thank you for being my inspiration, I love you to the moon and back.

I am very much indebted to my family, my dear brothers, Faisal, Abdulmjeed, and Abdulaziz, and my lovely sisters, Manal, Sara, and Hayfa, for their love, constant encouragement, and concern for me. They supported me in every possible way to see the completion of this work. Many thanks to my wonderful nephews and nieces for their love.

I would also like to express my thanks to all my research group members for their insightful comments, conversations, and advice, especially my co-authors, Dr. Ahmed Alamri, Dr. Filipe Pereira, Dr. Jialin Yu, Dr. Khulood Alharbi, Dr. Mohammad Alshehri, Dr. Olanrewaju Tahir Aduragba, and Dr. Tahani Aljohani.

Special thanks to my close friends, Dr. Aisha Alsehaim, Dr. Arwa Alsaqaabi, Dr. Latifah Abduh, and Dr. Muna Almushyti, Dr. Nada Almani, Tumadher Alharigy, for all their help, emotional support, and care they provided throughout my PhD journey. Also, I would like to thank my friends and everybody for helping me get through the difficult times.

Finally, I want to convey my appreciation to my University, King Abdulaziz University and Saudi Cultural Bureau in Britain (SACB) for giving me this opportunity to complete my PhD. Also, I would like to thank all the members of the Department of Computer Science at Durham University for all their help in facilitating this research.

Thank you all very much.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Acc Accuracy

ACM Association for Computing Machinery

AI Artificial Intelligence

AIED International Conference on Artificial Intelligence in Education

ANN Artificial neural network

BA Balance Accuracy

BERT Bidirectional Encoder Representations from Transformers

Bi-GRU Bidirectional-GRU

Bi-LSTM Bidirectional-LSTM

Bi-RNN Bidirectional-RNN

BJET British Journal of Educational Technology

BoW Bag of Word

CAERS Conversational Agent in an Educational Recommender System

cMOOCs connectivist Massive Open Online Courses

CNN Convolutional Neural Network

CRF Conditional Random Fields

DL Deep Learning

DNNs Deep Neural Networks

EDM Educational Data Mining

F1 F1-Measure

FN False Negatives

FP False Positive

FUMA Framework for User Modelling and Adaptation

GRU Gated Recurrent Unit

HF High Frequency

IEEE Institute of Electrical and Electronic Engineers

IJAIED International Journal of Artificial Intelligence in Education

JEDM Journal of Educational Data Mining

JLA Journal of Learning Analytics

LA Learning Analytics

LAK Learning Analytics and Knowledge

LDA Latent Dirichlet Allocation

LIME Local Interpretable Model-agnostic Explanations

LIWC Linguistic Inquiry and Word Count

LR Logistic Regression

LSTM Long Short-Term Memory

L@S Learning at Scale

ML Machine Learning

MLM Masked Language Model

MLP Multiple Layer Perceptron

MOOCS Massive Open Online Courses

NB Naive Bayes

NLP Natural Language Processing

NNs Neural Networks

NSP Next Sentence Prediction

P Precision

PDTB Penn Discourse Treebank

PLSDA Part-of-speech-focused Lexical Substitution for Data Augmentation

PR Precision verse Recall

PRISMA Preferred Reporting Items for Systematic Review and Meta-analysis

PRO Probability

PSO Particle Swarm Optimisation

QA Question/Answer

R Recall

RB ReaderBench

RBF Radial Basis Function

RCNN Recurrent Convolutional Neural Network

ReLU Rectified Linear Unit

RF Random Forest

RNN Recurrent Neural Network

SEANCE Sentiment Analysiss and Cognition Engine

SHAP SHapley Additive exPlanations

SLR Systematic Literature Review

SVC Support Vector Classifier

SVM Support Vector Machine

TAACO Tool for the Automatic Analysis of Cohesion

TAALES Tool for the Automatic Analysis of Lexical Sophistication

TAs Teaching Assistants

TF Term Frequency

TF-IDF Term Frequency Inverse Document Frequency

TN True Negative

TP True Positive

t-SNE t-Distributed Stochastic Neighbour Embedding

WAT Writing Assessment Tool

WoS Web of Science

XAI Explainable Artificial Intelligence

XGBoost Extreme Gradient Boosting

xMOOCs extended Massive Open Online Course

CHAPTER 1: INTRODUCTION

An introduction to the topic of this thesis is provided in this chapter and has been organised as follows. Firstly, it presents a brief introduction to MOOCs (Section 1.1), followed by an outline of the research problem (Section 1.2), the motivations for this work (Section 1.3), an explanation of the research scope (Section 1.4), the research questions (Section 1.5), the research objectives (Section 1.6), and an outline of the research contributions (Section 1.7). Finally, it offers a comprehensive outline and structure (Section 1.8) of the current thesis.

1.1. Introduction

Massive open online courses (MOOCs) are a subset of information systems known as *open distance online learning environments* with large-scale enrolment (Arguello and Shaffer, 2015). In recent years, Coursera[2], edX[3], Udacity[4], and FutureLearn[5] have emerged as popular platforms (Joseph, 2020). Since their emergence as a popular global mode of learning in 2012 (Yan *et al.*, 2019), MOOCs have been providing substantial support to learners by offering and delivering global, high-quality, education via a wide variety of online courses across numerous domains and subjects. MOOCs are provided by numerous universities, institutions, companies, and ventures (Chaturvedi, Goldwasser and Daumé III, 2014) to cater for a wide range and unlimited number of learners. Most of these courses are offered at no cost (free) or extremely cheaply (Yang *et al.*, 2017; McAuley *et al.*, 2010) and some have no prerequisite for enrolment

[2] www.coursera.org
[3] www.edx.org
[4] www.udacity.com
[5] www.FutureLearn.com

and low access barriers; as a result, these courses have helped attract a large learner cohort and reach hundreds of thousands of learners (Wise, Cui and Vytasek, 2016); at the end of 2021, about 220 million people were enrolled on MOOCs (Shah, 2021) to improve their lifelong learning in flexible way and improve their knowledge at their convenience (Yang *et al.*, 2015). Further, MOOC learners live all over the world and come from diverse knowledge backgrounds and education systems, and have a huge range of abilities, goals, and motivations.

1.2. Research Problem

MOOC courses offered by leading universities are playing an increasingly vital role in education; this was compounded during the recent COVID pandemic and lockdown as most educational institutions around the world turned to online study (Soni, 2020). MOOCs continue to grow and proliferate dramatically; however, the completion rates for MOOC courses are extremely low (Crossley *et al.*, 2016): only 3–5% on two MOOCs on the University of Melbourne platform (Coffrin *et al.*, 2014) and just 10% on the FutureLearn platform (Alamri *et al.*, 2019) — which is low enough to be a serious problem. There are several ongoing educational debates about the reasons for these low completion rates. One of the most critical factors identified is missing real-time direct interaction in terms of face-to-face communication, support, and collaboration, which leads online learners to feel isolated and suffer from a lack of meaningful human interaction compared with other educational environments. Because of this, some learners may feel stuck, confused, need clarification, and may struggle to stay on their course; if these issues are not addressed, they can ultimately lead to dropout (Yang *et al.*, 2015; Kizilcec and Halawa, 2015). Also, (Gütl *et al.*, 2014; Onah, Sinclair and Boyatt, 2014a) found that this issue was related to the lack of sufficient learner support and interaction with course instructors.

The primary way for learners to communicate, interact and express their feelings about MOOC content, learning progress, and highlight concerns, questions, and desire for help is via forum posts (comments). It needs to be noted that here the terminology *posts* is used interchangeably with *comments* in the scope of this thesis (for more details see Section 2.2.1.3); the two terms are used interchangeably in the literature to represent indirect interaction on asynchronous online discussion forum platforms (Chen *et al.*, 2019). Posts connect learners to learners, or learners to instructors. In general, such communication can have significant learning impacts (Ntourmas *et al.*, 2019); learners who participate in forum discussions are more likely to finish (parts of) a course (Klusener and Fortenbacher, 2015). However, a lack of

receiving responses and feedback on their problems from instructors or peers may cause learners to develop negative feelings about their studies, thus hindering their learning progress and ultimately causing them to drop out of the course (Yang *et al.*, 2015; Park and Choi, 2009), which therefore highlights the need for urgent instructor intervention (Almatrafi, Johri and Rangwala, 2018). Instructor intervention is one critical solution for reducing learner dropout. In addition, some types of learner queries and requests for support can only be answered by an instructor (Macina *et al.*, 2017).

From an instructor's perspective, intervention to address learners' questions in online learning is a central and essential teaching activity (Chandrasekaran *et al.*, 2017) and could make the difference between a learner completing the course or not. While instructors have limited time and bandwidth (Chandrasekaran *et al.*, 2015b), they try to assist, encourage, motivate, and support learners and tend to respond to their questions as much as possible. However, due to the tremendous number of learners enrolled on these platforms and the extremely high ratio of learners to instructors, it is very hard for instructors to commit to monitoring all learners' textual forum posts and determine when to intervene (Wei *et al.*, 2017). Therefore, instructors need to be selective in their interventions (Chaturvedi, Goldwasser and Daumé III, 2014). In addition, the massive amounts of posts on MOOCs, most of which are general discussion and forging social connections that do not involve any urgent issues or require intervention, mean that it is difficult and time-consuming (effort-intensive) for instructors to effectively manage to monitor and review all existing posts, which may number in the millions, and find cases where it is necessary to engage in meaningful interactions to resolve issues and provide feedback. Also, such intervention is often preferred to be performed in real-time (Chandrasekaran *et al.*, 2015b). This challenging research problem has encouraged research on instructor intervention in MOOC discussion forums.

1.3. Research Motivations

On MOOC platforms, struggling learners often describe their need for help via forum posts. However, the often-huge numbers of posts on forums make it unlikely that instructors can capture these posts and respond to all learners; many of these urgent posts are overlooked or discarded.

Natural language processing (NLP) which began in the 1950s (Kalyanathaya, Akila and Rajesh, 2019) is an exciting research direction in computer science for analysing large datasets

such as those associated with MOOCs. The main goal of NLP is processing 'natural language' using computers to analyse the data, extract information or even represent information in different ways (Conneau *et al.*, 2016). NLP is a prominent area of both the fields of *computational linguistics* and *artificial intelligence* (AI) (Garousi, Bauer and Felderer, 2019). NLP has been revolutionised with the emergence of *machine learning* (ML), *neural networks* (NNs) and *deep neural networks* (DNNs), particularly in relation to DNNs due to their significant performance and fewer requirement for engineered features (Yin *et al.*, 2017). Deep learning (DL) is an efficient approach for use in NLP (Wei *et al.*, 2017). One significant topic for NLP is *text classification*, which assigns an unstructured text to predefined categories (Zhang, Zhao and LeCun, 2015); it is considered a supervised machine learning model (Agarwal and Mittal, 2014). Text classification is an important area that renders itself appropriate for the problem addressed in the current thesis.

Since MOOCs contain a massive amount of data (big data) produced by huge numbers of learners, they are appropriate for study with *learning analytics* (LA) (Khalil and Ebner, 2017). LA is a crucial field of technology-enhanced learning (Ferguson, 2012) and defined in (Slade and Prinsloo, 2013) as student-generated, actionable data that are gathered, analysed, used, and properly disseminated with the aim of providing learners with appropriate administrative, cognitive, and instructional support. Therefore, the development of LA methods is also useful in addressing intervention issues to investigate learner data and how learners behave to improve and optimise instructor interventions.

Thus, the aim of this research project is to reduce the effort required by MOOC instructors and support them by (i) automatically exploring whether MOOC posts need urgent instructor attention and intervention, (ii) extracting language that highlights the need for urgent MOOC instructor intervention, and (iii) discovering when MOOC learners tend to drop out. In addition, the current thesis sought to expand the identification of urgent posts by proposing three research directions: (i) offering an automated intervention priority model built on learners' histories, (ii) solving the imbalanced data issue related to MOOCs and presenting an automatic intervention detection method to flag up when instructor assistance is required based on user modelling, and (iii) employing XAI to improve general instructor intervention in MOOC environments as well as annotators. This was achieved by analysing the textual content of learners' MOOC posts using NLP techniques as a text classification task, and clustering and proposing, designing and developing robust and useful supervised and unsupervised ML algorithms and models. In supervised algorithms, traditional ML, and DL were applied.

The main motivation of this research is to help instructional staff to better utilise their time by easily identifying and filtering learner posts from discussion forums in MOOCs, extracting urgent language terms that may help instructors to improve the quality of their interventions, and recognising dropout learners so that intervention can be provided before dropout using an automated system based on learner posts. The filtering of posts categorises interventions into *urgent* and *non-urgent* while of learners into *dropout* or *completers*. Then, assigning priority and adding adaption to the interventions to improve the quality of such interactions. In addition, applying XAI to assist both instructors and annotators and thus improving the process of instructor intervention in MOOC environments.

1.4. Research Scope

Researchers, MOOCs designers, universities, and educational institutions have begun to pay more attention to instructors' presence and interventions in online discussion forums (Mazzolini and Maddison, 2007) and MOOC-based environments specifically. To date, many researchers in this area (Khodeir, 2021; Sun *et al.*, 2019; Guo *et al.*, 2019) and others as discussed later in Chapter 2, have focused on the instructor intervention in MOOC environments based on posts without any focus on learners and their behaviours. However, this thesis tackled the problem of the instructor intervention task in MOOCs beyond that by identifying three key perspectives as follows:

- Posts: building supervised prediction models to predict urgent posts.
- Topics: building unsupervised prediction models to identify topics based on posts and correlate such topics with urgent posts.
- Learners: building supervised prediction models to predict learners who need urgent intervention by utilising the temporal history of learner post content.

Next, the task was to expand the efficacy with which urgent posts are identified by incorporating learner- and instructor-based models. Lastly, the task was to use an XAI approach to understand the supervised classification model and employ XAI to assist both instructor and annotators.

To conduct the above experiments, two dataset sources were included: (i) the FutureLearn platform (a course with 5790 forum posts) and the Stanford (11 courses with about 29,604 forum posts) (Agrawal *et al.*, 2015; Agrawal and Paepcke, 2019). Both datasets were annotated manually by human coders (for more details see Chapter 3).

Please also note that learners who need intervention but do not use forum posts as a communication means are not the target of this research project; identifying such learners would require an alternative approach.

1.5. Research Questions

The aim of this thesis was to address the problem of when instructor intervention is required based on MOOC learners' posts and address the gap in the existing literature on this area as described in Chapter 2; this was achieved by defining the following umbrella research question (RQ):

- ***RQ:*** *How can the need for urgent instructor intervention be automatically and realistically detected based on learner posts in MOOC environments?*

To help with answering this wide research question, the following sub-RQs were formulated:

- ***RQ1:*** *What are the most appropriate choices when classifying urgent posts that need instructor intervention in terms of: (i) their various dimensions; (ii) deep learning approaches; (iii) word- or word-character-based approaches?*

This RQ represents the first attempt to better understand how the problem of classifying urgent posts is to be tackled, and what methods can be used. This RQ is answered in Chapter 4.

- ***RQ2:*** *Can the language of urgency be detected from learner posts and can it be visualised simply and intuitively?*

After finding performant models as a result of ***RQ1***, the next step was to look more deeply into the language contained in the posts themselves in Chapter 5.

- ***RQ3:*** *Can learners who may drop out be predicted from the history of their most recent posts?*

Dropout is one of the main causes of the need for instructor intervention in MOOCs. This research question sought to analyse this cause in a more in-depth way (further details are provided in Chapter 6).

- ***RQ4:*** *Can the behaviour of learners who need an urgent intervention be analysed to lead to an effective intervention priority framework?*

This RQ seeks to analyse another aspect related to the urgent instructor intervention need problem: the relationship with learner behaviour, which brings the work one step closer to the intervention, by creating an intervention priority framework (details in more depth in Chapter 7).

- ***RQ5:*** *How can the prediction model for urgency detection be further improved to be applicable to adaptive intervention?*

This RQ attempts the final (and comprehensive) modifications to the prediction model for urgent instructor intervention need by addressing other issues found such as data imbalance (in Chapter 8).

- ***RQ6:*** *How can a transparent XAI model be constructed to detect urgent intervention need to support instructors' decisions to intervene in posts as well as improve human annotators' decisions on urgent posts intervention?*

The research on urgent intervention need detection on MOOCs has shown that understanding posts urgency is challenging both for machines and humans. To further help the process of urgent intervention, as presented in Chapter 9.

1.6. Research Objectives

This study aimed to investigate the possibility of developing ML models that predict the need for instructor intervention based on learners' posts on MOOC discussion forums. To achieve this and address the above RQs, the following research objectives (ROs) were formulated:

- **RO1:** To systematically review current models that predict instructor intervention need based on learner posts in MOOCs, analyse the findings for in-depth understanding of this topic, clarify the limitations of these models, and suggest some areas for development. The main (umbrella) RQ was formulated based on the findings related to this objective (Chapter 2).
- **RO2:** To create a new corpus for instructor intervention in MOOC forum discussion posts derived from the FutureLearn platform which is manually annotated by experts in the domain. This is important to represent different types of platforms and address the current shortage in the available dataset Stanford MOOCPosts (Chapter 3).
- **RO3:** To classify learners' urgent posts that need instructor intervention based on two methods: (i) using several dimensions to analyse posts as features in addition to textual

data, and (ii) using different levels and representation of textual inputs via the use of different DL approaches. This addresses RQ1 (Chapter 4).

- **RO4:** To extract from posts language that highlights the need for urgent instructor intervention using topic modelling and visualisation. This objective addresses RQ2 (Chapter 5).
- **RO5:** To examine the capability to predict which learners may drop out and need instructor intervention in MOOCs based on their textual post history using ML approaches. This objective is achieved by addressing RQ3 (Chapter 6).
- **RO6:** To enhance instructor intervention in relation to learner posts by introducing an effective priority intervention model based on learner behaviour. This addresses RQ4 (Chapter 7).
- **RO7:** To solve the highly unbalanced data issue in MOOCs datasets by using NLP approaches and automate the urgent-post identification process based on learner modelling to provide automatic and adaptive recommendations to instructors. This objective is achieved by addressing RQ5 (Chapter 8).
- **RO8:** To apply XAI techniques to interpret a MOOC intervention model to support instructors' decisions to intervene in posts and improve human annotators' decisions on urgent posts intervention processes. RQ6 addresses this objective (Chapter 9).

1.7. Research Contributions

This thesis provides several novel contributions to the instructor intervention problem in MOOC discussion forums by proposing different models (post-based: 2 models; topic-based: 1 model; learner-based: 1 model; post- and learner-based (adding priority): 1 model; post-based and user modelling (solving the imbalanced): 1 model; XAI: 1 model).
The key contributions are as follows:

- Systematically reviewing the available research on instructor intervention in MOOC discussion forums using preferred reporting items for systematic review and a meta-analysis (PRISMA) protocol (Chapter 2).
- Creating a Gold-standard corpus for instructor intervention on MOOC discussion forum environments (posts gathered from the FutureLearn platform) which is annotated by expert human annotators and analysed in terms of intervention (Chapter 3).

- Building a novel classifier for this problem based on a DL model that incorporates different dimensions of MOOC posts (i.e., numerical data in addition to textual data) to classify urgent posts (Chapter 4).
- Constructing different simple and hybrid deep learning models by applying plug & play techniques with various types of inputs and representing words to establish good combinations in terms of performance (Chapter 4).
- Extracting the language highlighting the need for urgent intervention in MOOCs by analysing text posts and proposing visualisation tools (Chapter 5).
- Identifying MOOC learners who may dropout and need instructor intervention by using their historical online forum posts as data and constructing a multi-input approach for siamese and dual BERT with binary text classification, with the resulting integrated networks being termed multi-siamese BERT and multiple BERT, respectively (Chapter 6).
- Proposing a novel priority model to identify posts that need urgent intervention based on learner histories; namely, past urgency, sentiment analysis, and step access (Chapter 7).
- Applying different data-balancing techniques for traditional and deep ML to identify instances when urgent instructor intervention is required in MOOC environments and proposing several new pipelines to generate more data for text augmentation (Chapter 8).
- Creating the first learner, instructor, and adaptation models to support instructors to deal with urgent posts in MOOCs (Chapter 8).
- Applying text classification explainability (XAI) to an instructor intervention model. Also, connecting the AI prediction error to a lack of human confidence, to be used for annotator support for creating high-standard corpora (Chapter 9).

1.8. Thesis Outline

The thesis is structured and organised into eleven chapters as follows:

- Chapter 1: Introduction: This chapter presents the research problem, motivations, scope, RQs, ROs, and contributions.

- Chapter 2: Background and Literature Review: This chapter introduces the background to the main topics and systematically reviews the literature on instructor intervention related to learner posts in MOOC environments.
- Chapter 3: Methodology: This chapter outlines the methodology used to address the RQs of this thesis. This discusses several data sources and how different datasets can be formulated, as well as performance evaluations, ethical issues, and overall experiments.
- Chapter 4: Intervention Prediction: Post-Based Model: This chapter describes the two experiments (a multidimensional deep learning model and plug & play model with DNNs) used to identify posts that need urgent intervention. In each experiment, the related works are provided. It then describes the models and discusses the results.
- Chapter 5: Analysing Text Posts using Modelling to Extract Urgent Language: This chapter provides an analysis of learner's text posts. It discusses the related research before describing the methodology and presenting the results.
- Chapter 6: Intervention Prediction: Learner-Post-Based Model: This chapter illustrates the experiment designed to detect learners at risk of dropping out by discussing the related research, methodology, and results.
- Chapter 7: Intervention Prediction: Post- and Learner-Based Model (Adding Priority in Intervention): This chapter explains the framework used to add priority for instructor intervention. It provides the methodology and discusses the results.
- Chapter 8: Intervention Prediction: Posts-Based and User Modelling (Solving the Imbalanced Data Issue): This chapter explains the experiment designed to solve the problem of imbalanced data and propose an adaption model. It describes the related work, methodology, and results.
- Chapter 9: An Explainable Artificial Intelligence (XAI) Approach for Urgent Instructor Intervention Models: This chapter clarifies how XAI can be employed to the instructor intervention task. It presents the related work followed by the methodology and results.
- Chapter 10: Discussion: This chapter discusses the important topics of the current thesis in relation to the obtained results; it also outlines the limitations and future research avenues.
- Chapter 11: Conclusion: This chapter concludes and summarises the key contributions and findings.

The thesis summary workflow is shown in Figure 1.1 (below).

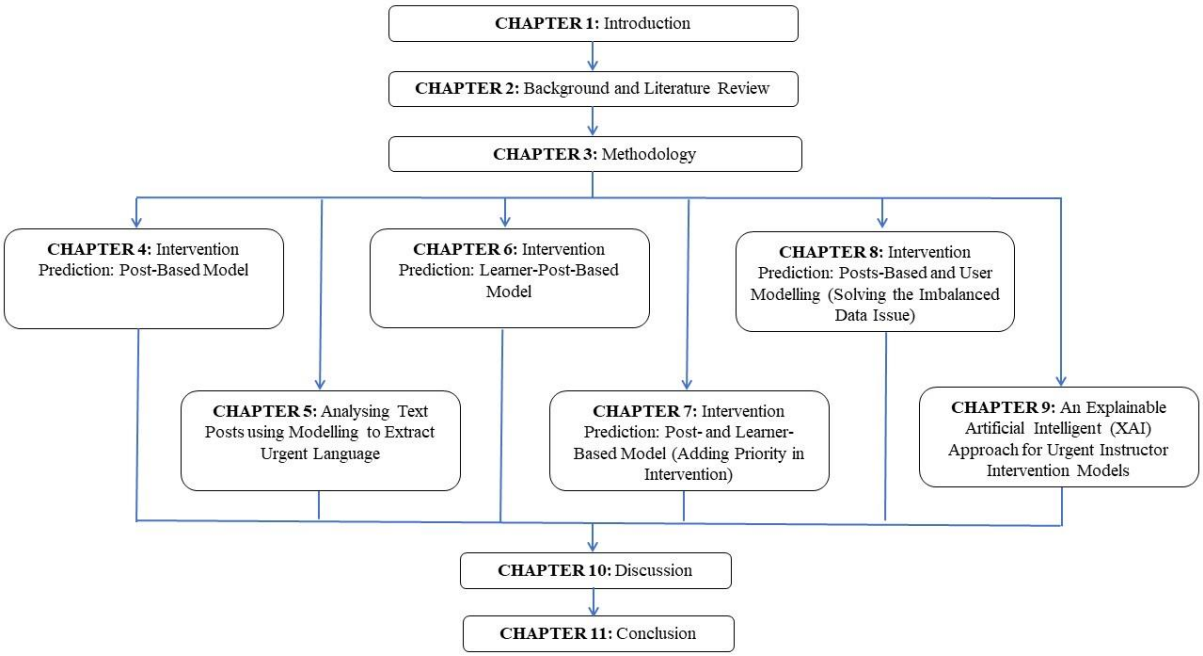

Figure 1.1: Thesis summary workflow.

The next chapter provides background on the thesis's main topics including MOOCs, NLP, and ML. In addition, it reviews the literature on instructor intervention related to learner posts in MOOC contexts.

CHAPTER 2: BACKGROUND AND LITERATURE REVIEW

2.1. Prologue

It is noteworthy that in any research project, the background and literature review is a crucial stage since it helps the researcher understand the subject, include important ideas for consideration, and identify the gaps in previous works. The literature review marks the starting point of a research project and identifies research gaps and questions that the thesis could address (Kraus, Breier and Dasí-Rodríguez, 2020). Thus, the purpose of this chapter is to introduce the thesis topic by outlining and understanding the background around the three main topics of this research project: (i) MOOCs, (ii) NLP, and (iii) ML (see Section 2.2). In addition, it reviews the literature covering these topics (see Section 2.3). Moreover, it provides a systematic review of instructor intervention need in MOOC discussion forums with the goal of analysing the extant research (see Section 2.3.4). The results of the systematic literature review will help to identify gaps in the current literature and aid the design, creation, and development of models that can address these research gaps.

2.2. Background

This section outlines the main theoretical background and provides brief definitions of the concepts related to the three main topics: (i) MOOCs (and a variety of similar platforms); (ii) NLP approaches; (iii) ML methods relevant to this thesis.

2.2.1. MOOCs

In 2008, the term MOOC was coined for the first time by George Siemens and David Cormier to describe an open online course: Connectivism and Connective Knowledge (CCK08) (Downes, 2008) launched by Siemen and Stephen Downes at the University of Manitoba, Canada (Liyanagunawardena, Adams and Williams, 2013; De Notaris, 2019). MOOCs then received great media coverage when the New York Times announced that 2012 was 'The Year of the MOOC' (Pappano, 2012). MOOCs are defined according to EU-funded MOOC projects and OpenupEd as

> Online courses designed for large numbers of participants, that can be accessed by anyone and anywhere as long as they have an internet connection, are open to everyone without entry qualifications, and offer a full/complete course experience online for free (Jansen and Schuwer, 2015).

(Khalil and Ebner, 2017; Wulf *et al.*, 2014) define the four words which comprise the term MOOC as:

- *Massive (M):* representing that there are significantly more learners enrolled than in usual distance-learning courses.
- *Open (O):* explaining the concept of *openness* as there tend to be no (or very few) requirements to participate and (mostly) free access is provided to everyone.
- *Online (O):* referring to the fact that courses are delivered across the global Internet and are not location-specific.
- *Courses (C):* constituting structured learning content according to the concept that primarily takes the form of articles, interactive social media channels, and video lectures.

Following that, MOOCs can be divided into two main learning paradigms: cMOOCs and xMOOCs: the former are connectivism-based while the latter lie closer to more traditional behaviourist models (Kesim and Altınpulluk, 2015). Most common MOOC platforms belong to xMOOCs where the responsibility of the participants is minimal and limited to their contributions to a forum (Borrás-Gené, 2019). xMOOCs differ significantly from cMOOCs as they focus on participation in online discussions (Daniel, 2012).

Meanwhile, there is a lot of interest in online courses as the majority of them do not require prerequisite qualifications to participate. These courses are taught by academics at top-ranking

universities since they offer a full distance learning environment with videos, assignments, presentations, and other course materials (Kesim and Altınpulluk, 2015) and are constantly growing, as shown in Figure 2.1 (below) (Shah, 2021). These online MOOCs courses are frequently created independently by academics and published by third-party online venues (Baturay, 2015).

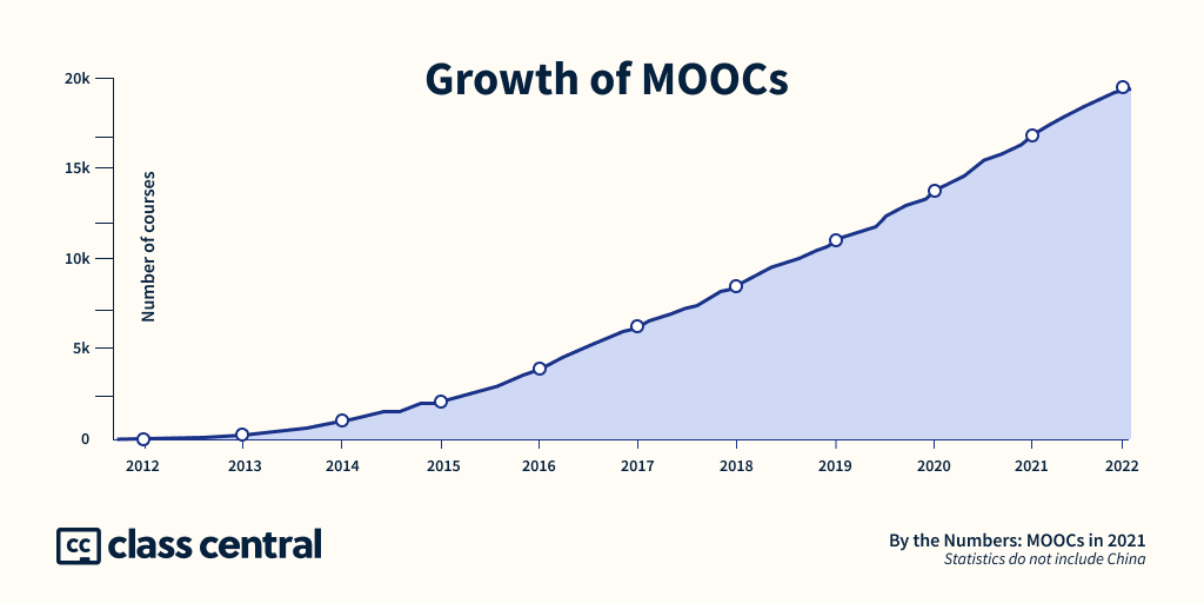

Figure 2.1: The number of courses in MOOCs during the years (Shah, 2021).

According to (Brahimi and Sarirete, 2015), there are many MOOC providers from different countries, including the United States (Coursera, Edx, Udacity), Europe (FUN, Iversity), the United Kingdom (FutureLearn), the Middle East (Rawaq, Edraak), and Australia (Open2Study). In recent years, millions of learners around the world have used MOOCs as a result of the increased popularity of these types of courses (Ipaye and Ipaye, 2013). In 2021, about 220 million learners joined MOOC platforms (Shah, 2021). Although each platform has its advantages and disadvantages, they all provide a wide range of academic courses from providers around the globe (Reutemann, 2016). Different platforms provide different structures of courses and discussion forum posts. For example, the accepted length of each post varies with different platforms settings. The next sections introduce the different platforms related to this thesis.

2.2.1.1. Stanford University Online

Stanford University, an elite, cutting-edge institution, created an educational initiative called Stanford Online which provides a variety of professional and academic educational

opportunities. Over 10 million learners in 190 countries are able to take more than 200 free and open online courses through Stanford Online using its Open edX technology. In 2011, in the early stages of MOOC at Stanford University, 160,000 learners from more than 190 nations enrolled on a course on artificial intelligence that was introduced by Peter Norvig and Sebastian Thrun (Voudoukis and Pagiatakis, 2022). Similarly, in October 2011, Andrew Ng, another Stanford University professor, conducted an online course with 100,000 learners (Herman, 2012). This clarifies the huge numbers of learners studying on MOOCs even at the beginning of the era. At the time of writing (2023) nearly 78 free online courses are available from Stanford University.

2.2.1.2. FutureLearn

Futurelearn is a European online learning information system that facilitates remote and online learning from top universities; it is similar to the American platform, Coursera (Reutemann, 2016), and offers free (or partly free) learning. According to the number of learners registered, FutureLearn in 2016 is the fourth largest MOOC provider in the world with 5.3 million learners (Shah, 2016). In 2012, FutureLearn began as a collaboration between numerous leading UK universities, the BBC, and the British Library; this platform later expanded to include courses from other schools, NGOs, and companies (Cristea *et al.*, 2018). To offer online courses and degrees, FutureLearn collaborates with around 300 leading international institutions and specialised organisations[6]. FutureLearn courses cover a wide range of topics and many of them are offered periodically as iterations (referred to as "runs"). The structure of these courses is hierarchical, with weeks and steps as shown in Figure 2.2 (Chitsaz, Vigentini and Clayphan, 2016). Each week contains several steps that represent a single learning unit including videos and assignments, etc. The step type can be recognised by its title (Chitsaz, Vigentini and Clayphan, 2016). Most courses run for six to ten weeks; others run for only two or three weeks (*Using FutureLearn*, 2020). In November 2019, FutureLearn announced on its website that ten million learners have officially studied on FutureLearn on over 2,400 courses.

[6] Current partners - FutureLearn

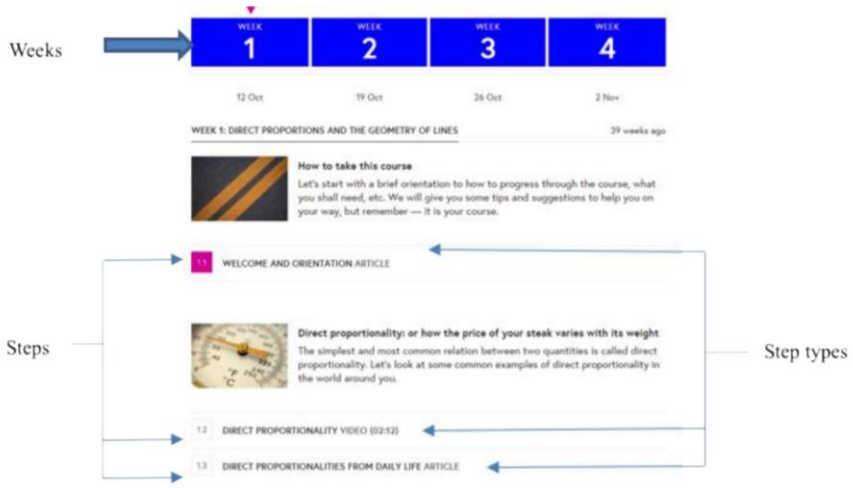

Figure 2.2: An overview of the FutureLearn course structure (Chitsaz, Vigentini and Clayphan, 2016).

2.2.1.3. Discussion Forums in MOOCs

Since the early 1990s, discussion forums have been utilised as online learning tools (Onah, Sinclair and Boyatt, 2014b). In MOOCs, discussion forums are one of the important components (Bonafini, 2017); these forums are crucial channels that allow learners to interact with their peers and instructors asynchronously. These interactions represent a form of social networking that include requests for help, clarification about course topics, as well as general communication.

Different MOOC platforms implement different structures and settings (Gardner and Brooks, 2018). For instance, on the FutureLearn platform, which follows a social-constructivist pedagogy, the discussion forum is a highly important element (Ferguson and Clow, 2015); thus, the discussion forums on the FutureLearn platform feature *comments* from participants under every step in a week; each comment posted can be replied to by any participant, except in the case of quizzes and exercises (Chua *et al.*, 2017), for formal discussion and commenting (Vigentini, León Urrutia and Fields, 2017). Thus, such forums allow participants to post at any step and at any time. The comments posted in such forums tend to be brief with a 1200-character limit and can be categorised into two types: (i) new posts as comments and (ii) replies to other posts.

Meanwhile, on the Coursera platform, which is the largest MOOC platform (Wu, 2021), discussion forums are constructed using special sections (Drobot, 2023). The instructor can divide the forums listed in the discussion forum tab into sub-forums to organise discussions (Chandrasekaran *et al.*, 2015a). Discussion forums on Coursera are composed of three levels:

(i) threads, (ii) posts, and (iii) comments, structured as follows. The forum contains several threads; each learner can create new threads or add content to pre-existing threads. These threads consist of one or more posts, arranged in temporal order (a new post is added to the thread under the most recent). The person who created the thread writes the initial post. A participant can reply to a thread (add a new post) or reply to a post (add a new comment). Some researchers propose that posts and comments can be used interchangeably (Brinton *et al.*, 2014), while others distinguish between posts and comments (Rossi and Gnawali, 2014).

The OpenEdX platform, which some related work used data from, is structured in a three-level hierarchy (Ntourmas *et al.*, 2019) as Figure 2.3 (below) shows. *Discussions-responses-comments* is the terminology utilised on this platform while in Coursera, as mentioned, the three levels are referred to as *threads-posts-comments.*

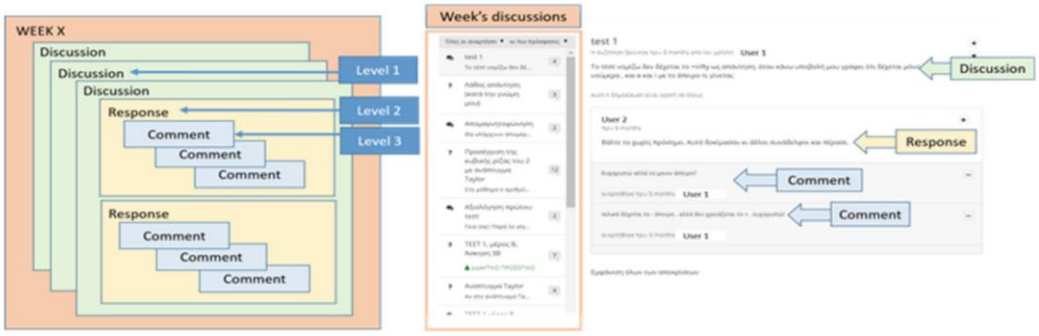

Figure 2.3: An overview of the OpenEdX discussion forum structure (Ntourmas *et al.*, 2019).

The focus of the current thesis is to identify posts on The Stanford MOOCPosts dataset in terms of two levels (*commentThread or comment*); on the FutureLearn, two levels (*comment or reply*) are used to identify if there is a need for intervention or not. As shown, different platforms use different terms; thus, here for simplicity and consistency, the term *post* is used to define such comments and replies on all platforms as the goal here is to identify posts that require instructor intervention regardless of if they are threads, posts, comments, or replies.

2.2.1.4. Instructor Intervention in MOOCs Corpora

The current interest in the instructor intervention problem in MOOCs led researchers of previous works to use different corpora. For example, the Anonymized Coursera Discussion Threads dataset (Rossi, 2023; Rossi and Gnawali, 2014) collected data from 60 courses on the Coursera platform with about 100,000 threads. Despite being huge in size, it does not contain the textual content of posts; it only provides metadata for linguistic analysis (about posts) such

as num_words (number of words). Another available dataset is the Stanford MOOCPosts dataset (Agrawal *et al.*, 2015; Agrawal and Paepcke, 2019) which is available on request for researchers; it contains 29,604 learner forum posts that include the text of posts. Although this dataset is large and widely used in researching the instructor intervention problem in MOOCs, it is limited to the textual content of posts along with some metadata; however, it does not include information about learners that can be used to study behaviours or dropout rates. This thesis applied experiments such as text classification and topic modelling using The Stanford MOOCPosts dataset. Then, a new dataset was created that gathered posts from the FutureLearn platform to conduct different experiments related to learner behaviour to identify urgent posts requests and MOOC learners at risk of dropping out.

2.2.2. NLP

NLP is a common topic in computer science literature. The study of NLP began in the 1950s (Kalyanathaya, Akila and Rajesh, 2019) from the intersection of linguistics and AI (Nadkarni, Ohno-Machado and Chapman, 2011). NLP uses computer methods, algorithms, and tools to learn, understand, process, and create human language content (Hirschberg and Manning, 2015). NLP's main goal is to process text computationally so that it can be analysed in terms of natural language data which can then be represented in different ways (Conneau *et al.*, 2016). The use of NLP has developed over recent years into research and technology domains because of its computational power, its capacity to process large amounts of linguistic data, and its ability to create effective ML models to help understand the structure of human language use (Hirschberg and Manning, 2015).

One significant topic and application of NLP is text classification which assigns an unstructured text according to its content to predefined categories (Zhang, Zhao and LeCun, 2015). Various ML models have been applied to text classification with a wide range of applications including sentiment analysis (Dawei *et al.*, 2021) and spam detection (Sharmin and Zaman, 2017).

Numerous methods exist for extracting information from unstructured text input and using it to train classification models. To represent words as numerical vectors, techniques include the Bag-of-Words (BoW) model, the Word Embedding model, and state-of-the-art Language models. The BoW model is relatively simple; it relies on the presence or absence of a word in a document by generating a vocabulary from a corpus of documents; it then tracks the

frequency of the use of each target word. The dimension of vector is fixed with the same length of the vocabulary featured in all the target documents. The *term frequency* (TF) representation counts words by giving words different weights based on how frequently they occur in each document. In the TF case, each word is counted as equally important. Thus, there are common words with the highest frequencies of use but that are assigned little importance (e.g., and, is, that, etc.). Meanwhile, *term frequency-inverse document frequency* (TF-IDF) is used to display the importance of a word in relation to a particular document in a collection or corpus. In vector representation, a word's value increases proportionally to its count, but a word's importance is inversely proportional to its frequency in the corpus. In other words, more frequently used words are weighted more lightly and less frequently used words are weighted more heavily.

The term *word embedding* was first used in 2003 by Bengio et al. (Bengio *et al.*, 2003) to explain the process by which words are mapped to fixed-length vectors of real numbers. The probability distribution for each word appearing before or after another is used to calculate these vectors. Thus, words associated with particular related contexts normally appear together in a corpus and thus they will be closer to one another in their vectors. Different popular word embedding techniques include word2vec (Mikolov *et al.*, 2013), GloVe (Pennington, Socher and Manning, 2014), and Fast Text (Bojanowski *et al.*, 2017). In *language models* such as BERT, which uses contextual dynamic embedding, each word has a representation that is a function of the entire input sequence. As a result, depending on the context, a word may have different vectors.

In this thesis word2vec and BERT were used as word embedding tools in different experiments. Therefore, the difference between them must be clarified. Both are used for generating vector representations of words. However, while word2vec vector representations are static, they capture contextual information about a particular word in relation to the corpus used to train them. Also, word2vec vector representations are context-independent: a word's embedding will be the same regardless of the context in which it was used as each word has a single vector (numeric) representation. BERT, meanwhile, is based on the context of a given word: it creates context-aware embeddings that enable each word to have various representations (each representation in this case is a different vector).

2.2.3. Machine Learning

ML is a field of computer science and AI that focuses on developing models that can learn independently from data and make decisions and predictions (Naqvi *et al.*, 2023). ML is defined by Arthur Samuel as a field that enables computers to learn without being explicitly programmed (Mahesh, 2020). *Supervised, semi-supervised,* and *unsupervised techniques* are the three main approaches used in ML-type learning algorithms. One might be selected for use over another depending on the nature of the research problem and the data available for analysis. Also, ML algorithms are employed in many different applications such as computer vision and NLP. The two main types of ML models are *traditional machine learning* and *neural networks*. The theoretical backgrounds of ML algorithms employed in this thesis are clarified in the following sections.

2.2.3.1. Traditional Machine Learning

In traditional ML approaches, features must be manually extracted by subject matter experts through a procedure called *feature engineering*. These features are fed to simple-structure ML algorithms. Feature engineering is crucial in the beginning to allow an algorithm to decide on outputs depending on what it has discovered from the given features. A brief overview of the traditional ML models used in this thesis is provided below; all of them are based on supervised learning. In Chapter 8 to classify urgent posts, the traditional ML used were naive Bayes, logistic regression, support vector machine, random forest, and boosting (extreme gradient boosting).

2.2.3.1.1. Naive Bayes

Naive Bayes (NB) (*independent Bayes*) *classifiers* (also referred to as *probabilistic classifiers*) were developed based on applying Bayes' theorem. NB is the simplest form of Bayesian network models (Jiang, Zhang and Cai, 2008). In Bayesian networks, features are assumed to be independent of one another; the naive classifier assumption holds that the presence of a feature in one category has *no* relevance to the presence of other features. NB models are easy to create; at the same time, NB models offer good performance and are particularly useful for use in classification tasks with large data sets to provide effective text classification.

2.2.3.1.2. Logistic Regression

Logistic regression (LR) (or the logit regression model) has been used since 1845 when population growth in the period was being studied mathematically (Cokluk, 2010). LR is a widely used statistical model used to calculate the probability that a given instance belongs to a certain class. Although LR is called regression, it is not regression per se; rather, LR is a classification algorithm which is based on the logistic function that has an output value of between 0 and 1. LR is commonly employed to address binary classification tasks.

2.2.3.1.3. Support Vector Machine

The *support vector machine* (SVM) (or support vector network) is widely employed for solving classification problems such as text classification. SVM is based on statistical learning theory and the structural risk minimisation principle (Kamath, Bukhari and Dengel, 2018). Finding a hyperplane in an N-dimensional space (N denotes the number of features) that categorises the data points (support vectors) clearly into distinctive groups in regions (one for each class) is the objective of SVM. There are numerous possible hyperplanes that might be chosen to split two groups of data points but the objective is to find the maximum distance between data points of both classes (the margin). Figure 2.4 (below) shows an overview of SVM with support vectors, hyperplanes, and margins.

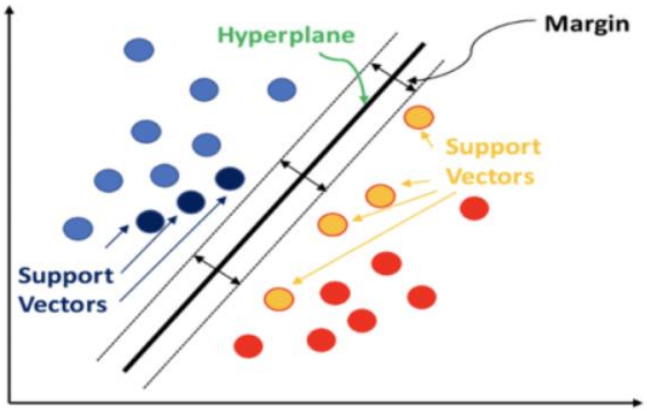

Figure 2.4: An overview of SVM - support vectors, hyperplanes, and margins.

2.2.3.1.4. Decision Tree

A *decision tree* (DT) operates on a hierarchical tree-like model of decisions and is used for both classification and regression tasks. The success of the DT method can be primarily attributed

to its simplicity of use and capacity to produce precise predictive models with comprehensible and interpretable structures (Yu *et al.*, 2010).

The structure of a DT consists of three major types of nodes: (i) the root node (the starting node of the tree that represents the data and is divided into different distinctive sets depending on specific features), (ii) the leaf node (the node at the end of the chain that represents the final outcome), and (iii) the internal node (a node other than a leaf node that represents a "decision") (Ali *et al.*, 2012). The pathways from root to leaf represent classification rules; to find the optimal split points inside a tree, a so-called greedy approach is conducted. The splitting process continues until the predefined stopping criteria are met. Random forest (RF) (discussed later) is an extended version of DT.

2.2.3.1.5. Random Forest

The *random forest* (RF) algorithm is a collection of decision trees known as forests as each tree is dependent on a random vector's values (Breiman, 2001). RF is an ensemble method that combines the results of many various decision tree classifiers to arrive at a single result rather than depending on one decision tree. The class that most of the trees chose (votes) is the output of the RF model in classification problems. Figure 2.5 (below) illustrates a RF voting system and an RF structure, where n is the number of trees. To achieve good performance, the number of trees should be selected through trial and error since there is no optimal number of trees that applies to all models.

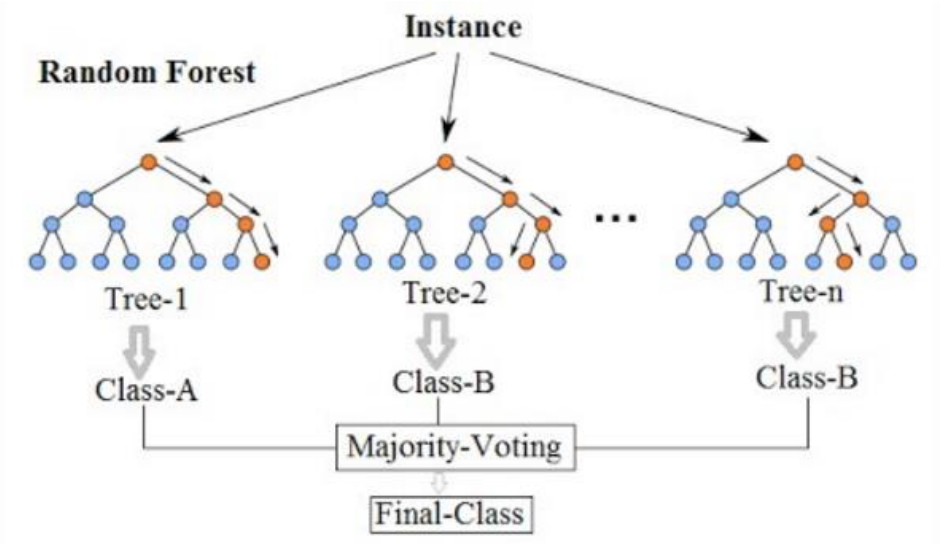

Figure 2.5: An overview of RF.

2.2.3.1.6. Boosting Model — Extreme Gradient Boosting (XGBoost)

XGBoost (which stands for *extreme gradient boosting*) was initially created in 2014. It is an extended, optimised, distributed, flexible, and fast technique for gradient boosting (Chen and Guestrin, 2016). XGBoost is a class of ensemble machine learning techniques (tree-based) that can be applied to classification or regression predictive modelling tasks. XGBoost is regarded as one of the best algorithms for supervised learning (Osman *et al.*, 2021). XGBoost supports parallel processing which is more advantageous than traditional boosting algorithms (sequential) and better results are obtained with sparse data. The main benefit of using XGBoost is that it successfully prevents overfitting (Zhang *et al.*, 2023) by offering a number of parameters of regularisation, such as gamma, alpha, and lambda.

2.2.3.2. Deep Learning

Deep learning (DL) is the most recent achievement of the ML era; DL is based on neural networks with complex structures. The main significant distinction between traditional ML and DL is the method of extracting features from data inputs. In DL, features are extracted by the algorithm itself from vast amounts of data (big data) with a lesser need for feature engineering (Whang *et al.*, 2023). Figure 2.6 (below) provides an overview of the differences between these two models.

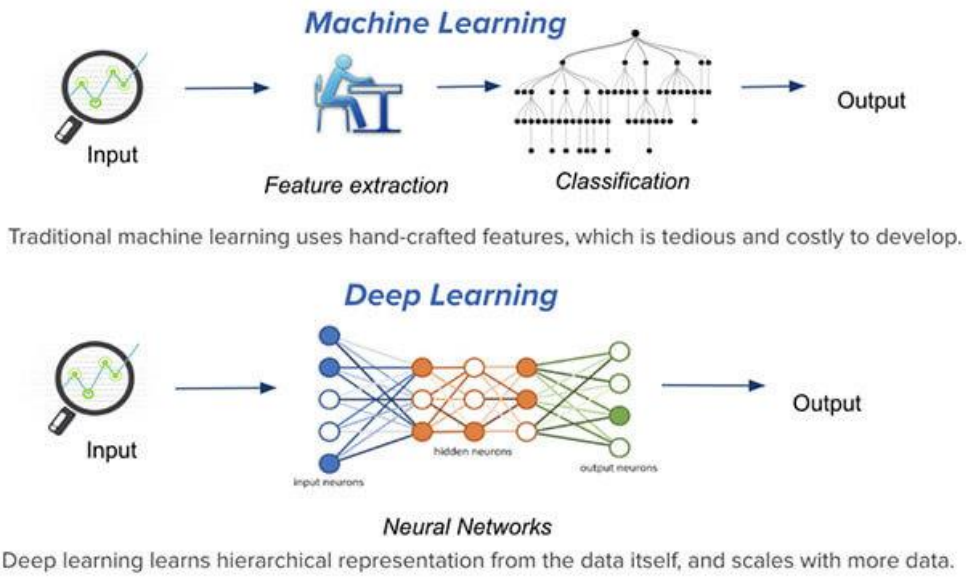

Figure 2.6: An overview of the difference between extracting features in traditional ML and DL.

*Artificial neural network*s (ANNs) are the backbone of deep learning algorithms; DL contains ANNs with many layers. Hence in the following sub-sections, the ANN and the major DL architectures used in this thesis: multi-layer perceptron (MLP), convolutional neural network (CNN) and recurrent neural network (RNN) are provided.

2.2.3.2.1. Artificial Neural Network

The primary goal of an ANN construction was to mimic the functioning of the human brain (Bashar, 2019) and understand how information is processed by biological nervous systems composed of neurons. ANNs have artificial neurons that are linked to one another in various layers of networks like a human brain that has neurons interconnected to each other. In the training of ANNs, the input is provided and the network is 'told' what the output should be. The network then assigns various weights between its neurons.

The training phase in a single neuron can be clarified as shown in Figure 2.7 (below) as a set of input values (*X1, X2, X3, ...Xn*), associated weights (*W1, W2, W3, ...Wn*), and a function that sums the weights, adds a bias (constant) parameter, and maps the results through the activation function to an output (*y*).

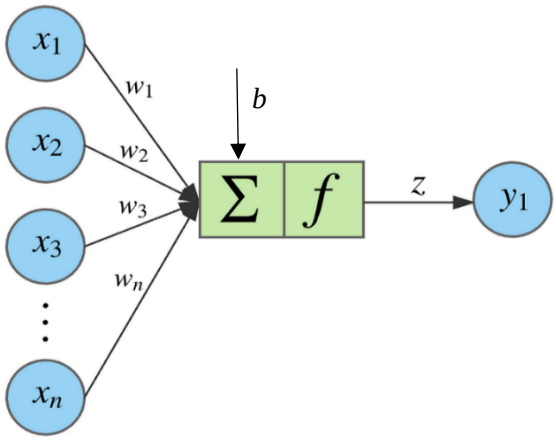

Figure 2.7: An overview of a artificial neural network neuron in the training phase.

This is described mathematically as follows:

$$s = \sum_{i=1}^{n}(w_i\, x_i) + b \tag{2.1}$$

2.2.3.2.2. Multi-Layer Perceptron

The *multi-layer perceptron* (MLP) is a typical example of feedforward ANN class that is fully connected with multi-layered. MLP is simple and one of the most widely used DL algorithms. The three essential layers used to define MLP architectures are the *input, one or more hidden, and output* layers as the output of one layer serves as the input for the next layer as depicted in Figure 2.8 (below). Each neuron is 'fully connected': nodes in each layer are connected to all other nodes in the previous and next layers as shown in Figure 2.8 (below).

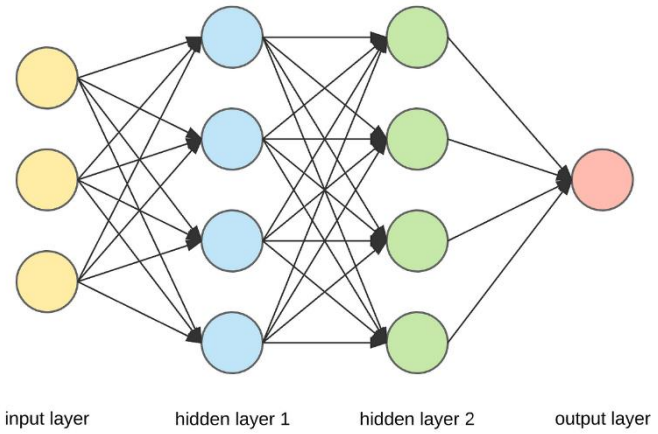

Figure 2.8: An overview of a multi-layer fully connected MLP.

Three main steps are involved in MLP training: (i) forward pass and prediction, (ii) a loss function is used to compare the prediction with the actual value, and (iii) backpropagation which determines the gradients for each node in the network, using the provided error value. The gradient is the quantity utilised to modify the internal weights of the network, thus enabling 'learning' to occur. The MLP used in this thesis is further explained in Chapter 4, in the first experiment (*multi-dimensional deep learning model*) to train the numerical data (multiple dimensions) as the sub-model from the overall model.

2.2.3.2.3. Convolutional Neural Networks

Convolutional neural networks (CNNs) (LeCun *et al.*, 1998) are one of the most important ANNs and an impressive form of DL; CNNs include neurons with their own respective biases and weights in several layers. The term CNN derives from a mathematical linear procedure between matrixes called *convolution* (Albawi, Mohammed and Al-Zawi, 2017) which is a process involving the *sliding* or *convolution* of a predetermined window of data. The network consists of multiple weights and biases in layers, but the features exist in the form of spatial structures. Three layers make the fundamental structure of a CNN: (i) a convolutional layer,

(ii) a pooling layer, and (iii) a fully connected layer. An overview of the structure of a CNN is shown in Figure 2.9 (below). The filters in the convolutional layer transform the large amount of data into *feature maps*. The pooling layer then processes these feature maps, reducing the parameters. The fully connected layer is processed using the output features from the pooling layer (see Figure 2.9 below). In NLP, the motivation for employing CNNs is to locate more complex features from constituting words or n-grams (Young *et al.*, 2018), and to obtain the hierarchical, high-level semantic representation of the selected input words (Gu *et al.*, 2018; Guo *et al.*, 2019).

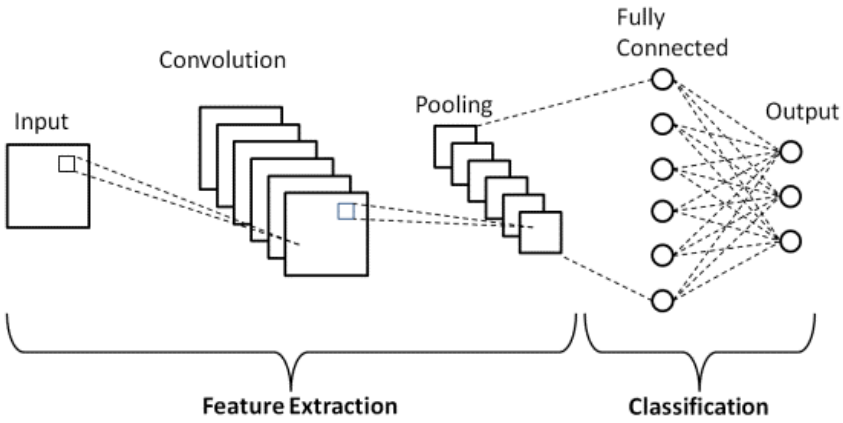

Figure 2.9: An overview of CNN structure.

CNN architecture was used in this thesis in three different experiments: in Chapter 4 in both experiments (i) the *multi-dimensional deep learning model*, (ii) *plug & play with deep neural networks* to predict urgent posts from text, and (iii) in Chapter 6 to predict dropout learners.

2.2.3.2.4. Recurrent Neural Networks

Recurrent neural networks (RNNs) (Elman, 1990) are a class of ANNs and one of the most widely used DL architectures. RNNs are capable of modelling sequential data or time series data types. RNNs handle inputs differently with a state that is not 'lost' (as in a normal ANNs) as the state is lost when an input is processed. Each current state's output is processed and concatenated with the input of the following step in sequence to extract information. In NLP, the motivation for using RNNs is their ability to capture text sequentially (Young *et al.*, 2018).

In RNNs, the gradient value is transmitted to an earlier state; however, this raises a problem known as the *vanishing gradient problem,* which gradually vanishes (i.e., gets smaller) as the number of time steps in the series increases. *Long short-term memory* (LSTM) and *gated recurrent unit*s (GRUs) are the two main varieties of RNNs that have emerged as a result of the

development of gating mechanisms to address some of RNN's limitations, as explained in the next sections.

2.2.3.2.4.1. Long Short-Term Memory

The *Long short-term memory* (LSTM) (Hochreiter, urgen Schmidhuber and Elvezia, 1997) is an enhanced RNN architecture that deals with dependence sequences of data with feedback connections; it has attained unprecedented performance in several fields. LSTM was developed to overcome the exploding/vanishing gradient problem faced by RNNs when learning long-term dependencies in datasets (Van Houdt, Mosquera and Nápoles, 2020) by adding a new component known as *cell state* and internal mechanisms called *gates*, namely: *forget, input and output gates* with a range of functions. The cell state recalls values over time intervals while gates control the flow of information into and out of the cell. The information flow in an LSTM cell is illustrated in Figure 2.10 (below).

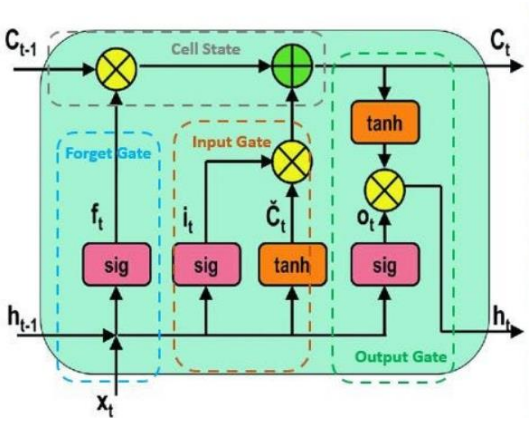

Figure 2.10: An overview of LSTM cell.

The flow of information in a LSTM cell is as follows:

- Forget gate: which information should be forgotten in the current cell state is decided by this gate. Information is derived from both the previous hidden state h_{t-1} and the current input x_t through the sigmoid function (σ) that generates an output value of f_t between 0 and 1. To *forget* means getting closer to 0, and to *keep* means getting closer to 1. This is illustrated in the following formula (2.2):

$$f_t = \sigma\left(w_f . [h_{t-1}, x_t] + b_f\right) \tag{2.2}$$

- Input gate: which new input data should be included in the new cell state is decided by the input gate; this is done by following two steps. First, the previous hidden state h_{t-1}

and the current input x_t are passed through the sigmoid function (σ), which decides which values will be updated by generating values between 0 and 1. 1 denotes importance and 0 indicates it is non-importance. Second, the hidden state h_{t-1} and the current input x_t are also passed through the tanh function which generates values ($\tilde{C}_t$) of between -1 and 1 to help regulate the network. This is mathematically expressed as follows (2.3) and (2.4). Then the output from tanh is multiplied (element-wise multiplication) by the output from sigmoid.

$$i_t = \sigma\left(w_i . [h_{t-1}, x_t] + b_i\right) \quad (2.3)$$

$$\tilde{C}_t = tanh\left(w_c . [h_{t-1}, x_t] + b_c\right) \quad (2.4)$$

- Cell state: which updates the previous cell state, C_{t-1}, to the present cell state, C_t. First, the previous cell state is multiplied (element-wise multiplication) by the forget value. Then the output from $i_t \times \tilde{C}_t$ is added (element-wise addition) following Equation (2.5):

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (2.5)$$

- Output gate: which information should be passed to the next hidden state; it manages the output flow to other cells or as the final results. The previous hidden state h_{t-1} and the current input x_t are passed into the sigmoid function (σ) as shown in Equation (2.6). Then the newly modified cell state is then passed to the tanh function. Then the output from tanh is multiplied by the output from sigmoid (element-wise multiplication) to obtain a final decision as shown in Equation (2.7).

$$o_t = \sigma\left(w_o . [h_{t-1}, x_t] + b_o\right) \quad (2.6)$$

$$h_t = o_t \times tanh(C_t) \quad (2.7)$$

This model is applied in Chapter 4 in the second experiment (*plug & play with deep neural networks*) to predict posts.

2.2.3.2.4.2. Gated Recurrent Unit

The *gated recurrent unit* (GRU) (Cho *et al.*, 2014) is another type of RNN and represents the most up-to-date generation introduced in 2014. GRU is an updated version of LSTM with a

simplified configuration and fewer parameters which leads to faster processing by combining both the forget gate and the input gate into a single unit gate called an *update gate* as shown in Figure 2.11 (below).

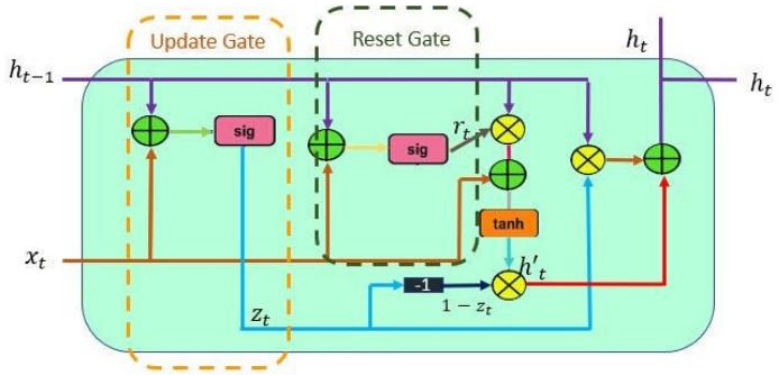

Figure 2.11: An overview of GRU cell.

GRU contains two gates (an *update gate* and a *reset gate*); it does not contain a cell state as in LSTM. Instead, GRU uses hidden states to transport information. The following explains the processing of a single unit in GRU and the basic workflow:

- Update gate: like the forget and input gates, the update gate decides which information should be forgotten and which new input data should be included in the current cell state. The update gate is calculated using the formula (2.8):

$$z_t = \sigma\left(w_z . \left[h_{(t-1)}, x_t\right] + b_z\right) \tag{2.8}$$

- Reset gate: how much of the past information from the previous hidden state to forget is decided by reset gate as in Equation (2.9).

$$r_t = \sigma\left(w_r . \left[h_{(t-1)}, x_t\right] + b_r\right) \tag{2.9}$$

- Current memory: the rest gate is used to keep the relevant past information. The respective equation is as follows:

$$\tilde{h}_t = tanh\left(w_{\tilde{h}_t} . \left[r_t \times h_{(t-1)}, x_t\right] + b_{\tilde{h}_t}\right) \tag{2.10}$$

- Final memory: to calculate the information for the current hidden state (h_t) to pass it on, as shown in Equation (2.11).

$$h_t = z_t \times h_{(t-1)} + (1 - z_t) \times \tilde{h}_t \tag{2.11}$$

This algorithm is used in Chapter 4 in the second experiment (*plug & play with deep neural networks*) to meet the objective of predicting urgent posts.

2.2.3.2.4.3. Bidirectional RNNs

Bidirectional RNNs (Bi-RNNs) (Schuster and Paliwal, 1997) are an expanded version of unidirectional RNNs which uses a combination of two standard RNN layers. Bidirectional RNNs process the input sequences on both sides, past to future (forward) and on the opposite direction, future to past (backward) as shown in Figure 2.12 (below) to predict the data using contextual information. In NLP, the motivation for utilising Bi-RNNs is that it can understand a comprehensive sentence and has complete sequential knowledge of all words occurring before and after each word in a given sentence.

Two types of such Bi-RNNs were used in this thesis, namely, Bi-LSTM and Bi-GRU, in the second experiment (*plug & play with deep neural networks*) in Chapter 4 to classify posts (see Chapter 4) and in Chapter 6 to identify dropout learners.

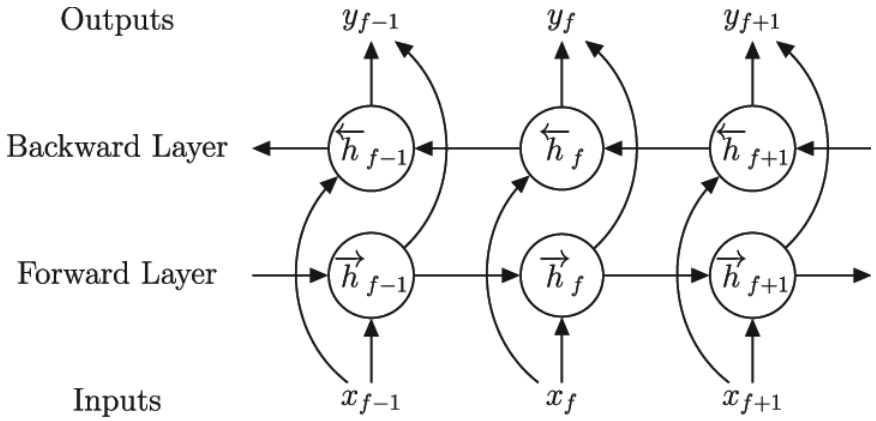

Figure 2.12: An overview of the Bi-RNN model.

2.2.3.2.5. Transformer Architecture

A recent development in attention mechanisms is the invention of *Transformer architecture* (Vaswani *et al.*, 2017) which has emerged as a promising method and a powerful DL algorithm. Transformer architecture is supported by an attention mechanism based on an encoder-decoder type of architecture. The encoder, which is the encoding layer, converts a series of input data into an abstract continuous representation. The decoder (the decoding layers) then takes the continuous representation and generates a single output. Each encoder and decoder layer employs an attention mechanism to assess the relative weight of each individual input item of

data. Transformers have extremely long-term memory which is made possible by the attention mechanism and processing all inputs simultaneously. Transformers are context-dependant and can understand the context given to each word with its meaning in the text.

BERT is the most powerful Transformer model and has been extremely popular, being applied in text classification models with high performance, such as in (Fonseca *et al.*, 2020) and (Pereira *et al.*, 2021). According to (Rogers, Kovaleva and Rumshisky, 2021), in 2020, BERT established itself as a common baseline in NLP experiments, with over 150 research publications analysing and enhancing the model.

2.2.3.2.5.1. BERT

BERT (Devlin *et al.*, 2018) which was launched in 2018 by Google AI researchers based on the Transformer architecture developed for NLP is an advanced language representation model pre-trained on a large amount of textual data. BERT uses context-dependent embedding to find the relationships between words and understand sentences. It can be used as pre-trained or fine-tuned by customising it to the specific NLP task as illustrated in Figure 2.13 (below). Two versions of BERT architecture are available (BERTbase and BERTlarge). The settings of BERTbase version are as follows: (layers=12, hidden states=768, heads=12, and parameters =110M) while in the BERTlarge version they are as follows: (layers=24, hidden states=1024, heads=16 and parameters =340M).

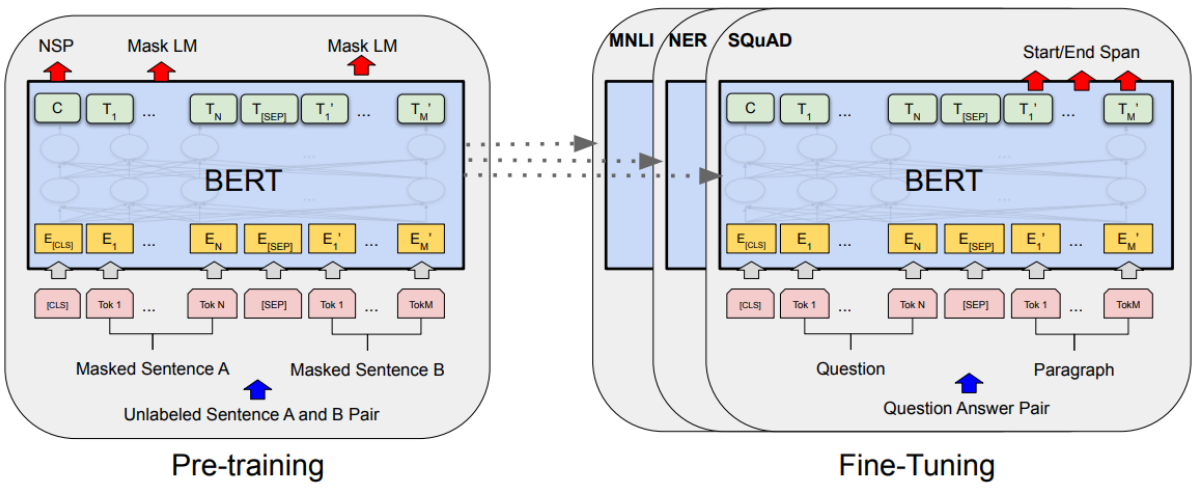

Figure 2.13: General pre-training and fine-tuning procedures for BERT (Devlin *et al.*, 2018).

BERT is based on a multi-layer approach and works as an attention mechanism that learns the contextual relationships between words or sub-words in a text. The Transformer encoder reads the entire sequence of words at once, in contrast to directional models, which read the text input

sequentially (from right to left or left to right). BERT uses a combination of two unsupervised training techniques: *masked language modelling* and *next sentence prediction.*

In the *masked language model* (MLM) approach, random word sequences are changed with a [MASK] token for 15% of the words in each sequence before being fed into the BERT. Based on the context offered by the other, non-masked, words in the sequence, the model then tries to predict the original value of the masked words.

In the *next sentence prediction* (NSP) mechanism, during the training phase, the model learns to predict whether the second sentence in a pair will come after another in the original document by receiving pairs of sentences as input. During training, 50% of the inputs are pairs in which the second sentence is the next one in the original text, and in the remaining 50%, the second sentence is a randomly selected sentence from the corpus. The underlying assumption is that the second phrase will not be connected to the first. This captures more long-term information.

BERT is used in this thesis in different experiments depending on the experimental settings: as embedding, in Chapter 4 with the second experiment (*plug & play with deep neural networks*) and in Chapter 6 which proposed multi-siamese BERT and multiple BERT, that accept more than two posts' inputs, which are discussed in detail in Chapter 6. Moreover, BERT is used to predict urgency in Chapters 7, 8, and 9.

2.3. Literature Review

MOOC providers, MOOC courses, MOOC learners, and even MOOC researcher numbers have grown dramatically in recent years. There has been a huge recent research interest in addressing the challenges faced by MOOCs; one of the central challenges is the point at which instructor intervention is needed. (Chaturvedi, Goldwasser and Daumé III, 2014) observed that, after instructor intervention in a thread, the thread increased its posting/viewing, indicating the importance of the instructor's intervention. However, given the huge number of posts on MOOCs, varying from 103 to 9300 (as shown in Figure 2.14 below) (Rossi and Gnawali, 2014), as well as the enormous learner-to-instructor ratio in MOOCs, it would be time-consuming and often impossible for an instructor to read all posts and then determine which posts required attention (Litman, 2016).

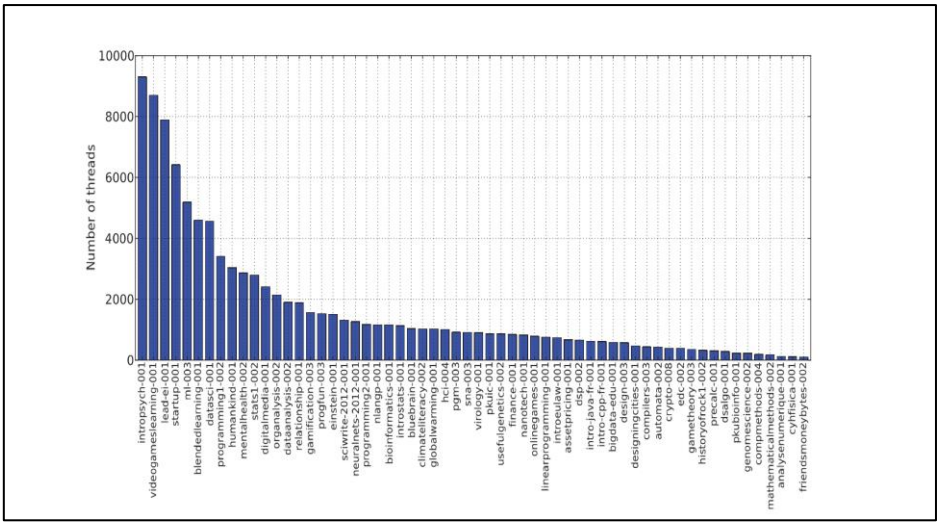

Figure 2.14: Number of threads versus course identifiers (Rossi and Gnawali, 2014).

The above challenge has been confirmed by an instructor from Vanderbilt University who claimed it is impossible to navigate such forum discussions after three days and suggested using NLP to reorganise forums (Hollands and Tirthali, 2014). Also, as another instructor, Dr. Williams (Zheng *et al.*, 2016), mentioned,

> Although it's impossible to take care of every student in my course given the huge population, I still feel bad when I cannot finish going through all the discussion on the forum or answering all the questions they raised. I know I shouldn't feel guilty. But as a teacher, I feel like it's my responsibility to help everyone in my class.

This clarifies how it is a challenge for instructors to identify the need for urgent instructor intervention in MOOC environments. Therefore, automatic identification of urgent cases to address instructor overload in MOOCs contexts is required.

This section gives a brief review of related works linked to the current thesis. First, it describes studies that have been performed to solve some of the challenges of MOOCs using NLP. Then, it describes previous studies on categorising posts on discussion forums. Following this, it explains the types of intervention possible, such as via automatic tools or human intervention. Finally, it systematically reviews previous studies on instructor intervention models.

2.3.1. MOOCs and NLP

NLP appeared in the 1950s; from its inception, researchers focused on different applications including educational ones, such as automatically scoring student texts and text-based dialogue

tutoring systems. Later, they concentrated on spoken language technologies. Although NLP and these applications are still attracting many researchers, recent phenomena have appeared, such as big data, and MOOCs. Researchers have focused on the related challenges and some even on combining AI approaches with NLP to study the big data produced by MOOCs (Litman, 2016).

Analysis of discussion forum content data from MOOC platforms offers the opportunity to understand learners and their behaviours (Gardner and Brooks, 2018) including *sentiment* (Wen, Yang and Rose, 2014) and *confusion* (Agrawal *et al.*, 2015) and so on. These methods include statistical analysis, and traditional ML and DNNs. There has been considerable research on MOOCs that use NLP as an analysis tool; for example, (Crossley *et al.*, 2016) used some NLP tools, such as the Writing Assessment Tool (WAT), Tool for the Automatic Analysis of Lexical Sophistication (TAALES), Tool for the Automatic Analysis of Cohesion (TAACO), ReaderBench (RB), and Sentiment Analysis and Cognition Engine (SEANCE) to understand learner completion. Other research has focussed on the prediction of course completion and success (Robinson *et al.*, 2016); here, the authors established NLP models using *unigram* and *bigram* features to parse text and make predictions.

The creation of *sentiment* recognition was explored by (Liu *et al.*, 2016) selected features for a prediction model using multi-swarm optimisation to recognise online course reviewers in MOOCs. Another study in sentiment analysis was conducted by (Wen, Yang and Rose, 2014), who mined learner opinions about a course and found a correlation between the sentiment ratio and learner dropout. In addressing *confusion* from posts, (Yang *et al.*, 2015) used NLP features (Linguistic Inquiry and Word Count (LIWC) and several question marks and sentences beginning with a confusing expression) and user click patterns, which were applied in a classification model. In addition, (Agrawal *et al.*, 2015) investigated identifying confusion by using three inputs, namely, post content (bag of words – BoW), post metadata, and classifier combinations.

(Wise *et al.*, 2017) proposed that NLP analysis in general and classification of posts in particular appear to be robust methods of solving the problem of instructor overload in MOOC environments. They applied 2410 features, including linguistic features (unigrams and bigrams) and the number of views and votes features to automatically predict content-related posts. (Arguello and Shaffer, 2015) incorporated linguistic and other features to predict speech acts in MOOC posts. They used 201 features including LIWC (60 features), sentiment (4

features), unigram (100 features), text similarity (6 features), temporal features (3 features), sequential correlation (7 features), author (1 feature), links (1 feature), modals (2 features), position (2 features), post comments (1 feature), punctuation (1 feature), and votes (1 feature).

Previous studies have shown that NLP techniques can be used to address different challenges facing MOOCs to enhance learner performance and learning outcomes.

2.3.2. Categories of Posts on Discussion Forums in MOOCs

Different previous studies and efforts have addressed discussion forum data in MOOC environments; posts have been categorised in different ways for different purposes. For example, (Stump *et al.*, 2013) classified every learner post in relation to the topic of posts and the role of posters into a manageable number of categories. They classified topics into *content, other coursework, social/affective, course website/technology, course structure/policies, other, missing data and non-English*, and classified the role of posters (*help-seeker (or information-seeker), help-giver (or information giver) and other)* to develop a framework to classify posts. Meanwhile, (Rossi and Gnawali, 2014) classified thread-based discussions into the types of communications between users such as *social talk, open-ended topics, (un)resolved issues,* and *course logistics*, etc., by utilising five types of language-independent features except word count.

(Arguello and Shaffer, 2015) classified posts into speech acts categorised to describe the purpose of a MOOC post to predict instructor intervention, assignment completion and assignment performance. They focused on seven speech act categories (*question, answer, issue, issue resolution, positive acknowledgement, negative acknowledgment* and *other*). An alternative approach by (Wise *et al.*, 2017) attempted to classify starting posts according to whether they were content-related posts (binary dimension: if the content of the post is related to the course material or not); their aim was to support instructors and learners in finding appropriate posts. The study also revealed that using *quantity of views* and *votes* as features was useless for identifying content-related posts.

(Yang *et al.*, 2015) and (Agrawal *et al.*, 2015) categorised posts based on confusion. Yang et al. (2015) classified each post according to different levels of confusion on a four-point Likert scale defined as *no confusion, slightly confused, moderately confused,* and *seriously confused.* In contrast, Agrawal et al. (2015) categorised every post in relation to the six following dimensions: *question, answer, opinion, confusion, sentiment and urgency*. The first

three dimensions were binary (0 or 1), while the last three dimensions took a discrete value from a scale (0–7).

Several studies, and the current thesis, built models based on Agrawal et al.'s (2015) data which is the Stanford MOOCPosts dataset. As shown later in the systematic literature review, various studies used The Stanford MOOCPosts dataset to classify different dimensions or only *urgent* dimension.

2.3.3. Categories of Types of Interventions

Prior efforts to solve the problem of determining when instructor intervention is required in MOOC contexts fall into two main categories (Wise, Cui and Vytasek, 2016): (i) automatic tools to provide specific solutions which are a type of recommendation system, and (ii) predicting the need for instructor intervention automatically with the goal of supporting and facilitating human interaction. In providing a specific solution using the first type of intervention, Agrawal et al. (2015) recommended showing video clips to confused learners as an intelligent intervention. Other research (Rossi *et al.*, 2022) used the Conversational Agent in an Educational Recommender System (CAERS) which, when a post on a discussion forum shows that pedagogical intervention is necessary, the topic of such a message is recognised and then suitable educational content is suggested to answer learner queries. Although recommendation systems provide learners with automated interventions and solutions, there are cases that require direct interaction by humans, hence the importance of this thesis. The next section provides a systematic review of studies on the intervention problem in MOOCs related to the second approach: predicting the need for instructor intervention automatically.

2.3.4. Systematic Literature Review of Identifying Instructor Intervention Need in MOOC Discussion Forums

While the literature on the proposing an intervention model that helps instructors to decide when intervention is needed has been continuously proliferating over the past few years, to date and to the best of the author's knowledge, no studies designed to survey these works have been conducted. A systematic literature review (SLR), as the name infers, provides a collection, evaluation, integration, and presentation of findings from various research on a particular research topic (Pati and Lorusso, 2018) through transparent and reproducible methods (Clark *et al.*, 2020) — unlike traditional literature reviews (Kraus, Breier and Dasí-Rodríguez, 2020).

This work aims to categorise academic studies on instructor intervention in MOOCs based on discussion forum posts using a systematic analysis of the available peer-reviewed works. The preliminary screening covered 414 abstracts published from 2014 — when the first abstract screened was published — until the end of 2022. The PRISMA protocol (Moher *et al.*, 2009) was followed to identify relevant papers to ensure rigour in the study selection and the reporting of findings. To the best of the author's knowledge, identifying the need for instructor intervention in MOOCs has not yet been addressed in extant MOOC-related SLRs, making the present SLR both timely and significant.

2.3.4.1. Systematic Literature Review Motivations

Although MOOCs have been around for a decade (Pappano, 2012), several platforms are noticeably lagging in dealing with the massive numbers of learners they attract; especially in terms of learner-instructor interaction and the provision of timely intervention. Thus, the current SLR is essential as it highlights the gaps in the extant research on these issues through surveying the literature and exploring how the surveyed studies have identified, modelled, and provided results and recommendations on MOOC discussion forum-based instructor intervention. While these studies are initially expected to share heterogeneous definitions of intervention, datasets, methodologies, and approaches for reporting results, an explicit exploration of the extent to which the surveyed models are similar/different based on the type of intervention(s) addressed is considered appealing.

2.3.4.2. Previous Surveys on MOOCs

The present SLR study was attractive to the researcher due to the limitations of the current systematic research about instructor intervention in MOOCs. (Meet and Kala, 2021) SLR highlighted that just (7%) of the research on MOOCs was instructor focused, which indicates the need for additional research on the role of instructor intervention. Thus, the importance of the current research project lies in its novelty in analysing research devoted to the problem of instructor intervention in MOOCs. Previous SLR research has concentrated on a variety of literature on different aspects of MOOCs during given periods (see Table 2.1 below). Therefore, a systematic analysis of instructor intervention in MOOC studies to date is required. To the best of the author's knowledge, this SLR is the first attempt at a comprehensive evaluation of the literature on instructor intervention in MOOC discussion forums.

Table 2.1. Previous SLRs on MOOCs, distributed by publication year, aims, focus, and period covered.

Reference	Year published	Aims and focus	Period covered
MOOCs in general			
(Liyanagunawardena, Adams and Williams, 2013)	2013	Reported on the concepts, case studies, and educational theories of published MOOC literature and classified these studies into different categories.	2008–2012
(Kennedy, 2014)	2014	Investigated the characteristics of MOOCs from three perspectives: (i) definitions of openness, (ii) barriers to persistence, and (iii) a distinct structure of two pedagogical approaches.	2009–2012
(Raffaghelli, Cucchiara and Persico, 2015)	2015	Explored and examined the trends in methodological approaches used in MOOC research.	January 2008–May 2014
(Veletsianos and Shepherdson, 2015)	2015	Explored interdisciplinarity in MOOC research.	2013–2015
(Veletsianos and Shepherdson, 2016)	2016	Presented a comprehensive picture of the literature by examining geographic distribution, publication outlets, citations, data collection, analysis methods, and research strands of empirical research on MOOCs.	2013–2015
(Bozkurt, Akgün-Özbek and Zawacki-Richter, 2017)	2017	Explored the trends and patterns in research on MOOCs.	2008–2015
(Moreno-Marcos *et al.*, 2018b)	2018	Surveyed prediction in MOOCs via characteristics of the MOOCs used for prediction, prediction outcomes, classifying the prediction features, techniques used to predict the variables and metrics used to evaluate the predictive models.	No initial date–2017
(Sanchez-Gordon and Luján-Mora, 2018)	2018	Investigated research challenges in MOOCs.	2008–2016
(Joksimović *et al.*, 2018)	2018	Surveyed the approaches followed in relation to model learning and assessment in MOOCs and analysed learning-related constructs used in the prediction and measurement of student engagement and learning outcomes.	2012–2015
(Van de Oudeweetering and Agirdag, 2018)	2018	Investigated the accelerators of social mobility.	2013–2015
(Paton, Fluck and Scanlan, 2018)	2018	Evaluated engagement and retention in vocational education and training (VET) in MOOCs and online courses.	2013–2017
(Zhu, Sari and Lee, 2018)	2018	Studied publication outlets, research methods, and topics of empirical MOOCs.	October 2014–November 2016
(Wong *et al.*, 2019)	2019	Studied methods to support self-regulated learning. Also, the effect of human factors is examined.	2006–2016
(Lee, Watson and Watson, 2019)	2019	Reviewed research on self-regulated learning in MOOCs.	2008–2016
(Zhu, Sari and Lee, 2020)	2020	Studied the research methods, topics, and trends of empirical MOOC research.	2009–2019
(Palacios Hidalgo, Huertas Abril and Gómez Parra, 2020)	2020	Surveyed MOOCs' origins and definition, their typologies and platforms, strengths and limitations, the concept of specialisation courses, and their didactic applications for foreign language learning.	2012–2019
(Jarnac de Freitas and Mira da Silva, 2020)	2020	Explored gamification in MOOCs.	2014–July 2019
(Lambert, 2020)	2020	Investigated how MOOCs contribute to student equity and social inclusion.	2014–2018
(Meet and Kala, 2021)	2021	Surveyed the trends and future prospects of MOOC research.	2013–2020

(Mehrabi, Safarpour and Keshtkar, 2022)	2022	Determined the global MOOC dropout rate and the variables influencing this frequency.	2000–2021
(Sallam, Martín-Monje and Li, 2022)	2022	Explored the current published research on Language MOOCs (LMOOCs), outlining the types of papers, countries where studies were performed, and institutions devoted to this field.	2012–2018
(Najmani *et al.*, 2022; Sallam, Martín-Monje and Li, 2022)	2022	Reviewed MOOCs recommender systems.	2012–2022
(Badali *et al.*, 2022)	2022	Determined the roles of motivating factors and theories that affect participant retention in MOOCs.	2015–2020
MOOCs + discussion forums			
(Almatrafi and Johri, 2018)	2018	Descriptive analysis and content analysis of discussion forums in MOOCs.	2013–2017

However, to the best of the author's knowledge, synthesising previous studies on instructor intervention — with a major focus on discussion forum-based works — has not yet been carried out. Since the emergence of MOOCs and their unprecedented proliferation over the past decade, different studies have dealt with estimating (or at least providing an insight into) the need for instructor intervention in MOOCs, and consequently deserve a separate survey to explore and synthesise these works. Thus, the present SLR contributes to the current literature by providing a promising synthesis of extant state-of-the-art studies on MOOC discussion forum-based instructor intervention by assessing the methodologies of the surveyed works from the data pre-processing stage to the performance metrics reported and highlighting some research opportunities and directions.

2.3.4.3. Survey Methodology

This SLR primarily covers the previous works on instructor intervention in MOOC discussion forums published since the emergence of MOOCs in 2011 (Ng and Widom, 2014) up until the end of 2022. Additionally, the inclusion and exclusion criteria were defined (see Section 2.3.4.3.2), describing the standards upon which the decision for including a given study was made. The inclusion criteria are intended to be as inclusive as possible while conducting the present survey to investigate the included works while keeping in mind the need to exclude any irrelevant prior work that does not meet the inclusion criteria.

2.3.4.3.1. Surveyed Resources

Two frequently used databases (Scopus[7] and Web of Science (WoS)[8]) were adopted as they are the two most comprehensive abstract and citation bibliographic databases of peer-reviewed scientific journals and conference proceedings. These databases contain over three billion cited references combined (Zhu, Sari and Lee, 2020; Pranckutė, 2021), including major publishers in the area of educational technology and e-learning. This includes the Association for Computing Machinery (ACM), Taylor & Francis Group, ELSEVIER, the Institute of Electrical and Electronic Engineers (IEEE), Xplore, ERIC, and Springer. Many typical venues for MOOC instructor intervention are also indexed within the above two databases including the British Journal of Educational Technology (BJET), the Journal of Learning Analytics (JLA), the Journal of Educational Data Mining (JEDM), the International Journal of Artificial Intelligence in Education (IJAIED), the International Conference on Learning at Scale (L@S), the International Conference on Learning Analytics and Knowledge (LAK), the International Conference on Artificial Intelligence in Education (AIED), and the International Conference on Educational Data Mining (EDM).

2.3.4.3.2. Eligible Studies: Inclusion and Exclusion Criteria

This survey includes works that meet certain requirements, such as: (i) being authored in English only, (ii) being peer-reviewed to ensure research rigour, and (iii) providing a sufficient elaboration on the methodology followed. The latter includes explaining the data used, the feature engineering approach followed, the learning algorithms adopted, and the results achieved. The present study disregards non-peer-reviewed types of publications, e.g., book chapters, magazines, and pre-print works.

The keywords and the Boolean operators that were used to search for the surveyed studies are as follows: (*massive* AND *open* AND *online* AND *course** OR *mooc**) AND (*interven** OR *urgen**) per appearance within the titles, abstracts, or keywords. Since the searched databases (Scopus and WoS) are case-insensitive, terms such as '*MOOCs'* and '*moocs',* were treated alike. Wildcards like the asterisk (*) were used after the root forms of the search terms to include any possible forms of the search terms. Additionally, parentheses were used to prioritise the order of precedence and search executions accordingly. The functionality of these

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

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

wildcards is typically standard within the databases used for retrieving surveyed works in the present review *(ACM Advanced Search; Springer Link Search Tips; IEEE Explore Search Tips; Web of Science Core Collection: Search Tips; Scpus: Tips and Tricks).*

The initial search retrieved 673 studies (352 from WoS and 321 from Scopus). However, removing duplicates (223), non-English-authored (16) and non-peer-reviewed (20) studies resulted in a total of 414 abstracts for screening.

2.3.4.3.3. Screening Process

During the abstract screening process, three PhD-holding independent annotators, all with previous research experience, labelled the shortlisted abstracts (as included or excluded) based on the following criteria:

2.3.4.3.3.1. Inclusion Criteria

- Written in English.
- Appears in a peer-reviewed journal article/conference paper proceedings to guarantee the highest levels of rigour.
- Focuses on research related to identifying at least one of: posts/comments, learners (dropout), topics that need instructor intervention or urgent intervention based on learners' text inputs in MOOC discussion forums.
- In terms of identifying urgent posts, it can be a standalone task or one of a set of different tasks.

2.3.4.3.3.2. Exclusion Criteria

- Research on irrelevant topics that do not meet the inclusion criteria.
- Some types of publications, including books, book chapters, magazines, and pre-print works.

These criteria were considered when the studies were retrieved from the databases and then were checked by the reviewers. Figure 2.15 (below) illustrates the outcomes of the screening conducted by each annotator along with the number of sessions and the time taken using the Rayyan[9] platform, which is an interactive AI-based website for abstract screening in a blind-

[9] https://rayyan.ai

reviewing environment (Ouzzani et al., 2016). It is a free online tool for academics to help them perform systematic reviews by significantly speeding up the selection and screening of research.

The three reviewers worked independently (blind) to reduce bias; each abstract was triple-evaluated and checked. After finishing the screening process of the 414 articles by each annotator, the blinding status was changed to blind-off which allows reviewers to see the decisions of the other reviewers. Next, the annotators discussed any conflicts using Rayyan's reviewing chat; this resulted in more agreement between annotators on some conflicting studies.

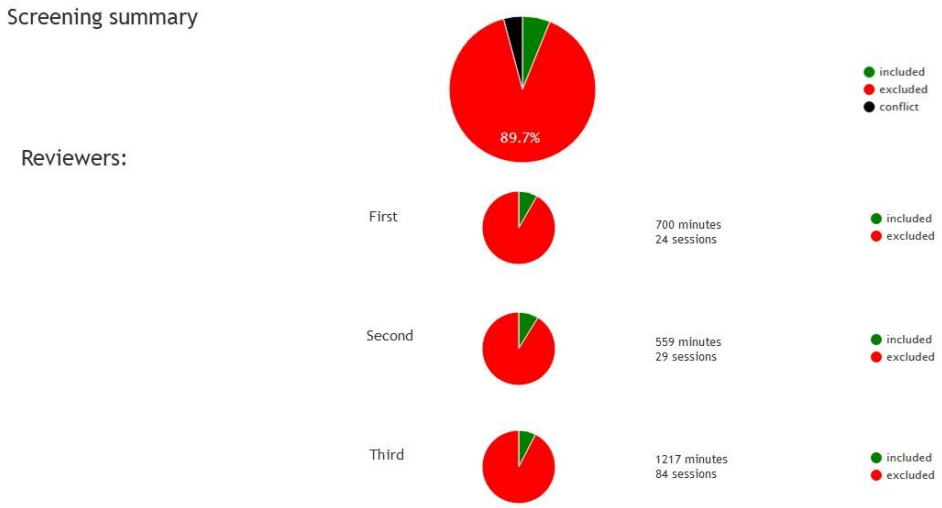

Figure 2.15: Summary of study screening conducted by the three annotators.

This step resulted in the exclusion of 372 abstracts and the inclusion of 25 abstracts unanimously based on the annotators' agreement, whereas 17 abstracts were further revised due to conflict between annotators.

This review was guided by the stepwise PRISMA[10] framework (Moher *et al.*, 2009), the most frequently used, prominent, and most-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). A PRISMA flowchart illustrating the sequential process of applying inclusion and exclusion criteria of each stage was used to produce a final number of studies for systematic review analysis which illustrates the exploration and screening of potentially suitable research studies (Harris *et al.*, 2014). It contains four successive phases (identification, screening, eligibility, included) to increase the transparency

[10] http://www.prisma-statement.org

and quality of the systematic review's reporting (Liberati *et al.*, 2009). Figure 2.16 (below) shows the protocol's four phases along with the outcomes of each stage.

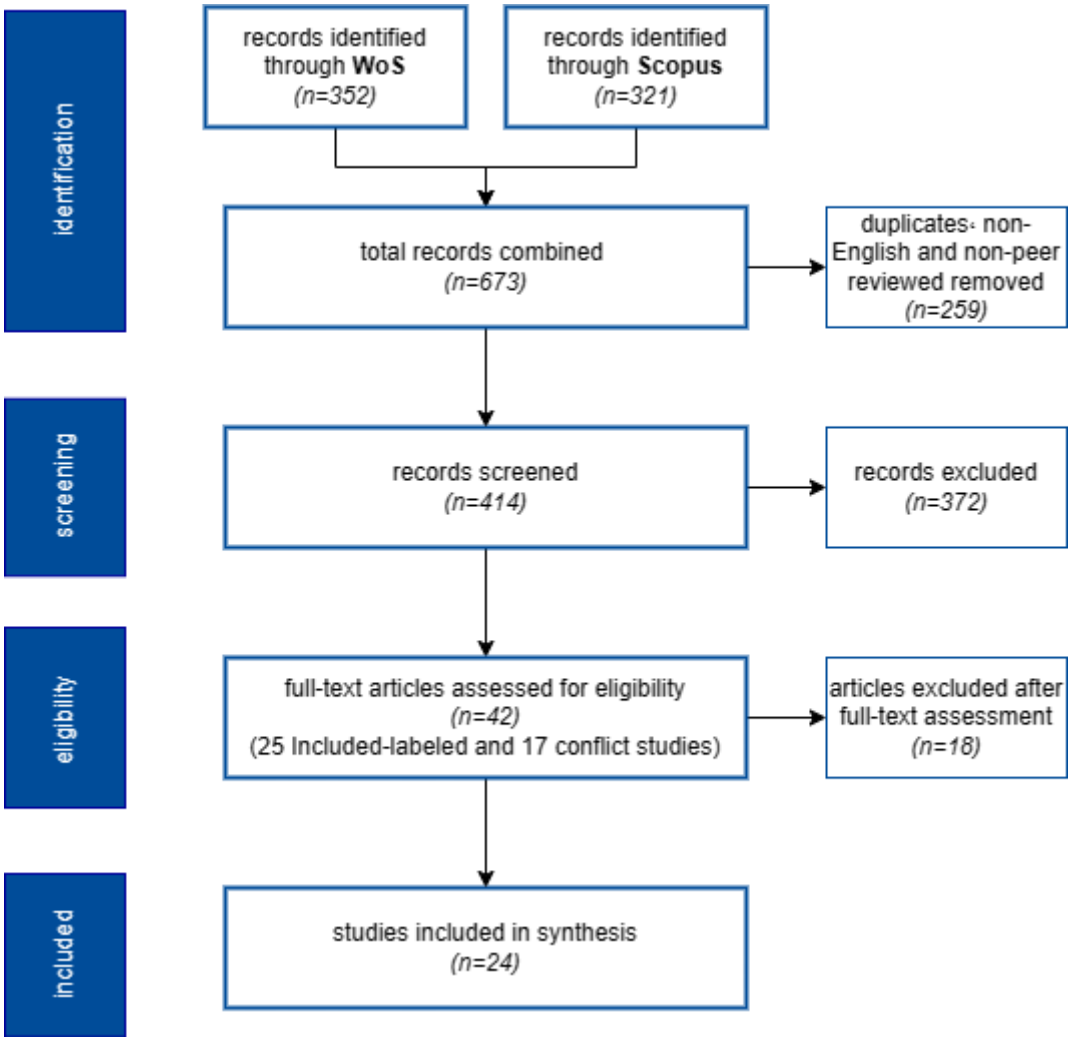

Figure 2.16: PRISMA flowchart diagram.

2.3.4.3.4. Excluded Studies

A further full-manuscript reading through of the conflicting studies was conducted for definite inclusion (or exclusion) of these studies. Out of the total number of 17 conflicting studies, 15 were excluded whereas two studies were included, rendering the final total number of studies selected for inclusion (n = 27). Then, from these included studies, three studies were excluded after reading the full articles because they were not related to identifying instructor intervention need. Thus, the final number of included studies for synthesis was (n = 24).

There were various reasons why some studies were excluded. This included not being based on discussion forums (Borrella, Caballero-Caballero and Ponce-Cueto, 2022; Sciarrone and Temperini, 2019; Sun *et al.*, 2021; Meier *et al.*, 2015; Kurtz *et al.*, 2022; He *et al.*, 2015;

Haniya, 2019; Itani, Brisson and Garlatti, 2018; Yang, 2022; Amarasinghe and Hernandez-Leo, 2019). Other reasons included investigating how to change learner behaviours rather than tackling intervention need when learners faced any questions (Schutzberg, 2019). This study used 'tough love' to guide learners to engage with help resources such as forums and learn how they seek answers to their questions independently. Another study (Alshehri and Cristea, 2022) predicted learners' certificate purchasing decisions using data from FutureLearn MOOC discussion forums and therefore was excluded. Another article (Sinha T, 2015) tracked learner interaction over time through video viewing and navigational and forum posting to predict the series of grades achieved by a learner in different MOOCs by using the probabilistic framework of conditional random fields (CRF). Regarding learning interaction and the development of social knowledge in MOOCs, (Chen and Yeh, 2021) used three types of role-assignment strategies to study knowledge construction and interaction patterns in asynchronous MOOC discussion forums. Another study (Koné *et al.*, 2020) offered a novel methodology to compute a collective activity indicator to address the issue of identifying and displaying the collective dynamics resulting from the interactions in MOOC forums. This would assist instructors to intervene in the course structure or course design, but it does not help them to intervene on posts or learners. Thus, based on all these reasons, all the above studies were excluded.

The three included studies by the same authors (Ntourmas *et al.*, 2018; Ntourmas *et al.*, 2019; Ntourmas *et al.*, 2022) do not intend to propose an approach to identify interventions; rather, they focus on instructor intervention from other perspectives. In the first study (Ntourmas *et al.*, 2018), the goal was to examine the characteristics of teaching assistant interventions and compare two MOOCs on different subjects (technology and humanities). The study aims to provide crucial information on the behaviour of teaching assistants (TAs) in online discussion forums and inspire the creation of efficient support and automatic responses to learners in the future. They used a Greek MOOCs platform (mathesis.cup.gr) which is based on OpenEdX. The findings showed some quite variations in the language employed, message lengths, response times, and discourse lengths between the two courses' TA interventions. In the second study (Ntourmas *et al.*, 2019), the authors evaluated a discussion forum design and found some issues. A mixed-methods study was conducted on two MOOCs offered via the OpenEdX platform. The findings of this study show that there are several usability problems with the OpenEdX forum design that negatively impact learners' support and therefore the course designers need to take this into account. Also, the study found that intervention is correlated with the number of participants as they revealed that instructors evolved more

complex tactics to support learners when the number of learners in the forum increased. In the last and very recent study (Ntourmas *et al.*, 2022), instructional strategies used by instructors (TAs) in two MOOCs covering various subject areas were investigated using a mixed-methods approach. The study aimed to assess the pedagogies used by TAs to support learners using a widely used methodology for evaluating the instructional quality of MOOCs. The TAs' intervention strategies were explored through content analysis and interviews, and the results were enhanced by linguistic and social network analysis. The findings of this study indicate some significant limitations in instructional design and emphasise the importance of making learner facilitation a central component of MOOCs' instructional design. The findings also highlight the requirement for guidelines which TAs should follow to take the most proper intervention decisions. The creation of tools is also discussed as being necessary for MOOC instructors to assist them in improving and maintaining the instructional design of their courses.

2.3.4.4. Instructor Intervention in MOOCs

The surveyed research concerning instructor intervention in MOOCs was categorised based on three main axes: identifying (i) posts, (ii) learners (dropouts), and (iii) topics. Then further sub-categorisation was performed on post-based studies based on the platform from which the studies were sourced as shown in Figure 2.17 (below). The surveyed studies were published between 2014–2022 and are chronologically ordered in the figure below.

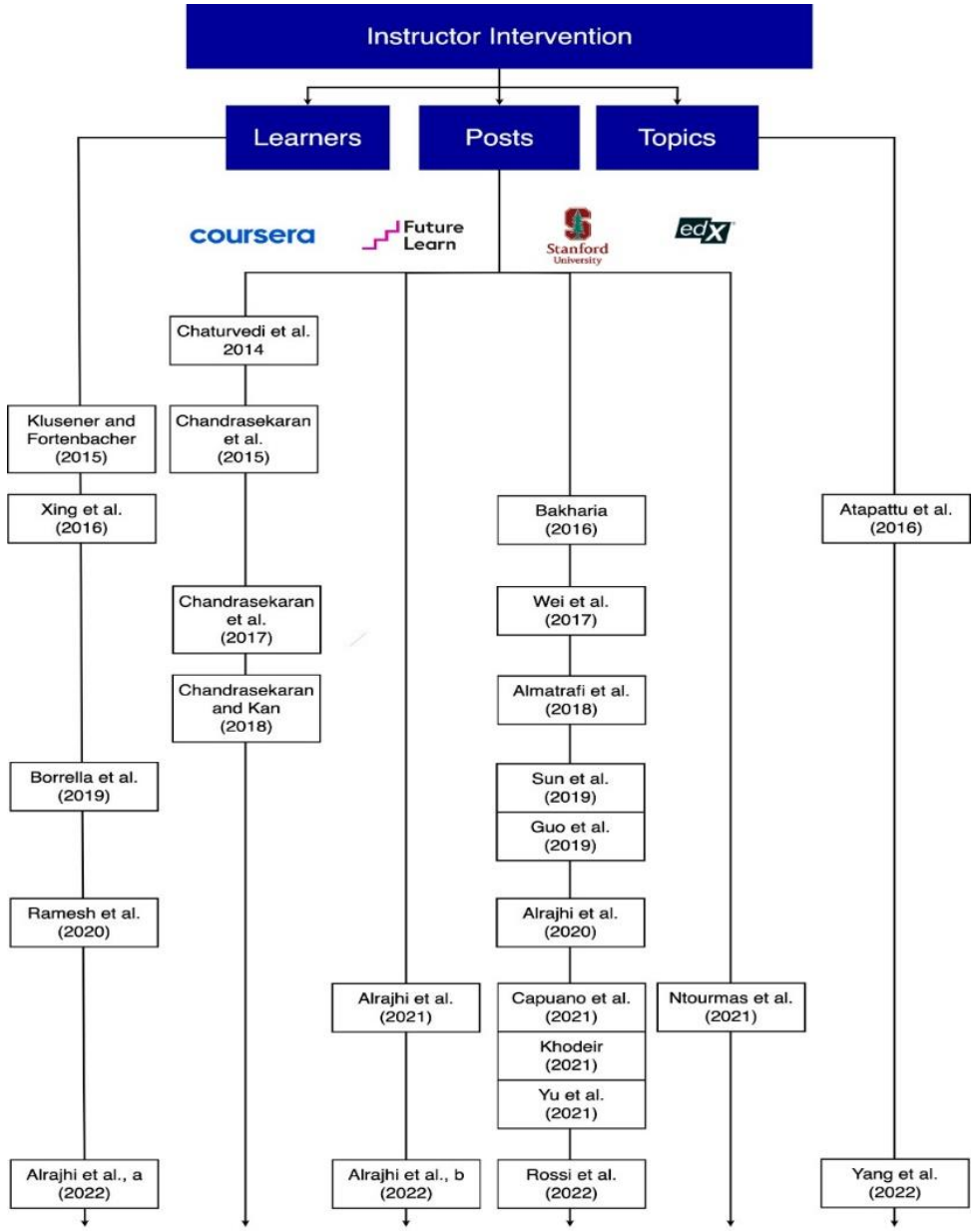

Figure 2.17: Categories of the surveyed studies.

2.3.4.4.1. Post-Based Identification

Most research on instructor intervention in MOOCs is based on identifying the need for intervention via posts, which account for about 70% of the surveyed studies. It is noted that the development of studies and their association with each other is within the same platforms and databases. Thus, it further categorises research into different MOOC platforms from which the datasets are sourced. In this section, all the included studies that used ML to predict posts that need instructor intervention were reviewed and clarified based on the following: MOOC platforms, intervention labels, classification models, performance evaluation and data splitting mechanisms, as described below.

2.3.4.4.1.1. Coursera

Various studies used datasets which are collected and offered by Coursera; the largest MOOC platform (Wu, 2021). This includes (Chaturvedi, Goldwasser and Daumé III, 2014) who were among the first researchers to present the problem of prediction for instructor intervention in the MOOC forum environment using course information, forum structure, and post content from two Coursera MOOCs from various fields (science and humanities). Their models use latent categories to abstract the contents of individual posts into threads. The problem to address was a binary prediction task based on instructors' intervention histories; they labelled data automatically, as follows: positive (1) if instructors had posted replies; negative (0) otherwise. The three employed models (logistic regression, linear chain Markov model, and global chain model) to determine whether or not an instructor would intervene in threads or posts. They reported the precision, recall and F-measure of the positive class to measure the performance of their models. They split and evaluated data using ten-fold cross-validation. They concluded that it is important to use a thread structure in predicting instructor intervention behaviour. However, they did not directly use the posts (textual inputs) for training their predictive model.

Several research studies on instructor intervention in MOOC discussion forums have been authored by one research team. (Chandrasekaran *et al.*, 2015a) proposed a taxonomy of pedagogical feasible instructor interventions for automated assistance on when and how to intervene in discussion forums that would maximally benefit learners' studies. They used 61 courses from Coursera encompassing different academic fields including sciences, humanities, and engineering. Their label is a type of intervention that promotes learner learning by annotating the contents of discussion forums. To predict intervention, they used CRF which is a probabilistic model. As the research is still in progress, they will measure learner performance in terms of quantitative and qualitative measures.

The next research study (Chandrasekaran *et al.*, 2017) investigated discourse relations and used Penn Discourse Treebank (PDTB) based features to predict the need for instructor intervention. Data from 14 Coursera MOOCs across 7 courses were used as the corpus for this study. They set labels wherein intervened threads are considered as positive and non-intervened threads as negative. They used three different models: (i) a maximum entropy classifier as a baseline with a set of features; (ii) only PDTB discourse relations as features; (iii) baseline + PDTB. They discussed the models' performance in terms of recall, precision, and F1 for the

positive class using two assessment schemes of in-domain and out-of-domain schemes. They show that using PDTB relation-based features resulted in better classifier performance compared to the baseline model.

For position bias in the intervention setting, (Chandrasekaran and Kan, 2018) used a corpus that includes discussion forum threads from 14 MOOCs on Coursera. The dataset used covers a range of subject areas and courses offered by different universities across the world and taught by instructor teams of diverse sizes. In this study, they showed that there *is* strong position bias in instructor intervention based on where the thread appeared on the forum at the time the intervention occurred as they are ordered by their *last update time*. When bias from the training data was removed, the performance of the de-biased intervention classifier improved.

2.3.4.4.1.2. Stanford

In 2015, the Stanford MOOCPosts dataset was made available for researchers upon request which was manually labelled by nine consultants (three per field) and published by (Agrawal *et al.*, 2015). The dataset includes about 29,604 learners' posts in total, gathered from 11 online courses offered by Stanford University in the humanities/sciences, medicine, and education fields. Six categories, including confusion, sentiment, urgency, question, answer, and opinion were used to classify each post on these courses. Urgency was defined as: How urgent is it that this post should be seen by an instructor?; this dimension was used to describe instructor intervention in this survey. The following studies used this dataset to identify urgency, either as one of a set of different tasks or as a standalone task.

The review began with studies that detected urgency within different tasks. In terms of the transfer learning model and cross-domain MOOC forum post classification, (Bakharia, 2016) was the first to consider this matter, conducting preliminary research on cross-domain classification. This was achieved by training different classifiers to classify forum posts into three different categories, specifically, *confusion, urgency, and sentiment*. Next, validation of the classification was performed via different unseen domain areas. The author constructed labels by transforming a 7-point scale of three different categories to a binary (Yes/No) classification, with values greater than 4 signifying the presence of a category. The classification algorithms adopted were naive Bayes, SVM (using different kernels: radial basis function (RBF) and linear), AdaBoost, and random forest. The findings indicate low cross-domain classification accuracy; nevertheless, the author stated that transfer learning should be given more consideration in the context of education.

In a follow-up study, along with transfer learning, (Wei *et al.*, 2017) proposed a model for classification and transfer learning approaches of cross-domain MOOC forum posts based on DNNs. The proposed model combined a CNN and LSTM (called ConvL) to identify three tasks: *confusion, urgency,* and *sentiment* in posts using textual data only. They fed the feature representation for each word as the local contextual feature via CNN; from these features, the posts' representations as semantic relationships of features were learnt via LSTM. They framed the label as posts with a score greater than 4 were positive and used accuracy as a performance metric.

(Capuano *et al.*, 2021) proposed a multi-attribute text categorisation tool based on attention hierarchical recurrent neural networks for word encoding and attention for word aggregation at different levels (sentence and document). Various attributes were used in text categorisation tool — among them *urgency*. To classify urgency, three levels were used (low, medium, and high) where scores below 3 were mapped to the negative/low class, scores above 5 were transferred to the positive/high class, and the remaining scores were mapped to the neutral/medium class. They reported the average results as follows: precision (%) recall (%) F-score (%) in addition to loss. Four-fold cross-validation was then carried out to evaluate model performance.

In terms of proposing an automatic solution rather than real instructor intervention, (Rossi *et al.*, 2022) suggested an architecture known as CAERS which is a combination of a conversational agent and an educational recommendation system. The conversational agent provides intervention to support learners and instructors to develop their knowledge through autonomous interference and resource recommendations. Posts are classified into three categories (*question, answer*, and *opinion*) and three parameters (*sentiment, confusion,* and *urgency*). However, they used confusion to train the predictive model.

The following studies deal with intervention as a standalone task; they used the *urgency* dimension to represent intervention: (Almatrafi, Johri and Rangwala, 2018) built a generalised model to identify reliably urgent posts regardless of the content of the course by implementing different linguistic features and metadata as features to train different traditional ML models. They inspected various feature sets (an NLP tool which features *LIWC* and three metadata which are *up_count, reads, post_type,* and *TF*) with different classification models (naive Bayes, SVM, random forests, AdaBoost (decision trees as base estimators) and logistic regression). They defined the labels as follows: posts with a score of 4 or higher were *urgent*;

otherwise, the post was *not* considered *urgent*. They used weighted F1 and Cohen's Kappa as evaluation metrics. Also, precision, recall and F1 were reported for each class.

In addition, (Sun *et al.*, 2019) distinguished potentially important urgent posts by presenting a DL model as an improved recurrent convolutional neural network (RCNN) method to obtain contextual information. They specified that posts with a score of 4 or more are deemed *urgent*, whilst posts with a value of less than 4 are regarded as *non-urgent*. Their model achieved higher performance in identifying urgent intervention-needed posts compared to other models (naive Bayes, SVM (RBF), random forest, CNN, RNN, LSTM, GRU, and RCNN). The evaluation metrics used include accuracy, precision, recall and macro F-score (F1).

(Guo *et al.*, 2019) is another study that used a combination of DL models (CNN + GRU) to extract semantic and structural information to detect posts that needed urgent responses. This was performed by applying *attention* to develop a hybrid character/word neural network. The Char-CNN was proposed to capture noise information. The course information associated with a given post was proposed for contextualisation. Posts with a score higher than 4 on the *urgency* dimension were considered urgent in this study. They calculated the weighted F-score to assess the performance of their model. Also, they reported precision, recall and F-score (F1) on both the *urgent* and *non-urgent* classes.

Another study by (Alrajhi, Alharbi and Cristea, 2020) focused on predicting urgent posts. They found significant correlations between different dimensions (sentiment scale, confusion scale, opinion value, question value, and answer value) and the need for urgent intervention. Thus, they constructed a multidimensional DL intervention model that combines different dimensions as numerical features with text. They trained the text data (learners' posts) with a CNN model and the numerical data (multiple dimensions) with a MLP model. They defined the urgent label when post scores were > 4 (required urgent intervention). They used the average accuracy, precision, recall and F1-score (F1) per class (0 as non-urgent; 1 as urgent) to measure performance. The results demonstrated that the combined, multidimensional features model outperforms a text-only model.

(Khodeir, 2021) developed a multi-layer Bi-GRU based on BERT as a pre-trained embedding layer to classify learners' urgent/non-urgent posts. The author used BERT for word embedding to represent words in their context considering that urgent posts scored 4 or above; otherwise, the post was deemed non-urgent. The performance metrics used were F1-weighted

and precision verse recall (PR) curves as metrics for model evaluation. In addition, precision, recall and F1-score (F1) for each class were also reported.

(Yu *et al.*, 2021) adopted Bayesian deep learning for the first time in MOOC forums to identify the need for urgent instructor intervention using two techniques: (i) Monte Carlo Dropout, and (ii) Variational Inference, as a novel approach to determine whether a learner's post requires instructor interventions. A threshold of 4 was used to categorise the need for intervention into two categories: (i) need for urgent intervention (value > 4) with label 1, and (ii) no need for intervention (values <= 4) with label 0. They applied RNNs with attention mechanisms as a baseline model. Following that, they presented two methods for using Bayesian DL with this baseline model: Monte Carlo Dropout and Variational Inference. They provided mean accuracy, F1 score, precision score, recall score under each class, and entropy based on the prediction layer. They ran two different sets of experiments. In the first, they divided the data into training and testing sets with a ratio of 80% and 20%, respectively, using stratified sampling. In the second experiment, they used a split of 40% and 60% for training and testing, respectively. The findings suggest that Bayesian deep learning provided a critical uncertainty measure that could not be obtained by traditional neural networks.

2.3.4.4.1.3. FutureLearn

The research below from the same authors used data from a FutureLearn platform which is manually labelled by domain experts, following Agrawal et al.'s (Agrawal and Paepcke, 2019) instructions.

(Alrajhi *et al.*, 2021) proposed a new automated intervention priority model for MOOCs based on learner histories in terms of *urgency*, *sentiment analysis*, and *step access*. They classified posts with a score of 4 and above as *urgent (*1) and *non-urgent* (0) otherwise. Their model contains two phases ((i) prediction, and (ii) intervention priority). In the prediction phase, they used BERT to classify urgent posts; in the intervention priority phase, they suggested a priority of intervention to help high-risk learners first. They reported results from the predictive phase model based on average accuracy over the two classes, recall, precision and F1-score for the minority (urgent class).

Next, a study by (Alrajhi *et al.*, 2022) used an EXplainable artificial intelligence (XAI) approach to develop an urgent instructor intervention model that can interpret the model outputs and potentially assist MOOC instructors to provide effective intervention. They

demonstrated how combining a predictor with the findings of XAI, particularly colour-coded visualisation, can be utilised to assist instructors in deciding on intervention as they used text classification explainability. The label is the same as in their previous work: a binary scale (1:3 to 0 and 4:7 to 1). Using BERT, they developed an automatic urgent intervention model. Then the Captum package was used to interpret the BERT model outputs (average accuracy, precision, recall, and F1-score) for each class which were employed to evaluate the classifier's performance. Using the *stratify* method, they divided the data into training and testing sets (80% and 20%, respectively). Then, they divided the training set once more, with 90% allocated to training and 10% to validation.

2.3.4.4.1.4. EdX

Within the surveyed studies, one research project employed data from a major Greek MOOC platform on Mathesis (mathesis.cup.gr) which is based on EdX (OpenEdX).

(Ntourmas *et al.*, 2021) proposed and evaluated an alternative method for developing two classifier forum posts models that identify the need for intervention by utilising the semantic similarity of the forum transcripts with training features from MOOC corpora. They centred their attention on the feasibility of transferring such support between two MOOCs in distinct academic fields (humanities and technology). The manual labelling of the starting posts for both courses was carried out by two coders with the guidelines for three categorisations ((i) problem related to the course material, (ii) problem related to the course logistics, and (iii) no action required: discussion related to community building). A support vector classifier (SVC) was the classification algorithm employed. They performed predictions on the beginning posts of the second course using the classifier (SVC model) from the first course, and vice versa. They used accuracy, precision, recall and F1 score for evaluation metrics. Using stratified sampling, the dataset was divided into 75% for training and 25% for testing.

2.3.4.4.2. Learner-Based Identification

Some studies on instructor intervention in MOOCs sought to identify learners at risk of dropping out. These learners are detected based directly on their posts on discussion forums or as a feature in addition to other features as follows:

The first study by (Klusener and Fortenbacher, 2015) predicted learner success based on three MOOC forum activities wherein unsuccessful learners are considered to be at-risk

learners. The features of successful learners were obtained from forum activities and blended into a learning profile based on Iversity's MOOC. Feedback could be produced for learners who are labelled as at-risk learners based on their learning profile. They established an analytics tool based on ML which can classify learners using different features such as the number of answers in a forum or the number of up-votes. At least 80% of all video lessons for a course must be watched by a learner for the learner to be considered successful. The algorithms used for identifying successful learners include decision trees, random forest, decision rules, step regression, and logistic regression. They used ten-fold cross-validation with accuracy and recall of both classes (unsuccessful and successful learners) to evaluate the model's performance.

(Xing *et al.*, 2016) identified learners who are at risk of dropping out by proposing a temporal modelling approach to predict learner dropout behaviour. Then, the historical features were appended, which outperform the simple temporal features. The data used in this study were collected from a specific course on the Canvas platform. The dropout label for each learner is determined by looking at whether they would be active in the coming week and any of the following weeks. As a result, the labels are generated thus: 0 denotes dropout and 1 denotes active. The general Bayesian network (GBN) and decision tree (C4.5) were the only two algorithms used in this study. The performance was calculated based on ten-fold cross-validation using area under the curve (AUC) and precision as performance measures.

(Borrella, Caballero-Caballero and Ponce-Cueto, 2019) developed a method for identifying learners who are at risk of dropping out of a course by targeting learners who skipped the midterm or final exam; they created and tested an intervention aimed at reducing that risk. To predict dropout, they used different clickstream features such as *clicks in the forums.* Data from the MITx MicroMasters MOOC were used along with random forest and logistic regression to create the predictive models used in their experiment. They tested these algorithms based on the recall values. In addition, they reported precision value. This study provided recommendations for MOOC designers and instructors on how to increase completion rates and improve learner motivation and engagement.

(Ramesh *et al.*, 2020) developed an interpretable statistical relational learning model that can understand learner participation in online courses using a combination of behavioural, linguistic, structural, and temporal features. The data used for building the model was gathered from seven Coursera courses. Various traditional ML models were trained including SVM, logistic regression, multi-layer perceptron, linear regression, and decision trees. They evaluated

their model using various metrics including area under the precision-recall curve for positive and negative labels and area under the ROC curve using ten-fold cross-validation.

(Alrajhi, Alamri and Cristea, 2022) predicted the need for intervention based on temporal history by combining the sequence of posts written by learners using data from FutureLearn. Other DL and Transformer techniques were adopted to train the model. They followed the approach of (Alamri *et al.*, 2021) in their definition of dropout: learners are likely to drop out if they did not access 80% of the topics in the following week. The metrics they used are accuracy and precision, recall, and F1-score (F1) for each class to evaluate the performance (in percentages) of the various models. The data were divided into training and testing (80% and 20%, respectively). The training data was then divided into training data (80%) and validation data (20%), respectively.

2.3.4.4.3. Topic-Based Identification

Extracting topics related to intervention may improve the intervention process. This section reviews two studies that are concerned with identifying topics that need intervention.

(Atapattu T, 2016) used three Coursera MOOCs featuring different disciplines (machine learning, statistics, and psychology) to classify, analyse, and visualise topics from MOOC discussion forums. They used LDA to detect topic clusters. By linking the topics with the relevant weekly lectures as a graph of connections between topics and threads, this study made it easier for instructors to identify and navigate the most significant topic clusters and discussions.

(Yang, Ren and Wu, 2022) proposed a new method that identifies topic attention based on the TEAM model that is combined with data characteristics regarding the behaviour and content present on a MOOC discussion forum. The dataset was gathered from a Chinese MOOC entitled Microcourse Design and Production.

2.3.4.5. Synthesis of the Surveyed Works

One of the most crucial aspects of conducting a SLR is data synthesis (Kraus, Breier and Dasí-Rodríguez, 2020). An analysis of the surveyed instructor intervention in MOOCs based on discussion forum models is presented in this section. For better categorisation and a synthesised analysis of the surveyed studies, the data sources (platforms and numbers of courses, threads, and posts) and adopted methodologies (data labelling ‘ground truth’, prediction models and

algorithms, training and testing splitting techniques, and performance metrics) were reported. In addition, the use of XAI is discussed where applicable.

2.3.4.5.1. Data Sources

2.3.4.5.1.1. Platforms

The sources of the data used in the existing intervention models can be better understood by looking at the platforms that were used to develop them. According to the analysis, the Stanford platform was the most prominent with 41.6% (n = 10) of the surveyed works using it (see Figure 2.18 below). This trend is likely because the Stanford MOOCPosts dataset is available for researchers upon request and is a valuable resource as it contains different courses (11 courses) from three different domains with manually labelled features. Regarding other platforms, Coursera comes second with six where in most of the studies the labelling is based on the actual intervention offered (as discussed later in Section 2.3.4.5.2.1). Following that, FutureLearn featured in three studies by the same authors. The remaining (less represented) platforms with only one study are OpenEdx, Iversity, MITx MicroMasters, Chinese University, and Canvas.

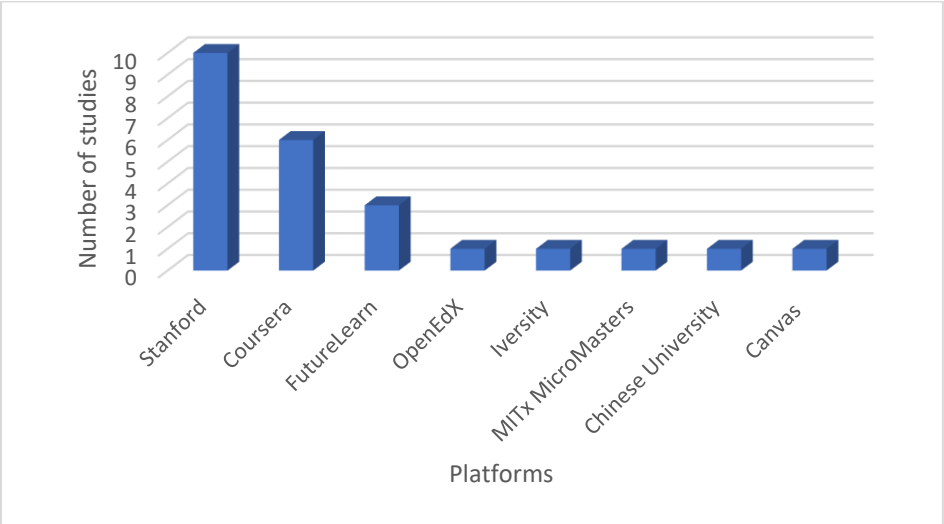

Figure 2.18: Number of surveyed studies across platforms.

2.3.4.5.1.2. Numbers of Courses, Threads and Posts

The datasets that were used within the surveyed studies included different numbers of courses varying from one course only up to 61 courses. The number of threads ranged from 1780–33665 whereas the number of posts ranged from 1977–29604 (as shown in Figure 2.19, Figure 2.20, and Figure 2.21, respectively). Please note that the above statistics were not mentioned

by all the surveyed studies. The interpretation of this large variation is associated with how the intervention is defined. For example, in post-based identification, while some studies predict intervention need from posts, they do not use extensive human effort for post labelling since they consider the real intervention by an instructor as the label. Also, in some studies, semi-supervised learning was employed to annotate the unlabelled data to enable the easier compilation of many courses and threads. One example of a study that followed this approach is (Chandrasekaran *et al.*, 2015a), which will be further discussed in Section 2.3.4.5.2.1.

In the Stanford dataset which contains 29604 posts from 11 courses, each researcher carried out some specific pre-processing which led to the exclusion of very few posts; in this review, however, it will consider that all studies were carried out under the same number (29604) as represented in the following figures for the Stanford platform (all studies). In contrast, some researchers used subsets of these data (e.g., (Bakharia, 2016; Wei et al., 2017; Sun et al., 2019)) as they selected for evaluation the three courses with the most forum posts from each domain area. Therefore, Figures 2.19, 2.20, and 2.21 (below) clarify these studies as Stanford (Bakharia, 2016; Wei et al., 2017; Sun et al., 2019).

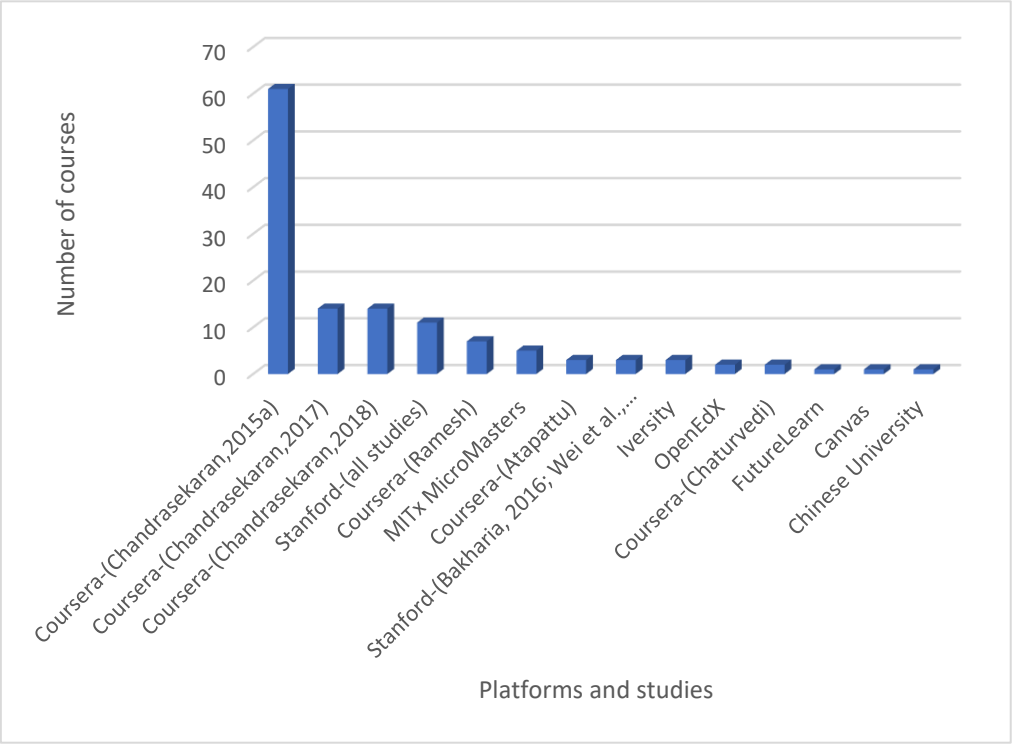

Figure 2.19: Number of courses across platforms and studies.

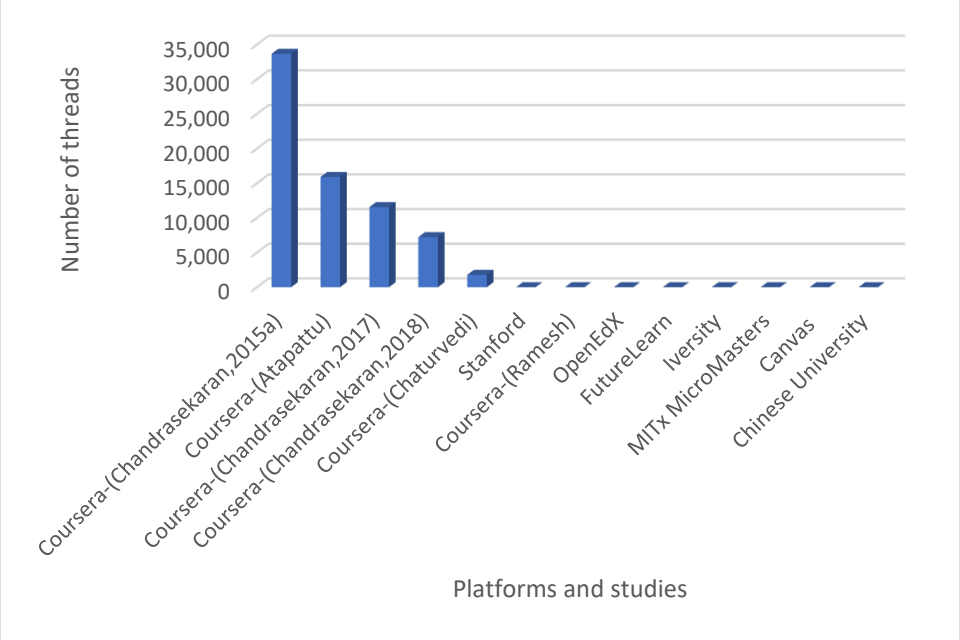

Figure 2.20: Number of threads across platforms and studies.

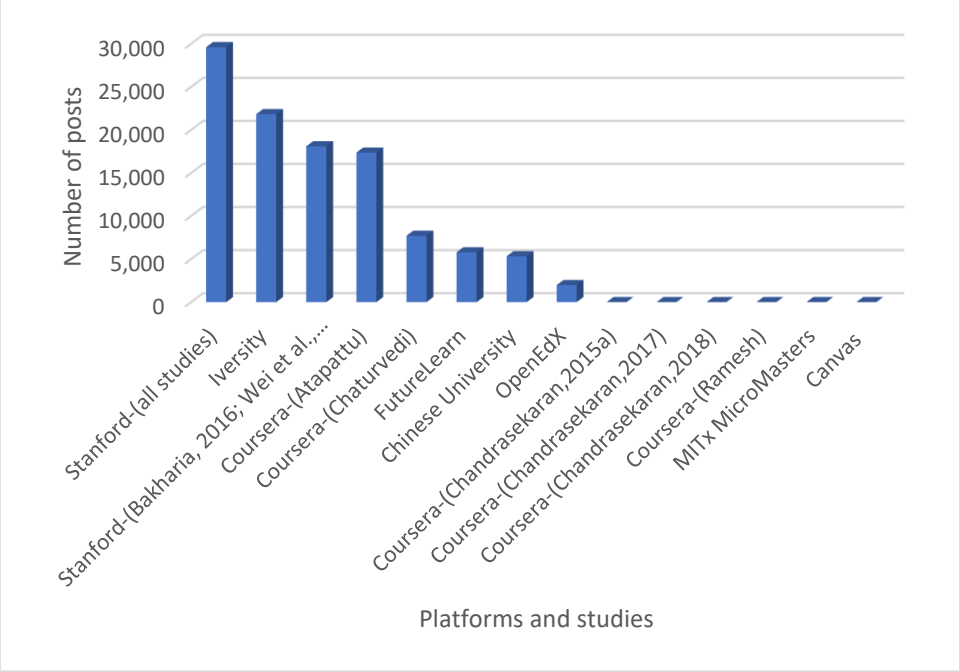

Figure 2.21: Number of posts across platforms and studies.

2.3.4.5.2. Adopted Methodologies

2.3.4.5.2.1. Data Labelling (Ground Truth)

The topic of instructor intervention in MOOCs is complex; as mentioned before, it involves the consideration of posts, learners, and topics. To identify posts and learners, researchers typically used supervised learning while unsupervised learning was used to identify topics. In

supervised learning, models are constructed with known labels, thus, labelled data is needed to predict posts and learners.

Researchers used two methods to define labels related to intervention decisions to identify posts. The first method was defining posts that need intervention as when instructors have intervened in posts in actual MOOCs, where each thread is labelled as *to be intervened* or *not intervened.* It should be noted that this labelling technique may be inaccurate due to annotators' potential subjectivity. Also, there may have been posts where the instructor needed to intervene but did not (either because they missed it or did not have enough time, etc.). Conversely, there may have been unnecessary interventions because instructors occasionally employ various teaching and intervention techniques. The second method is when the labels are set manually by humans to decide a ground truth as in the Stanford and FutureLearn datasets. Table 2.2 (below) categorises the surveyed studies based on the method of label definition mentioned above.

Table 2.2: Studies taken by each method for labelling data to identify posts.

Method to define label	Platforms	Studies
Real intervention	Coursera	Chaturvedi et al., 2014 Chandrasekaran et al., 2017 Chandrasekaran and Kan, 2018
Human coding	Coursera	Chandrasekaran et al., 2015a
	Stanford	Bakharia, 2016 Wei et al., 2017 Almatrafi et al., 2018 Sun et al., 2019 Guo et al., 2019 Alrajhi et al., 2020 Capuano et al., 2021 Khodeir, 2021 Yu et al., 2021 Rossi et al., 2022
	FututrLearn	Alrajhi et al., 2021 Alrajhi et al., 2022
	OpenEdX	Ntourmas et al., 2021

Table 2.2 (above) illustrates that the four platforms (Coursera, Stanford, Futurelearn, and OpenEdX) that the posts were driven from them, the researchers used human coding to define labels for classification. In Coursera, there is only one research study (Chandrasekaran *et al.*, 2015a) that manually annotated discussion forum via two groups, (1) crowdsourced human annotators (Amazon Mechanical Turk[11]) and also (2) physically via the use of on-site human

[11] Amazon Mechanical Turk (mturk.com)

annotators. As the annotation process is time- and effort-consuming, they used limited human annotation with a seed corpus and used semi-supervised learning to label unlabeled data in order to increase the size of the corpus.

In contrast, all the studies that used the Stanford or Futurelearn platforms followed Agrawal and Paepcke's instructions as clarified on their website (Agrawal and Paepcke, 2019). The urgency for intervention was rated on a Likert scale from 1–7, where 1 indicates that there is no need for the instructor to read the post and 7 indicates that it is extremely urgent to intervene. While posts from the Stanford and Futurelearn platforms were evaluated following the same instructions, they are different in terms of the number of coders (in the Stanford data, urgency evaluation was performed by two coders while in the FutureLearn data scoring was performed by three coders), calculating the urgency scores of the gold-standard datasets; thus different techniques were used to construct gold-standard corpora (for more details see Section 3.2).

In the Stanford dataset, different studies used different methods to classify posts, most of them deal with the problem as a binary task while others used multilabel tasks. In the former, a score of 4 was deemed to be neutral; some researchers set the threshold as >4 as needing urgent intervention and others set the threshold as >=4 as needing urgent intervention. While two studies from the same authors (Alrajhi *et al.*, 2021; Alrajhi *et al.*, 2022) used the FutureLearn dataset problem as a binary task with ground truth for intervention >=4. A summary of all the studies along with their respective techniques of defining labels is presented in Table 2.3 (below).

Table 2.3: Definition of intervention labels in the listed studies.

Type of classification	Categories	Threshold	Platform	Studies
Binary	Urgent and need intervention. Non-urgent and no need for intervention.	Urgent >4	Stanford	Bakharia (2016) Wei et al. (2017) Guo et al. (2019) Alrajhi et al. (2020) Yu et al. (2021)
		Urgent >=4	Stanford	Almatrafi et al. (2018) Sun et al. (2019) Khodeir (2021)
			FutureLearn	Alrajhi et al. (2021) Alrajhi et al. (2022)
Multi-Class	Low Medium High	Urgent <3 Urgent = 3,4,5 Urgent >5	Stanford	Capuano et al. (2021)

Using the OpenEdX dataset, (Ntourmas *et al.*, 2021) addressed a multi-class classification problem. Two coders categorised the starting post based on the following: no action required (NAR), content-related problem (CR) or logistics-related problem (LR). Several factors that may be more appropriate for identifying intervention need are connected to these categories. From these categories, CR signals that the instructor should intervene.

In contrast, research that identified at-risk learners who need intervention found it generally straightforward to set labels by identifying learner dropout. Researchers used different definitions to predict dropout as there is no formal definition of dropping out (Sunar *et al.*, 2016). (Xing *et al.*, 2016) determined the status of learners as *dropouts* (label = 0) or *active* (label = 1) by how active they will be throughout the upcoming week. Another study (Borrella, Caballero-Caballero and Ponce-Cueto, 2019) defined dropout based on learners discontinuing submitting graded problems on a course; corresponding to this definition, any learner who does not complete the midterm or final test is deemed to have dropped out at some point during the course. Another definition proposed by (Alrajhi, Alamri and Cristea, 2022) defined dropout as failing to access 80% of the topics in the next week, which considered such learners to be dropouts following the approach of (Alamri *et al.*, 2021).

Other researchers sought to predict *successful learners* in contrast to *at-risk learners.* (Klusener and Fortenbacher, 2015) assumed that at least 80% of all video lessons must be watched by a learner for the learner to be considered successful. Another study (Ramesh *et al.*, 2020), took into account two indicators to identify successful learners: (i) *performance* (whether the learner gains a certificate at the end of the course), and (ii) *survival* (whether the learner completes the course).

2.3.4.5.2.2. Prediction Models and Algorithms

Various supervised and unsupervised ML prediction models were adopted within the surveyed studies. The supervised approach involves both traditional and DL models as employed in the surveyed papers. In DL, there is a tendency to build hybrid models such as CNN+GRU with Char-CNN. Also, some used Transformer models (BERT). Figure 2.22 (below) provides a summary of the various predictive algorithms used in the surveyed studies.

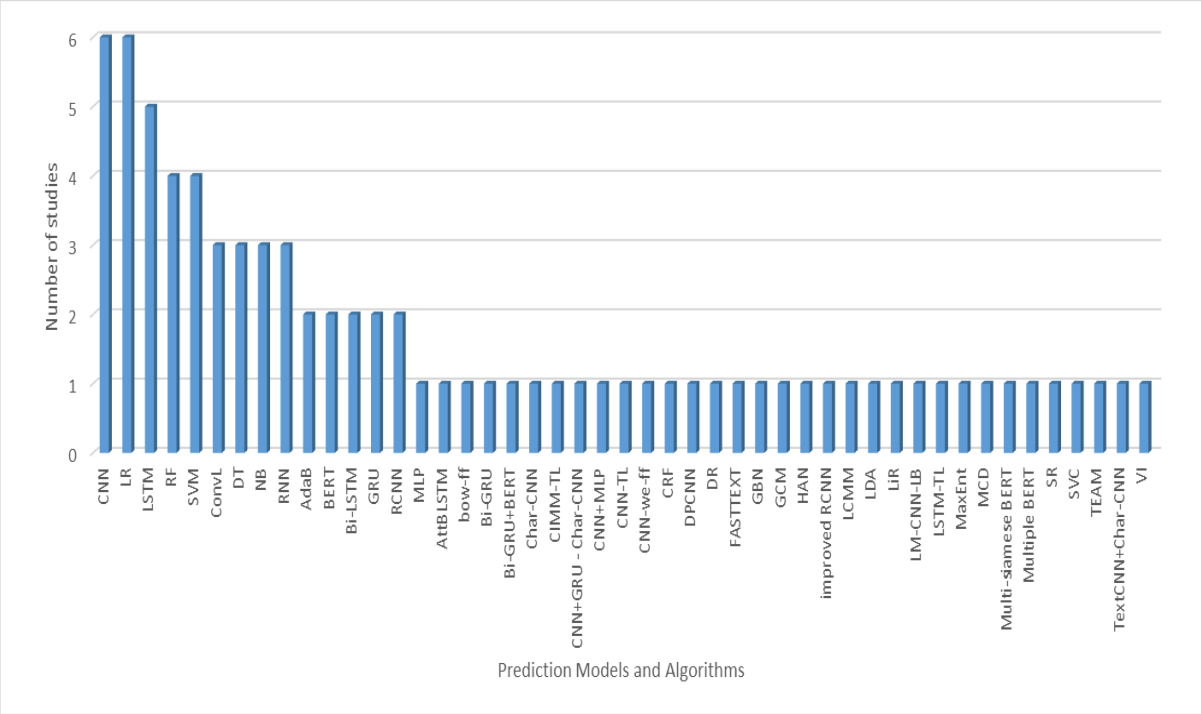

Figure 2.22: Number of surveyed studies across prediction models and algorithms.

2.3.4.5.2.3. Training and Testing Splitting Techniques

In ML, the data is typically split into a training set and a testing set to train and evaluate the model. As shown in Figure 2.23 (below), most of the surveyed studies used either the k-fold cross-validation or the percentage-splitting techniques to split the data. The reason why some researchers used percentage-splitting techniques rather than k-fold cross-validation is that with DL models that contain numerous parameters, k-fold cross-validation is unnecessary because it makes training more complex (Aljohani, 2022).

Using k-fold cross-validation techniques, eight out of 11 studies used ten-fold cross-validation (the most common splitting ratio) (Chaturvedi, Goldwasser and Daumé III, 2014; Bakharia, 2016; Almatrafi, Johri and Rangwala, 2018; Sun *et al.*, 2019; Klusener and Fortenbacher, 2015; Xing *et al.*, 2016; Ramesh *et al.*, 2020; Rossi *et al.*, 2022), whereas the remaining studies used lower number folds, specifically four (Capuano *et al.*, 2021) or five (Chandrasekaran *et al.*, 2017; Alrajhi *et al.*, 2021) folds. The reason for minimising the number of k and not using the frequently used ten-fold approach is because of the very low numbers in minority classes. Meanwhile, in the percentage-splitting technique, different percentages (training/testing) were applied such as 0.75/0.25 (Ntourmas *et al.*, 2021), 0.66/0.34 (Almatrafi, Johri and Rangwala, 2018; Guo *et al.*, 2019; Khodeir, 2021), 0.80/0.20 (Rossi *et al.*, 2022;

Alrajhi, Alharbi and Cristea, 2020; Yu *et al.*, 2021; Alrajhi, Alamri and Cristea, 2022; Alrajhi *et al.*, 2022), and 0.40/0.60 (Yu *et al.*, 2021). Some research mentions validation with percentages of 20% (Alrajhi, Alamri and Cristea, 2022) and 10% (Alrajhi *et al.*, 2022) for training. As MOOC data for posts and learners is unbalanced (please note that it is to be expected that any such dataset would be unbalanced, with the non-urgent posts and dropouts being the predominant class), some researchers (Chandrasekaran *et al.*, 2017; Ntourmas *et al.*, 2021; Almatrafi, Johri and Rangwala, 2018; Guo *et al.*, 2019; Alrajhi, Alharbi and Cristea, 2020; Khodeir, 2021; Yu *et al.*, 2021; Alrajhi *et al.*, 2021; Alrajhi *et al.*, 2022) adopted stratified techniques (Farias, Ludermir and Bastos-Filho, 2020) to guarantee that each set has the same sample of both positive and negative instances. Also, they applied the cross-domain approach (Bakharia, 2016; Wei *et al.*, 2017; Almatrafi, Johri and Rangwala, 2018; Guo *et al.*, 2019; Khodeir, 2021; Ntourmas *et al.*, 2021) where model training is performed in a specific domain and testing is performed in another domain. (Khodeir, 2021) used a validation figure of 30% from testing in both the percentage-splitting technique and the cross-domain technique. Another study that sought to identify learners (Borrella, Caballero-Caballero and Ponce-Cueto, 2019) used older runs of the course for training and recent runs for testing.

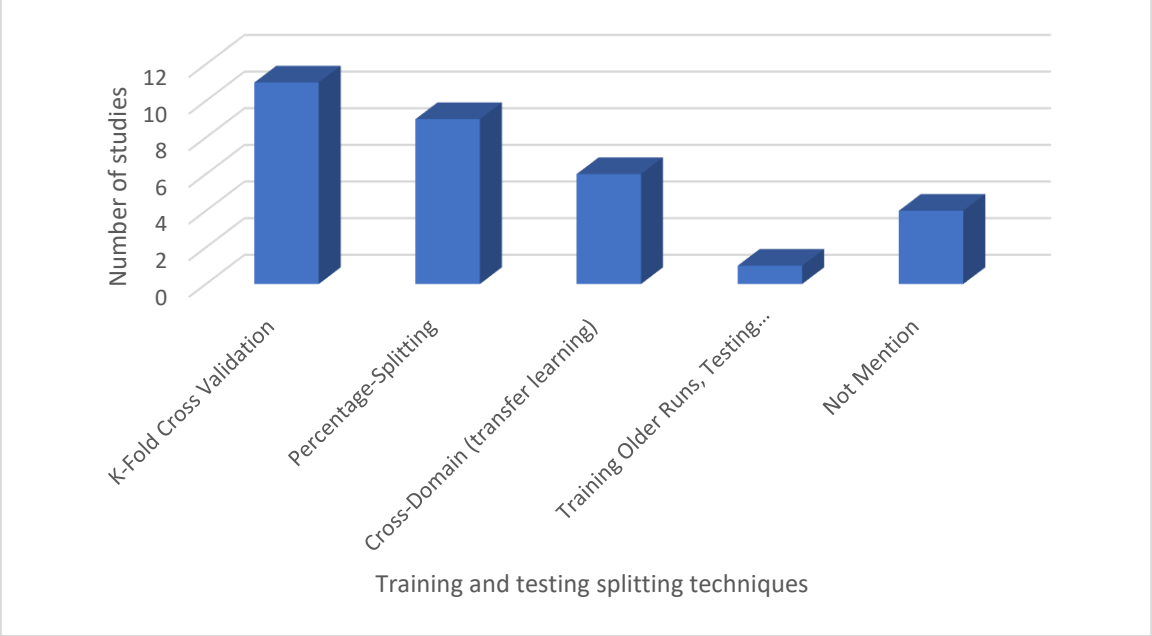

Figure 2.23: Number of surveyed studies across training and testing splitting techniques.

2.3.4.5.2.4. Performance Metrics

Researchers employed and reported various metrics to assess the model's performance and quality. Some researchers considered evaluating models and selecting the metrics as the MOOC

data for posts and learners are unbalanced. Please note that here all the metrics that the authors reported are considered, regardless of what metrics they used to assess their models' performance. Precision (P), recall (R), and F1 are the most commonly used metrics in the surveyed papers (Chaturvedi, Goldwasser and Daumé III, 2014; Chandrasekaran *et al.*, 2017; Ntourmas *et al.*, 2021; Capuano *et al.*, 2021; Almatrafi, Johri and Rangwala, 2018; Sun *et al.*, 2019; Guo *et al.*, 2019; Alrajhi, Alharbi and Cristea, 2020; Khodeir, 2021; Yu *et al.*, 2021; Alrajhi *et al.*, 2021; Alrajhi *et al.*, 2022; Alrajhi, Alamri and Cristea, 2022; Chandrasekaran and Kan, 2018; Klusener and Fortenbacher, 2015; Xing *et al.*, 2016; Borrella, Caballero-Caballero and Ponce-Cueto, 2019; Atapattu T, 2016) as these papers use all or some of these metrics to evaluate models (see Figure 2.24 below). These relate to gaining a more detailed understanding of a classifier's performance rather than focusing only on overall accuracy as the data are unbalanced. Some research reported values for these metrics as averages (Chandrasekaran and Kan, 2018; Ntourmas *et al.*, 2021; Bakharia, 2016; Wei *et al.*, 2017; Capuano *et al.*, 2021; Rossi *et al.*, 2022; Sun *et al.*, 2019; Xing *et al.*, 2016; Borrella, Caballero-Caballero and Ponce-Cueto, 2019; Atapattu T, 2016), others focused only on the important targeted class (Chaturvedi, Goldwasser and Daumé III, 2014; Chandrasekaran *et al.*, 2017; Alrajhi *et al.*, 2021) and more specifically reported on more details for each class (Almatrafi, Johri and Rangwala, 2018; Guo *et al.*, 2019; Alrajhi, Alharbi and Cristea, 2020; Khodeir, 2021; Yu *et al.*, 2021; Alrajhi *et al.*, 2022; Klusener and Fortenbacher, 2015; Alrajhi, Alamri and Cristea, 2022; Ramesh *et al.*, 2020). Some studies (Bakharia, 2016; Wei *et al.*, 2017; Rossi *et al.*, 2022) reported model accuracy only, but this is not appropriate for unbalanced data as it might assign high values to a weak classifier which is misleading. Other, less common metrics were used such as weighted-F1 (Almatrafi, Johri and Rangwala, 2018; Guo *et al.*, 2019; Khodeir, 2021), precision verse recall curves (PR) (Khodeir, 2021; Ramesh *et al.*, 2020), loss (Capuano *et al.*, 2021), Cohen's Kappa (Almatrafi, Johri and Rangwala, 2018), etc., as shown in Figure 2.24 (below).

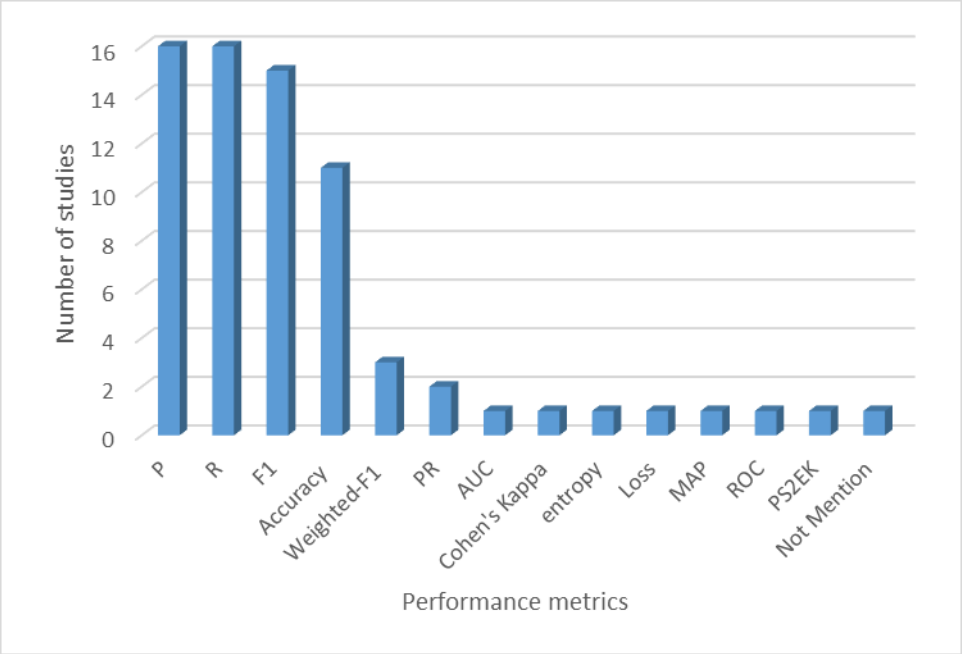

Figure 2.24: Number of surveyed studies across performance metrics.

2.3.4.5.3. EXplainable Artificial Intelligence

As shown before in Section 2.3.4.5.2.2, most papers adopted DL for prediction. The nature of these models is characterised as 'black-box' and their results are unclear to humans in terms of exploring how these internal models behave and explain their decisions due to the complexity of these models. Understanding such black-box approaches is currently an active and significant area of research. XAI aims to increase confidence in decisions made based on AI and more specifically ML models. Only one research intervention (Alrajhi *et al.*, 2022) applied the XAI approach to explain the model's outputs for instructors. They demonstrated how a good predictor could be used in conjunction with XAI, particularly colour-coded visualisation, to assist instructors in taking decisions on intervention. Table 2.4 (below) introduces a general outline of the surveyed studies.

Table 2.4: Outline of previous studies on MOOC instructor intervention among three main axes: identify (posts or learners or topics) (NA denotes missing; other abbreviations are explained in the following tables).

Identify	Author	Platform	#Courses/ #Threads/ #Posts	Intervention label	Type of classification	Data types as features	Models	Splitting techniques	Performance metrics
Posts	(Chaturvedi, Goldwasser and Daumé III, 2014)	Coursera	2/ 1780/ 7700	Real intervention	Binary	Linguistic (about posts) + metadata as numerical	LR, LCMM, GCM	10-fold cross-validation	P, R, F1, (Positive class)
	(Chandrasekaran *et al.*, 2015a)	Coursera	61/ 33665/ NA	Human coding	Multi-Class	Linguistic (about dialogue and discourse analysis)	CRF	NA	Research in progress (they will measure learner performance in terms of quantitative and qualitative)
	(Chandrasekaran *et al.*, 2017)	Coursera	14/ 11554/ NA	Real intervention	Binary	Linguistic (about posts) + post discourse from the PDTB-based discourse parser	MaxEnt	Stratified 5-fold cross-validation	P, R, F1, (Positive class)
	(Chandrasekaran and Kan, 2018)	Coursera	14/ 7219/ NA	Real intervention	Binary	Linguistic (posts contents as unigrams with Tf-idf score + about posts) + metadata as numerical	SVM	NA	P, R, F1, (Average)
	(Bakharia, 2016)	Stanford	3/ NA/ 18093	Human coding	Binary	Linguistic (posts contents as unigrams with Tf-idf score)	NB, SVM, AdaB, RF	10-fold cross-validation + transfer learning	Acc, (Average)
	(Wei *et al.*, 2017)	Stanford	3/ NA/ 18093	Human coding	Binary	Linguistic (posts contents)	CNN-NTL, CNN-TL, LSTM-NTL, LSTM-TL,	Transfer learning	Acc, (Average)

						CIMM-TL, LM-CNN-LB, ConvL-NTL, ConvL, ConvL-in domain		
(Capuano *et al.*, 2021)	Stanford	11/ NA/ 29604	Human coding	Multi-Class	Linguistic (posts contents)	bow-ff, cnn-we-ff, convL, HAN	4-fold cross-validation	P, R, F1, (Average) Loss
(Rossi *et al.*, 2022)	Stanford	11/ NA/ 29604	Human coding	NA	Linguistic (posts contents)	LR	Percentage-splitting (training 80%, testing 20%) + 10-fold cross-validation	Acc, (Average)
(Almatrafi, Johri and Rangwala, 2018)	Stanford	11/ NA/ 29604	Human coding	Binary	Linguistic (posts contents as unigrams with Tf score + LIWC) + metadata as numerical (up_count + reads + post_type)	NB, SVM, RF, AdaB, LR	10-fold cross-validation + stratified percentage-splitting (training 66%, testing 34%) + transfer learning (courses based) + transfer learning (domain-based)	P, R, F1, (Each class) Weighted-F1, Cohen's Kappa
(Sun *et al.*, 2019)	Stanford	3/ NA/ 18093	Human coding	Binary	Linguistic (posts contents)	CNN, RNN, LSTM, GRU, RCNN, improved RCNN	10-fold cross-validation	Acc, P, R, Macro F1, (Average)
(Guo *et al.*, 2019)	Stanford	11/ NA/ 29604	Human coding	Binary	Linguistic (posts contents) + metadata as	Char-CNN, TextCNN, LSTM, CNN +	Stratified percentage-splitting (training 66%, testing 34%) +	P, R, F1, (Each class)

					linguistic (course_display_ name)	LSTM, Bi-LSTM, RCNN, AttBLSTM, DPCNN, TextCNN + Char-CNN, CNN + GRU - Char-CNN	transfer learning (courses based) + transfer learning (domain-based)	Weighted-F1
(Alrajhi, Alharbi and Cristea, 2020)	Stanford	11/ NA/ 29604	Human coding	Binary	Linguistic (posts contents) + metadata as numerical (sentiment, confusion, opinion, questions, and answers)	CNN, CNN + MLP	Stratified percentage-splitting (training 80% + testing 20%)	Acc, (Average) P, R, F1 (Each class)
(Khodeir, 2021)	Stanford	11/ NA/ 29604	Human coding	Binary	Linguistic (posts contents) + metadata as linguistic (course_display_ name)	RNN, CNN, FASTTEXT, LSTM, Bi-GRU Based on BERT	Stratified percentage-splitting (training 66%, testing 34%) + transfer learning (courses based) + transfer learning (domain-based) In all validation 30% from testing	P, R, F1, (Each class) Weighted-F1, PR
(Yu *et al.*, 2021)	Stanford	11/ NA/ 29604	Human coding	Binary	Linguistic (posts contents)	RNN, MCD, VI	Stratified percentage-splitting (training 80%, testing 20%) + percentage-splitting (training 40%, testing 60%)	Acc, (Average) P, R, F1, (Each class) Entropy
(Alrajhi *et al.*, 2021)	FutureLearn	1/ NA/ 5786	Human coding	Binary	Prediction: linguistic (posts contents)	BERT	Stratified 5-fold cross-validation	Acc, (Average) P,

						Priority: numerical (urgency + sentiment analysis + step access)			R, F1, (Positive class)
	(Alrajhi *et al.*, 2022)	FutureLearn	1/ NA/ 5786	Human coding	Binary	Linguistic (posts contents)	BERT	Stratified percentage-splitting (training 80%, testing 20%) validation 10% from training	Acc, (Average) P, R, F1, (Each class)
	(Ntourmas *et al.*, 2021)	OpenEdX	2/ NA/ 1977-starting posts	Human coding	Multi-Class	Calculation of the semantic similarities with two corpora.	SVC	Stratified percentage-splitting (training 75%, testing 25%) + transfer learning	Acc, P, R, F1, (Average)
Learner	(Klusener and Fortenbacher, 2015)	Iversity	3/ NA/ 21825	Successful learners	Binary	14 features as numerical such as (number of answers + up-votes)	DT, RF, DR, SR, LR	10-fold cross-validation	Acc, (Average) R, (Each class)
	(Xing *et al.*, 2016)	Canvas	1/ NA/ NA	Dropout	Binary	6 features as numerical such as (number of discussion posts + number of forum views)	GBN, DT(C4.5)	10-fold cross-validation	AUC, P, (Average)
	(Borrella, Caballero-Caballero and Ponce-Cueto, 2019)	MITx MicroMasters	5/ NA/ NA	Dropout	Binary	14 features as numerical such as (number of clicks in the course and discussion forum)	RF, LR	Training older runs, testing recent runs	R, P, (Average)
	(Ramesh *et al.*, 2020)	Coursera	7/ NA/ NA	Successful learners	Binary	Different features as numerical such as (forum content and interaction: linguistic as	SVM, LR, MLP, LiR, DT	10-fold cross-validation	PR, (Each class) ROC

						sentiment analysis and structure)			
	(Alrajhi, Alamri and Cristea, 2022)	FutureLearn	1/ NA/ 5786	Dropout	Binary	Linguistic (posts contents)	CNN, LSTM, Bi-LSTM, GRU, Bi-GRU, Multi-siamese BERT, Multiple BERT	Percentage-splitting (training 80%, testing 20%) validation 20% from training	Acc, (Average) P, R, F1, (Each class)
Topic	(Atapattu T, 2016)	Coursera	3/ 15894/ 17362	NA	NA	Linguistic (posts contents)	LDA NB	NA	F1, (Average) MAP
	(Yang, Ren and Wu, 2022)	Chinese University	1/ NA/ 5325	Topic attention	NA	Linguistic (posts contents + behaviour about posts)	TEAM	NA	PS^2EK

Table 2.5: List of features abbreviations and acronyms.

Abbreviation /Acronym	Description
LIWC	Linguistic Inquiry and Word Count
PDTB	Penn Discourse Treebank
Tf	Term Frequency
Tf-idf	Term Frequency - Inverse Document Frequency

Table 2.6: List of models' abbreviations and acronyms.

Abbreviation /Acronym	Description
AdaB	AdaBoost
AttBLSTM	Attention Bidirectional Long Short-term Memory
BERT	Bidirectional Encoder Representations from Transformers
Bi-GRU	Bi-directional Gated Recurrent Unit
Bi-LSTM	Bidirectional Long Short-term Memory
Bow-ff	Bag of Words- feed forward neural network
Char-CNN	Character-level Convolutional Neural Networks
CIMM-TL	Consumption Intention Mining Model- Transfer Learning
CNN	Convolutional Neural Network
CNN-TL	Convolutional Neural Network- Transfer Learning
Cnn-we-ff	Convolutional Neural Network- feed forward neural network
ConvL	Convolutional Neural Network Long Short-term Memory
CRF	Conditional Random Fields
DR	Decision Rules
DT	Decision Tree
GBN	General Bayesian Network
GCM	Global Chain Model
GRU	Gated Recurrent Unit
HAN	Hierarchical Attention Network
LCMM	Linear Chain Markov Model
LiR	Linear Regression
LR	Logistic Regression
LSTM	Long Short-term Memory
LSTM-TL	Long Short-term Memory- Transfer Learning
MAP	Mean Average Precision
MaxEnt	Maximum Entropy
MCD	Monte Carlo Dropout
MLP	Multi-Layer Perceptron
NB	Naive Bayes
RCNN	Recurrent Convolutional Neural Network
RF	Random Forests
RNN	Recurrent Neural Network
SR	Step Regression
SVC	Support Vector Classifier
SVM	Support Vector Machine
VI	Variational Inference

Table 2.7: List of metric abbreviations and acronyms.

Abbreviation /Acronym	Description
Acc	Accuracy
AUC	Area Under Curve
F1	F1-score
P	Precision
PR	Precision verse Recall curves
R	Recall
ROC	Receiver Operating Characteristic Curve

2.3.4.6. Critical Evaluation and Limitations

Based on the current systematic review, the limitations and potential suggested improvements for the present MOOC instructor intervention models and several research directions that are presented in this thesis to fill the identified gap in current models are highlighted in this section. Please note that the surveyed research includes the experiments done by the author of this thesis, which were published up until the end of 2022, but it is considered to fill the gap in the currently available research.

As mentioned before, the surveyed research was categorised into three identification intervention models: (i) posts, (ii) learners, and (iii) topics; thus the limitations in this section may be observed for all kinds of models or in relation to the specific categories used.

The most emphasised limitation in the surveyed research on identifying posts is the lack of data available for researchers. More data would enable a higher level of model generalisation and further validate the achieved results. It is challenging to consider the results of the present models as generalisable because, as shown earlier in Figure 2.18, a significant number of research studies that identify posts used only one source of data, namely the Stanford dataset. Even though this dataset is a good resource because it contains 11 courses with 29604 posts from different domains, it still only represents one platform. As a result, the literature must be expanded to include additional platforms to adequately represent the huge variety of real-world MOOC environments that exist today. This is because different platforms have different structures and allow different word counts for posts. Additionally, due to the nature of the Stanford dataset, which only contains learner posts, studying further behaviours related to the need for intervention such as dropout is not possible. Therefore, constructing a new model using a more diverse dataset would improve the generalisability and reliability of the results. This was improved in this thesis by creating a new corpus from another platform (FutureLearn) which is further clarified in Chapter 3.

Another challenge is the difficulty of comparing the results of the present studies at the level of performance achieved. This is due to the inconsistent identification of posts that need intervention introduced by each study. Moreover, each study has different methodologies such as pre-processing, number of posts analysed, how labels are constructed, and splitting data into training and testing sets. To make a comparison of the literature findings, the same data with the same characteristics should be used. Therefore, the data proposed for researchers should be split into training and testing sets to allow comparison between different proposed models.

Therefore, with the available data, the suggestion for any experiment is to apply a baseline model and make a comparison with the proposed model; it is believed that this is the best solution for conducting a reliable comparison. This was achieved in different experiments in this thesis. The only research that the author was able to compare with existing research (Guo *et al.*, 2019) is the *plug & play with deep neural networks* experiment (Alrajhi and Cristea, 2023). This is because the Stanford data was used in this experiment, and the same method of splitting the data was applied, as clarified in Chapter 4, Section 4.3. Therefore, it is possible to compare the same data in the two training and testing sets.

In terms of identifying posts, recent models use DL based on text-only without combined mixed data such as text data with metadata. Thus, this thesis proposed an experiment using a novel multi-dimensional DL model as clarified in Chapter 4 in the first experiment: the *multi-dimensional deep learning model*. Also, different current DL models using hybrid models with different levels of input, word-based or character-word-based and different embeddings such as BERT or word2vec were used. However, no research has examined character-word-based embedding with BERT to represent words. Thus, in Chapter 4, the second experiment *plug & play with deep neural networks* was performed to examine this assumption.

The research presented in the field of identifying topics is very scant: only two research papers address this issue (Atapattu T, 2016; Yang, Ren and Wu, 2022). Thus, there is a need for more research in this field and linking it to intervention by extracting language which indicates a request for intervention. Also, adding visualisation to topics of posts is another promising approach considering that most individuals are visually oriented. This was implemented in the experiment presented in Chapter 5 of this thesis.

To identify at-risk learners, among the surveyed studies, there is no research dealing with text directly and studying the historical content of learner posts: all the proposed models used numerical features about posts and do not focus on post content. This thesis proposed an experiment on intervention prediction based on learners' posts as explained in Chapter 6.

Another limitation in identifying posts is that there is no research focusing on identifying posts that relate to learners' behaviour or adding any priority for intervention. In this thesis, this was proposed in the experiment in Chapter 7 as learner histories were studied based on three features for assigning priority (past urgency, sentiment analysis, and step access). Then, an automated intervention priority model was proposed based on these variables.

The other unexplored research direction that requires attention is the fact that most datasets sourced from MOOCs tend to be imbalanced. This imbalance can be shown in terms of the number of posts that need intervention. Of the surveyed papers, none tackles the imbalanced data issue. Some studies (Almatrafi, Johri and Rangwala, 2018; Khodeir, 2021) briefly consider some common techniques such as splitting data and metric selection but do not address the issue of improving data to overcome the data-imbalance problem. Therefore, there is a need to deal with imbalanced data by applying data balancing techniques such as oversampling and undersampling. This limitation has been tackled in the experiment for solving the imbalanced data issue in Chapter 8.

Regarding XAI (as it is useful in deep learning to understand model decisions), there is no research applied to the instructor intervention task. However, the experiment presented in Chapter 9 applied XAI in a novel way to fill the gap in the literature by assisting both instructors and annotators.

2.4. Epilogue

Researchers have attempted to address the issue of instructor intervention in MOOCs using techniques from both NLP and ML. This chapter provided background information that relates to this thesis on MOOCs, NLP, and ML. In addition, a literature review of MOOCs and NLP was presented. Specifically, the SLR identified and reviewed instructor intervention need in MOOC discussion forums. For the first time, it gathered several research studies on instructor intervention in MOOCs based on discussion forums by identifying three different perspectives: (i) posts, (ii) topics, and (iii) learners. The current thesis helps to focus on the unexplored and unfilled gaps in the literature to advance the field of instructor intervention in MOOC discussion forums.

The following chapter provides the methodology for how this thesis contributes to the field of instructor intervention in MOOC discussion forums.

CHAPTER 3: METHODOLOGY

3.1. Prologue

This thesis sought to enhance the quality of determining instructor intervention needs in MOOCs and investigate the feasibility of using NLP and ML to predict when instructor intervention in MOOC environments is necessary based on asynchronous discussion forums using learner language from posts. To address and answer the RQs and achieve the specified objectives of this thesis, the research methodologies are outlined and explained in detail in this chapter. This involves presenting several datasets used in the experiments featured in this thesis which were collected and extracted from posts on discussion forums of different MOOC platforms (Section 3.2) and how different corpora can be created based on these datasets. Additionally, the overall process and experimental framework are described and summarised to enhance the detection of the need for instructor intervention involving posts, topics, and learners by implementing different ML approaches as clarified in Section 3.3 (for more details see the main chapters featuring the experiments: Chapters 4–9). Also, the performance evaluations are provided in Section 3.4. Moreover, Section 3.5 discusses the ethical issues considered in the current research project.

3.2. Datasets

Nowadays, public datasets with labelled MOOC forum discussion posts that contain learners' textual data for use in solving the instructor intervention task are quite limited (Guo *et al.*, 2019) (to the best of the author's knowledge, there is only one suitable dataset, The Stanford MOOCPost dataset; notably, most of the studies surveyed in Chapter 2 used this dataset). This

thesis involved different experiments to address the RQs and objectives by using four corpora to conduct the analysis and identify the need for instructor intervention. It is better to use different datasets that employ diverse datasets to contribute to the robustness, generalisation, bias, and reliability of ML models as the results are based on data-driven findings. The data source originates from two educational MOOC platforms (Stanford and FutureLearn) to represent different types of platforms. As mentioned previously, different platforms have various discussion forum structure formats and different numbers of words (allowed) per post. In addition, the Stanford dataset lacks data about learner behaviours (e.g., step access). Thus, it is better to create another dataset to fill this gap.

In particular, one corpus (the Stanford MOOCPost dataset) retrieved from the Stanford platform is available for researchers on request and three new corpora derived from FutureLearn which are manually annotated and built were used. Two FutureLearn corpora are similar in purpose and the difference is in the methods used to obtain posts urgent class labels as one of the challenges of creating an instructor intervention need dataset is how to define an urgent intervention label. The third corpus, meanwhile, relates to learner dropout or completion and intervention need. Understanding the nature of the data is essential since it influences the research plan. In the following sub-sections further explanations of the datasets are provided.

3.2.1. Stanford MOOCPost Dataset

This research project used the Stanford MOOC benchmark posts dataset (Agrawal and Paepcke, 2019), which is available to academic researchers on request. It covers three different areas with a variety of courses: education (1 course), humanities/sciences (6 courses), and medicine (4 courses), resulting in a total of 29,604 anonymised learners' forum posts that are spread across 11 Stanford courses (Agrawal *et al.*, 2015). Each post was manually coded and labelled by three independent human consultants' coders (ODesk) to create a gold-standard dataset. Each post was evaluated against six categories/dimensions (*sentiment, confusion, urgency, opinion, question,* and *answer*). *Opinion, question, and answer* were assigned binary values while *sentiment, confusion* and *urgency* were rated values based on a scale of 1–7. To explain, for *sentiment*, 1 = *extremely negative* and 7 = *extremely positive*; for *confusion*, 1 = *extremely knowledgeable* and 7 = *extremely confused*; for *urgency* (which describes how urgent the post is with respect to a required response (intervention) from the instructor), 1 = *no reason to read the post a*nd 7 = *extremely urgent*, the instructor definitely needs to reply. For more detail on the scales of urgency see Figure 3.1 (below).

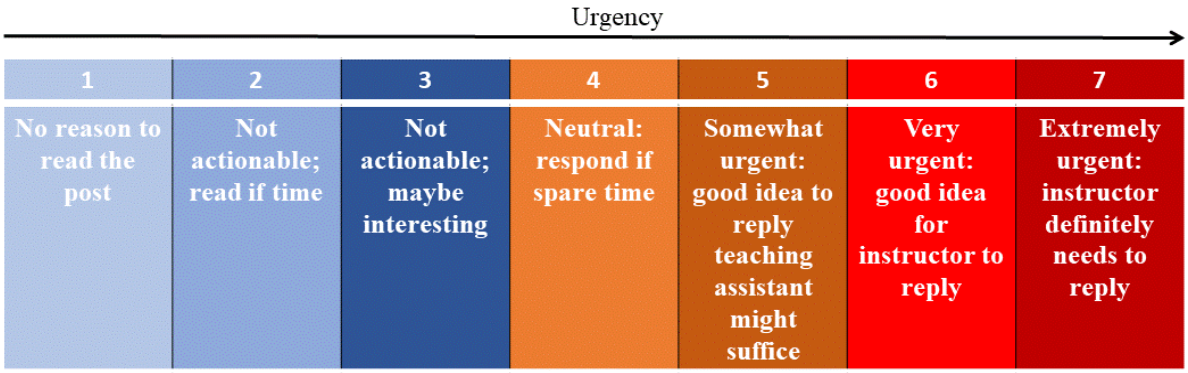

Figure 3.1: The scale of urgency applied (1–7).

The final gold-standard dataset contains, in addition to textual posts, a column for each dimension, based on computing scores between optimal coders and other metadata. The label scores are computed as the average between the optimal agreement combination of coders. For more additional information about the coding method and the creation of the gold-standard dataset see the website of (Agrawal and Paepcke, 2019) [12].

In terms of urgency, agreement was calculated between the optimal coders (number of coders = 2) and in combination with the Likert variables (1–7). Krippendorff alphas were computed; their results in each domain were:

- Medicine: 0.625
- Education: 0.142
- Humanities/Sciences: 0.517

To create a gold-standard dataset for urgency of instructor intervention, the urgency score was computed as an average between two coders. Thus, the results contain the following values: 1/1.5/2/2.5/3/3.5/4/4.5/5/5.5/6/6.5/7. In this thesis, the author believes that two coders are not enough to make a label decision. As (Snow *et al.*, 2008) found, an average of four non-experts are equivalent in quality to one expert-level annotator in labelling data. Thus, to create new corpora for this thesis as explained next in Section 3.2.2, the decision to intervene is based on experts and more than two coders to improve the quality of the annotation task. Table 3.1 (below) shows some randomly selected samples of post content along with their urgency ratings.

[12] https://datastage.stanford.edu/StanfordMoocPosts/

Table 3.1: Examples of postings' content and their ratings for urgency.

Sample	Urgency Ratings
Great	1
Interesting! How often we say those things to others without really understanding what we are saying. That must have been a powerful experience! Excellent!	1.5
Sometimes parents and teachers also expose children to negative messages about math.	2
Mistakes give our brains something to CHEW on!	2.5
Great ideas. Asking students to illustrate how mistakes can lead to their learning is normalizing and encouraging to others.	3
What is \Algebra as a Math Game\" or are you just saying you create games that incorporate algebra."	3.5
I have tried to submit a document form of my response, but still nothing happens...	4
I'm Brazilian, but I read very well in English, so it would be good that all videos have subtitles. thank's [sic] [13]	4.5
Session 2 is not working for me (progress) - none of my assignments have shown up as complete.	5
What happened to sessions 5 & 6? I finished 5 yesteday [sic], now can't find it or 6 (that I wanted to work on today)?	5.5
Pls help!!! I clicked on \submit\" by mistake and now it has been sent to peer assessment!! My answer is completely blank!!! how do I undo it??"	6
Anybody from staff course could provide us a response?	6.5
I hope any course staff member can help us to solve this confusion asap!!!	7

Although the original dataset is multivalued, in order not to add additional complexity, an instructor's decision whether to intervene or not is a binary one; thus the seven-point scale is superfluous. Moreover, other text classification research on identifying urgent learner posts has often converted the scales used to a binary categorisation (Almatrafi, Johri and Rangwala, 2018; Guo *et al.*, 2019) as explained in detail in Section 2.3.4.5.2.1 on SLR in Chapter 2. Thus, the approach of (Guo *et al.*, 2019) was followed in the current thesis to structure the problem of detecting urgent posts as a binary classification task by converting the 1–7 scale into binary values:

- Urgent intervention required > 4 ⇒ Need for urgent intervention (1)
- Otherwise ⇒ No need for intervention (0)

As clarified in literature review, different researchers use different thresholds to construct urgent scales (urgency>4 or urgency>=4). In the Stanford dataset used in the current thesis, the decision to intervene was set at >4, this is further supported by the analysis findings (Chapter 4, Section 4.2.3.1) and in (Alrajhi, Alharbi and Cristea, 2020) as a correlation was found between specific values (4 and 4.5) for the *sentiment* and *confusion* scales.

[13] Sic, adverb, used in brackets after a copied or quoted word that appears odd or erroneous to show that the word is quoted exactly as it stands in the original.

Ultimately, the experimental data was prepared by excluding posts that contained unmeaningful content, as in (Wei *et al.*, 2017) and (Almatrafi, Johri and Rangwala, 2018), (e.g., only numbers or the automated anonymisation of signs was used (<redacted>); thus, across the whole dataset, non-urgent cases represented 81% (23991 posts) and urgent cases accounted for 19% (5606 posts) (with *urgent* posts having *urgency>4*), as shown in Figure 3.2 (below). In general, urgency data are notoriously skewed (with urgent posts being significantly fewer than non-urgent ones).

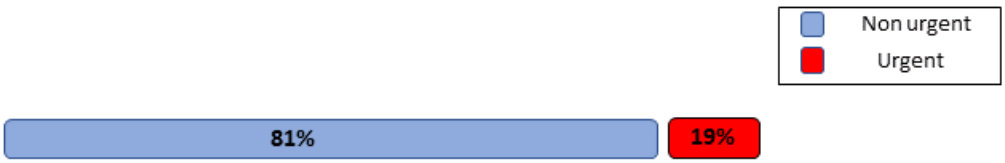

Figure 3.2: The distribution of the two classes (urgent, non-urgent) in the Stanford dataset.

To inspect the distribution of each class on each field and course, 13 posts with an empty course name in the Humanities/Sciences course type 'Course display name' were removed. As in some experiments in this thesis 'course_display_name' was used as metadata in the classification model following (Guo *et al.*, 2019). Then, each class was represented as shown in Figure 3.3, Figure 3.4, and Figure 3.5 (below).

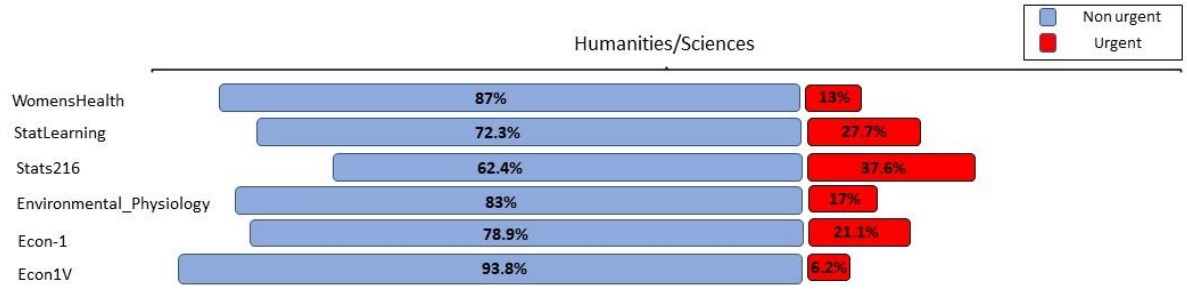

Figure 3.3: The distribution of the two classes (urgent, non-urgent) in the Stanford dataset (Humanities/Sciences) field.

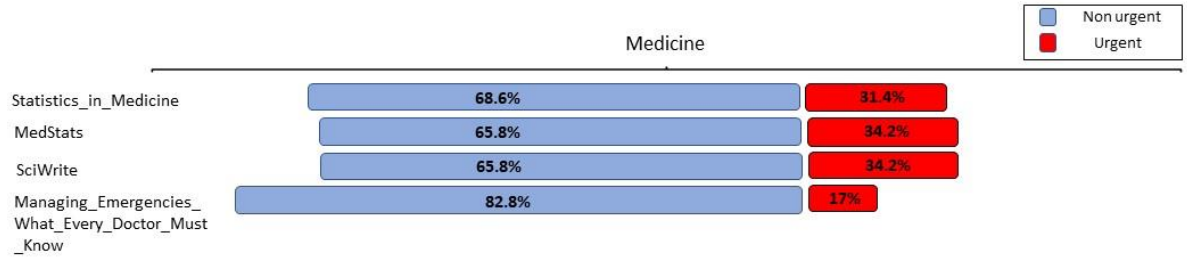

Figure 3.4: The distribution of the two classes (urgent, non-urgent) in the Stanford dataset (Medicine) field.

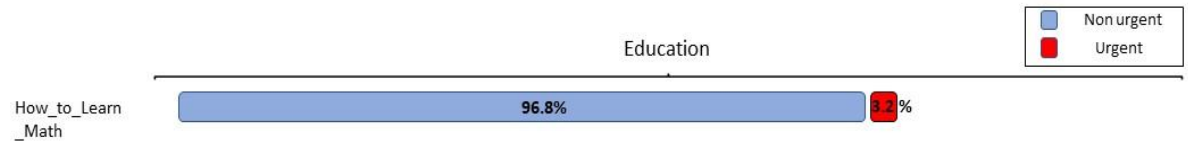

Figure 3.5: The distribution of the two classes (urgent, non-urgent) in the Stanford dataset (Education) field.

The vast majority of recently published research on urgent post classification employed the Stanford MOOC Post dataset as the data source as clarified in Chapter 2. This dataset was used in different thesis experiments (Chapter 4) as all the experiments were related to identifying posts only; in Chapter 5 the experiment was to identify topics in specific courses which contained a high percentage of urgent posts. Also, this dataset was used in Chapter 8 to further validate the proposed solution as clarified in detail in each chapter. However, even though this dataset is an excellent resource, it still represents just one platform; hence other platforms need to be investigated to represent the current wide range of real-life MOOC environments. This is because different platforms have different structures and (minimum) numbers of words per posts. In addition, the nature of the data does not allow the study of learner behaviour in terms of step access or predicting dropout. To address these research gaps and investigate other data sources, creating new data with the required information is still necessary to ensure that a diversity of data is used, broad coverage of different platforms is achieved, and knowledge of this field is enriched. Thus, the present thesis provides an analysis of the FutureLearn platform (which requires additional effort to complete the manual annotation) as discussed in the next sub-section.

3.2.2. FutureLearn Dataset

The raw benchmark corpus dataset utilised consisted of real MOOC forum posts (textual data) for a specific course. In addition to the textual data, data on learner behaviour features (step, visit time, etc.) were used as they are associated with individual learner IDs. Data from FutureLearn's MOOC platform course entitled Big Data (Run 2) was produced and provided by the University of Warwick, UK. This course was selected because it is rich in posts (comments), was popular and dealt with a novel subject. These characteristics mean that this dataset would likely include an adequate number of urgent posts, as big data is arguably a challenging topic. In addition, it contains a high percentage of learners who dropped out (Alamri *et al.*, 2020). The course was conducted in 2016 over a nine-week period; it contains 8263 English-language text posts. The objectives of this research project were to classify urgent

posts in MOOC discussion forums gathered during the first half of the course; this was because other previous research indicated that most learners who dropped out were likely to do so in the early stages (Cristea *et al.*, 2018; Alamri *et al.*, 2019); therefore, intervention, if it was required, would be likely be needed early on before dropout. In this regard, the following steps were taken to select suitable instances from the original data and prepare them for the annotation process. Learner posts within the first half of the course (weeks 1–5) of the long course were extracted and prepared, representing approximately half of the nine-week course (≈ 50%). After this point, all instructors' posts were excluded. This resulted in a total of 5790 posts.

Considering the hardship of using manual annotation to obtain posts, 5790 posts was considered sufficient for this research project. These collected text posts were prepared and manually labelled to assign urgency and annotated by domain experts. This task proved to be quite challenging even for human annotators, which confirms the findings of previous researchers (Chandrasekaran *et al.*, 2015b) who noted that it is difficult for humans to create such a gold-standard data set via manually labelling individual cases requiring instructor intervention since different instructors have varying preferences and strategies for responding to their learners' questions in practice.

The annotation process was performed independently and manually by four computer science experts; of them, three are instructors at the Department of Computer Science at the University; in addition, one is the author of this thesis. In labelling the MOOC urgency corpus, as in the Stanford dataset, (Agrawal and Paepcke, 2019) instructions were given to annotators (as shown in Appendix A), who were asked to manually classify each learner post using the seven-point Likert urgency scale (1–7), representing the range of urgency level (1: *no reason to read the post* – 7: *extremely urgent: instructor definitely needs to reply*) as clarified in Section 3.2.1. Nevertheless, determining which posts require urgent responses is difficult, as selection can be a subjective issue; for example, at present, instructors tend to rely on their own judgement, however, this approach may omit potentially urgent posts thus reducing the effectiveness of the support offered (Chandrasekaran *et al.*, 2015b). After completing the annotations, as a cleaning process (four) posts containing anything other than values from 1–7 were excluded.

Then, the quality of the manually labelled posts was validated and evaluated by using the Krippendorff's alpha (Hayes and Krippendorff, 2007). The resulting agreement between all annotators was low (Krippendorff's alpha=0.33); meanwhile, the Stanford dataset suffered

partially from similar issues as explained in Section 3.2.1; the agreement between the optimal coder combination for the Likert variables (1-7) varied considerably per domain (Education: 0.14; Humanities/Sciences: 0.52; Medicine: 0.63).

Therefore, to address this problem, firstly the 1-7 scale was converted into a simplified scale (1–3), as per Figure 3.6 (below). This meant, for example, mapping 1, 2, and 3 as non-urgent together — as they all are non-actionable, into (1).

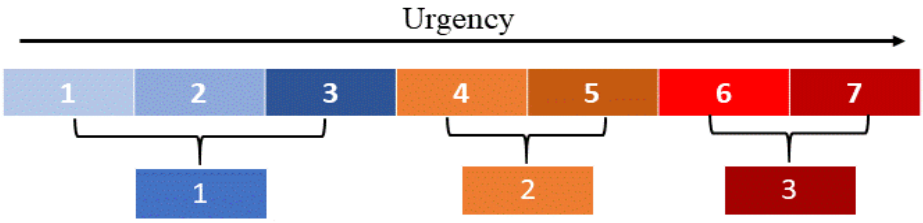

Figure 3.6: Dimensionality reduction: converting the (1-7) scale into a (1-3) scale.

From the obtained annotated data, two datasets were created as follows: (i) the **U**rgent i**N**structor **I**n**TE**rvention (UNITE), and (ii) the Gold-standard corpus. The differences in their respective data creation strategies are explained in the following sub-sections.

3.2.2.1. Urgent iNstructor InTErvention (UNITE)

To create the UNITE dataset and be able to use the data reliably, identifying a dependable sub-set was decided; this sub-set was selected by including only posts that have a level of agreement between annotators of >75%; in other words, at least three annotators (out of four) must have agreed on the post's label. Thus, a voting method was used, which is considered the most appropriate way to integrate different opinions about the same task (Troyano *et al.*, 2004). In this case, only 4622 reliable posts could be included in the gold-standard dataset (approximately 80% of the original data).

The aim here was to obtain as many potentially urgent posts as possible, thus, the problem was framed as a binary classification problem with outputs *urgent* and *non-urgent*, by converting and ranking the gold-standard labels as:

- Scale = 2 or 3 → Urgent.
- Scale = 1 → Non-urgent.

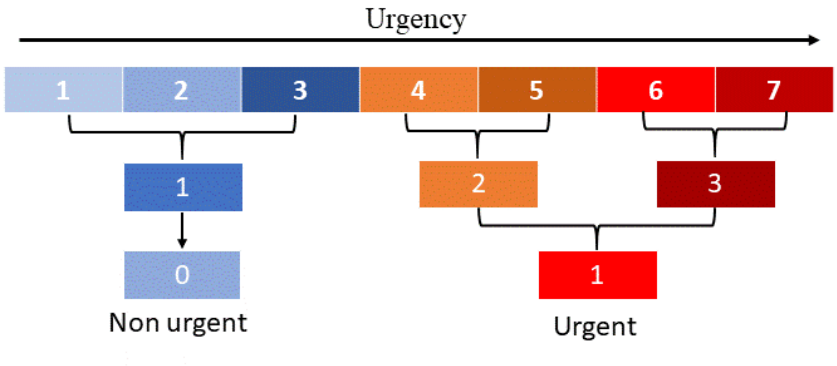

Figure 3.7: Final gold-standard labels for the UNITE corpus.

Figure 3.7 (above) depicts the final gold-standard labels generated for this corpus. Please note that it erred on the side of caution in this final step by including neutral posts (*urgency* = 4) as urgent. This is because, for the Stanford data (Section 3.2.1), while some researchers supposed that urgency≥4 represents *urgent* posts (Almatrafi, Johri and Rangwala, 2018), others regard urgency>4 as *urgent* (Guo *et al.*, 2019). As here it was only working with integer values for labels, it was deemed that a value of 4 and above signifies *urgent.* This is also in line with the researcher's protocol on favouring recall and false positives (FP).

Unsurprisingly, for the UNITE dataset, this division still resulted in a very high proportion of posts being categorised as *non-urgent* (93%, i.e., 4,292 posts with only 330 urgent posts: 7%, see Figure 3.8 below), thus illustrating a high degree of imbalance. For this reason, this dataset data was employed in Chapter 8 in dealing with solving the imbalanced data issue.

Figure 3.8: The distribution of the two classes (urgent, non-urgent) in the UNITE dataset.

3.2.2.2. Gold-Standard Corpus

To build the Gold-standard corpus and be able to increase the reliability of the data based on the raw data, an annotator who disagreed strongly with other annotators was dropped. From the remaining three annotators, a label value was calculated by converting the scale to binary (1–3→0, 4–7→1). Then, a voting process was applied between the three remaining annotators, resulting in a binary-class label as: 0 → *non-urgent*; 1 → *urgent*. This resulted in 5786 posts in

a class size of 5786 (‘0’ non-urgent → 4903 (84 %), ‘1’ urgent → 883 (15 %)) as shown in Figure 3.9 (below). This dataset was used in Chapter 7 because the proposed priority model was built on learner histories and this aspect can be studied based on these data. Also, this dataset was used in Chapter 9 because it allowed the study of agreement between human annotators. This information is not available in the Stanford dataset.

Figure 3.9: The distribution of the two classes (urgent, non-urgent) in the gold standard corpus dataset.

3.2.2.3. Dropout

This corpus dataset was created to identify learners at risk of dropping out and who may need instructor intervention based on their forum posts. The dropout data was collected in the same way as in the FutureLearn (UNITE and gold-standard datasets) corpora (i.e., during the first five weeks). Upon exploring the data, it included about 871 active learners, who were defined as those who participated in the discussion forums and had written at least one text post (Yang *et al.*, 2015; Wen, Yang and Rosé, 2014) from a total of 11281 *enrolled learners* and 4683 *accessed learners*. *Enrolled learners* refers to those who registered on the course; *accessed learners* are those who both enrolled *and* accessed the course at least once during the first five weeks (Alamri *et al.*, 2021).

To create a corpus for all commenters, the histories of learner posts were collected (their most recent posts made during the first five weeks). Then, learners needing intervention were defined as those who dropped out after week 5. Dropout was defined following the approach of (Alamri *et al.*, 2021) on their weekly prediction of dropout: they supposed that learners are considered to have dropped out if, in the following week, they did not access 80% of the available topics. Therefore, for each learner, dropout was defined as accessing less than 80% of the available topics in week 6, therefore, the dropout rate was 66% (574 learners needed instructor intervention) while 34% of learners (297) completed the course; this distribution is illustrated in Figure 3.10 (below).

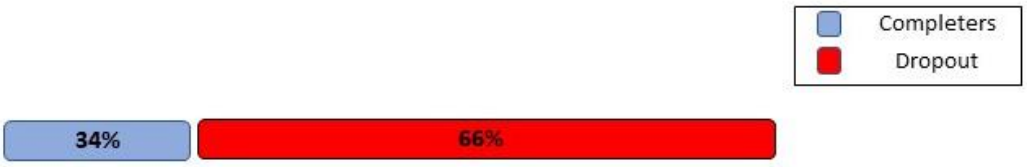

Figure 3.10: The distribution of the two classes (completers, dropout) in the Dropout dataset.

3.3. Experiments Architecture

The methodology was designed to address the problem of identifying when an instructor needs to intervene as instructors are primary target users based on posts written by learners (learners are potential secondary target users). The focus was on three main different aspects: *posts, topics, and learners.* Then, this was expanded to identify posts considering two attributes: (i) learner modelling and (ii) user modelling (instructors and learners) as shown in Figure 3.11 (below).

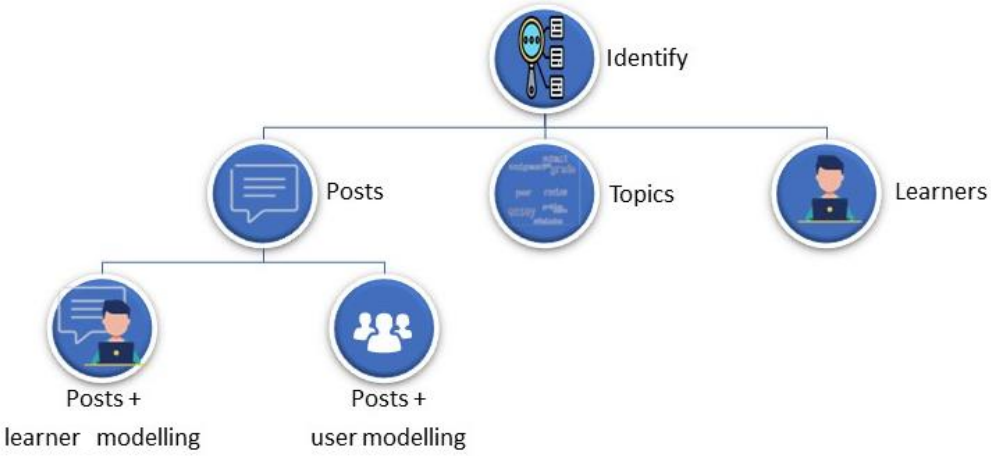

Figure 3.11: Different aspects of instructor intervention in the thesis.

In addition, using XAI to assist instructors in making decisions on when urgent intervention is required, also supports annotators in creating high-quality, gold-standard datasets for the urgent intervention problem.

Thus, the experiments performed in this thesis were based on different levels of identification and XAI, as discussed in the following sub-sections.

3.3.1. Posts (Chapter 4)

To determine if a MOOC forum post highlighted a need for urgent instructor intervention, different binary classifiers were developed.

- Analysis combining several different dimensional values of learners' posts with their textual data and proposing a multidimensional deep learning model (CNN+MLP) to predict posts that require urgent intervention in a MOOC environment.
- Constructing hybrid neural networks using different deep learning models for word-based and word-character-based inputs, comparing word2vec with BERT as an embedding approach to provide a more comprehensive overview of the construction of models to predict the urgency of intervention in MOOC forums.

3.3.2. Topics (Chapter 5)

To explore the urgent-like language that learners use to express their need for immediate intervention:

- Create a text post analysis framework using topic modelling via LDA to identify urgent language and visualise it with the aim of supporting instructors and learners.

3.3.3. Learners (Chapter 6)

To identify learners who need intervention and may drop out and alert instructors about them:

- Propose the prediction model architecture to identify learners' need for instructor intervention based on learners' posting history by integrating the most recent sequence of posts written by learners. Other deep learning architectures and Transformer models were constructed; for the Transformer model specifically, the siamese and dual temporal multi-input approaches were proposed.

3.3.4. Posts + Learner Modelling (Chapter 7)

To add priority in posts requiring intervention based on learners' modelling and their behaviours:

- Propose a new intervention framework designed to add the feature of prioritising urgent posts based on learners' history to assist instructors in making effective decisions to intervene, and ability to adapt their interventions.

3.3.5. Posts + User Modelling (Chapter 8)

To solve the imbalanced data issue and propose adaptive intervention models based on user modelling:

- Automatically classify if a MOOC learner's post is urgent using traditional ML algorithms and Transformers (BERT) requires flagging for instructor intervention based on the instructor and learner models to drive such recommendations to instructors and tackle the imbalance problem.

3.3.6. EXplainable artificial intelligence (XAI) (Chapter 9)

To understand the 'black-box' models of urgent instructor-intervention models in MOOCs:

- Provide an explanation of ML decisions using the Captum tool that explains individual predictions in the urgent intervention task in a MOOC environment to support both instructors and annotators.

3.4. Performance Evaluations

An evaluation metric is a tool for measuring how well a classifier and model perform when tested on unseen data (testing set) (Hossin and Sulaiman, 2015). It is essential for the creation, evaluation, and selection of ML models and there are several ways to evaluate performance.

In this thesis, as the different data considered were highly unbalanced, using accuracy (*Acc*) is ineffective for measuring performance as the data could be biased towards the majority class in the imbalanced class dataset (Gong, 2021). Thus, to achieve more accurate results, other metrics were used to measure the performance of the models per class (negative and positive) that represent *non-urgent* and *urgent* and *completers* and *dropouts*, such as precision (*P*), recall (*R*) and F1 score (*F1*) derived from the true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negatives (*FN*) of the confusion matrix which was used as the basis of these different metrics (see Figure 3.12).

		Predicted Values	
		Negative	Positive
Actual Values	Negative	True Negative TN	False Positive FP
	Positive	False Negative FN	True Positive TP

Figure 3.12: Confusion matrix.

To explain in more detail, TP = predicted positive and is actually positive; TN = predicted negative and is actually negative; FP = predicted positive and is actually negative; FN = predicted negative and is actually positive.

Precision refers to analysing the proportion of positively predicted cases, whereas *recall* assesses how effectively a model predicts positive instances. Thus, the metrics (P, R, F1, and Acc) are calculated as the following equations:

$$P = \frac{TP}{TP + FP} \tag{3.1}$$

$$R = \frac{TP}{TP + FN} \tag{3.2}$$

$$F1 = \frac{2 \,.\, P \,.\, R}{P + R} \tag{3.3}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{3.4}$$

Here, a discussion of what constitutes a good classifier of urgent intervention in posts is required, in terms of the best trade-off between false positives (incorrectly identifying posts requiring urgent intervention) and false negatives (failing to identify posts that require urgent intervention). This was interpreted as giving priority to intervention in urgent cases; hence, false negatives were more problematic than false positives. Thus, *recall* (*R*) was prioritised to ensure the capture of all urgent cases as well as balance accuracy (*BA*) (Brodersen *et al.*, 2010) (which refers to the average recall for each class and is equal to the mean of the *sensitivity* (true positive rate) and *specificity* (true negative rate)); its use is especially useful when the classes are imbalanced. It can be calculated as:

$$BA = \frac{TP/\,(TP+FN) + TN\,/\,(TN+FP)}{2} \tag{3.5}$$

3.5. Ethical Considerations

The data used in this research project is intended for research purposes only. In addition, the necessary usage permits required for this thesis were all obtained. This included permission to use FutureLearn data as during registration, learners give their permission and consent to the use of their information for research purposes. In addition, the annotation process for labelling posts has received ethical approval from the University.

3.6. Epilogue

Numerous research has investigated instructor intervention in MOOC forum posts. However, the literature has failed to consider aspects such as the topics and language used by learners to express their urgency, learners' post history to identify potential dropouts, prioritise intervention based on learner modelling and behaviour, solving the imbalanced data issue, and automating the urgent-posts-identification process based on user modelling and the use of XAI to comprehend model decisions to support both instructors and annotators. This thesis covers instructor intervention in MOOC forums from different perspectives. It also covers the identification of urgent posts as a main objective. Therefore, different datasets were used to complete these objectives. One of them is available on request (Stanford) and the other was derived from the FutureLearn platform which required additional effort for labelling.

The methodology and architecture of the experiments to fill the mentioned gaps in the literature were provided in this chapter; this covered the different classification and clustering models. It also covered classifying posts as urgent or non-urgent, clustering and analysing topics in learners' posts, and classifying learners as completers or dropouts based on their post history. Then, it covered expanding the classification of posts to add priority for intervention based on learners' behaviour and classifying posts by considering user modelling and solving the imbalanced data problem. In addition, it covered using XAI to help instructors as well as annotators by identifying urgent posts to create high-quality datasets. The following chapters discuss these experiments in more detail.

CHAPTER 4: INTERVENTION PREDICTION: POST-BASED MODEL

4.1. Prologue

Most posts in asynchronous MOOC discussion forums feature general communication; among them, some posts request help from instructors. Instructor intervention is important to address struggling learners' needs by replying to their questions and requests for help as learners often describe feelings of confusion and express their need for help via forum posts. However, the often-huge numbers of posts on forums present in MOOCs make it unlikely that instructors can monitor all posts and respond to those requesting help which means that many of these urgent posts are overlooked or discarded. This is exacerbated by the high ratio of learners to instructors in MOOC environments. Thus, capturing target posts that need intervention is a critical yet challenging task. To overcome this, the best solution is to propose classification models that identify posts that require urgent instructor intervention. The main aim of this chapter is to recognise urgent posts on MOOCs by constructing two novel experiments on this domain as shown in Figure 4.1 (below) using the Stanford dataset as explained in Section 3.2.1. This can help guide instructors to identify posts that need intervention.

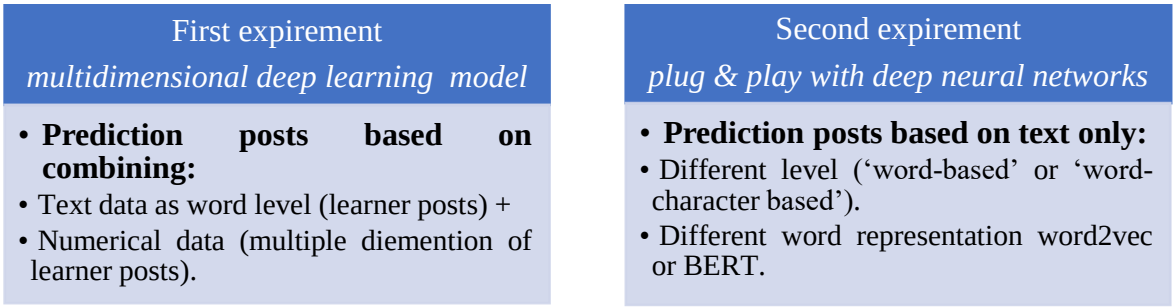

Figure 4.1: Urgent posts prediction experiments.

The first experiment *(multidimensional deep learning model)* aimed to construct a classifier to identify the need for instructor intervention on posts based on DL models that integrates different aspects of MOOC posts (*sentiment, confusion, opinion, question,* and *answer*) with text to classify urgent posts as numerical data and textual data.

The second experiment *(plug & play with deep neural networks)* sought to discover what the preferable combination is between different (hybrid) DL models to construct the best predictor model for classifying posts that need instructor intervention. It applies the plug & play technique for word-based and word-character-based input, as it is expected that adding additional character-sequence information may increase performance. These models were constructed based on different embeddings (word2vec or BERT) to represent the words used in posts.

4.2. A Multidimensional Deep Learning Model

Mining raw data on MOOC learners' posts may provide a helpful way of classifying posts where learners require urgent intervention from instructors. In this experiment, a method based on the correlations of five different dimensions of learner posts (*sentiment, confusion, opinion, question,* and *answer*) to determine the need for urgent intervention was proposed. Then, a multidimensional DL model was developed which contributes to the intervention task based on the above five dimensions in addition to learners' posts texts to determine the need for urgent instructor intervention.

This model is a novel classifier for this area; many recent studies have focused on detecting struggling learners' posts using different methods as clarified in Chapter 2. Some of these approaches use features extracted from the properties of posts (Chaturvedi, Goldwasser and Daumé III, 2014) while others are based on text-only features with DL (Guo *et al.*, 2019; Sun *et al.*, 2019). However, few studies have combined mixed data such as text data with metadata

(Chandrasekaran *et al.*, 2015b; Almatrafi, Johri and Rangwala, 2018); such studies are limited as they are all based on traditional ML only. Thus, this experiment to classify urgent posts is based on a DL model containing sub-models to investigate different aspects of MOOC posts (*sentiment, confusion, opinion, question,* and *answer*) as numerical data and textual data.

Thus, the first goal of this chapter is to show how MOOC posts can be mined and create urgent instructor intervention prediction models based on correlations of different dimensions of learners' posts in an attempt to answer the following two RQs:

- ***RQ1.1:*** *Is there a relationship between the various dimensions of the learners' posts and their need for urgent instructor intervention?*
- ***RQ1.2:*** *Does using several dimensions as features in addition to textual data increase the model's predictive power for identifying posts that require the need for urgent instructor intervention when using deep learning?*

The following are the main contributions of this experiment: (i) exploring the statistical analysis of different dimensions of MOOC learners' posts in relation to non-urgent and urgent posts, and (ii) building a novel classifier for this area based on DL models that incorporates different dimensions of MOOC posts to classify urgent posts, i.e., numerical data in addition to textual data.

4.2.1. Related Work on MOOC Analysis

Recently, data from MOOC discussion forums has been subject to significant research efforts to study, analyse, and evaluate different learners-related aspects including *sentiment* (Wen, Yang and Rose, 2014), *confusion* (Agrawal *et al.*, 2015), and *the need for urgent intervention* (Almatrafi, Johri and Rangwala, 2018) to improve the educational quality of MOOC environments and improve the overall educational outcomes of MOOC learners.

Researchers have employed sentiment analysis for different purposes; for instance, they used it to predict *attrition* (Chaplot, Rhim and Kim, 2015), *performance and learning outcomes* (Tucker, Pursel and Divinsky, 2014), *emotions* (Moreno-Marcos *et al.*, 2018a) and *dropout* (Wen, Yang and Rose, 2014) by using different ML approaches. These methods include statistical analysis and traditional ML and DL. A growing number of researchers have studied *confusion*; For example, (Yang *et al.*, 2015) explored click patterns to identify the impact of confusion on learner dropout; (Agrawal *et al.*, 2015) attempted to assist confused learners by

developing a tool that recommends relevant video clips to learners who had submitted posts that indicated learner confusion.

However, while all these studies focused mainly on employing learner sentiment and confusion to achieve different goals, they do not exploit *sentiment* and *confusion indicators* to predict urgent instructor intervention. Therefore, the current research project seeks to use these aspects as metadata in addition to other aspects such as *opinion, question,* and *answer* to predict the urgency posts, which represents a new model in the MOOC instructor intervention prediction field.

4.2.2. Methodology

The main aim of this experiment was to analyse the effect of combining several different dimensions with textual data to predict posts where learners require urgent intervention in a MOOC environment. In the next sub-sections, the pre-processing techniques used with the dataset for this experiment are introduced. In addition, the exploratory statistical analysis and models' architectures are discussed.

4.2.2.1. Dataset

In this study, the Stanford MOOC benchmark posts (Section 3.2.1) dataset was used; it is a large data size featuring 11 courses. The experimental data was prepared as follows: noisy data was cleaned up by removing automated anonymisation (e.g., <nameredac>, <phoneredaci>, <zipredaci>) and removing punctuation and hyperlinks as in (Wei *et al.*, 2017). Case-folding and lemmatisation were also applied (Guo *et al.*, 2019). However, the stopwords were kept, as recommended by (Wise *et al.*, 2017) to improve performance.

4.2.2.2. Exploratory Statistical Analysis

The relationship between the ratio number of non-urgent and urgent posts using the five dimensions (*sentiment, confusion, opinion, question,* and *answer*) for these posts was calculated. For the first two dimensions (*sentiment* and *confusion*), the values to integers were rounded down merely for visualisation purposes (e.g., 1 and 1.5 to 1; 2 and 2.5 to 2; etc.). Then, the mean value (μ) was calculated for each of the different aspects (*sentiment* for non-urgent versus urgent posts; *confusion* with urgency and without; etc.). This aimed to discover if the data were normally distributed; the commonly used Kolmogorov-Smirnov (K-S) test was

applied. As the data were not normally distributed, a Mann-Whitney U test was used to check if the differences were significant (Massimiani *et al.*, 2019). Then, the Bonferroni correction was calculated as multiple comparisons were conducted. Finally, Pearson product-moment correlations were calculated (Cohen *et al.*, 2009) between non-urgent and urgent posts and were then measured over the other dimensions. For the correlation between non-urgent/urgent posts with *sentiment* and *confusion* values, the scale was converted to positive/negative: positive if the value was > 4 and negative otherwise.

4.2.2.3. A Multidimensional Deep Learning: Predictive Intervention Models

The first step was to develop a basic model based on text-only data and then incorporate other dimensions (*sentiment scale, confusion scale, opinion value, question value,* and *answer value*) as numerical features. In general, the text data (learner posts) was trained with a CNN model and the numerical data (multiple dimensions) with a MLP model (see Figure 4.2 below). CNN was selected to classify text by following (Guo *et al.*, 2019) as they reported that TextCNN outperforms LSTM. Note though that the goal was to show the power of the multidimensional approach and not optimise the individual parts of the classifier.

Figure 4.2: Different types of data with different networks.

The data were divided into two distinct sets: one for training and the other for testing (80% and 20%, respectively) using stratified sampling (Farias, Ludermir and Bastos-Filho, 2020). This was to ensure that the training and testing sets had approximately the same distributions of the different classes (non-urgent and urgent), although the dataset has many non-urgent posts. The training set was split into two sets: training and validation (80% and 20%).

4.2.2.3.1. Text Model

As shown in Figure 4.3 (below), in the text model, the first layer is the input layer, with a maximum length = 200, as each post was padded out to a predetermined length (200 words) by following (Guo *et al.*, 2019) to control the length of the input sequence to the model. Then, the

embedding layer reused the pre-trained word embeddings (Word2vec-GoogleNews-vectors-negative300) (Mikolov *et al.*, 2013) and was fine-tuned during training. Word2vec was selected as the pre-trained model, as (Guo *et al.*, 2019) showed that it outperformed GloVe on urgency classification tasks. Next, for the CNN layer, 1D Convolution was applied (128 filters, kernel size of {3,4,5} and Rectified Linear Unit (ReLU) as the activation function) as in (Guo *et al.*, 2019) to derive interesting features, followed by 1D global max pooling to produce final features. Then, for the drop-out layer, a drop-out rate of 0.5 was used as in (Guo *et al.*, 2019) to prevent overfitting. Then, the fully connected layer with the sigmoid as an activation function was used to classify output I as: 1- needs urgent intervention or 0 – no intervention required:

$$I = \begin{cases} 1, & if > .5 \\ 0, & if \leq .5 \end{cases} \tag{4.1}$$

After constructing the model, it was trained using the Adam optimisation algorithm as in (Guo *et al.*, 2019). Binary cross-entropy was used as a loss function because this problem involves binary decisions, and the popular metric *accuracy* was used to report performance. In addition, for a more comprehensive result and to deal with potential majority class bias, the *P*, *R* and *F1* for each class were calculated. In addition, due to the class imbalance, the *BAs* were measured for model evaluation.

4.2.2.3.2. Overall Model (Text Model + Other Dimensions Model)

The overall model is a general model that contains mixed data to predict urgent learner posts. Here, numerical data as features were added in addition to text. As an initial study, the text data was combined with meta-data in one single model; however, the model's performance was unsatisfactory. As the model combines multiple inputs and mixed data, therefore two different sub-models were constructed (see Figure 4.3 below), with the first sub-model being the text-only model.

The second sub-model is a MLP neural network with five inputs that represent the five dimensions (*sentiment, confusion, opinion, question,* and *answer*). Then, these features were added one by one to the MLP model as single inputs (one dimension at a time) to check the individual effect of each particular dimension. The next layer is a hidden layer with 64 neurons. This is followed by a fully connected layer with the sigmoid as an activation function to classify the posts as in the text model. The outputs from these two sub-models were combined via

concatenation to construct the overall model. Finally, a fully connected layer that consists of one neuron with the sigmoid activation function was used at the end of the network to classify the output as in the sub-models.

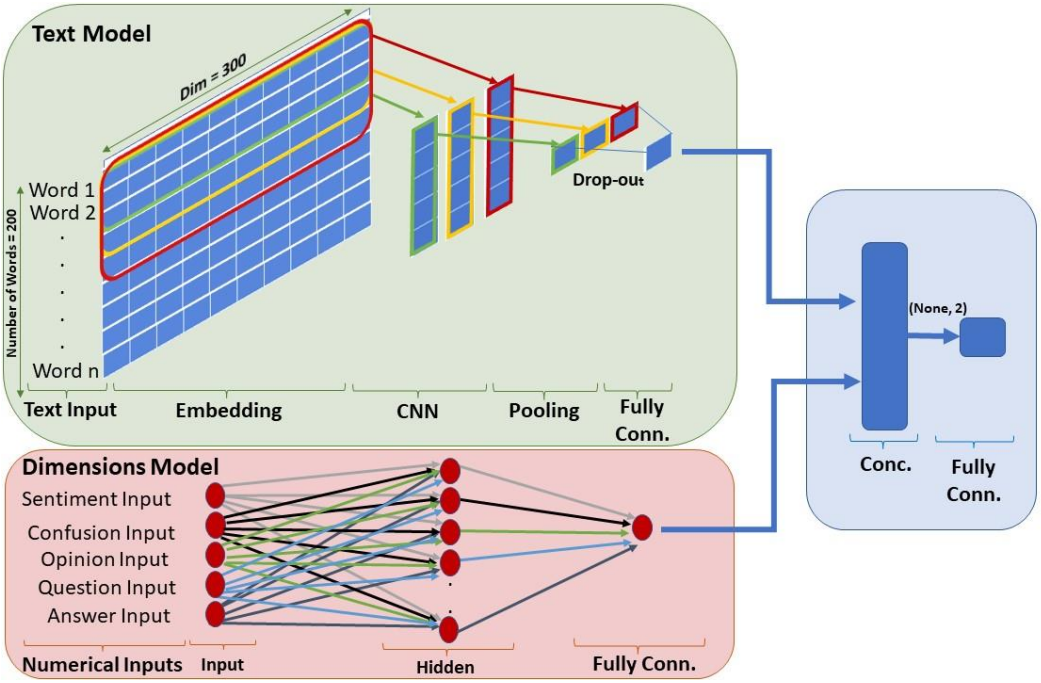

Figure 4.3: Overall model.

After training, McNemar's statistical hypothesis test (McNemar, 1947) was applied to check if the observed differences between any two classifiers were statistically significant. Also, the Bonferroni correction was applied to compensate for multiple comparisons.

4.2.3. Results and Discussion

In this section, the charts and results of the analysis of the relations between non-urgent and urgent posts with different dimensions are presented to address ***RQ1.1***. Then, the results obtained after training each model were reviewed (text only and overall model) to address ***RQ1.2***.

4.2.3.1. Statistical Analysis

The relationship between the rates of non-urgent/urgent posts across the five different dimensions was analysed. As shown in Figure 4.4 (below; left: *sentiment* (1–7)), the number of urgent posts exceeded the number of non-urgent posts in the negative sentiment scale (1–3) and vice-versa: the number of urgent posts was less than that of non-urgent posts on the positive

sentiment scale (5–7). *Sentiment* (4) was interpreted as neutral. To reach this conclusion, the values of (4) and (4.5) on the sentiment scale were compared; there were a higher proportion of non-urgent with a sentiment of (4.5). Figure 4.4 also shows that (right: *confusion* (1–7)) the ratio of non-urgent posts was higher than that of urgent posts for *non-confused* posts, i.e., with a confusion value of between 1–3 in contrast to *confused* posts (5–7). The values (4) and (4.5) for *confusion* were compared as well; here, unlike for *sentiment*, the results showed a higher number requiring urgent attention for the (4.5) value.

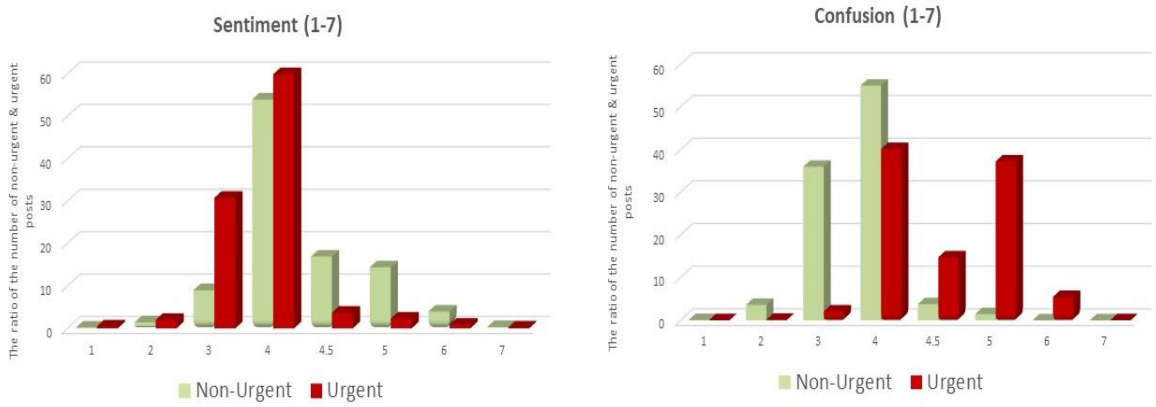

Figure 4.4: The relationship between the ratio of the number of (non-urgent & urgent) posts and sentiment scale (1-7) (left), confusion scale (1-7) (right).

A similar analysis for the remaining dimensions was performed (*opinion, question,* and *answer*), which are binary (Figure 4.5). For *opinion*, most of the posts were non-urgent. For *question*, there were more urgent posts; this highlights that *questions* often represent posts where learners require urgent intervention. In *answer*, in general, most posts are not answered, indicating that most learners do not like to answer their peer's questions; this highlights the importance of instructor intervention. *Answer* posts, as expected normally represent non-urgent posts.

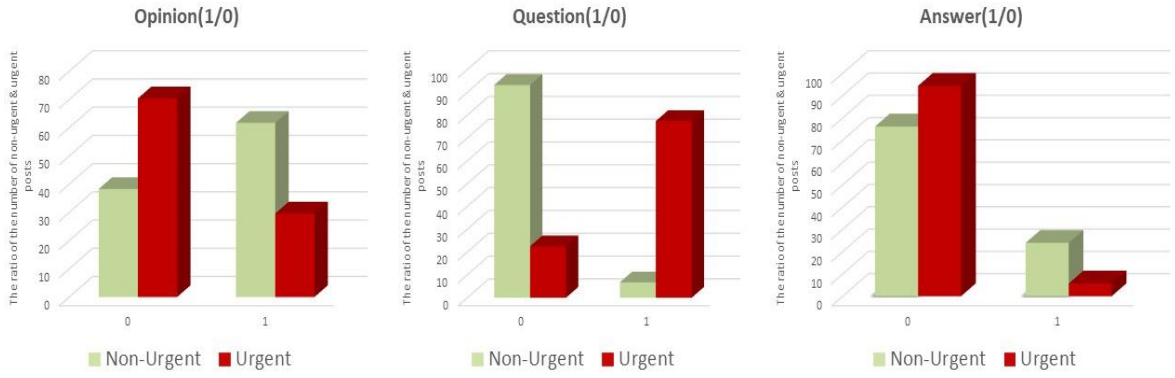

Figure 4.5: The relationship between the ratio of the number of (non-urgent & urgent) posts and opinion (1/0) (left), question (1/0) (middle) and answer (1/0) right.

Next, the averages of the sentiment dimensions were computed: the mean of the *urgency sentiment* was 3.83 and the mean of *non-urgency sentiment* was 4.25 (see Table 4.1 below). Importantly, this difference is statistically significant (Mann-Whitney U test: $p < 0.05$). Then, the same steps were repeated for all dimensions, as shown in Table 4.1. Then, a Bonferroni correction was applied ($p < 0.01$) indicating that the set of all comparisons is significant.

Table 4.1: Average different dimensions with (non-urgent/urgent).

Dimension	Mean (non-urgent)	Mean (urgent)	P
Sentiment	4.25	3.83	$p < 0.01$
Confusion	3.75	4.59	$p < 0.01$
Opinion	0.61	0.29	$p < 0.01$
Question	0.06	0.77	$p < 0.01$
Answer	0.23	0.05	$p < 0.01$

Next, as explained in the methodology, the dimensions were compared. Correlation results are shown in Table 4.2 (below), suggesting a strong correlation between *urgency* and *confusion* and between *urgency* and *question*.

Table 4.2: Correlations between non-urgent/urgent posts reflected on different dimensions.

Dimension	Non-urgent/urgent
Sentiment	-0.244
Confusion	0.571
Opinion	-0.253
Question	0.691
Answer	-0.177

4.2.3.2. A Multidimensional Deep Learning: Predictive Intervention Models

Table 4.3 (below) reports the performance of every trained model as a comparison between different inputs. The average *Acc* and *P*, *R,* and *F1* per every class category (0 as non-urgent) and (1 as urgent) were calculated. In addition, as the dataset for this study is imbalanced, *BAs* were calculated to measure the performance of models. The results revealed that adding all features as other dimensions (*sentiment scale, confusion scale, opinion value, question value*

and *answer value*) in addition to texts increases classifier performance in *BA* for classifying urgent posts. In addition, *R* for the minority class as '*Recall'* is a more crucial metric because it ensured the capture of all urgent posts.

Table 4.3: The performance results for different inputs (Acc, P, R, F1 %), **Bold:** best performance of BA and best performance of R for class 1 (Urgent).

Inputs	Acc	Non-urgent (0)			Urgent (1)			BA
		P	R	F1	P	R	F1	
Text	0.88	0.90	0.95	0.93	0.73	0.56	0.64	0.76
Text + all features	0.91	0.93	0.97	0.95	0.84	**0.67**	0.74	**0.82**
Text + sentiment	0.88	0.91	0.95	0.93	0.73	0.57	0.64	0.76
Text + confusion	0.87	0.90	0.95	0.92	0.73	0.52	0.61	0.74
Text + opinion	0.87	0.90	0.95	0.92	0.71	0.57	0.63	0.76
Text + question	0.90	0.91	0.98	0.94	0.86	0.59	0.70	0.78
Text + answer	0.89	0.92	0.95	0.93	0.73	0.64	0.69	0.79

Next, the differences were checked if these were statically significant (McNemar's test: $p < 0.05$) as shown in Table 4.4 (√ indicates a statistically significant difference in the disagreements between the two models while × signifies a statistically non-significant difference in the disagreements between the two models). The results confirm that there are differences between the (text + all features) model and the other models as they have different proportions of errors. Then a Bonferroni correction was used between (text + all features) and different models; and found that $p < 0.008$, meaning the set of all comparisons is significant.

Table 4.4: McNemar's test results between models.

	Text	Text+ all features	Text+ sentiment	Text+ confusion	Text+ opinion	Text+ question	Text+ answer
Text							
Text+all features	√						
Text+ sentiment	√	√					
Text+ confusion	×	√	√				
Text+ opinion	×	√	×	×			
Text+ question	√	√	√	√	√		
Text+ answer	√	√	√	√	√	√	

4.3. Plug & Play with Deep Neural Networks

The second experiment proposed a classification model for identifying when a given post needs intervention from an instructor based on a hybrid DNN with different levels of inputs. These deep models cannot deal directly with word content, working instead on word embeddings, to produce word vectors (Rani and Kumar, 2019) as clarified in detail in Section 2.2.2. Currently, BERT has become popular as a word embedding tool because it produces word embeddings based on their context — unlike existing word embedding models, which embed each word in a single vector, without taking into account the different contexts of use (Mazari, Boudoukhani and Djeffal, 2023). In addition, (Khodeir, 2021) showed that using BERT to represent words improved performance in detecting urgent posts.

This research project was inspired by the work of two researchers: (Guo *et al.*, 2019) and (Khodeir, 2021). Guo et al., (2019) used DL models combining (CNN + GRU) that extract semantic information and structural information to detect posts that needed urgent responses by applying attention to develop a hybrid character/word neural network. Meanwhile, Khodeir (2021) utilised a multi-layer (Bi-GRU) based on BERT as an embedding layer to classify learners' urgent posts. She used BERT as a word embedding to represent words in context. However, the novelty of this research project is the application of using a plug & play approach in a DL model based on word2vec or BERT for different word-based or word-character-based inputs to provide a more comprehensive view of constructing models designed to predict the urgency of intervention need in MOOC forums.

Thus, constructing different hybrid (deep) neural networks that integrate various DNNs with *word-based* or *word-character-based inputs* using two different methods to represent word input (word2vec or BERT) addresses the second goal of this chapter as illustrated in the following RQs:

- ***RQ1.3:*** *What is the preferable combination between different deep learning models to construct the best predictor model amongst them to identify posts that need instructor intervention?*
- ***RQ1.4:*** *Do word-character-based approaches outperform word-based approaches for the post urgency problem and is this different when using BERT for word embedding, compared to more traditional models (e.g., word2vec)?*

The key contributions of this experiment are the following: (i) analysing and exploring for the first time MOOC post data in terms of length (number of words and characters per post), (ii)

constructing different simple and hybrid deep learning models by applying plug & play techniques to establish good combinations in terms of performance, (iii) applying an attention mechanism that considers word-based input only by using a separate attention score for every word according to their importance, (iv) for the first time, showing the quality of BERT and its sufficiency when using word-based input only without adding word-character-based input.

4.3.1. Related Work on Towards Plug & Play: Combinations in Deep Learning

For text classification, CNNs are known to be better at extracting local and position-invariant features while RNNs are effective at modelling units in sequence due to the latter's different architecture — since CNNs are hierarchical, while RNNs are sequential (Yin *et al.*, 2017).

Whilst DL has been proven to be performant, recently, a great amount of research has focused on combining two or more types of DNNs to produce a more effective combined model. This is specifically prominent in the computer vision field (Zhao, Han and Xu, 2018; Ullah *et al.*, 2017; Tsironi *et al.*, 2017). Recently, many researchers have also applied combinations of different DNN models to text analysis and classification tasks. For example, (Wang, Jiang and Luo, 2016) introduced a technique to combine CNN and RNN models to perform sentiment analysis on short texts; their results showed that this approach leads to improvements in accuracy. (Lai *et al.*, 2015) proposed a RCNN which involved applying a RNN to capture contextual information followed by a CNN to obtain the final representation for sentence classification.

Another study by (Zhang, Robinson and Tepper, 2018) combined CNN and GRU to detect hate speech on Twitter. Their model outperforms existing models on six out of seven datasets with F1 scores of between 1%–13%. Also, as mentioned in Chapter 2, (Wei *et al.*, 2017) proposed a framework for transfer learning based on CNN and LSTM and showed the effectiveness of their model on the Stanford MOOCPosts dataset.

All previous studies have combined different types of DNNs for word-level inputs only. In addition to combining different layers at the word-level, other researchers have combined characters with words as input. For example, (Liang, Xu and Zhao, 2017) used word-level and character-level representation as input to classify informal text. Their results are competitive in relation to other studies on the SemEval-2010 Task8 and outperform existing models on the KBP-SF48 dataset by achieving better learning of character features. Also, (Yenigalla *et al.*,

2018) proposed a method to integrate both character- and word-based models for text classification to address the problem of unseen words in word-based models. Their results showed that this approach resulted in accuracy being improved.

(Guo *et al.*, 2019) proposed an attention-based model that concatenates word-level and character-level representation to extract semantic and structural information. They clarified that the MOOC posts contain a lot of noise; this problem can be overcome by adding character-level input to capture this special information.

As mentioned earlier, this experiment was built based on Guo et al.'s (2019) research; however, they used semantic and structural information to classify posts that need intervention. Semantic information was learned by applying a CNN while structural information was learned by using the last hidden state of the GRU. Then, they used an attention mechanism to learn the weights of the word-character representations. In contrast, in the current experiment, a CNN was used to extract local features and investigate different types of RNNs (plug & play) to model units in sequence by returning all the hidden states to the attention mechanism to allocate weights to every word. The attention mechanism was applied only to word-based input before character-based input was added to improve noisy data such as misspellings. In addition, the words were represented using two methods: (i) BERT as in (Khodeir, 2021) in contrast to (ii) (Guo *et al.*, 2019) who used word2vec (google-news Vectors).

4.3.2. Methodology

This experiment seeks to identify posts that need urgent intervention by using a plug & play technique with a multi-layered DNN. The pre-processing of the dataset for this experiment is introduced in the following sub-section. This is followed by an exploration of the dataset in terms of length (number of words and characters per post). Finally, it proposes the architectures of the predictive intervention models.

4.3.2.1. Dataset

In this experiment, the Stanford MOOC posts dataset (see Section 3.2.1) was also used. Further pre-processing was applied as mentioned above, including data cleaning and all automated anonymisation tags were removed (e.g., <zipredaci>, <phoneredaci>) (Wei *et al.*, 2017). Next, the text was converted to lowercase. As in the previous experiment, the stopwords were kept because, as (Wise *et al.*, 2017) noted, model performance can improve if stopwords are

included. Next, the final input was prepared by adding the name and the domain of the course to the text input. This approach followed (Guo *et al.*, 2019) who argued that to understand the information contained in posts, one should connect the course and domain information of the post to the text of the post. Thus, 13 posts with an empty course name were removed, leaving 29,584 posts.

4.3.2.2. Exploring the Dataset

As an essential step, the dataset was explored and analysed to understand the data. Here, the analysis focused on the number of words and characters in all posts; the use of DL models requires that the length of the input sequence to these models should be specified. As mentioned in Section 4.3.2.1, the input to these models was the text in the posts and the name and domain of the course; therefore, this information had to be considered in the following calculations. As shown in Figure 4.6 (below), A is the distribution of the number of words per post (mean = 60.36 words, minimum = 2 words; maximum = 498 words). B is the distribution of the number of characters per post (mean = 380.28 characters, minimum = 29 characters; maximum = 2556 characters).

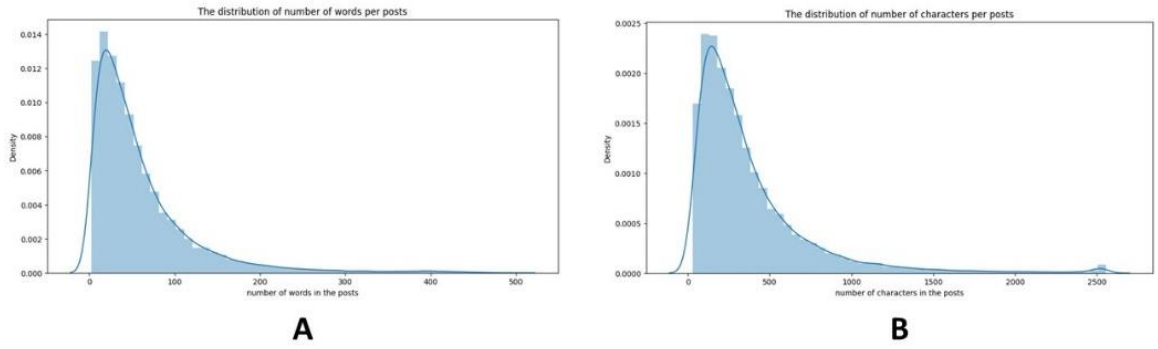

Figure 4.6: Distributions of posts: A = number of words per post – B = number of characters per post.

To understand the behaviour of learners and how many words they write when they need urgent intervention (Label = 1) or not (Label = 0), the representation of the number of words per label was visualised (see Figure 4.7 below). To check if any statistically significant differences between the two populations were present, the Mann-Whitney U test was used. It was found that $p < 0.05$, meaning that statistically significant differences were evident in terms of the length of posts (number of words).

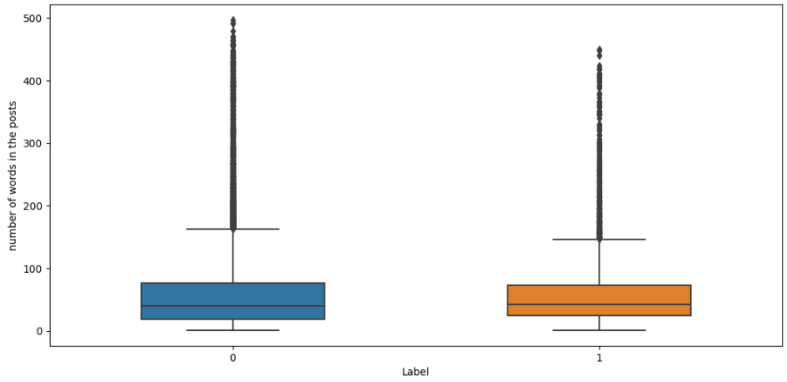

Figure 4.7: Box plot for number of words per post written by learners needing intervention (Label = 1) or not needing intervention (Label = 0).

Following this, posts written by learners who need urgent intervention were analysed. As depicted in Figure 4.8 (below), A is the distribution of the number of words per urgent post (mean = 59.63 words; minimum = 2 words; maximum = 450 words). On the right side, B shows the distribution of the number of characters per urgent post (mean = 372.12 characters; minimum = 32 characters; maximum = 2556 characters).

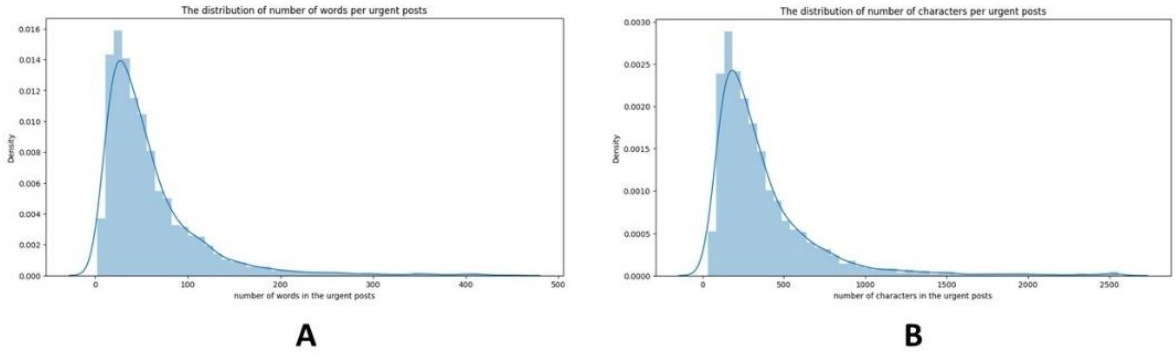

Figure 4.8: Distributions of urgent posts: A = number of words per urgent posts – B = Number of characters per urgent posts.

Finally, to discover which words were most frequently used by learners in urgent posts, the top 30 most frequent words in these posts were calculated after removing stopwords; see Figure 4.9 (below).

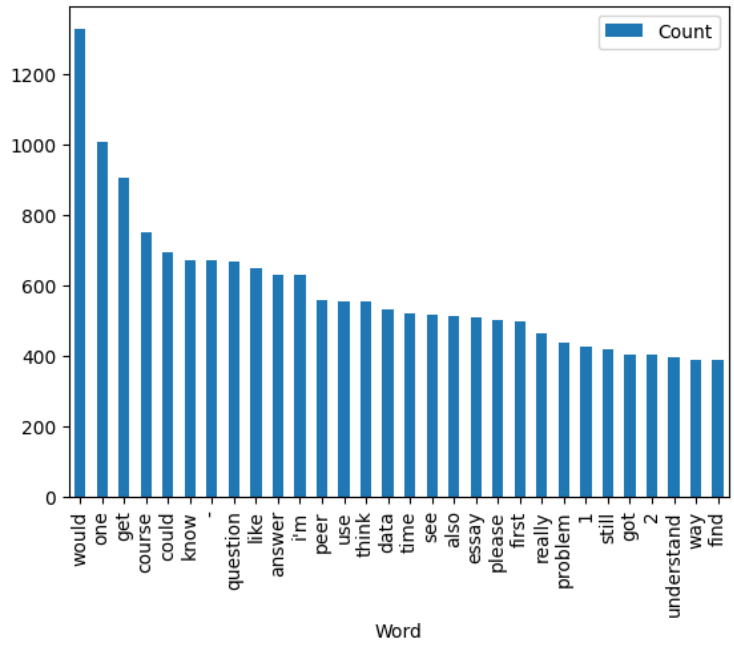

Figure 4.9: The top 30 frequency words in urgent posts.

4.3.2.3. Plug & Play: Predictive Intervention Models

The input of these models is posts from each learner; the output is the classification (if the post needs intervention or not) according to its urgency (binary prediction). Words were presented in numerical form (word embedding) using (i) *word2vec* as in (Guo *et al.*, 2019), which converts words into vectors that depict semantics; and (ii) *BERT* as in (Khodeir, 2021), which generates contextual representations for each word. Two different training models were implemented: (i) using word-based input, and (ii) using character-based input in addition to word-based input to configure what is called in this research project *word-character-based input.* Figure 4.10 (below) illustrates the general architectures of these two cases (word-based input and word-character-based input). Word2vec or BERT were selected as a word embedding tool, a CNN was used to extract local complex context features, a RNN was used to model units in sequence and learn feature structures, and an attention mechanism was used to give higher weight to keywords. In word-character-based models, in addition to the DNN layers for words, a CNN was applied to select the features for characters.

The dataset was split into training, validation, and testing as follows: training and testing (80% and 20%, respectively); it was divided by using stratified sampling (Farias, Ludermir and Bastos-Filho, 2020) to select a sample that is representative of different classes (*urgent intervention needed* and *no intervention needed*). After that, the training data was split into training and validation sets (80% and 20%, respectively).

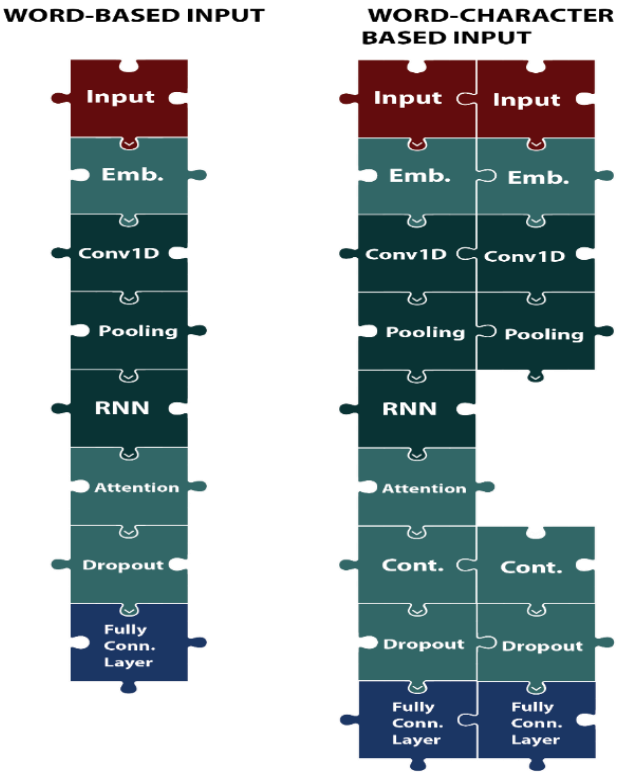

Figure 4.10: Deep learning as a puzzle: general architectures for two cases (word-based input and word-character-based input).

4.3.2.3.1. Word-Based Input

Different basic and combined DL models were constructed to select features by applying plug & play approaches (i.e., CNN, RNN (LSTM, Bi-LSTM, GRU and Bi-GRU)) as they represent modifications of RNN. These models are based on word2vec or BERT. In the following, the general word-based architecture is explained; however, during the implementation, some layers were removed and other layers were added, such as different types of RNN and attention layers.

In word2vec, (Word2vec-GoogleNews-vectors-negative300) was applied as in (Guo *et al.*, 2019) as mentioned before, they showed this renders better results than using GloVe on the Stanford dataset. In addition to pre-trained embedding, the word embeddings were trained during the NN training (fine-tuning).

In BERT, the BERT tokeniser was used to tokenise sentences into tokens using *bert-base-uncased*; this means that there is no difference between a letter written as a capital or lowercase. The sentence is split into tokens which represent the original words; BERT's tokenizer uses a WordPiece model, which breaks down words into subwords. Also, special tokens [CLS] and [SEP] were added; [CLS] is inserted at the start of the text and [SEP] is inserted at the end; or if there are more sentences, [SEP] are used to separate these sentences (Clark *et al.*, 2019).

The maximum length of each post was constrained by following (Khodeir, 2021) to 512 tokens since 512 tokens is the maximum model input size for BERT. Thus, sequences of less than 512 were padded out with zeroes and the rest which contained more than 512 tokens were trimmed down to ensure that each sequence has the same length.

As Figure 4.10 (above) shows, the first layer (the input layer), followed by the embedding layer which maps the words onto vectors. The output from these layers is passed onto the 1D convolution layer as an input (with 128 filters, a kernel size of {3,4,5}, and ReLU as the activation function) as in (Guo *et al.*, 2019) to derive interesting features. Then, the produced features are further compressed by using a pooling layer (max pooling). These features feed into one of a set of different RNN layers (LSTM, Bi-LSTM, GRU and Bi-GRU) with 128 hidden units, which helps to determine the relationship between words.

The next layer is the attention layer, which learns the weighting for each word. The attention with a context mechanism (Yang *et al.*, 2016) was used. That is, through a series of mathematical formulas, a context vector is randomly initialised and multiplied by each word, to generate the importance score.

Next, for the following drop-out layer, a drop-out rate of 0.5 was used as recommended by (Guo *et al.*, 2019) by randomly dropping out nodes during the training phase and using an early stopping mechanism to alleviate overfitting. Then, the fully connected layer is used to classify the output (1 = needs urgent intervention or 0 = no intervention needed) by calculating the probability (*PRO*) thus:

$$PRO = \begin{cases} 1, & if > .5 \\ 0, & if \leq .5 \end{cases} \tag{4.2}$$

After the model was created based on different word embeddings, it was trained by using the Adam optimiser, as in (Guo *et al.*, 2019). Binary cross-entropy was used as a loss function because the problems involve binary decisions. The batch size was set to 64 and automatic early stopping was employed to stop training after five epochs when there no progress in performance was evident.

4.3.2.3.2. Word-Character Based Input

In the second case, characters were added in addition to words. The length of each input post was set at 1024 as recommended by (Guo *et al.*, 2019) because most posts had fewer than 1024 characters. Next, the characters were encoded and character embedding was applied as per

Johnb30/py_crepe code on GitHub, which is a reimplemented version of the Crepe character-level convolutional neural net model originally shown in (Zhang, Zhao and LeCun, 2015).

To select the most important features, a standard 1D convolutional and pooling layer was selected, respectively. Convolution with a filter = *128* and a kernel size of *{5,7,9}* and the ReLU activation function were used, following the recommendations of (Guo *et al.*, 2019). The selected features were concatenated with the features from the word-based input and the rest followed the processing approach explained in the word-based input section.

4.3.3. Results and Discussion

This section presents the results obtained after training every model to answer both ***RQ1.3*** and ***RQ1.4***. The average *ACC* and *P*, *R* and *F1* for each class and the *BA* are reported for the word-based approach using the various DL models (first row); then it reports the combined results (word- and character-based results) (second row) as shown in Table 4.5 (below). In this research project, the models were compared based on their respective *BA* scores as this is a widely used metric for binary classification in imbalanced datasets (Alamri *et al.*, 2021).

From these results, the *BA* scores for models using BERT for word embedding outperformed all the models based on word2vec. That means it is better to represent words using BERT. The best value from all the models in terms of *BA* is **0.875** and for *R* for (1) class is **0.81** in CNN + LSTM + Attention model based on BERT at word-level. The interesting point is that a Bi-LSTM performed worse than an LSTM; this is dependent on the nature of the data and task as utilising bidirectional information might introduce noise and thus hinder performance.

In terms of the word-based vs word-character-based approaches, it was observed that if word2vec is used as a word embedding, the word-character method often outperforms word-based ones (these improvements are statistically significant using McNemar's test: $p < 0.05$, as shown in Table 4.5 below). In contrast, for models using BERT for word embedding, there is no improvement between the use of the different approaches i.e., word-only and word-character-based. Also, the difference between these models is not always statistically significant which means BERT is good enough to represent words without any support.

Table 4.5: The performance results of word2vec and BERT for word embedding for word-based and word-character-based approaches for the different models (Acc, P, R, F1, BA %) and P.V value, **Bold:** best performance of BA and best performance of R for class 1 (Urgent), *Italic:* statistically significant.

Model	Word Embedding	Level	Acc	Non-urgent (0)			Urgent (1)			BA	P.V
				P	R	F1	P	R	F1		
CNN	Word2vec	Word	0.87	0.89	0.96	0.92	0.74	0.51	0.60	0.732	*P ≤ 0.05*
		Word+Char	0.90	0.92	0.95	0.94	0.77	0.66	0.71	0.807	
	BERT	Word	0.91	0.93	0.96	0.95	0.82	0.69	0.75	0.826	P > 0.05
		Word+Char	0.91	0.92	0.97	0.95	0.84	0.66	0.74	0.815	
CNN + GRU	Word2vec	Word	0.87	0.90	0.95	0.92	0.72	0.55	0.62	0.749	*P ≤ 0.05*
		Word+Char	0.91	0.93	0.96	0.95	0.82	0.68	0.74	0.823	
	BERT	Word	0.91	0.95	0.94	0.95	0.76	0.79	0.77	0.865	*P ≤ 0.05*
		Word+Char	0.92	0.95	0.96	0.95	0.80	0.76	0.78	0.860	
CNN + Bi-GRU	Word2vec	Word	0.88	0.92	0.93	0.93	0.68	0.67	0.68	0.798	*P ≤ 0.05*
		Word+Char	0.90	0.92	0.96	0.94	0.80	0.65	0.71	0.803	
	BERT	Word	0.92	0.94	0.96	0.95	0.82	0.74	0.78	0.851	*P ≤ 0.05*
		Word+Char	0.93	0.94	0.97	0.95	0.84	0.75	0.79	0.856	
CNN + GRU + Attention	Word2vec	Word	0.88	0.92	0.93	0.93	0.69	0.67	0.68	0.800	*P ≤ 0.05*
		Word+Char	0.91	0.93	0.96	0.95	0.81	0.70	0.75	0.829	
	BERT	Word	0.92	0.95	0.96	0.95	0.80	0.76	0.78	0.859	P > 0.05
		Word+Char	0.92	0.94	0.97	0.95	0.83	0.74	0.79	0.854	
CNN + Bi-GRU + Attention	Word2vec	Word	0.88	0.90	0.95	0.93	0.74	0.56	0.64	0.755	*P ≤ 0.05*
		Word+Char	0.91	0.93	0.96	0.94	0.80	0.67	0.73	0.816	
	BERT	Word	0.92	0.95	0.94	0.95	0.77	0.80	0.78	0.872	P > 0.05
		Word+Char	0.92	0.95	0.96	0.95	0.81	0.78	0.80	0.868	
CNN + LSTM	Word2vec	Word	0.81	0.81	1.00	0.90	0.00	0.00	0.00	0.5	*P ≤ 0.05*
		Word+Char	0.91	0.92	0.97	0.95	0.82	0.66	0.73	0.814	
	BERT	Word	0.92	0.95	0.94	0.95	0.77	0.79	0.78	0.869	*P ≤ 0.05*
		Word+Char	0.92	0.95	0.95	0.95	0.80	0.78	0.79	0.869	
CNN + Bi-LSTM	Word2vec	Word	0.88	0.89	0.96	0.93	0.77	0.51	0.61	0.738	*P ≤ 0.05*
		Word+Char	0.90	0.93	0.95	0.94	0.77	0.69	0.73	0.821	
	BERT	Word	0.92	0.94	0.96	0.95	0.81	0.76	0.78	0.857	P > 0.05
		Word+Char	0.92	0.95	0.96	0.95	0.81	0.77	0.79	0.865	
CNN + LSTM + Attention	Word2vec	Word	0.88	0.92	0.94	0.93	0.71	0.65	0.68	0.795	*P ≤ 0.05*
		Word+Char	0.90	0.92	0.97	0.94	0.83	0.62	0.71	0.794	
	BERT	Word	0.92	0.95	0.94	0.95	0.77	**0.81**	0.79	**0.875**	*P ≤ 0.05*
		Word+Char	0.92	0.95	0.95	0.95	0.80	0.79	0.80	0.874	
CNN + Bi-LSTM + Attention	Word2vec	Word	0.88	0.91	0.95	0.93	0.73	0.61	0.66	0.777	*P ≤ 0.05*
		Word+Char	0.91	0.92	0.97	0.95	0.83	0.66	0.74	0.815	
	BERT	Word	0.92	0.95	0.95	0.95	0.79	0.78	0.78	0.863	P > 0.05
		Word+Char	0.92	0.95	0.95	0.95	0.78	0.79	0.78	0.868	

To evaluate the proposed model and compare it with others in the literature, the best model performance was compared with the state-of-the-art model **(Guo *et al.*, 2019)** with the same label (urgent>4). Guo et al. (2019) used the Stanford dataset and applied three different methods to split the data into training and testing sets. However, the author believes that it is not valid to compare the current model with those proposed by other researchers even if the same data with a random split is used: the same data in both the training and testing sets should be used. Thus, one of the methods called Group C that Guo et al. (2019) used to split data is suitable for comparison with the current model which is drawn from a specific MOOC domain (the humanities) for testing while the data from the other two MOOC domains (medicine and education) are kept for training. This ensures that the same data occurs in both datasets. Also, using this method of performing cross-domain (i.e., not testing a model on the domain in which it was trained) is a useful way of evaluating models.

Therefore, this research project conducted the experiment with the best performing model **CNN + LSTM + Attention** based on BERT at word-level by splitting the data following the Group C as in Guo et al. (2019). Table 4.6 (below) reveals that the proposed model (**CNN + LSTM + Attention**) outperforms the state-of-the-art model in *R* and *weighted-F1*. Here, the *weighted-F1* value was calculated for comparison purposes.

Table 4.6: Comparison between the proposed model (CNN + LSTM + Attention (word)) and Guo et al.'s (2019) state-of-the-art model. **Bold:** Best performance in BA and best R for class 1 (Urgent).

Model	Non-urgent (0)			Urgent (1)			Weighted F1
	P	R	F1	P	R	F1	
Guo et al.'s (2019)	0.907	0.945	0.926	0.807	0.731	0.767	0.884
CNN + LSTM + Attention (word)	0.93	0.94	0.93	0.75	**0.74**	0.74	**0.894**

4.4. Epilogue

Identifying the need for instructor intervention is a crucial issue in MOOC environments. Many researchers have tried to predict when an intervention is needed in MOOC post forums by implementing different prediction models which have rendered different levels of performance. In this chapter, the problem of identifying when instructors should intervene in a particular post has been tackled by implementing two different experiments. In the first experiment, a multidimensional *post*-based learner model was developed by exploring DL approaches. Specifically, it compared text-based models with enriched models with five different dimensions (*sentiment, confusion, opinion, question, and answer*). The relationships between

urgent post rates and these dimensions were also observed. The results show that learners' negative feelings, misunderstandings, lack of desire to express an opinion, number of questions, and decreasing number of answers increase in learners in need of urgent intervention, possibly due to the psychological effects of stress. The contributions of this research project include showing that adding these dimensions as features, in addition to text, leads to better predictive performance in DL models. Moreover, a new architecture based on sub-models was constructed to train this multidimensional, mixed data.

In the second experiment, this research project has explored MOOC posts needing urgent instructor attention and intervention (or not), by analysing the textual contents of learners' posts and information about the related MOOC courses. To reach this goal, the current study attempts to discover the best way of constructing DL models, by using different inputs (word-based or word-character based) based on word2vec or BERT as word embedding. Then, a combination of models was presented by applying the plug & play technique. This concretely means adding different inputs, stacking multiple layers, connecting layers, etc. The conclusion is that using BERT for word-embedding is more effective as a stand-alone method without the need for any addition of character-based input.

The next chapter analyses text posts to identify topics and extract urgent language. This is the first time the language of the need for urgent intervention in MOOCs was obtained.

CHAPTER 5: ANALYSING TEXT POSTS USING TOPIC MODELLING TO EXTRACT URGENT LANGUAGE

5.1. Prologue

Discussion forums on MOOCs are a major communication tool between learners and instructors (Onah, Sinclair and Boyatt, 2014b), generating large amounts of posts which are exchanged as unstructured textual content. With the increasing number of text posts from learners communicating in MOOCs (Tucker, Pursel and Divinsky, 2014), it is very challenging, effort-intensive, and time-consuming for an instructor to monitor all the available posts and then detect and respond to those learners who need urgent intervention (Almatrafi, Johri and Rangwala, 2018). Moreover, learners may inadvertently make posts that may appear to be more urgent than they are. As clarified previously recent solutions based on supervised ML (Khodeir, 2021; Guo *et al.*, 2019; Sun *et al.*, 2019) have achieved remarkable performance. However, whilst these approaches have focused on identifying urgent posts, they have not sought to extract urgency-related topics that learners mention in their posts or identify urgent language in posts.

This chapter aims to analyse learner posts in MOOCs forums from the perspective of urgency, and, importantly, to extract the language used by learners to express urgency via the use of an automatised approach. In this research project, urgent language was defined as the most frequently encountered words and phrases that learners use in their posts to signal the need for urgent attention and intervention. To identify words (topics), topic modelling using the widely used LDA (Blei, Ng and Jordan, 2003) was applied.

The aim was to not only detect urgent intervention but also to establish if there is a way to make language signalling the need for urgent intervention explicit by providing a visual representation for instructors. As most people are visual, creating visual aids for instructors (or

learners) is expected to help instructors decide *when* and *where to intervene*, and help learners to potentially use language that clearly signals their need for assistance. Thus, the following RQs were formulated:

- ***RQ2.1:*** *Can the language of urgency be detected from learners' posts?*
- ***RQ2.2:*** *Can the language of urgency be visualised simply and intuitively?*

The main contributions of this chapter are: (i) to the best of the researcher's knowledge, this is the first work to automatically detect the language of urgency in MOOC posts by modelling text posts and relating them to urgent posts; (ii) showing that the majority of urgent posts for specific urgent topics begin with new threads; (iii) designing an urgency visualisation tool for instructors, and providing suggestions for learners on how to signal the need for intervention.

5.2. Related Work

To date, topic analysis, modelling, and visualisation in the context of instructor intervention in MOOCs has received little attention from researchers. In the next sub-sections, the most important studies on MOOC-related topic modelling and visualisation will be presented and discussed.

5.2.1. Topic Modelling in MOOCs

With the emergence of topic modelling, several studies have focused on modelling posts from discussion forums in MOOCs using LDA. (Atapattu and Falkner, 2016) used LDA to identify the main weekly topics of discussion in MOOCs and labelled them to provide a framework that can be effectively used to locate and navigate informational need. (Ezen-Can *et al.*, 2015) applied an unsupervised algorithm to group similar posts, and then found the top topic words using LDA to better support learner outcomes. (Robinson, 2015) used LDA to extract topics that learners mentioned in their discussions on Cartograph. They revealed the most popular places learners talked about in class to improve the future development of the course. In fact, LDA has become one of the most popular and widely used topic modelling tools.

Thus, the current study also used LDA, albeit to identify the words learners use when they need urgent intervention, as further explained in Section 5.3.2.

5.2.2. Visualisation in MOOCs

In many recent visualisation-based works related to MOOC discussion forums, researchers aimed at assisting instructors. For example, recently, (Almatrafi and Johri, 2022) proposed an experimental approach to improve MOOCs based on summaries of learners' opinions about the course extracted from discussion forums. The visual results were meant to allow both expert and non-expert gain an understanding of different aspects of the course. (Wong, 2018) constructed a visual analytics tool (MessageLens) using different visualisation tools to assist MOOC instructors in better understanding forum discussions from three perspectives: discussion topics, learner attitudes, and learner communication.

Here, visualisation for instructors was employed to help them understand topics that learners use in their discussions on a specific course and colour-code posts based on these topics to assist both instructors and learners (see Section 5.3.3).

5.3. Methodology

In this section, as shown in Figure 5.1 (below), the framework design of an analysis model to identify urgent language is explained and visualised. The dataset used in this research project sourced from the Stanford MOOC platform was analysed to select the appropriate course as this research project seeks to identify topics based on the course level. Then, how the textual data of posts were processed to provide input for the analysis model is described. After that, the unsupervised approaches used are discussed to analyse and mine learners' textual posts and extract useful urgent-language patterns. Also, several visualisation aids are introduced, mainly for instructors, but also for learners.

Figure 5.1: An analysis model of urgent language and an instructor visualisation aid for learner posts.

5.3.1. Dataset

The Stanford MOOCPost dataset was used for this research project. As an initial step, 13 posts with an empty course title in the humanities course type in the dataset were removed. The number of posts with the percentage of urgent and non-urgent numbers of posts respectively in each course are reported in Table 5.1 (below).

Table 5.1: The number of urgent and non-urgent posts for all courses in the Stanford MOOCPosts dataset. **Bold:** Large number of posts and large percentage of posts that represent urgent intervention.

	Course	Posts number	Non-Urgent number	Urgent number
Humanities/Sciences	WomensHealth	2141	1863 (87.0%)	278 (13.0%)
	StatLearning	3029	2191 (72.3%)	838 (27.7%)
	Stats216	327	204 (62.4%)	123 (37.6%)
	Environmental_Physiology	2467	2048 (83.0%)	419 (17.0%)
	Econ-1	1583	1249 (78.9%)	334 (21.1%)
	Econ1V	160	150 (93.8%)	10 (6.2%)
Medicine	Statistics_in_Medicine	3320	2276 (68.6%)	1044 (31.4%)
	MedStats	1218	802 (65.8%)	416 (34.2%)
	SciWrite	**5181**	3407 (65.8%)	1774 (**34.2%**)
	Managing_Emergencies_What_Every_Doctor_Must_Know	279	231 (82.8%)	48 (17.2%)
Education	How_to_Learn_Math	9879	9559 (96.8%)	320 (3.2%)

Next, a particular course (SciWrite) in the medical field was chosen as a case study for analysis. The reason for selecting this course is because it contains a large number of posts (5181) and a high percentage of posts that require urgent intervention (34.2%) compared to the rest of the courses as shown in Table 5.1. Therefore, it provides good data on urgent posts.

To prepare the SciWrite course dataset, various preprocessing was performed. This included splitting and tokenising the sentences into a list of words, then performing cleaning, such as removing unnecessary parts, including those that may lead to the identification of the learners (emails, some characters, quotes, anonymisation). Next, tokens were converted into lists. Then, stopwords were removed. Afterwards, the phrase models based on bigrams (two words often appearing together in the post) and trigrams (three words appearing together) were built, after which lemmatisation was applied; finally, stopwords were removed once more after lemmatisation.

5.3.2. Extracting Urgent Language

To extract urgent language as a starting point for the automated language analysis (text-document modelling), words from forum posts were clustered into different topics based on the unsupervised statistical model (LDA), as explained in Section 5.3.2.1. This was followed by associating topic lists and trending terms within urgent posts as potentially useful indicators for identifying and giving an overview of urgent language. The next sub-sections explain the follow-up steps.

5.3.2.1. Topic Modelling (LDA) Setup

In NLP, topic modelling is an unsupervised technique commonly used for analysing a collection of textual documents (Xiong and Litman, 2013). It offers a convenient way to classify, extract, and discover hidden topics from the keywords associated with each topic (Wang, Wen and Rosé, 2016) and recognise latent patterns from unstructured text (Sharma and Sharma, 2017; Jacobi, Van Atteveldt and Welbers, 2016).

The generative probabilistic topic-modelling model LDA has emerged as one of the most popular algorithms; it is utilised for modelling texts to extract topics from unlabelled texts (Geng *et al.*, 2020) as a set of documents. LDA was originally proposed by (Blei, Ng and Jordan, 2003) to overcome some limitations of prior models, as mentioned in (Huang and Wang, 2021) such as semantic ambiguity.

The LDA model assumes that each document features a mix of different topics; a topic is a theme comprised of a collection of words that frequently appear together (Nanda *et al.*, 2021). To explain, the model processes a document term matrix by supposing that each document (d) contains different topics (t) as a probability distribution $p(t|d)$. In turn, each topic (t) contains different words (w), with t a probability distribution over w $p(w|t)$ (Prabhakar Kaila and Prasad, 2020; Nikolaev *et al.*, 2019). The input of this model is a bag-of-words model (Curiskis *et al.*, 2020) and the output is represented by different topics, each with lists of terms (words) which are ordered from having the highest relevance to the topic to the lowest (Wong, Wong and Hindle, 2019).

In this research project, the Gensim package written in Python was used to train the LDA model. This model only needs feeding with the number of topics (k), which is a free parameter that can be tuned (Abebe *et al.*, 2019) and can be considered as a hyperparameter. It is a challenging task (Ni Ki *et al.*, 2021) as there is no optimal way to choose this number. Choosing a low number of topics tends to produce more general output whereas using a higher number of topics provides more detailed output (Asmussen and Møller, 2019). Many researchers proposed different techniques to select k:

i. *Topic coherence*, which is one of the main techniques used to find the number of topics.
ii. *LDA visualisation tool pyLDAvis*, a web-based exploration tool for interactive topic modelling visualisation (Onah and Pang, 2021) which applies a different number of topics and compares the results.
iii. *Human interpretation and judgment* are used as criteria.

In this research project, a coherence metric was applied by computing c_v coherence (Syed and Spruit, 2017) for various numbers of topics to obtain a number close to the optimal one. Thus, several models were built with different k, starting from 2–20 at intervals of 1; all the parameters of the models were kept at their default values. Then, based on the coherence score, the best value was selected. Next, parameters (passes and iterations) were tuned to achieve the best topics, where *passes* refers to the total number of passes through the corpus during training, and *iterations* refers to controlling the maximum number of iterations through the corpus when inferring the topic distribution of a corpus. These two parameters were set and tuned to passes = 50 and iterations = 200.

5.3.2.2. Extracting Urgent Language via LDA

To inspect and provide an overview of the terms (words and phrases), the top ten terms on each topic were presented. After that, t-distributed stochastic neighbour embedding (t-SNE) (Van der Maaten and Hinton, 2008) was used, which is a cutting-edge unsupervised technique for dimensionality reduction to visualise clusters with high dimensions in 2D space. The final aim was to reveal and capture the key language MOOC learners use to express their need for urgent intervention. To reach this primary goal, the most dominant topic for every post was found (as every post is composed of a mixture of words) and each word was drawn from one topic, as shown in Figure 5.2 (below). This enabled the identification of the most representative post for each topic as an example to understand each topic.

Figure 5.2: Each post is a collection of words that belongs to a specific topic.

Next, to find the most discussed topics in the posts for the whole course, the number of posts by dominant topic was plotted.

Finally, for each topic, the percentage of posts with predominantly urgent posts was calculated, and the same with non-urgent posts. A threshold of more than 80% inclusion of dominant topics was set; under the assumption of this ensuring that they were the most representative posts of that particular topic.

5.3.3. Instructor Visualisation Aid

To further support the instructor intervention task, different visualisation aids were proposed as potentially powerful tools to allow instructors to become aware of learners' use of particular topics in the discussion forums as well as to help learners to become aware of their own use of language in MOOC forum posts to signal the need for instructor intervention. Thus, based on

the results of LDA analysis, the instructor can focus on specific topics that represent the language of urgency. Specifically, three different aids in this work were applied as follows:

i. *Wordclouds:* the top ten terms in each topic were visualised using wordclouds, which is a visual representation of topics in a cloud-shaped format that are depicted in different sizes based on the probability of each term (word) (instructors only).

ii. *pyLDAvis*: used to represent, distinguish, and interpret topics (instructors).

iii. *Coloured posts*: each token in the post was coloured with the topic colour to help instructors or learners to determine urgent words.

5.4. Results and Discussion

In this section, the overall results are presented and interpreted, taking into account the LDA results and exploring how urgent language can be extracted via LDA to answer ***RQ2.1***. Additionally, the varying visualisation aid was shown to answer ***RQ2.2***, to help instructors develop an intuitive sense about urgent posts language and help learners to become more conscious of their language use.

5.4.1. Topic Modelling (LDA)

As discussed in Section 5.3.2.1, the optimal number of topics (k) was estimated based on the coherence score (see Figure 5.3 below); the number of topics (num topics) was based on the coherence score. Selecting six topics rendered the highest coherence score on the y-axis.

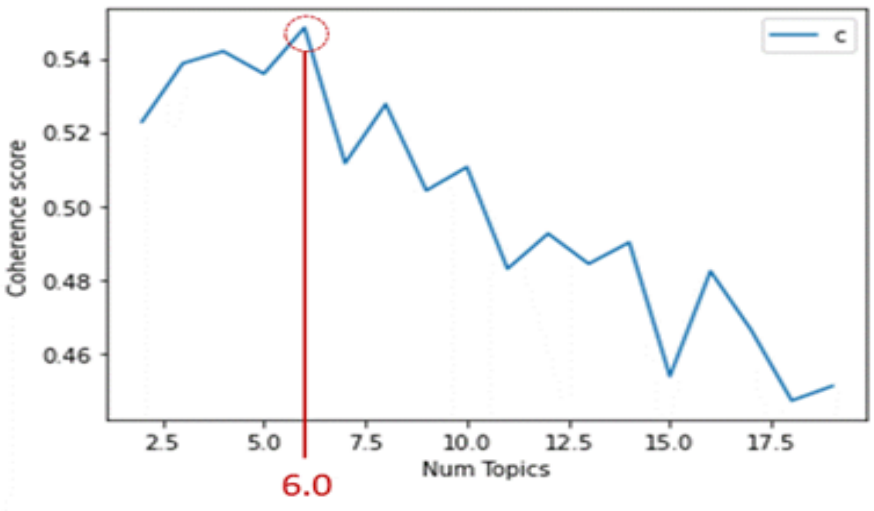

Figure 5.3: Selecting the optimal number of LDA topics.

5.4.2. Extracting Urgent Language via LDA

The results of the ten most relevant terms generated and the highest probability for every topic output as per the LDA with k = 6 in this experiment are presented in Table 5.2. Identifying these terms is sufficient to understand the terms that belong to a specific topic. Some words appeared in different topics: *course* appeared in topic 0, topic 3, and topic 4.

Table 5.2: Most relevant terms with their probability distribution over topics for the six topics identified by LDA.

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0.021*course[14]	0.008*change	0.023*use	0.042*thank	0.022*reviewer	0.076*essay
0.013*time	0.007*risk	0.021*write	0.033*course	0.021*get	0.060*grade
0.011*certificate	0.007*use	0.015*paper	0.022*answer	0.020*peer	0.044*submit
0.008*page	0.007*com	0.013*sentence	0.014*question	0.020*score	0.037*peer
0.007*hope	0.007*think	0.011*word	0.013*homework	0.020*review	0.035*review
0.007*continue	0.006*study	0.010*writing	0.012*get	0.018*course	0.029*assignment
0.007*take	0.006*different	0.009*think	0.010*quiz	0.017*comment	0.018*submission
0.006*good	0.005*effect	0.008*example	0.010*post	0.016*give	0.017*problem
0.006*hour	0.005*mean	0.008*scientific	0.010*video	0.012*grade	0.016*score
0.006work	0.005*include	0.007*make	0.009*problem	0.011*think	0.014*get

Figure 5.4 visualises the higher dimensional data in lower dimensions using the t-SNE algorithm for topic-based exploration as explained in Section 5.3.2.2. Here, the different topics are mapped onto two dimensions and each topic is denoted by a specific colour label.

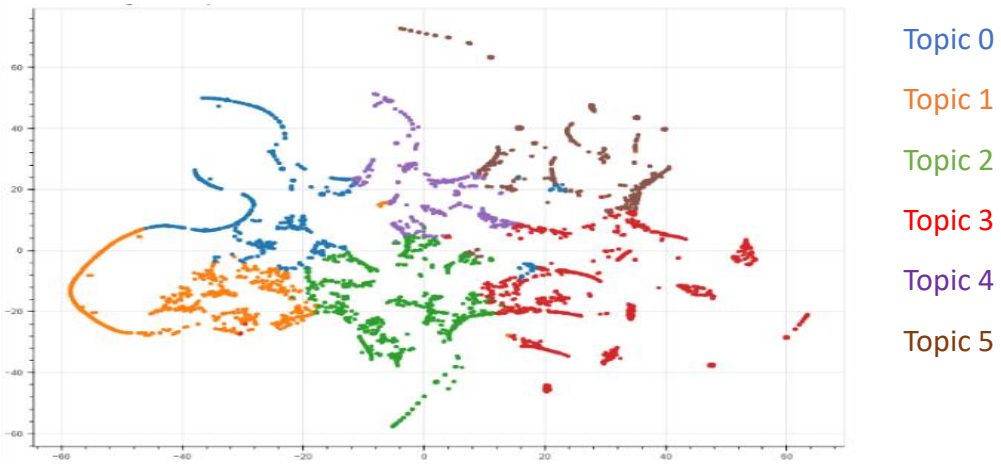

Figure 5.4: t-SNE clustering of six LDA topics.

[14] For the "course" term, it has a probability of 0.021 distribution over topic 0.

As an example of the most dominant topic, Table 5.3 (below) shows the dominant topic as a percentage, with the terms (keywords) for the first ten posts in the corpus. For all these ten posts, topic 3 was the dominant one, with different contributions counting towards it.

Table 5.3: Dominating topic for the first 10 posts.

Dominant topic	Topic percentage contribution	Keywords	Tokens of posts
3	0.53	thank, course, answer, question, homework ...	[hope, useful, place, discuss, related, course...
3	0.47	thank, course, answer, question, homework …	[video, unit, work, however, one, work, perfec...
3	0.49	thank, course, answer, question, homework ...	[think, question, ask, profession, radio, butt...
3	0.42	thank, course, answer, question, homework …	[head, use, cap, answer, quiz, type, lowercase...
3	0.93	thank, course, answer, question, homework ...	[know, contact, get, mark, ill, even, email, s...
3	0.58	thank, course, answer, question, homework ...	[open]
3	0.83	thank, course, answer, question, homework ...	[thank, link, able, view, first, video, rest, ...
3	0.61	thank, course, answer, question, homework ...	[video, lecture, show, unavailable]
3	0.72	thank, course, answer, question, homework ...	[access, youtube]
3	0.72	thank, course, answer, question, homework ...	[sorry, trouble, video, incorrect, correct, av...

Table 5.4 (below) shows the most representative tokens from posts for each topic as a sample of what a topic is about. The minimum contribution was about 0.97, which shows that these tokens represent the topic almost perfectly.

Table 5.4: The most representative tokens of posts for each topic.

Topic number	Topic percentage contribution	Keywords	Tokens of posts
0	0.97	course, time, certificate, page, hope, continue, take, good, hour, work	[make, follow, revision, dedicated, prosthesis, allow, sprinter, run, low, metabolic_cost, ...
1	0.99	change, risk, use, com, think, study, different, effect, mean, include	[immortality, alluring, concept, scientist, believe, possible, upload, mind, recreate, ...
2	0.99	use, write, paper, sentence, word, writing, think, example, scientific, make	[note, necessarily, right, way, way, protective, occurrence, inhibit, reoccurrence, estimate, ...
3	0.97	thank, course, answer, question, homework, get, quiz, post, video, problem	[subtitle, video, available, download, soon, meanwhile, view, video, course, webpage, youtube, ...
4	0.98	reviewer, get, peer, score, review, course, comment, give, grade, think	[thank, elfatih, point, review, student, paper, helpful, receive, excellent, feedback, first, ...
5	0.99	essay, grade, submit, peer, review, assignment, submission, problem, score, get	[dear, problem, want, post, rd, assignment, order, take, look, feedback, essay, go, section, ...

Figure 5.5 (below) depicts the distribution of posts by dominant topic. Topic 3 (thank, course, answer, …) has the highest number of posts as a dominant topic; followed by topic 2 (use, write, paper, …), and topic 5 (essay, grade, submit, …).

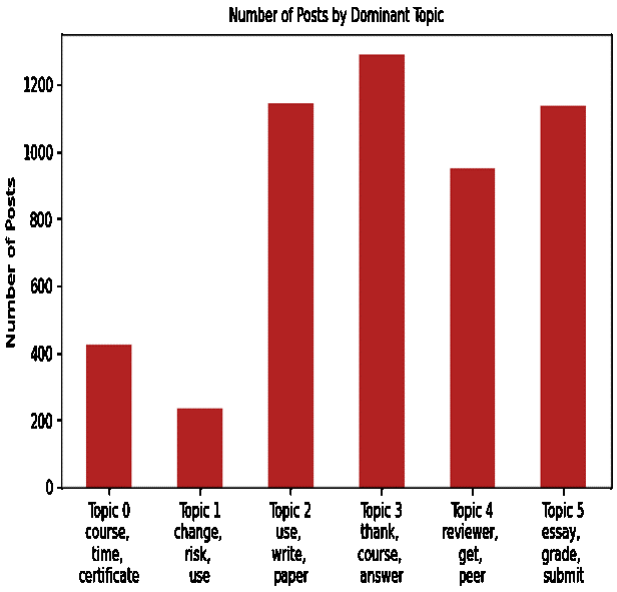

Figure 5.5: Number of posts by dominating topics.

Table 5.5 (below) shows the percentage of urgent and non-urgent posts where the contribution of the dominant topics was more than 80%. The total number of posts with a topic percentage

contribution of > 80% was 1218 posts. Topic 5 covered about 58% of urgent posts. That means the most important words that learners used and expressed in texts when they need urgent intervention can be found in topic 5. These include: *essay, grade, submit, peer, review, assignment, submission, problem, score, get etc*. Analysing them, these terms make sense as illustrators of urgent language, as an imminent test can provoke a sense of urgency. Also, despite topic 2 and topic 3 being associated with many posts, the urgent posts only accounted for 26% and 32% of posts, respectively.

Table 5.5: The percentage of urgent and non-urgent posts for each topic where the dominant contribution was more than 80%. **Bold:** large percentage of posts that represent urgent intervention.

Topic number	Posts number	Urgent number	Urgent %	Non-urgent number	Non-urgent %
0	44	13	30 %	31	70 %
1	55	11	20 %	44	80 %
2	298	76	26 %	222	74 %
3	320	102	32 %	218	68 %
4	171	42	25 %	129	75 %
5	330	190	**58 %**	140	42 %

Next, the posts were manually inspected in which the dominant topics belonged to topic 5. For further illustration of why dominant topics for topic 5 only accounted for 58% of the urgent posts, the post type was reviewed. In MOOC discussion forums, as clarified in Section 2.2.1.3, there are two types of posts in the Stanford dataset: *commentThread* (the first post) and *comment* (a reply to a specific post). (Chaturvedi, Goldwasser and Daumé III, 2014) supposed that the first post tends be a question and the reply might be the answer to the question or a comment about the question. The same scenario was assumed in the present study where the commentThread tended to be urgent while the comment was likely non-urgent. Therefore, urgent and non-urgent posts were analysed in topic 5 from the point of view of the post being a commentThread or a comment. As shown in Table 5.5 (above), the number of posts in topic 5 was 330. These posts were classified based on post type (commentThread or comment); 101 posts were commentThreads, with 96% urgent posts; 229 were comments, with just 40% urgent posts, as per Table 5.6. This further explains why the language used in these non-urgent comments imitates urgent language: when replying to threads, learners used similar terms and language as that of the original thread, the writer of which may have been in urgent need of

intervention; however, their reply (comment), albeit written in similar language, did not need urgent intervention.

Table 5.6: The percentage of urgent and non-urgent posts for thread and comment in Topic 5.

Type of posts	Posts number	Urgent number	Urgent %	Non-urgent number	Non-urgent %
CommentThread	101	97	96 %	4	3.9 %
Comment	229	93	40.6 %	136	59 %

5.4.3. Instructor Visualisation Aid

To enlighten instructors as to when intervention is required and provide an overview of the different topics and the probability of each term occurring in each topic, wordcloud visualisations that represented each topic in a distinct colour were created and each term appeared in a different size representing the probability of each term (word) appearing, as shown in Figure 5.6 (below).

Figure 5.6: Word cloud visualisation (top ten terms) for each topic.

To enable instructors to interpret the topics simply and interactively, pyLDAvis was used (see Figure 5.7). Every topic is represented as a bubble, the size of which represents the percentage of the number of posts about this topic. The largest bubble means that it contains the highest percentage of posts about this topic. The distance between the centre of the bubbles indicates

the similarity between the topics. The bar chart illustrates the top 30 terms for specific topics. For example, in Figure 5.7, these terms are the most useful for the current topic selected.

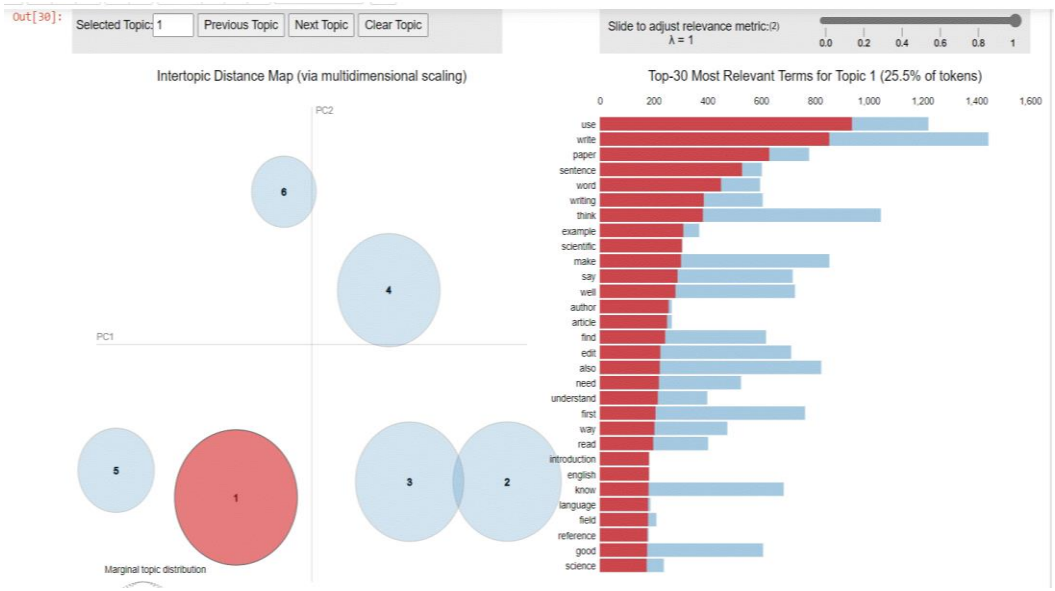

Figure 5.7: pyLDAvis - top 30 terms for each topic.

To further help both instructors and potentially learners, the tokens of the posts were additionally coloured with the topic colour, as illustrated in Figure 5.8 (below). For example, post 1 (first post) contains different colours (red, purple, and green) belonging to topic 3, topic 4, and topic 2, respectively. Therefore, if the instructor finds a brown colour (topic 5) and it is a thread, this indicates that this post needs urgent intervention.

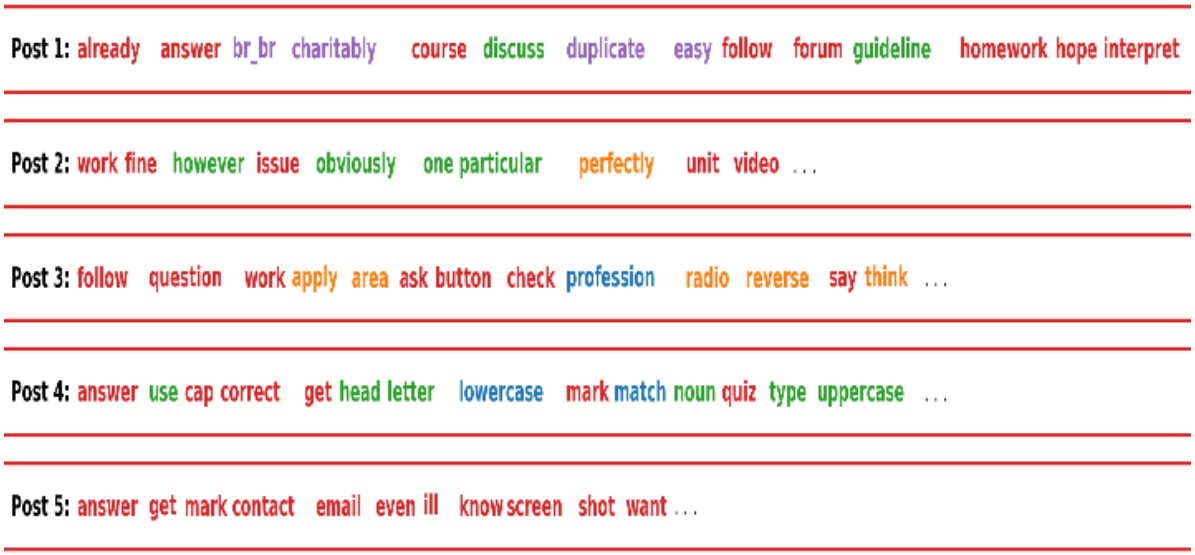

Figure 5.8: Topic colouring for the first 5 posts tokens.

5.5. Epilogue

The main challenge for instructor intervention in MOOCs is the nature of discussion forums as they contain many posts, and a low number thereof require urgent intervention. This chapter showed that learners often express their need for urgent intervention via discussion forums using special linguistic terms. It is possible to extract this language to help instructors decide when intervention is required as well as learners when writing posts to request intervention. In addition, visualisation can be employed to aid in the comprehension of a learner's language, allowing the instructor to potentially intervene more effectively. In this research project, learners' posts were analysed to explore the language used to express their need for urgent intervention using a course from the Stanford dataset as a case study. The analysis shows that some words are related to each other and express the need for urgent intervention, especially in posts as a thread type.

Importantly, for the first time, this study has proposed a context-dependent urgency language criteria using language highlighting the need for urgent intervention in a MOOC environment and showed some straightforward and easily reproducible ways to extract and visualise the need for such intervention.

The next chapter seeks to identify at-risk MOOC learners who may need instructor intervention in a new way by using their historical online forum posts as data and a novel multi-input approach for other deep learning architectures and Transformer models.

CHAPTER 6: INTERVENTION PREDICTION: LEARNER-POST-BASED MODEL

6.1. Prologue

The problem of high learner dropout rates in online MOOC-based education contexts is one of the most long-standing challenges of such learning environments (Cristea *et al.*, 2018). This topic has attracted many researchers to explore this problem and propose different intervention models. This chapter seeks to address the problem of high learner dropout rates in MOOCs, which can reach 90% (Rivard, 2013) by proposing a learner-post-based model. Interaction with an instructor is considered one of the most important factors for mitigating dropout among MOOC learners (Hone and El Said, 2016). Thus, it might be helpful to consider the sequence of learners' textual posts to identify learners who require instructor intervention to reduce dropout rates and improve the quality of the interventions offered. Therefore, the proposed learner-post-based intervention model is described in this chapter.

The aim of this study was to develop a ML model to identify learners who require intervention by an instructor based on the sequence of learner posts to predict and mitigate learner dropout on MOOC-based courses. This is because an absence of interaction and feedback by instructors on discussion forums has been associated with increased dropout rates (Hone and El Said, 2016; Wei *et al.*, 2017). Thus, this chapter proposes an intelligent intervention model to help instructors. This challenge was formalised as a text classification problem by developing and employing a supervised binary classification model with multiple text inputs based on learner posts.

The input consists of learners' most recent posts (as further defined in Section 6.3.2) and the output is the predicted dropout. Two recent popular types of classifiers: other DL (Young *et al.*,

2018) and Transformer (Vaswani *et al.*, 2017) were applied and trained, and examined various numbers of inputs for prediction. Therefore, the following RQs were investigated:

- ***RQ3.1:*** *Which multi-input models (processing several recent posts) are useful for predicting learners who may drop out (thus may need instructor intervention)?*
- ***RQ3.2:*** *Does clustering learners based on their number of posts prior to the prediction step improve prediction outcomes?*

The key contributions of this chapter are as follows. (i) to the best of the researcher's knowledge, the current literature does not investigate the history of learners' written MOOC posts; this is the first study to attempt to identify MOOC learners who may need instructor intervention by using their historical online forum posts as data. (ii) the other contribution of this work is the use of a multi-input approach for siamese and dual BERT with binary text classification, with the resulting integrated networks being termed multi-siamese BERT and multiple BERT, respectively.

6.2. Related Work on Dropout

The issue of intervention to help prevent learners from dropping out of MOOC course environments is an interesting area for many research communities (Whitehill *et al.*, 2015; Cobos and Ruiz-Garcia, 2021; Xing and Du, 2018; Borrella, Caballero-Caballero and Ponce-Cueto, 2019) and an important research direction. In prior literature as shown in SLR in Chapter 2, instructor intervention in MOOCs has been demonstrated and studied from two main perspectives: (i) posts on discussion forums, and (ii) learners.

The use of posts on discussion forums for intervention prediction has received considerable focus as discussed in detail in the SLR in Section 2.3.4.4.1; researchers have attempted to establish and provide different intervention models as a text classification task (Sun *et al.*, 2019; Almatrafi, Johri and Rangwala, 2018; Khodeir, 2021; Guo *et al.*, 2019), or used posts features as an input for the classifier (Chandrasekaran *et al.*, 2015b; Chaturvedi, Goldwasser and Daumé III, 2014).

From a learner perspective, prevalent studies have addressed intervention and dropout rates using learner characteristics, learning activity or clickstream data, such as predicting dropouts per week based on the weekly history of the learner (Kloft *et al.*, 2014). Also, (Xing and Du, 2018) created a similar weekly prediction mechanism by applying a DL approach.

In contrast, there is limited research on intervention based on the posts of learners who are likely to drop out (Prenkaj *et al.*, 2020). This is due to the low percentage of learners who enrol on a MOOC course and write posts (only around 5–10%) (Rose and Siemens, 2014). For example, (Gitinabard *et al.*, 2018) showed that out of 55,013 and 10,190 learners who had registered and enrolled on courses, only 750 and 519 engaged with discussion forums by making posts, respectively. Among the few pieces of research on this topic, (Crossley *et al.*, 2015) used NLP tools to predict learners who completed a MOOC course with an accuracy of 67.8 %. Other researchers combined clickstream data with discussion forum data. For example, (Crossley *et al.*, 2016) predicted learner completion by employing clickstream data and language in a discussion forum with a 78% accuracy rate. In addition, some additional research has been discussed before in detail in the SLR.

Furthermore, using sentiment analysis gathered from learners' posts, (Chaplot, Rhim and Kim, 2015) predicted attrition based on different features including sentiment analysis using a neural network and achieved 72.1% accuracy. Using the same method, (Mrhar, Douimi and Abik, 2021) predicted dropout rates based on sentiment analysis and clickstream data. Also, (Wen, Yang and Rose, 2014) found a significant correlation between sentiment and attrition.

As previously stated, this study aimed to develop an intelligent intervention system to reduce learner dropout in MOOC courses. The proposed model offers a novel approach that predicts learner dropout (need for intervention) based on learner post history as a multi-input text classification task to improve instructor intervention and reduce dropout rates.

6.3. Methodology

The core aim of this experiment was to construct a temporal multi-input approach using other deep learning architectures and Transformer models. In the next sub-sections, the dataset used in this research and the intervention models (other deep learning architectures and Transformer) will be discussed in detail.

6.3.1. Dataset

The dataset investigated for this research project was the Dropout dataset (clarified in Chapter 3, Section 3.2.2.3). These data include the history of learners' posts in discussion forums and flags if they (i) complete the course (no need for intervention) or (ii) dropout (need intervention). The number of posts written by these learners varied from 1–209. To explore the

number of words featured in each post, Figure 6.1 shown the distribution of the number of words per post (mean = 53.21 words, minimum = 1 word; maximum = 226 words).

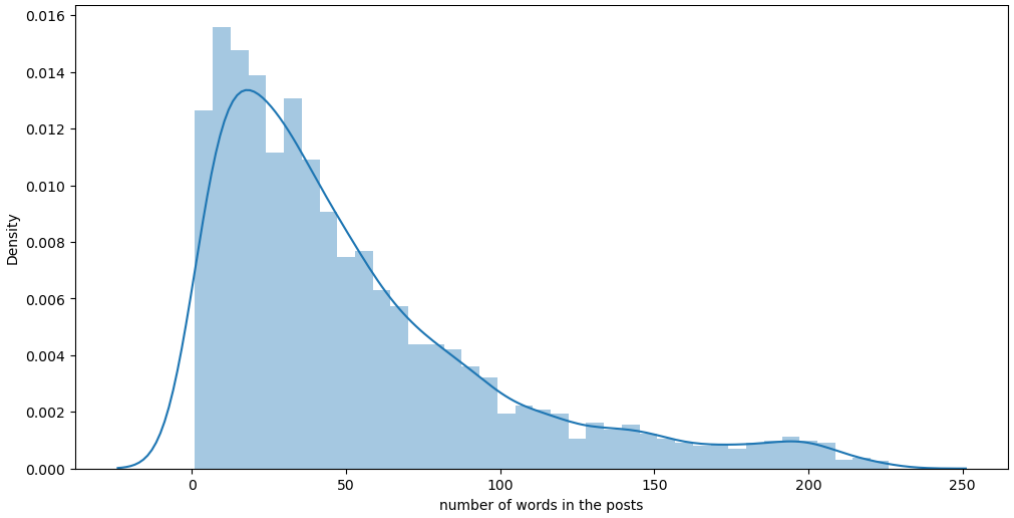

Figure 6.1: Distribution of number of words per post.

6.3.2. Intervention Models

To identify the learners' need for instructor intervention, the general architecture of the prediction model was proposed as shown in Figure 6.2 (below). The input of this model is the most recent sequence of learner posts while the output is the prediction of if a learner needs instructor intervention (dropout) or not (completer).

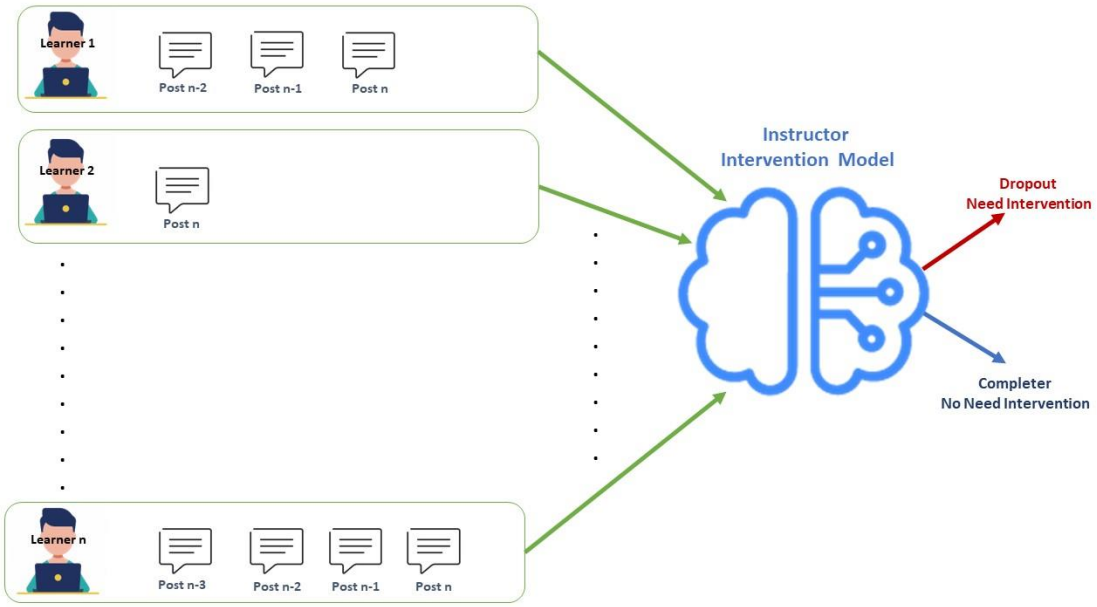

Figure 6.2: Architecture of the intervention prediction model.

For the number of inputs, i.e., the most recent sequence of learner posts, it was assumed that the learner writes multiple posts and that the number of such posts is an unknown value and may differ from one learner to another. Therefore, in this experiment, an incremental number of posts ranging from 3–7 was examined; however, as mentioned, the total number of posts ranged from 1–209. Hence, further experimentation was needed to cluster learners based on their respective number of posts to analyse if grouping as a pre-processing step improved the results.

Thus, the commenters were clustered into three groups (which identified that the optimal number of clusters was three using the silhouette method (Rousseeuw, 1987)), based on the number of posts written, using the Fisher Jenk algorithm (North, 2009), as shown below in Table 6.1. Next, the focus was on group 1, as it contained the highest number of learners (797 commenters) and was thus the most representative of the average number of posts written by learners. Of these learners, 557 (69.8% ≈ 70%) dropped out and 240 (30%) completed the course. After that, the same experiments were repeated for the best intervention models using group 1 (797 commenters) with the mean input rounded up from *3.66* to 4 and excluding the other two groups. Please note that group 1 also had the smallest standard deviation (Std) of 3.43.

Table 6.1: Statistics of each cluster group.

Group	Count	Mean	Std	Minimum	Maximum
1	797	3.66	3.43	1	16
2	65	28.89	12.24	17	62
3	9	108	43.40	71	209

The prediction models were developed based on two main types of algorithms: other deep learning architectures and Transformer. The reason for using these models is because they represent the cutting-edge in NLP and eliminate the need for specific feature engineering because they can extract features. The two types are illustrated in the following sub-sections.

6.3.2.1. Other Deep Learning Architectures

The two cutting-edge DL algorithms were applied: CNN and RNN. For RNN, Bi-LSTM and Bi-GRU were used. Each input was treated as a sub-model before these sub-models were concatenated to build the main model. The prediction for the output of the final/output layer

was a probability value where if the value was equal or larger than 0.5, it was deemed positive (1). The outcomes of (1) indicate a potential dropout and that urgent intervention is required; an outcome of (0) represents a completer and that no intervention is required.

The general architectures are the same for all models. As a pre-processing step to prepare the data for input, a dictionary for each input that contains unique vocabulary words was built. To specify the length of the word sequences, (Guo *et al.*, 2019) approach was followed, constraining the length of each input to 200 words; also the fact that most posts were ≤200 words was explored in Figure 6.1, which means just 1.3% of posts were affected by truncation. The shortest sequence was padded by 0 and posts >200 words were trimmed. The next layer after the input layer is the embedding layer. This layer obtains dense vector representations for words, which was used and fine-tuned during training, starting with pre-trained word embedding using word2vec (Mikolov, Le and Sutskever, 2013) (Word2vec GoogleNews-vectors-negative300). Then, the following layers differed according to the different networks (CNN and RNN).

The data were split into training data and testing data (80% and 20%, respectively, equivalent to 696 and 175 learners, respectively). Then, the training data was split into training data and validation data (80% and 20%, respectively, equivalent to 556 and 140 samples, respectively). Lastly, the model was trained using the Adam optimiser (batch size = 64; epochs = 20).

6.3.2.1.1. CNN

The general architecture is shown in Figure 6.3. In the convolutional layers, for each input, three Conv1Ds were applied with 128 units and different kernel sizes (3, 4, and 5) following (Guo *et al.*, 2019). These layers go through a (ReLu) activation, followed by a max-pooling layer to further compress features. Then, the output from each input was concatenated. Next, all the outputs for all the inputs were concatenated. This is then passed to the dense layer with 64 neurons and ReLu activation. Then, a dropout layer is employed to avoid overfitting (Otter, Medina and Kalita, 2020) as a regularisation technique. Finally, the output layer has 1 unit with a sigmoid activation function because it performs a binary classification task.

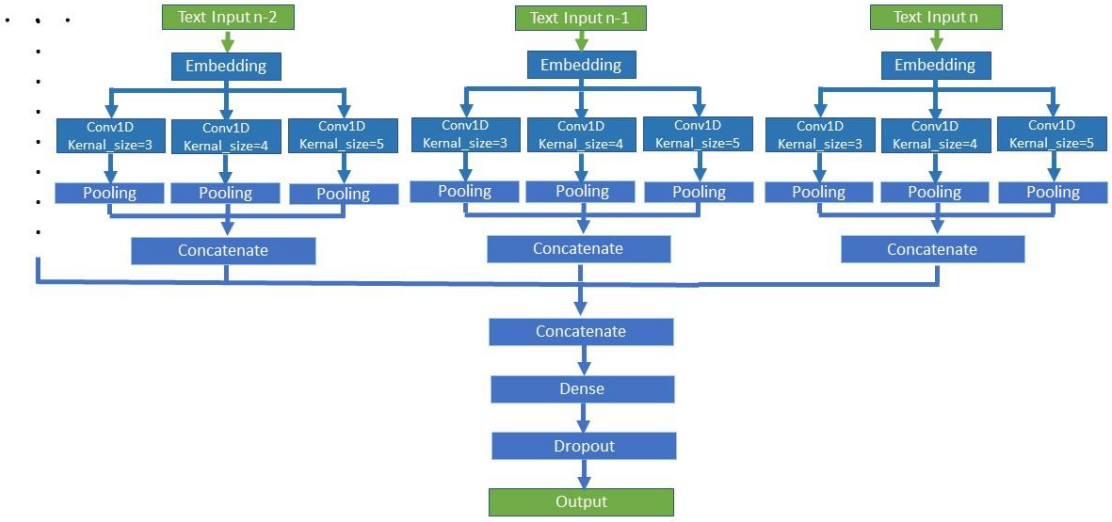

Figure 6.3: The general architecture of the CNN with multi-input.

6.3.2.1.2. RNN

These two different networks share the same architecture (see Figure 6.4 below), the Bi-LSTM and Bi-GRU contain two layers which were trained by adding another hidden layer to reverse to the first layer as discussed in detail in Chapter 2. Thus, as the next layer after the embedding layer, the RNN layer is Bi-LSTM or Bi-GRU (128 units). Afterwards, the output for each input was concatenated. Then, a dense layer and dropout layer were added as in the CNN. Finally, to obtain the classification, the sigmoid as an activation function was applied to the output layer.

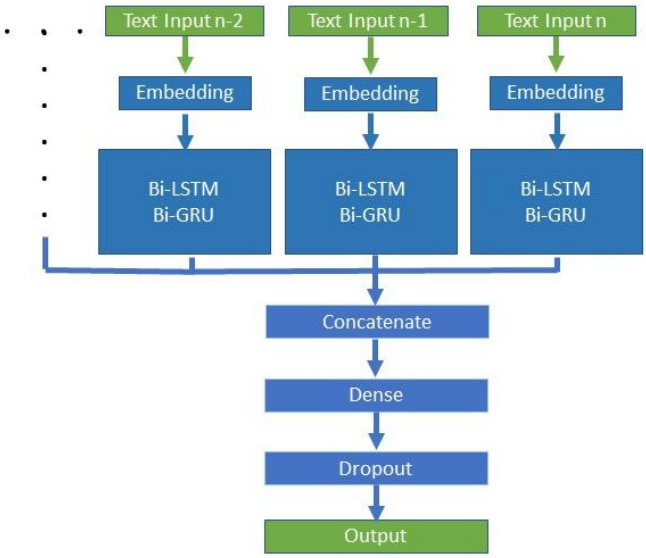

Figure 6.4: The general architecture of the RNN with multi-input.

6.3.2.2. Transformer

Two different models were developed and built upon the siamese and dual Transformers BERT networks to enable the insertion of more than two inputs into the BERT model. The inspiration to use these two techniques was Marco Cerliani's code on GitHub, which was consequently modified and to which more than two inputs were added (3–7 inputs); additionally, these two multiclass classification models were converted into two binary classification models: multi-siamese BERT and multiple BERT. The structure of these models is presented in Figure 6.5 (below). Each text input was converted to Transformer inputs with the special tokens ([CLS] and [SEP]). Then, BERTbase was utilised, as the training time for this version is less than for the BERTlarge. The same training and testing data as in the DL models were used. Then, these models were trained using the Adam optimiser, batch size = 6 and epochs = 3. The same went for the DL model: the prediction was calculated, where if the value was equal or larger than 0.5, it was deemed positive (1). The (1) denotes a potential dropout who needs urgent intervention and (0) denotes a completer and no intervention is required.

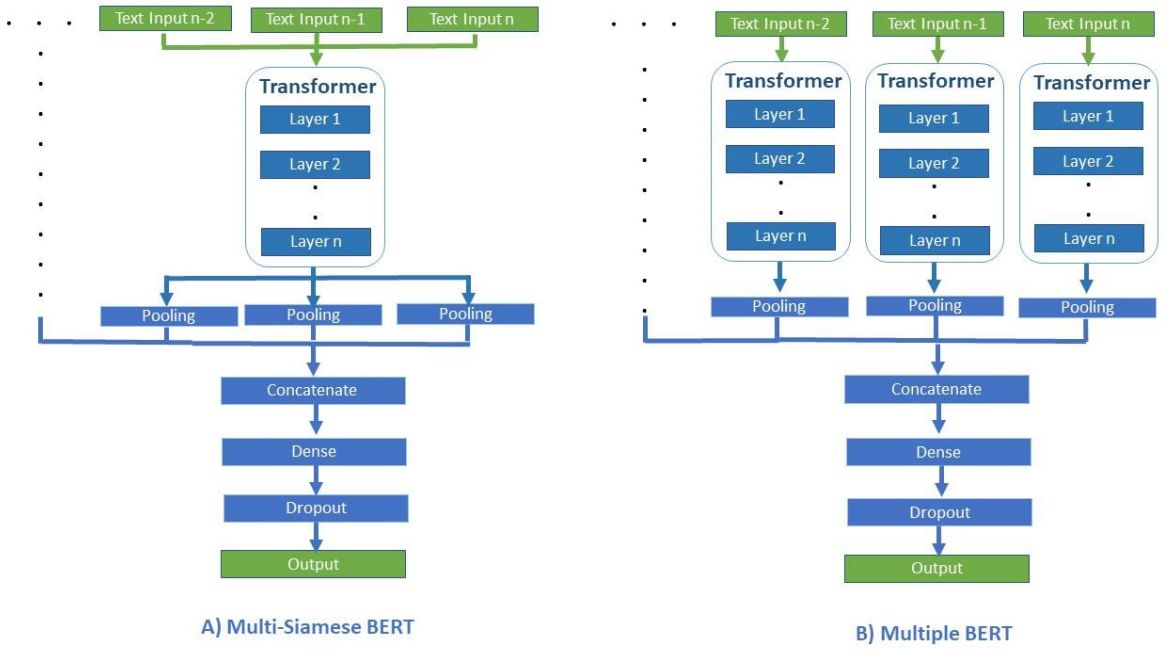

Figure 6.5: The general architecture of a) multi-siamese BERT and b) multiple BERT.

6.3.2.2.1. Multi-Siamese BERT

In this model, the different text input passes to the same Transformer. Then, the output is compressed with a global average pooling. After that, the outputs were concatenated and passed to the dense, dropout, and output layers as in DL.

6.3.2.2.2. Multiple BERT

In this model, each input passes to different Transformers and is reduced via average pooling; then, all the outputs of the global average pooling were concatenated; after that, as in the multi-siamese BERT, the output is passed to the dense, dropout, and output layers.

6.4. Results and Discussion

The experimental results of the multi-input model predictions to address ***RQ3.1*** are presented in Table 6.2 (below). In addition to accuracy *Acc*, *P*, *R* and *F1* metrics for each class and *BA* for both classes were also used to comprehensively assess the performance (in percentages) of the different models.

In this experiment, the models were evaluated and compared based on *BA* as the data classes are imbalanced. The results reveal that multi-siamese BERT and multiple BERT outperform all the DL models in *BA* (0.698, 0.696 respectively ≈ 70.0).

Table 6.2: The performance results of the different multi-input models with different inputs (all learners), **Bold:** best performance of BA, *Italic:* optimal number of inputs per model based on BA.

Type	Input	Acc	Completer (0)			Dropout (1)			BA
			P	R	F1	P	R	F1	
Other Deep Learning Architectures									
CNN	3	0.66	0.44	0.43	0.44	0.76	0.76	0.76	0.598
	4	0.69	0.49	0.43	0.46	0.77	0.80	0.78	*0.618*
	5	0.67	0.31	0.08	0.12	0.70	0.93	0.80	0.500
	6	0.66	0.42	0.32	0.37	0.73	0.81	0.77	0.566
	7	0.64	0.40	0.38	0.39	0.74	0.75	0.74	0.598
Bi-LSTM	3	0.65	0.41	0.36	0.38	0.74	0.78	0.76	0.568
	4	0.73	0.61	0.26	0.37	0.74	0.93	0.82	0.595
	5	0.67	0.44	0.36	0.40	0.74	0.80	0.77	0.580
	6	0.71	0.53	0.32	0.40	0.75	0.88	0.81	*0.598*
	7	0.65	0.39	0.28	0.33	0.72	0.81	0.76	0.547
Bi-GRU	3	0.67	0.44	0.36	0.40	0.74	0.80	0.77	0.580
	4	0.63	0.39	0.38	0.38	0.73	0.75	0.74	0.561
	5	0.67	0.45	0.43	0.44	0.76	0.77	0.76	0.602
	6	0.63	0.42	0.51	0.46	0.76	0.69	0.72	0.598
	7	0.69	0.49	0.47	0.48	0.77	0.79	0.78	*0.629*
Transformer									
Multi-Siamese BERT	3	0.71	0.52	0.58	0.55	0.81	0.76	0.78	0.673
	4	0.63	0.44	0.75	0.56	0.85	0.58	0.69	0.668
	5	0.69	0.49	0.72	0.58	0.85	0.68	0.75	***0.698***
	6	0.65	0.45	0.75	0.56	0.85	0.60	0.70	0.676
	7	0.65	0.45	0.72	0.55	0.83	0.61	0.71	0.665
Multiple BERT	3	0.67	0.47	0.55	0.50	0.79	0.73	0.76	0.638
	4	0.67	0.47	0.68	0.55	0.83	0.66	0.74	0.671
	5	0.65	0.46	0.81	0.59	0.88	0.58	0.70	***0.696***
	6	0.71	0.52	0.60	0.56	0.81	0.75	0.78	0.678
	7	0.67	0.46	0.49	0.48	0.77	0.75	0.76	0.622

To address ***RQ3.2***, how the best-performing algorithms (Transformers) performed in the given group (group 1), Table 6.3 (below) shows that the performance of all commenters outperforms group 1 in the multi-siamese BERT in *BA* as well as in the multiple BERT. Therefore, it provided negative values on prediction outcomes, contrary to expectations, especially in *R* in class (0).

Table 6.3: Comparison between the performance results of different multi-input transformer models with 4 inputs (all learners and group 1), **Bold:** best performance in BA.

Type	Group	Acc	Completer (0)			Dropout (1)			BA
			P	R	F1	P	R	F1	
Transformer									
Multi-siamese BERT	All	0.63	0.44	0.75	0.56	0.85	0.58	0.69	**0.668**
	Group 1	0.68	0.59	0.24	0.34	0.70	0.91	0.79	0.575
Multiple BERT	All	0.67	0.47	0.68	0.55	0.83	0.66	0.74	**0.671**
	Group 1	0.69	0.64	0.25	0.36	0.70	0.92	0.80	0.589

6.5. Epilogue

Although MOOCs have had a significant impact on facilitating learning, they suffer from unacceptable dropout rates. Previous studies have explored how to identify when learners need intervention based on learner behaviour to estimate the risk of dropping out. This chapter investigated instructor intervention by attempting to predict dropouts from learners' most recent posts to enable instructors to better identify learners requiring assistance and intervene effectively. It established various ML models including other deep learning architectures and Transformer with multi-input. The Transformer models were developed based on siamese and dual BERT to insert more than two inputs for the Transformer models. The multi-input consists of the most recent learner posts.

The results indicate that the intervention model can predict dropout and the need for intervention with more accuracy and better detect at-risk learners with the Transformer models. However, contrary to expectations, grouping learners before prediction might harm prediction outcomes, particularly in the minority class.

The following chapter proposes a novel priority model for the urgency of intervention based on learner histories.

CHAPTER 7: INTERVENTION PREDICTION: POST- AND LEARNER-BASED MODEL (ADDING PRIORITY IN INTERVENTION)

7.1. Prologue

Usually, researchers attempt to focus on identifying learners' posts using different methods as a solution to the instructor intervention problem. However, such approaches have not yet considered the study of learner behaviour in the context of instructor intervention prediction in MOOCs. To improve instructor intervention, the nature of an intervention can be tailored based on learner behaviour. On this basis, the main aim of this chapter is the development of intervention prediction models that focus on both posts and learner behaviours by creating an automated intervention priority model.

Analysing and mining learners' urgent posts is a fundamental step towards understanding learners' need for instructor intervention in MOOC environments. Additionally, it is conjectured that it is essential to understand learners' behaviours before proposing a particular intervention to ensure that the latter is the most appropriate. Thus, in this research project, the distribution of posts that need intervention were analysed. Then, the relationship between high-frequency commenters and their behaviours in terms of number of posts written by them, their access rates, and completion rates were explored. High-frequency (HF) commenters were defined as learners who make many posts that need intervention. The end goal was to propose an automated intelligent intervention priority model built on learner histories — past urgency, sentiment analysis, and step access. The research questions formulated in this chapter are as follows:

- ***RQ4.1:*** *Is there a relationship between the number of posts written by learners who need urgent intervention and the average number of posts?*
- ***RQ4.2:*** *Is there a relationship between high-frequency (HF) commenter learners who require urgent intervention and their average number of step access instances?*
- ***RQ4.3:*** *Is there a relationship between the number of HF commenter learners and completion-rates?*
- ***RQ4.4:*** *How can an intervention priority framework based on behaviour be designed?*

The main novel contributions of this study, to the best of the researcher knowledge, are the filtering system and the priority-in-intervention approach shown in this chapter.

7.2. Methodology

The main objective of this study was to propose an automated intervention priority model based on learner behaviour. Therefore, this section presents the statistics on the dataset used in this chapter, explores urgency and learner behaviour, and offers a novel priority-in-urgent-intervention framework.

7.2.1. Dataset and Statistics

The raw corpus dataset utilised here was provided by the FutureLearn platform (Gold-standard corpus) as explained in Section 3.2.2.2. The 5786 learner posts were created in 5 weeks. The number of steps (that represent a single learning unit including videos and assignments, etc.) and posts per week appear in Table 7.1 (below).

Table 7.1: Weekly statistics on the Gold-standard corpus.

Week	# of steps	# of posts	# of active learners	Average posts per learner
1	11	2130	749	2.84
2	12	1600	419	3.81
3	15	1123	236	4.75
4	11	753	180	4.18
5	4	180	92	1.95

Figure 7.1. (below, left) illustrates the number of posts written over five weeks. This number decreased gradually, dropping to 180 posts in the last week, from 2130 posts in the first week

(-99.9%). Every week has a different number of steps for learners to complete. Thus, the number of posts per step were also represented (Figure 7.1, right, below) on the temporal axis. These numbers oscillate more — showing that some topics trigger more posts than others — although the overall numbers follow a downwards trend.

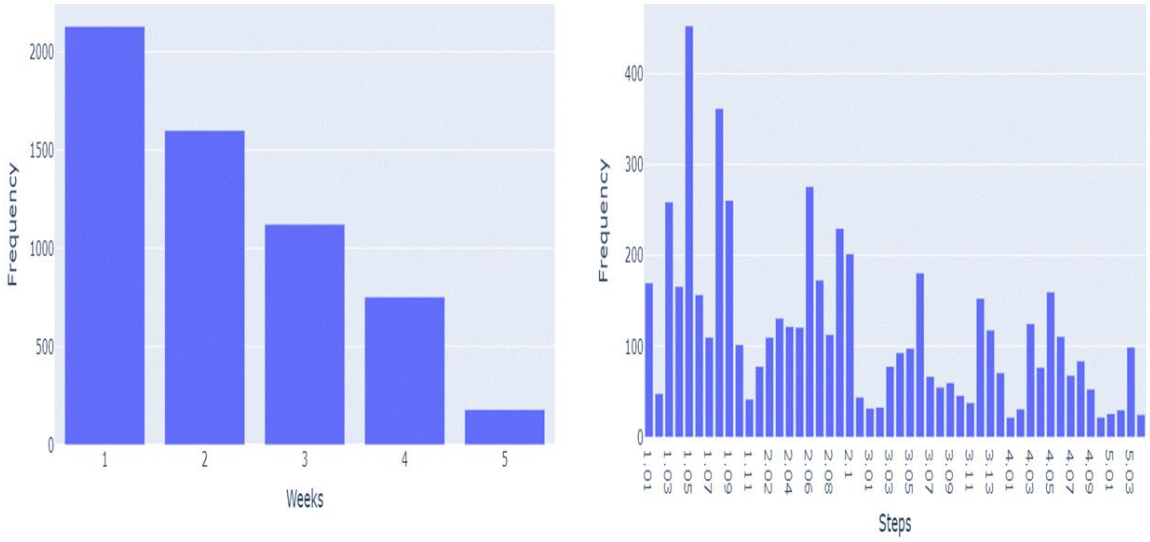

Figure 7.1: The number of posts in every week (left) and in every step (right).

Who is, however, writing these posts? The distribution of the number of *active learners* (commenters) is examined in Figure 7.2 (below); it provides a visual representation of active learners every week and at every step.

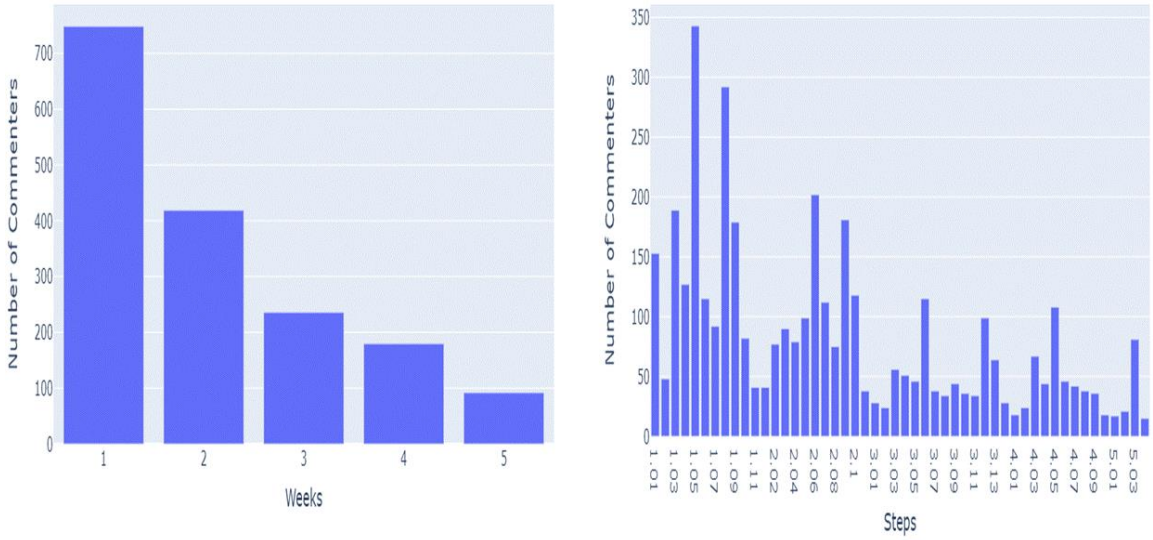

Figure 7.2: Active learners (commenters) in every week (left) and in every step (right).

Next, the posts that need urgent intervention were observed, to focus on this trend. Hence, a line graph over the five weeks was visualised to explore how urgency changed over time

(Figure 7.3, left, below). Overall, the first few weeks had a higher percentage of posts needing intervention (Figure 7.3, left, below), drawn from a higher number of posts (Figure 7.1, left, above). The fluctuation from week 4 to 5 is due to the drastic drop in overall posts. Also, the percentages of urgent posts for every step were visualised, (Figure 7.3, right, below), which showed high fluctuation. Further graphic comparisons of the results between the number of urgent and non-urgent posts across weeks and steps are shown in Figure 7.4 (left, right, below).

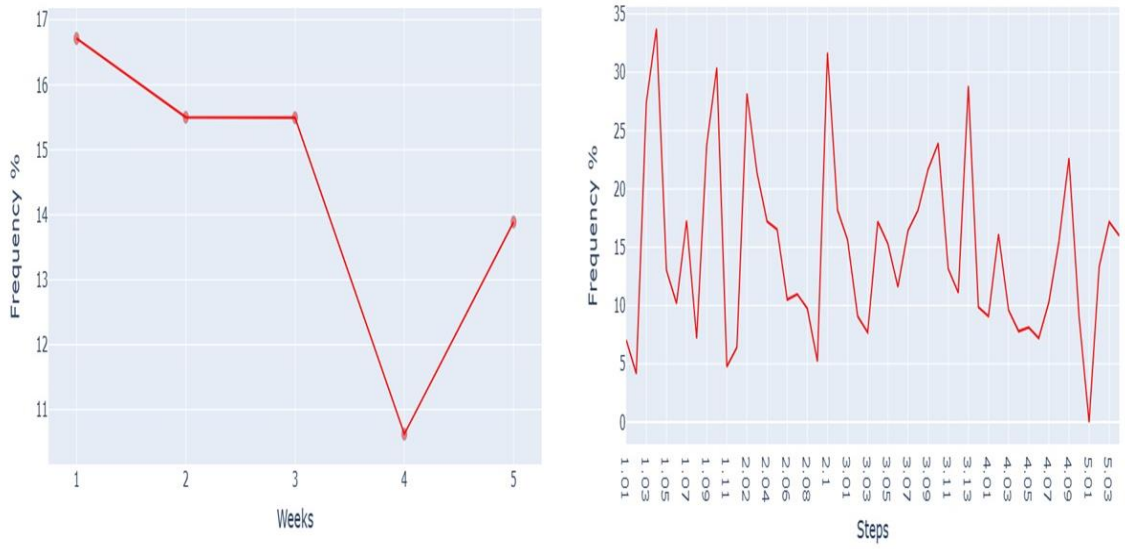

Figure 7.3: The percentage of urgent posts for every week (left) and for every step (right).

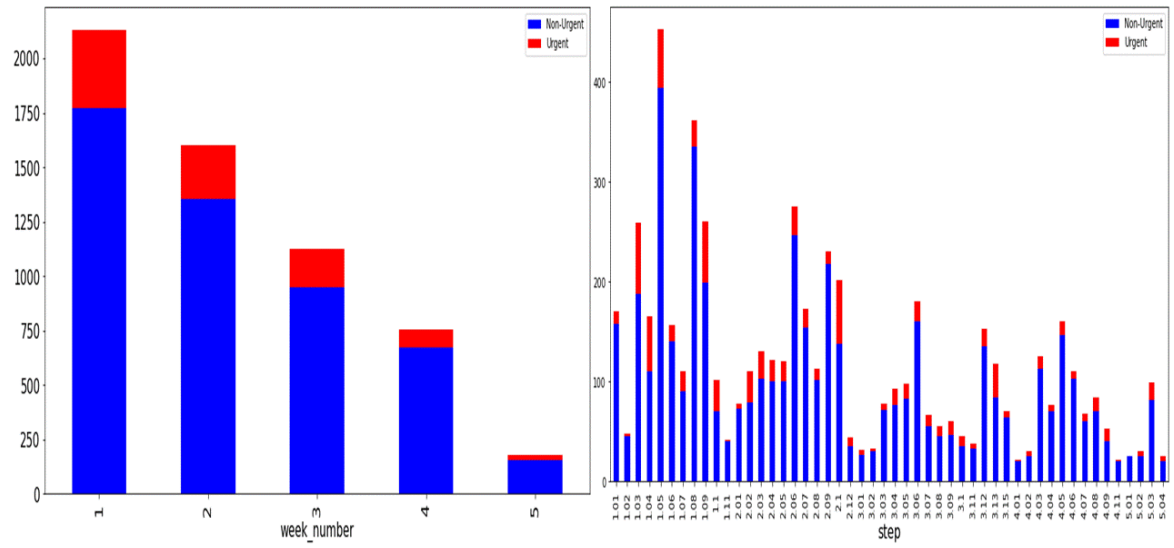

Figure 7.4: Comparing urgent and non-urgent post numbers for every week (left) and every step (right).

7.2.2. Exploring Urgency and Learner Behaviour

As an initial step, to understand learner behaviour in writing MOOC posts, the relationship between the number of posts written by learners who need urgent intervention with the average number of posts were explored. Then, to study the effect of urgency on learner behaviour, the relationship between HF commenters and their learning behaviour were explored — here, this involved simply comparing the number of step accesses made by learners. A learner who needs urgent intervention (HF commenters) was defined as per equation 7.1; let n = number of posts, u(p) = urgent posts, and p = a post.

$$HF\ Commenters = \frac{\sum_{n=1}^{\infty} u(p)}{\sum_{n=1}^{\infty}(p)} = 100\% \tag{7.1}$$

The average number of step accesses for each group (non-urgent) and HF commenters (urgent) were calculated to track how every group behaves on the platform.

Finally, completers were addressed with respect to their need for intervention. Completers were defined according to equation 7.2, where total access steps = number of total access per learner, total course steps = total number of steps in a course.

$$Completer = total\ access\ steps \geq total\ course\ steps * 0.80 \tag{7.2}$$

Completers were defined as in Eq. 7.2 because, in spite of the large number of previous studies, a formal definition of learner dropout is lacking (Sunar *et al.*, 2016). Therefore, the definition in (Alamri *et al.*, 2019) was applied, namely, that completers were defined as learners who accessed a number of steps equal to or higher than 80%.

7.2.3. Priority in Urgent Intervention

In this study, a new intervention framework designed was proposed to add prioritisation to urgent posts based on learner history to assist instructors' decisions to intervene and optimise their time and ability to adapt their intervention as needed. It begins by supposing that, when an instructor intervened, some of these posts were potentially urgent. Then, for these potentially urgent instances, priority (high-, mid-, or low-) was added, depending on the learner risk level.

The idea is to focus on learners, understand their behaviours, and perform segmentation based on three variables (urgency, sentiment analysis, and number of step accesses).

The model includes two phases (see Figure 7.5 below), In the first phase (prediction phase), a supervised classifier is used to predict if posts need a response urgently or not. In the second phase (intervention priority phase), the output of the previous phase (urgent posts) is taken as input. Then, priority is added to these posts based on the history of the learners who wrote these posts using unsupervised ML (clustering). Therefore, based on these groups, different priorities were assigned to posts.

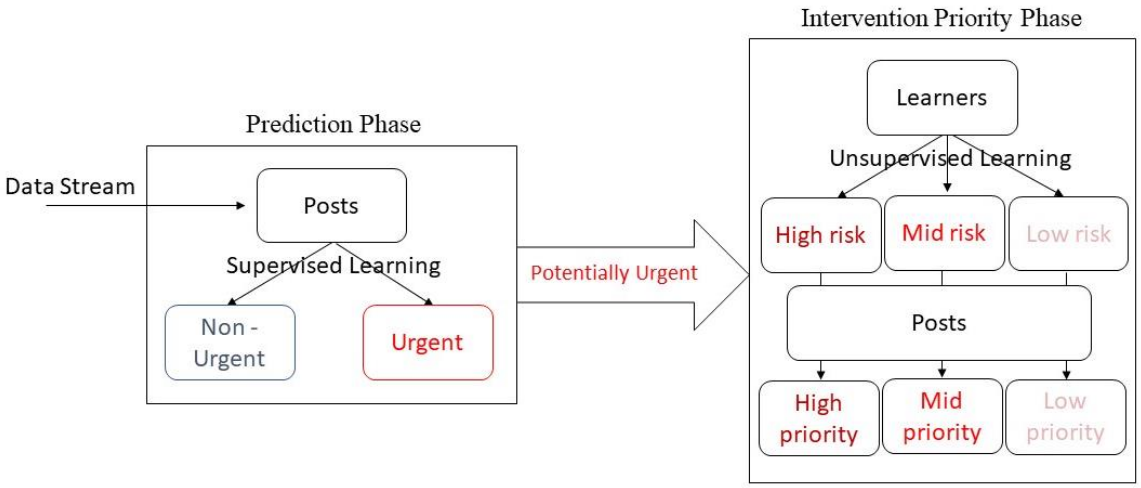

Figure 7.5: Priority in urgent intervention framework.

7.2.3.1. Prediction Phase

In this phase, the state-of-the-art text classification (BERT) model at the time of writing was applied to predict urgency. The 'bert-base-uncased' (L=12, H=768, A=12, Total Parameters=110M) version was used and the model was trained by setting batch size = 32 and epochs = 4.

7.2.3.2. Intervention Priority Phase

The behaviour of learners based on three variables (urgency, sentiment analysis, and step access) were studied. These three variables were studied because they address ***RQ4.2*** and ***RQ4.3***. Moreover, as clarified in Section 4.2.3.1, a sentiment analysis study found a negative

correlation between urgency and sentiment analysis; meaning urgent posts correlate with negative sentiments. The processing was as follows:

i. *Urgency.* To find the learners for whom most of their posts need intervention, the number of urgent posts for each learner was calculated. After that, all the learners were clustered in an unsupervised manner into three groups by assigning each learner to a specific cluster based on their number of urgent posts.

ii. *Sentiment Analysis.* To extract sentiment polarity, every post was analysed into three sentiment categories (positive, negative, and neutral) using the VADER tool (Hutto and Gilbert, 2014). This tool was selected because it is well-known and some research has shown that VADER outperforms Text Blob (Loria, 2018) in social media sentiment categorisation (Bonta and Janardhan, 2019; Min and Zulkarnain, 2020). Then, the overall average value of sentiments for each learner was found and sentiment clusters were created where a low sentiment number indicates high-risk learners.

iii. *Step Access*. For each learner, the number of step accesses was calculated. Then, learners were clustered into three groups based on these values. A high step access number is an important indicator of learning activity, possibly connected to high motivation.

For every variable (urgency, sentiment analysis, and step access), all learners were clustered into three groups by applying natural breaks optimisation with the Fisher Jenks algorithm (North, 2009) as it works on one dimensional data. Therefore, every learner has three scores that represent the three clusters' variables (urgency, sentiment analysis, and steps access). An overall score for every learner was calculated as in Eq. 7.3.

$$Overall_{score} = urgency_{cluster\text{-}score} + sentimentAnalysis_{cluster\text{-}score} + stepAccess_{cluster\text{-}score} \quad (7.3)$$

Thus, the overall score will be between 0–6. Then, the overall score was mapped onto different levels of risks:

- Higher than 3→ High risk;
- Higher than 1 and lower than or equal to 3→ Mid risk;
- Others → Low risk.

Then, learners were segmented as below:

- *High-risk.* Learners who had a high overall score based on the three variables (urgency, sentiment analysis, and access steps).
- *Mid-risk.* Learners who had a medium overall score based on the three variables.
- *Low-risk.* Learners who had a low overall score based on the three variables.

Based on these levels of risks, the priorities for intervention for all potentially urgent posts were calculated — see Algorithm 7.1.

Algorithm 7.1. Priority of Intervention (P, U, S, A)

Input:

- P: Stream of potentially urgent *post* instances.
- U: Number of *urgent* posts for each learner.
- S: Average value of posts' *sentiment* for each learner.
- A: Number of steps *access* for each learner.

Output:

i) Urgent posts with the priority intervention results.

Method:

```
Build 3 learner clusters for Urgency.
Build 3 learner clusters for Sentiment Analysis.
Build 3 learner clusters for Steps Access.
Compute the Overall Score.
if Overall Score is higher than 3 then
    High risk learner.
    Urgent post = high priority intervention.
else if Overall Score is higher than 1 then
    Mid risk learner.
    Urgent post = mid priority intervention.
else
    Low risk learner.
   Urgent post = low priority intervention.
end if
End Algorithm
```

7.3. Results and Discussions

In this section, the charts and the results of exploring urgency and learner behaviours are depicted to answer the following research questions, ***RQ4.1***, ***RQ4.2*** and ***RQ4.3***. Then, the

results of the proposed priority framework along with its confirmed effectiveness are proposed to address ***RQ4.4***.

7.3.1. Exploring Urgency and Learner Behaviour

To inspect learners' writing behaviour, an average number of posts were transformed into an urgency bar chart (1 urgent post, 2 urgent posts, etc), as shown in Figure 7.6 (below). Interestingly, it was observed that, usually (but not always), if a learner writes more posts that need intervention, they tend to write more posts in total. This is useful in that they do not 'give up' and present for longer time which allows instructors a better opportunity to offer intervention.

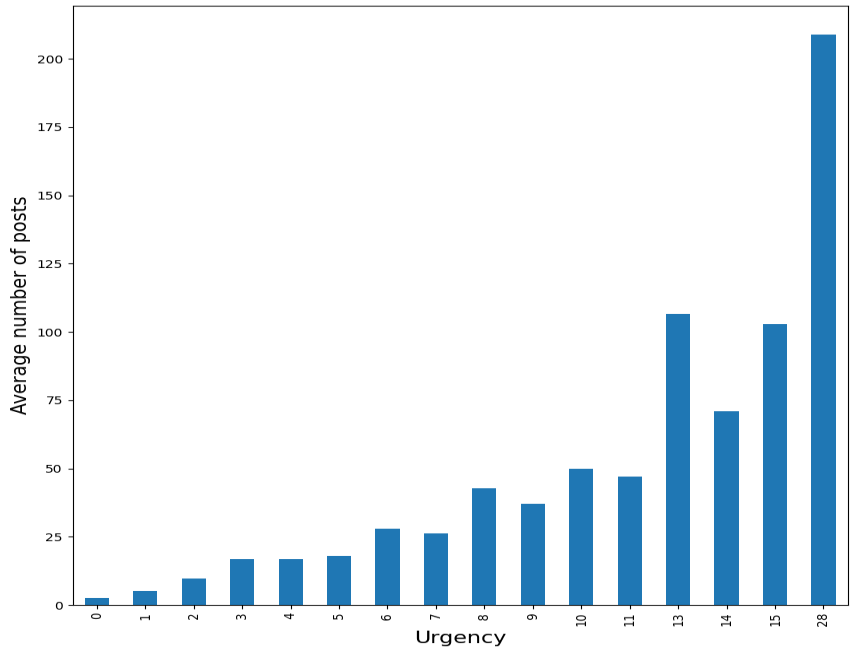

Figure 7.6: Relationship between urgent posts (urgency) and average number of posts.

For the relationship between HF commenters and the number of steps accessed, as (Figure 7.7, left, below) shows, the average numbers of steps accessed were calculated for HF learners (urgent group) and the non-urgent group. The findings were as follows: in general, both groups accessed learning materials, but the average number of step accesses in the urgent group was lower (33 steps). This difference is statistically significant (Mann-Whitney U test: $p < 0.05$). Consequently, the key observation here indicates that learners who make posts not needing intervention were likely to have potentially high levels of motivation and so engage in more learning activity.

The results on the relationship between urgency and completion is shown in Figure 7.7 (right, below). As seen, HF learners who require urgent intervention are less likely to complete the course (13 %). This difference is statistically significant (Mann-Whitney U test: $p < 0.05$). From this result, it can be concluded that learners who need intervention tend not to complete the course, which may be one of the reasons for the high dropout rate and so confirms the need for instructor intervention in urgent posts.

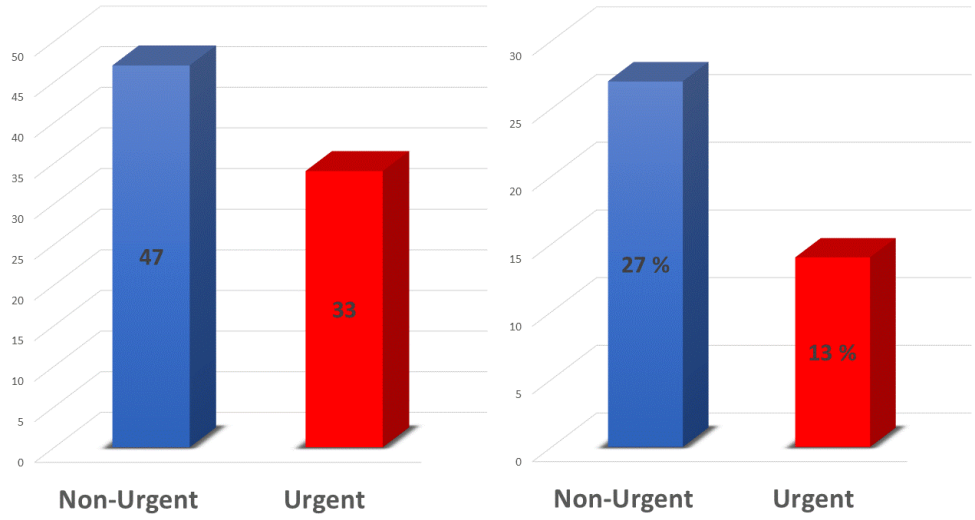

Figure 7.7: For each group: average number of steps accessed (left), completion rate (right).

7.3.2. Priority in Urgent Intervention

As per Section 7.2.3, a framework containing two phases was proposed. It supposed that the instructor can decide to intervene after five weeks (original data). In the prediction phase, a stratified five-fold cross-validation was used to estimate the performance of the classification model. To evaluate BERT, accuracy averaged over two classes was measured (urgent, non-urgent), recall, precision, and F1-score for the important and minority urgent classes (Table 7.2 below). The *recall* metric was prioritised that gave us the rare urgent cases rather than *precision* — preferring to ensure that all urgent cases were captured.

Table 7.2: The results of the BERT model Average Acc, P, R, F1 % for class 1 (Urgent).

Acc	P	R	F1
0.90	0.65	0.72	0.68

In the intervention priority phase, there were 387 commenters who had at least one post that needed urgent intervention. Table 7.3 (below) shows the minimum (min) and maximum (max)

for each variable in every cluster. For urgency labelling, the label resulting from the manual annotators with a voting mechanism was used, not the one predicted by a classifier, to increase accuracy.

Table 7.3: The minimum (min) and maximum (max) for each variable in every cluster.

Cluster	Urgency 'min : max'	Sentiment Analysis 'min : max'	Steps Access 'min : max'
0	'1 : 3'	'27 : 75'	'35 : 52'
1	'4 : 9'	'7 : 24'	'15 : 34'
2	'10 : 28'	'-3 : 6'	'0 : 14'

Finally, to further validate the effectiveness of this proposed model, the relationship between different risk groups of learners (high, mid, low) was identified and their completion-rates were computed. The distributions are visualised in Figure 7.8 (below). From this box plot, it can be noted that most completion-rates of high-risk learners are very low, whilst mid-risk learners have average completion-rates and those of low-risk learners are very high. This is further confirmation that the proposed risk model, based on data from the first half of the course, and refining the potential urgency model, can correctly predict learners at risk of not completing their course and separate them from the other two milder risk groups.

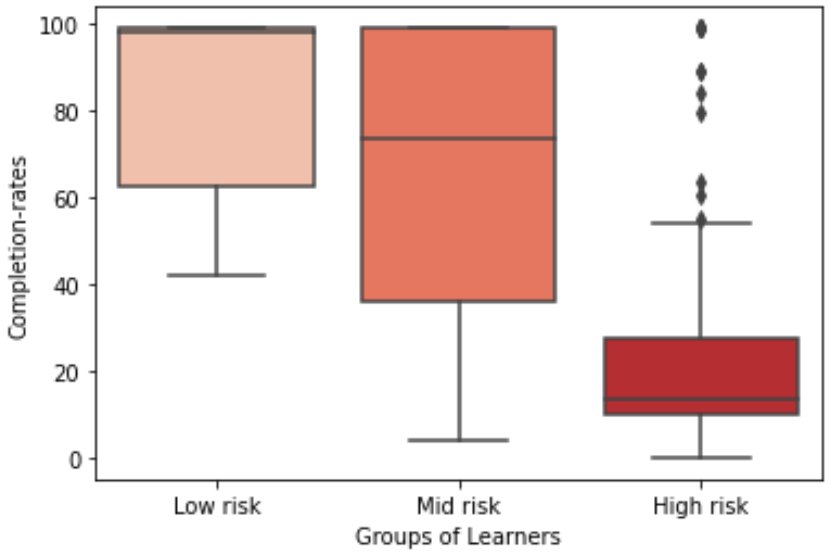

Figure 7.8: Box plot for groups of learners' risk and their completion-rates.

7.4. Epilogue

In this chapter the problem of making automatic, intelligent intervention in MOOCs posts for learners who require attention was addressed. The analysis of learner posts for urgency demonstrated that learners with high step access rates require less instructor intervention, whilst the step access of HF commenters is less than that of other commenters. This might be due to a decrease in learners' motivation to continue accessing the course material when they have many posts that need intervention. In addition, the results confirmed that most course completers did not need much intervention to their posts. Based on these findings, a framework and algorithm for prioritising intervention was constructed to encourage instructors to help their learners and support them by focusing on high-risk learners first to improve the potential outcomes of the intervention. This framework can be used as part of an intelligent system in MOOC environments.

The following chapter examines various data balancing methods to solve the imbalanced data issue and apply different traditional and Transformer models to identify urgent posts. In addition, it develops learner and instructor models to assist instructors in responding to urgent posts in MOOCs.

CHAPTER 8: INTERVENTION PREDICTION: POST-BASED AND USER MODELLING (SOLVING THE IMBALANCED DATA ISSUE)

8.1. Prologue

In MOOCs, identifying urgent posts on discussion forums is an ongoing challenge. Whilst urgent posts require immediate reactions from instructors to improve interaction with their learners, the task is difficult as truly urgent posts are rare. From a data analytics perspective, this represents a highly unbalanced (sparse) dataset.

This chapter aimed to automate the urgent-post identification process based on fine-grained learner modelling for use in generating automatic intervention recommendations for instructors. To showcase and compare these models, the models were applied to the first gold-standard dataset for Urgent iNstructor InTErvention (UNITE) which was created by labelling FutureLearn MOOC data (Section 3.2.2.1). Both benchmark traditional classifiers and Transformer were implemented.

The core problem with MOOC data is that it is intrinsically imbalanced; such datasets are characterised by a highly skewed class distribution due to the (naturally) small number of instances of urgent posts. In text classification tasks, performance often depends on the quality of the data (Wei and Zou, 2019). Therefore, to tackle the imbalanced data problem and improve the size and quality of the training data, the dataset was manipulated by comparing, for the first time for the unbalanced data problem, three data balancing techniques: (i) text augmentation, (ii) text augmentation with undersampling, and (iii) undersampling. Also, several new pipelines for combining different augmenters for text augmentation are proposed.

To illustrate the usage of the fine-grained learner models to provide adaptive support for instructor intervention in MOOC environments, an adaptation case is described where instructors can decrease their workload by using one of the proposed models. Also, an expanded model is illustrated that uses more extensive learner knowledge (based on the number of posts per learner) to discuss how such adaptation models can be further expanded.

This chapter addresses the following two RQs:

- ***RQ5.1:*** *How can the data imbalance issue (urgent versus non-urgent) in learners' posts be addressed?*
- ***RQ5.2:*** *What would an adaptive intervention model to assist instructors in dealing with urgent posts look like?*

The main contributions of this chapter are: (i) for the first time in the literature, applying data balancing techniques for traditional ML and Transformer models to identify instances when urgent instructor intervention is required in MOOC environments. These techniques include text augmentation, text augmentation with undersampling, and undersampling to overcome the imbalanced data problem and improve performance; (ii) proposing several new pipelines (3X and 9X) to generate more data for text augmentation by incorporating different NLP augmenters and providing a range of approaches; (iii) creating the first learner, instructor, and adaptation models to support instructors to deal with urgent posts in MOOCs; (iv) showcasing the challenges and difficulties involved in instructor-intervention decisions in MOOC environments by manually inspecting and analysing the (relatively small) set of errors generated by the best classifier, along with the best data balancing and text augmentation solutions.

8.2. Related Work

An obvious issue related to the instructor intervention problem in MOOC environments is that for urgent posts, imbalanced data is a characteristic of the data itself (as there are fewer urgent posts than non-urgent, normally). This fact has been largely overlooked in urgent post detection: the closest research to this (Almatrafi, Johri and Rangwala, 2018; Khodeir, 2021) which considered some standard techniques, e.g., data-splitting, model-training, and evaluation-metric selection; however, it failed to deal with improving the data imbalance. In addition, while available intervention models for urgent posts concentrated on classifying posts as clarified before in SLR Chapter 2 (post-based identification), they did not pay any attention to the behaviours of learners or designed adaptive instructor intervention models based on

learner (or instructor) models. Therefore, this section reviews the literature closest to this chapter's proposal: (i) the area of text augmentation, specifically for balancing data, and (ii) adaptive models in MOOCs.

8.2.1. Text Augmentation

This section presents related works on text augmentation in NLP. The aim of text augmentation is to expand data (Liu *et al.*, 2020) by providing and applying a set of techniques that create synthetic data from an existing dataset (Shorten, Khoshgoftaar and Furht, 2021). The performance of model predictions on a number of NLP tasks can be enhanced by text augmentation and this prevents overfitting (Li *et al.*, 2022). Text augmentation is used to alleviate the issue of limited or scarce labelled training data (Anaby-Tavor *et al.*, 2020), which leads to low accuracy and recall for the minority class (Liu *et al.*, 2020).

The existing literature shows that previous researchers utilised NLP augmentation approaches; for example, (Wang and Yang, 2015) applied text augmentation by performing synonym replacement and identifying similar words based on lexical and semantic embedding. Another study by (Kobayashi, 2018) proposed a new word-based approach for text augmentation based on contextual augmentation; they applied synonym replacement by using a bi-directional predictive language model. Next, (Wei and Zou, 2019) explored straightforward text editing techniques for augmentation using one of four simple techniques (synonym replacement, random insertion, random swap, and random deletion). Recent work (Xiang *et al.*, 2020) proposed a part-of-speech-focused lexical substitution for data augmentation (PLSDA) approach to generate more instances via word substitution. Another augmentation work was applied in translation: (Yu *et al.*, 2018) generated new data to enhance their training data using back-translation with two translation models: the first translates sentences from English to French, while the second translates from French to English.

Some researchers tackled augmentation by using text augmentation libraries (NLPAug) for specific tasks. (Jungiewicz and Smywiński-Pohl, 2020) used a range of augmentation techniques for sentiment analysis, including (NLPAug) based on BERT and WordNet. More recently, (Pereira *et al.*, 2021) used the same BERT-based library and contextual word embedding augmenter to generate more programming problem statements on a training dataset.

In this experiment, the text data were also augmented based on the NLPAug library. Unlike in prior research which usually focuses on the word-level for augmented data, several different levels (character, word, sentence) were used. Different techniques were applied based on word

embedding: word2vec (using words as a target), contextual word embedding: BERT, DistilBERT, RoBERTa, and XLNet (using words or sentences as a target), and OCR engine error (using characters as a target). In addition, various pipelines were created based on sequential flow. Three different approaches were constructed because, in textual augmentation, the best approach is based on the dataset; if any approach improved on performance for specific data, this may be detrimental to other data (Qiu *et al.*, 2020).

8.2.2. Adaptive Models in MOOCs

The other branch of prior research relevant to this chapter is adaptation and adaptive models implemented in MOOCs. As MOOCs are a rather recent addition as clarified in Section 2.2.1, with the term MOOC coined in 2008 (Stracke and Bozkurt, 2019), and their launch in 2012 (Jordan and Goshtasbpour, 2022), adaptation has been slow to be introduced to MOOC data, with most approaches still being designed using a one-size-fits-all basis (Shimabukuro, 2016; Rizvi *et al.*, 2022) to some extent, despite the decades of research in adaptive educational hypermedia (Ahmadaliev *et al.*, 2019), intelligent tutoring systems (Mousavinasab *et al.*, 2021; Hodgson *et al.*, 2021), and the like. Nevertheless, a few researchers have started proposing adaptation in MOOCs. For instance, (Alzetta *et al.*, 2018) designed a customised learning path in an interactive and mobile learning environment and MOOCs using a question/answer (QA) system. Another work on adaptive models in MOOCs (Lallé and Conati, 2020) created a framework for user modelling and adaptation (FUMA) to provide adaptive support for learners' during video usage. They used video watching and interaction behaviours as features to reveal inactive learners. Another very recent work proposes an optimal learning path to prevent MOOC learners from dropping out (Smaili *et al.*, 2022); they provide each learner with an adaptive appropriate path based on interaction with the environment using particle swarm optimisation (PSO).

In this research, unlike in previous research, the building of adaptive models was enabled based on learner posts with the aim of improving communication with instructors.

8.3. Methodology

This study aimed to automatically classify if a MOOC learner's post is urgent and so requires flagging for instructor intervention. This means modelling learner data (their posts) to recommend an action to the instructor (here, reply). This is called a fine-grained learner model,

as each learner is represented by the set of their posts. More formally, it can be written that for learner l_1, their learner model L between time points t_1 and t_2 is given by:

$$L(l_1,t_1,t_2)= \{F_{t(p)\in[t1,t2]} (\text{urgency } (l_1,p))\} \quad (8.1)$$

Where F(.) can be any function aggregating the urgency for a given interval (e.g. a sum of urgency), and urgency (l_1,p) represents the fine-grained learner information at the level of a single post p of learner l_1, made during the given time interval [t_1,t_2]. This learner model L(.) is used to generate recommendations for instructors (see Section 8.3.3). To achieve this objective, the FutureLearn corpus was manually annotated (as discussed in Section 3.2.2.1); additionally, the highly popular and well-used benchmark Stanford dataset was used (Section 3.2.1) to validate the best model, thus demonstrating the generalisability of the proposed approach and its applicability across courses and domains.

To determine the most appropriate method, NLP techniques were used to construct a diverse predictive model for text classification. Two main types of supervised classifiers were employed:

1. A traditional ML approach with handcrafted features as a baseline model; and
2. A fine-tuned version of BERT, representing the latest advance in NLP at the time of writing (2021) as a powerful supervised Transformer model, as discussed in Section 2.2.3.2.5.1.

To tackle the imbalance problem, several different techniques were employed (see Section 8.3.2.2). One technique that was considered is text augmentation; here, different approaches were relied upon (see Text Augmentation Section 8.3.2.2.2) and the minority-class data were augmented with various multipliers (such as 3x and 9x). The reason for using text augmentation is that it prevents overfitting; it is considered a crucial regularisation technique (Coulombe, 2018).

8.3.1. Dataset

This research was conducted on the FutureLearn and Stanford MOOC-based platform datasets. In FutureLearn, Urgent iNstructor InTErvention (UNITE) was used because it contains a very limited number of urgent cases (7%) which represent unbalanced data. Then, as mentioned, the best model from FutureLearn (UNITE) was validated using the Stanford MOOC-based dataset.

8.3.2. Experiments for Imbalanced Data

To achieve a comprehensive understanding of the best way to automatically identify the urgency of posts on MOOCs, as mentioned, two common supervised ML strategies were used (traditional ML and Transformer with BERT) to automatically classify posts. Additionally, as urgency-detection is a typically imbalanced data problem, hence any MOOC provider would need to take this data imbalance into account — various techniques to deal with input data were experimented on, as per Figure 8.1 (below).

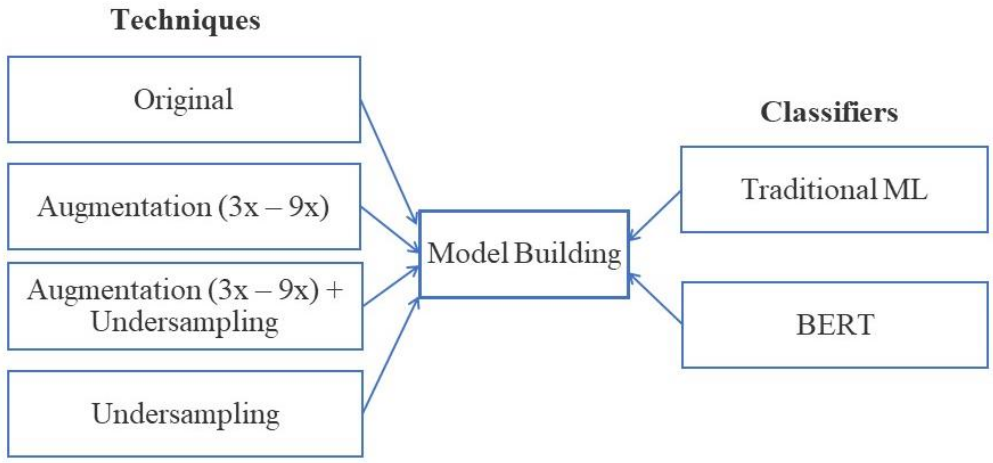

Figure 8.1: The proposed pre-processing (data balancing) and ML pipeline combinations.

First, several training models were applied to the original data on the UNITE corpus. Then, to improve performance, three solutions were designed and developed to handle imbalanced data: (i) text augmentation; (ii) text augmentation + undersampling; (iii) undersampling (for details see Section 8.3.2.2). Text augmentation involves using a range of approaches in different combinations to augment the minority class in the training data. In undersampling, randomly selected instances from the majority class were used, while in text augmentation + undersampling involves using a combination of the two previous techniques. All the experiments were conducted using a stratified 4-fold cross-validation approach to ensure representative results. The general architecture of the proposal classification model is shown in Figure 8.2 (below).

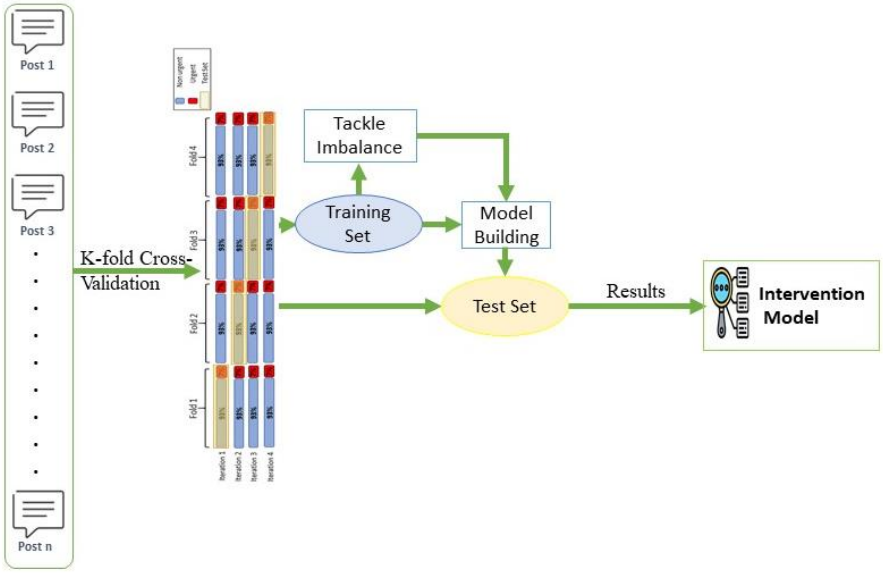

Figure 8.2: The general architecture of the classification model.

8.3.2.1. Classifiers

As mentioned, two major classification model types were compared to classify the posts: (i) traditional ML (a basic model typically used by ML algorithms), and (ii) BERT (one of the most popular Transformer models, as further explained).

8.3.2.1.1. Traditional Machine Learning

Several ML models were applied (see Figure 8.3 below) to the classification task, each with different fundamental mechanisms for feature engineering. This includes count vector and term frequency-inverse document frequency (TF-IDF) to find an adequate classifier to predict urgent posts. Different feature sets were extracted via four different classical methods: (i) count vector; (ii) TF-IDF vector (word-level); (iii) TF-IDF vector (n-gram word-level); and (iv) TF-IDF vector (n-gram character level). Then, different popular classifiers were built across these different sets of features (naive Bayes, logistic regression, support vector machine, random forest, and boosting model — extreme gradient boosting (XGBoost)), as displayed in Figure 8.3 (below).

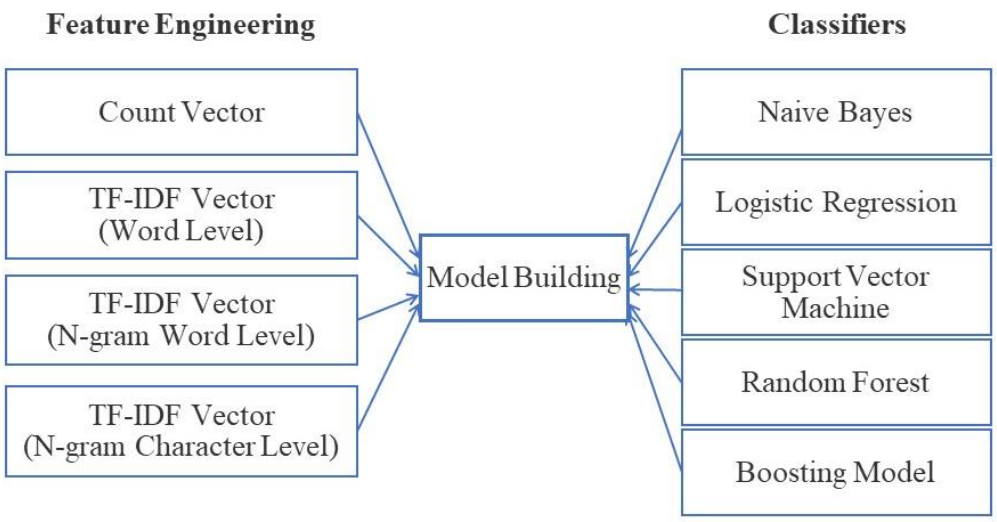

Figure 8.3: The framework of the traditional ML classifiers using different features.

Each post was represented with a specific vector; the count-vector counts the frequency of every given word in every post. TF-IDF calculates the score in the form of a numerical statistic to evaluate the degree of relatedness between a particular word and a specific post in a collection of posts; it thus represents a measure of how important a word is in a collection of posts. Three different levels of TF-IDF were considered as tokens: (i) word, (ii) n-gram word with range (2,3), and (iii) n-gram character with a range of (2,3) with the maximum number of features (5000 applied to each level).

8.3.2.1.2. BERT

For Transformers, the currently most popular and competitive approach in text classification tasks, BERT, was employed. Using BERT enabled feature engineering to be avoided as a well-known approach in deep learning. A pre-trained BERT was fine-tuned with one additional layer for the classification task. The version of BERT classifier used was the BERT-base-uncased (L=12, H=768, A=12, Total Parameters=110M) and it is the smaller model of the two available (as explained in Section 2.2.3.2.5.1) and it was selected due to shorter training time. For the BERT input, which is a sequence of tokens, each post was limited to the final 128 tokens. This decision on the final tokens and size was based on various pre-experiment trials (final/first tokens; different sizes) that rendered this number (128 tokens) as the most suitable. The Adam optimiser was used to tune BERT over four iterations.

8.3.2.2. Text Balancing Techniques

Several classifier models were developed based on different techniques for manipulating the data. First, each of the models was run using the original UNITE corpus. Then, to tackle the imbalance problem, the following approaches were independently applied: (i) text augmentation, (ii) combined text augmentation then undersampling, and (iii) resampling using undersampling.

8.3.2.2.1. Original Data Usage (UNITE Corpus)

As an initial experiment, all models were implemented directly with original UNITE data. The dataset was split into four groups using stratified k-fold cross-validation, choosing a value of k = 4 (four folds) as in (Capuano and Caballé, 2019; Capuano *et al.*, 2021). The k-fold cross-validation-run approach was chosen because it allowed results with less bias to specific data to be obtained (Berrar, 2019). Stratification in the dataset was used: the selection of data led to an equal distribution of every class in every set. Thus, every fold contained the same percentage of samples from each class (see Figure 8.4 below) as follows: training fold 3466 or 3467 samples (3219 as class 0, i.e. non-urgent, and 247 or 248 as class 1); testing fold 1156 or 1155 samples (1073 as 0 and 83 or 82 as 1) in each iteration (see Table 8.1 below). Please note that the more frequently encountered 10-fold validation was not used, as, due to the very low number of urgent cases, this would have resulted in a too-low value per stratum for efficient stratification.

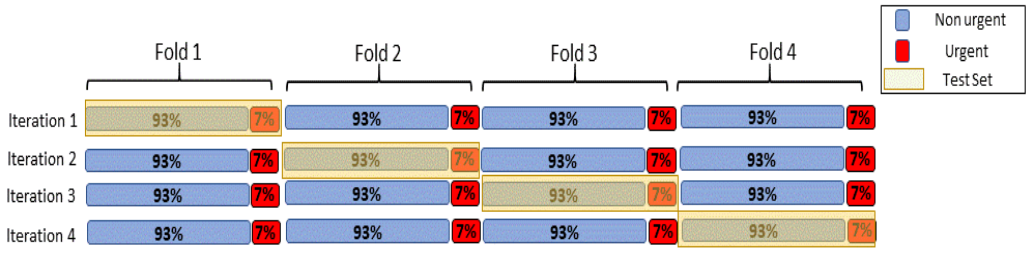

Figure 8.4: Splitting the data using 4-fold cross-validation and stratification.

Table 8.1: Number of cases for every class in (training, testing) sets in each iteration: original data.

# of iteration	Training set		Testing set	
	0	1	0	1
1	3219	247	1073	83
2	3219	247	1073	83
3	3219	248	1073	82
4	3219	248	1073	82

For the training with BERT, the training data were divided into 90% training (0=2897, 1=222 or 223) and 10% validation (0=322 1=25), as well as the use of stratification.

However, it was found that the results were unsatisfactory for the various classifiers (see Section 8.4) due to class imbalance. To overcome this issue and enhance prediction performance, alternative techniques were employed as discussed in the next sections.

8.3.2.2.2. Text Augmentation

To manage the class imbalance problem and boost performance, the data instead was pre-processed using artificial resampling (augmentation) to generate more minority-class cases for the training set of each fold, resulting in an almost balanced dataset. Every instance was augmented in the minority class into three and nine instances, respectively. These values were chosen based on literature reporting that for some databases, a low number of repetitions might not be sufficient to decrease the bias of the model in indiscriminately predicting the majority class; however, a higher repetition value might also render the data non-representative (Haixiang *et al.*, 2017; Madabushi, Kochkina and Castelle, 2020; Fonseca *et al.*, 2020), so experimentation was necessary. Thus, in this work, experimentation was performed at every iteration, with the number of items in the training and testing set for 3x and 9x augmentation, as shown in Table 8.2 (below).

Table 8.2: Number of cases for every class in (training, testing) sets in each iteration: text augmentation (3x – 9x).

Quantities	# of iteration	Training set		Testing set	
		0	1	0	1
3x	1	3219	988	1073	83
	2	3219	988	1073	83
	3	3219	992	1073	82
	4	3219	992	1073	82
9x	1	3219	2470	1073	83
	2	3219	2470	1073	83
	3	3219	2480	1073	82
	4	3219	2480	1073	82

To achieve the augmentation goal, common, easy-to-implement techniques for text augmentation were applied using the public NLPAug library. The text augmentation library (NLPAug) is a Python library dedicated to augmentation (Raghu and Schmidt, 2020). The simple code via the Edward Makcedward Github repository (Makcedward, 2020) was accessed. Three different hybrid approaches were used: (i) word-level with the same type (BERT), (ii) word-level with different types, and (iii) different levels (character, word, sentence), as shown in Table 8.3 (below).

Table 8.3: The approaches using different augmenters.

Approach	Level	Augmenter	Type	Action
1	Word	ContextualWordEmbsAug	BERT	Insert
				Substitute
			DistilBERT	Substitute
2	Word	WordEmbsAug	Word2vec	Substitute
		ContextualWordEmbsAug	BERT	Substitute
			RoBERTa	Substitute
3	Character	OcrAug	OCR	Substitute
	Word	ContextualWordEmbsAug	BERT	Substitute
	Sentence	ContextualWordEmbsForSentenceAug	XLNet	Insert

In the first, a hybrid approach consisting of three different actions 3x in a ContextualWordEmbsAug augmenter based on BERT was applied — by inserting and substituting with BERT and substituting with DistilBERT — to discover the most appropriate word for augmentation, as shown in Table 8.4 (below).

Table 8.4: An example of different augmenters for 3x in the first approach on a post in UNITE.

Type	Text
Original	I hope any course staff member can help us to solve this confusion asap!!!
BERT (insert)	i hope any course support staff member can come help enable us to solve this current confusion case asap ! ! !
BERT (substitute)	our trust one important staff member can help us to solve this confusion slowly ! ! !
DistilBERT (substitute)	i hope any course faculty member should teach us to alleviate problem confusion asap ! ! !

Then, the 3x method was built, and the number of instances was increased to 9x by generating an additional 3x more instances for every instance. This was achieved by constructing six sequential pipelines, each representing a multi-augmenter (bi- or tri-augmenter), as shown in Table 8.5 (below). Table 8.6 (below) provides examples of 9x augmentation. From this table, it can be noticed that there is an issue with the quality of augmented text on some pipelines. Despite this issue, the performance is improved, as discussed in the results section.

Table 8.5: Different pipelines to generate 9x in the first approach.

Pipeline	Type	Action
Pipeline 1	BERT	Insert
	BERT	Substitute
Pipeline 2	BERT	Insert
	DistilBERT	Substitute
Pipeline 3	BERT	Substitute
	BERT	Insert
	DistilBERT	Substitute
Pipeline 4	BERT	Substitute
	DistilBERT	Substitute
Pipeline 5	DistilBERT	Substitute
	BERT	Substitute
	BERT	Insert
Pipeline 6	DistilBERT	Substitute
	BERT	Insert

Table 8.6: An example of different augmenters for 9x in the first approach.

Type	Text
Original	I hope any course staff member can help us to solve this confusion asap!!!
BERT (insert)	i hope any acting course staff member can help us financially to solve these this ... confusion situation asap ! ! !
BERT (substitute)	i recommend a course staff member can help our all solve this confusion tonight ! ! !
DistilBERT (substitute)	i hope any helpful staff member may help us to unlock the mystery asap ! ! !
Pipeline 1	the four know some successful course change group member can even get us this solve this global confusion asap ! ! !

Pipeline 2	as i hope any course staff staff experienced can somehow help inspire us and suggest solving this puzzle asap ! ! !
Pipeline 3	i wonder if various further course instructors or volunteers could employ yo `u might ultimately solve this particular trouble indeed ! ! !
Pipeline 4	we hope only one staff volunteer to help us both solve their confusion immediately ! ! !
Pipeline 5	sincerely hope any new permanent staff department member cannot aid me by easily in solve this time at crisis ! ! !
Pipeline 6	and i hope for any facebook staff member member can persuade them to quickly solve this situation well together ! ! !

Next, the second approach was conducted, another augmentation procedure, by mixing several augmenter functions based on the word-level (see Table 8.3 above): WordEmbsAug (substitute word2vec) and ContextualWordEmbsAug (substitute BERT and substitute RoBERTa).

Lastly, as per Table 8.3 (above), the third approach was constructed, which is based on three different levels of augmenter (character, word, and sentence). For character-level, OcrAug (a substitute for OCR) was used. For word-level, ContextualWordEmbsAug (a substitute for BERT) was used. For sentence-level, ContextualWordEmbsForSentenceAug (insert XLNet) was used.

Then, the traditional ML and BERT models were applied, as explained in Section 8.3.2.1, based on 3x and 9x augmentations.

8.3.2.2.3. Text Augmentation + Undersampling

By creating nine new artificial instances in the training set, an almost-balanced dataset was obtained, albeit with a concern about its non-representativity. However, by creating three new instances, the data variation was moderately increased and a smaller move towards balancing the dataset was performed. Hence, the concern of minimising model errors was addressed by frequently predicting the majority class, achieving instead high accuracy yet low recall and precision for the minority class. These two concerns were dealt with by applying a hybrid resampling method combining this augmentation technique with undersampling.

In these experiments, the aim was to balance the datasets by combining both text augmentation and undersampling methods as follows. First, by increasing instances to 3x or 9x in the minority class. Second, in undersampling, randomly reducing the number of elements in the majority class to be equal to the minority class in every fold. Therefore, the numbers of

samples for each pipeline in the urgent and non-urgent classes were approximately 990 for 3x and 2475 for 9x.

8.3.2.2.4. Undersampling (Random)

To balance the class distribution in the original data, an alternative popular method was performed — the undersampling technique for imbalanced data classification — by randomly removing instances in the majority class. Thus, in this case, the numbers of samples for each class were 247 or 248, respectively.

8.3.2.2.5. FutureLearn and Stanford Datasets

As explained, the distribution of the urgent class in the FutureLearn dataset (7%) was different than in the Stanford dataset (19%). Therefore, the effect of these different techniques to handle imbalanced data were expected to affect the performance results of the different datasets. Figures 8.5 and 8.6 (below) show the distribution of every class in every fold for every method for UNITE and 3x for the Stanford dataset, respectively.

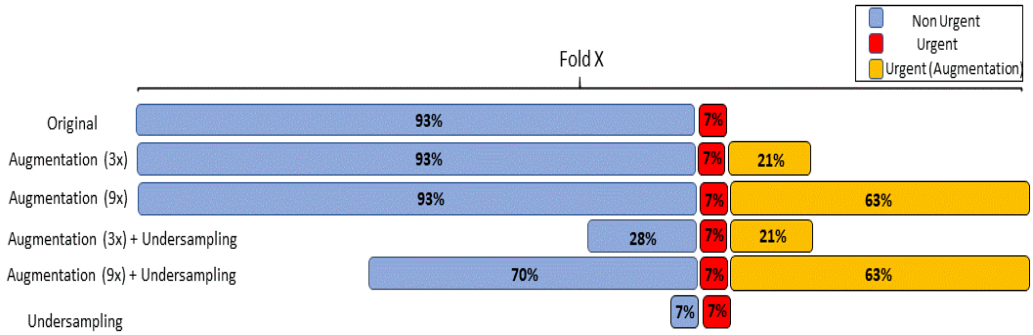

Figure 8.5: The distribution of every class in every fold in every method for UNITE: FutureLearn dataset.

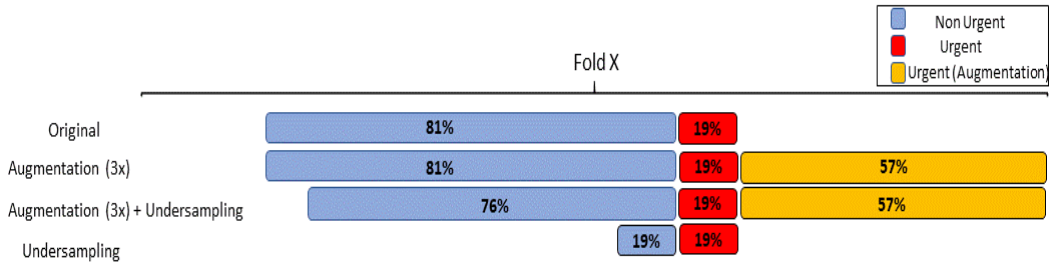

Figure 8.6: The distribution of every class in every fold for every method for the Stanford dataset.

8.3.3. Illustration of Adaptive Intervention Models

This section introduces the design of illustrative adaptive intervention models for instructor interaction based on the automatic urgency detection approach. These models showcase how the user model parameters proposed by this study can fit into simpler or, gradually, more complex user models; here, the term *users* means *instructors*, as the primary target users, and *learners,* as the potential secondary target users. Specifically, two practical scenarios for semi-automatic instructor intervention were provided: (i) semi-automatic intervention that tackles unbalanced data with a classification model, and (ii) filtering posts that improve instructor intervention by filtering the results based on learners, their number of posts, and time of posting.

8.3.3.1. Semi-automatic Instructor Intervention: Basic Scenario

The first scenario introduces an artificial support instructor model as a pipeline incorporating the classification model to represent the learner model using additional information on the instructor (the instructor model), as shown in Figure 8.7 (below).

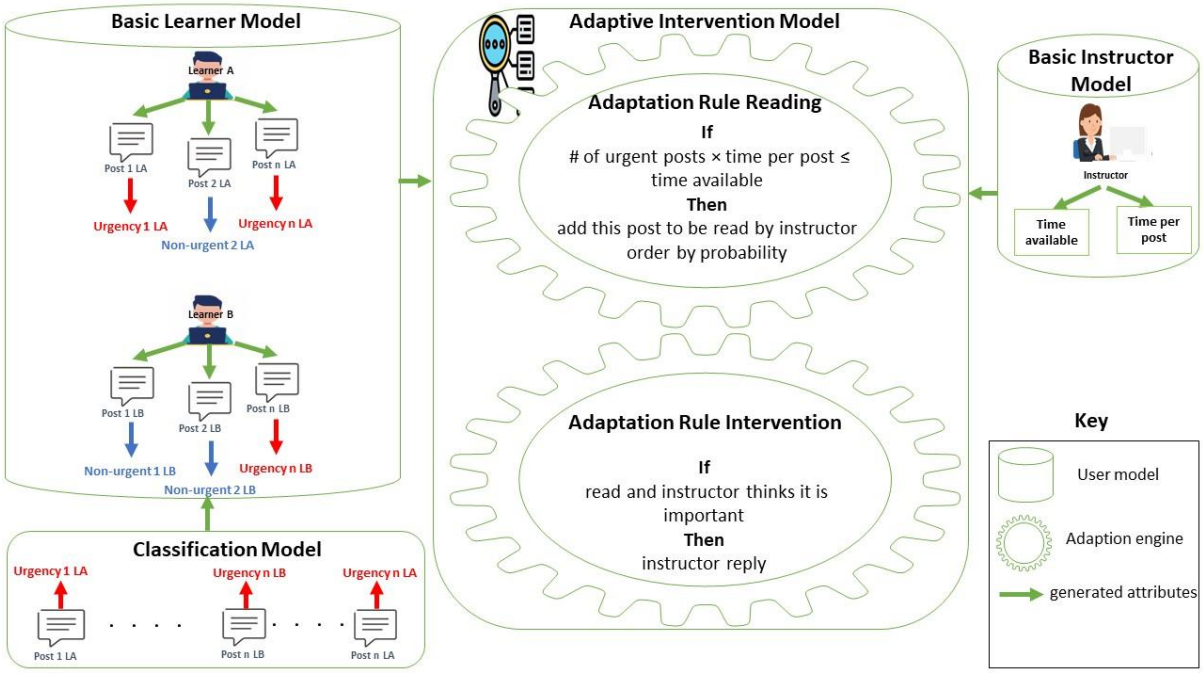

Figure 8.7: The adaptive intervention model based on learners' posts; note how the proposed predicted urgency becomes a (derived, fine-grained) learner model variable, together with the posts per learner.

A basic instructor model would minimally contain variables such as the available time instructors have for a specific session, and a time for reading per post, or, alternatively, a maximum number of posts to read in that session (hence, a simple two-variable user model for

the instructor). The learner model also contains two variables: (i) learner posts, and (ii) urgency of posts at post-level (fine-grained). Based on this information, the adaptive intervention model can automatically retrieve the top-most urgent posts, depending on their ranking (e.g., from a probability score given by a classification model), thus reducing overload on instructors.

For example, instructor Laura has answered all yesterday's posts from learners. She wishes to know if there are any urgent posts today as she has only 30 minutes, after which she needs to teach another course. All this information represents the instructor model. The MOOC webpage for today has three items, each has acquired an average total of 150 posts. She thinks that she would be able to answer a maximum of about 10–15 posts (and adds this information to her instructor model[15]). Thus, the artificial support instructor recommends Laura to answer the most urgent top five posts for each of the three items from today's class. This recommendation represents the adaptive model, which is the combination of the classification model and the proposed technique to deal with imbalanced data, which automatically classifies posts and detects urgent posts thus adapting to the instructor's needs, helping Laura to avoid reading all the posts, and improving her interaction with the learners.

8.3.3.2. Semi-adaptive Instructor Intervention: Expanded Scenario based on Coarse Granularity and Expanded Learner Models

The first scenario deals with the recommended urgent posts, as per the pipeline proposed in this research project. However, this model can be further improved. Next, how posts can be grouped to further refine the learner model and deal with urgency at (higher granularity) learner level (instead of the post level) is described. This may show if a learner is generally in trouble and needs support, which may make dealing with that learner more useful. This is consistent with findings of a study by (Alrajhi *et al.*, 2021) and clarified in Chapter 7, which showed that learners write more posts overall when they require urgent intervention.

For example, instructor John wishes to use Laura's system for classifying posts but has noticed that learners tend to either make many urgent posts when they are in trouble or are overall happy, and thus make fewer posts. He would like his workload reduced and hence avoid answering to seemingly urgent posts made by users with very few posts. Thus, he wishes learners to be grouped into urgent and non-urgent learners, as shown in Figure 8.8 (below).

[15] Alternatively, the system could automatically convert Laura's available time (of 30 minutes) into a number of questions to be answered.

John will now be able to answer urgent learners first, even if some of the non-urgent learners may have posted posts that sound urgent but who may have less need for intervention.

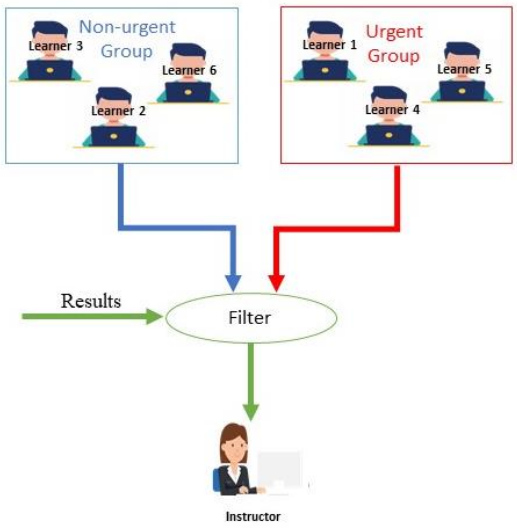

Figure 8.8: Refining the learning modelling of urgency based on two learner groups (non-urgent/urgent).

An extension to the learner model would be to add this coarse-grained, learner-level classification to the learner model, the number of posts, and then to further cluster them based on this. The correlation between the number of written posts per learner versus the number of posts made by those who need urgent intervention was computed using Pearson's correlation. Therefore, first, silhouette analysis was applied, to check the number of clusters, then the Fisher-Jenks algorithm was used (because it only works on one-dimensional data) to perform the clustering. These clusters were then merged into two groups that differentiate between learners with a high number of posts and learners with a low number of posts (urgent/ non-urgent learners).

In addition, the intervention was further adapted based on the time stamp of the posts of each of these learners to provide John with posts of the urgent learners, ordered on a first-come-first-served (FCFS) basis. Thus, number of posts and time stamps are variables added to the extended learner model in this example. The overall adaptive model is summarised in Figure 8.9 (below) using the same instructor model as previously, but here an expanded, three-variable learner model was used: (i) coarse-grained learner-level urgency, (ii) fine-grained post-level urgency, and (iii) learner posts.

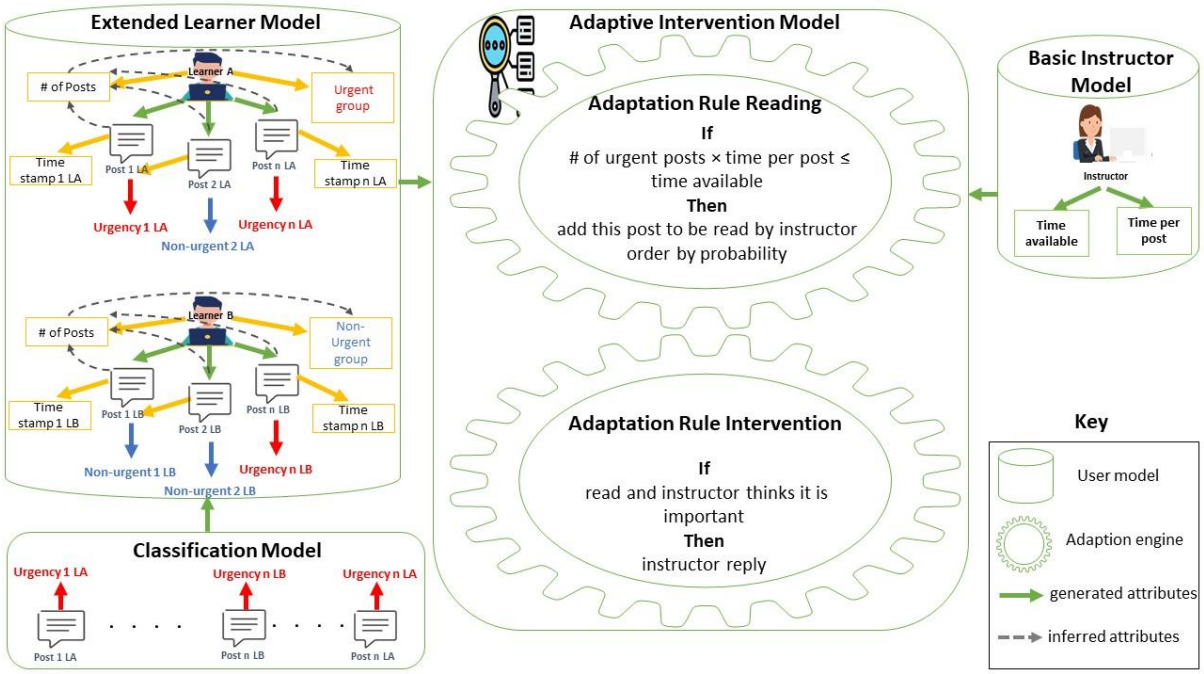

Figure 8.9: The adaptive intervention model based on coarse-grained, expanded learner modelling with two learner groups based on number of posts (low/high); here, the instructor model is the same as in Figure 8.7 but the learner model has been expanded with an additional variable (*coarse-grained learner-level urgency)*.

8.4. Results and Discussion

This section presents and offers an interpretation of the overall results considering the experiments for imbalanced data to address ***RQ5.1***. Additionally, adaptive intervention models were applied to two scenarios to address ***RQ5.2***.

8.4.1. Experiments for Imbalanced Data

This section provides a discussion of the experimental results for the two main types of classifiers (traditional ML and Transformer).

8.4.1.1. Traditional Machine Learning on the UNITE Dataset

In traditional ML, five different classifiers were tested with different types of feature engineering in three different augmentation approaches (Approach #1: word-level, with the same type (BERT), Approach #2: word-level with different types (word2vec, BERT and RoBERTa), and Approach #3: different levels (character, word, sentence) with different types (OCR, BERT and XLNet) as discussed in Section 8.3.2.2.2 Text Augmentation. Table 8.7 (below) shows the results of the comparison between the accuracy (*ACC*) of the basic classifier

(naive Bayes) with the count vector as features. Despite some of these models obtaining around 90% accuracy (see Table 8.7 below), this does not mean that they are good models; they could be biased towards the majority class on the imbalanced class dataset. Thus, to achieve more accurate results as explained before, the other metrics were used to measure performance, such as *P*, *R* and *F1* for each class.

The data in the testing set is a highly imbalanced with skewed class proportions to non-urgent (0) and a very small proration of urgent (1) (only 7%). Therefore, this research project aimed to correctly classify urgent cases represented by recall (*R*). It proposes to use recall as the main evaluation metric for urgent posts, as recall (the correct identification of most of the urgent cases, preferably all, allowing for false positives) ensures that all urgent cases have precedence, which is more important than precision (the correct identification of only urgent cases, but possibly missing some, thus allowing for false negatives). Specifically, it tries to improve the outcome of *R* for the positive class. In addition, it separately shows how filtering can be added as a process to retrieve the most urgent posts that obtained priority from their probability in the classification models. This approach potentially reduces the instructor effort required to review and read many posts (see Section 8.4.2.1).

Table 8.7 (below) shows the count-vector feature as a case study. The evaluation of *R* for class 1 (urgent), based on the original data was very low (0.05). Improved performance was achieved by applying different approaches to enhance the data and address the imbalance problem. The best result was obtained using undersampling (Under) (0.82), but the results dramatically decreased for class 0 (non-urgent) to 0.49 (from 0.99). In contrast, the performance of the manipulated data with 3x augmentation + undersampling achieved the best performance, achieving a balance between class 1 and class 0. Most of the three approaches for augmentation run on the same scenario, albeit with some exceptions which will be discussed later in this section.

The aim was to find the best techniques to deal with the imbalanced data problem between the three different approaches for augmentation (not to find the best feature engineering approach). The reason for using different features was to confirm which imbalanced data technique is better across all feature sets and to make the experiments more generalisable. Therefore, the findings can be generalised to (i) all approaches on specific features, (ii) all features on a specific classifier, and (iii) all classifiers, since the effectiveness of the proposed methods of data manipulation were similar for most classifiers (as shown in Appendix B). For

conciseness, it was decided to report and discuss only one of these classifiers (naive Bayes) with one feature (count vector); the results of the other types of classifiers are provided in the Appendix B. However, the exceptions are discussed in the next paragraph.

Table 8.7: The performance results of the naive Bayes model with count-vector feature engineering with original data, with three approaches to augmentation (see Table 8.3 above) using 3x and 9x (see Table 8.2 above) with and without undersampling and with undersampling without augmentation. Underlined: best performance of R for class 1 (urgent), **Bold:** best performance of R, balancing between class 1 (urgent) and class 0 (non-urgent) in the UNITE dataset.

Feature Engineering	Augmentation		Undersampling	Acc	Non-urgent 0			Urgent 1		
					P	R	F1	P	R	F1
Count vector	×		×	0.92	0.93	0.99	0.96	0.29	0.05	0.08
	Approach #1	3X	×	0.90	0.94	0.95	0.94	0.26	0.24	0.25
		9X	×	0.84	0.95	0.88	0.91	0.21	0.44	0.29
		3X	√	0.75	0.96	**0.76**	0.85	0.16	**0.57**	0.25
		9X	√	0.81	0.96	0.84	0.89	0.19	0.50	0.28
	Approach #2	3X	×	0.91	0.94	0.97	0.95	0.29	0.18	0.22
		9X	×	0.90	0.94	0.95	0.95	0.25	0.21	0.23
		3X	√	0.79	0.96	**0.81**	0.88	0.17	**0.51**	0.26
		9X	√	0.88	0.94	0.93	0.94	0.24	0.28	0.26
	Approach #3	3X	×	0.90	0.94	0.96	0.95	0.28	0.21	0.24
		9X	×	0.87	0.95	0.92	0.93	0.23	0.31	0.26
		3X	√	0.78	0.96	**0.80**	0.87	0.17	**0.55**	0.26
		9X	√	0.85	0.95	0.89	0.92	0.20	0.36	0.25
	×		√	0.52	0.97	0.49	0.65	0.11	0.82	0.19

Whilst most of the findings were the same, there were a few exception cases; for example, (i) the strongest predictors for recall were mostly those with undersampling (Under). However, some models (random forest and boosting (XGBoost)) with TF-IDF vectors (n-gram word-level) were better in other approaches for text augmentation than undersampling (see Table 8.8 below).

(ii) the best performance was often obtained from the data with 3x augmentation + undersampling, achieving a balance between class 1 and class 0 levels, but some models 9x augmentation + undersampling outperformed 3x augmentation + undersampling, as shown in Table 8.9 (below).

Table 8.8: Cases in which the results performance of R for class 1 (urgent) of the text augmentation techniques were higher than the results performance of R for class 1 (urgent) for the undersampling technique.

Classifier	Feature Engineering	Augmentation		Under	Acc	Non-urgent 0			Urgent 1		
						P	R	F1	P	R	F1
Random Forest	TF-IDF vectors (n gram word-level)	Approach #1	9X	×	0.91	0.95	0.95	0.95	0.36	0.39	0.38
			3X	√	0.90	0.95	0.94	0.95	0.33	0.40	0.36
			9X	√	0.89	0.95	0.93	0.94	0.30	0.41	0.34
		Approach #2	3X	√	0.89	0.95	0.93	0.94	0.30	0.39	0.34
			9X	√	0.88	0.95	0.93	0.94	0.27	0.36	0.31
		Approach #3	9X	×	0.91	0.95	0.95	0.95	0.38	0.41	0.39
			3X	√	0.90	0.95	0.94	0.95	0.34	0.41	0.37
			9X	√	0.90	0.95	0.93	0.94	0.33	0.42	0.37
		×		√	0.86	0.95	0.90	0.92	0.21	0.36	0.26
Boosting (XGBoost)		Approach #1	9X	√	0.70	0.95	0.71	0.81	0.12	0.53	0.20
		Approach #2	9X	√	0.66	0.95	0.67	0.79	0.11	0.55	0.19
		×		√	0.74	0.95	0.76	0.84	0.14	0.49	0.21

(iii) in terms of approaches, the goal of building more than one approach was to generalise the results of the technique used in data manipulation. Thus, the results of the different approaches revealed that no single approach can be considered to be the best approach. However, interestingly, Approach 3, which is based on different levels (character, word, sentence) always obtained the best results for *R* if TF-IDF vectors (n-gram character level) as a feature were used, across all experiments (as shown in Appendix B).

Table 8.9: Cases in which the results performance of R for class 1 (urgent) of the 9x augmentation + undersampling were higher than the results performance of R for class 1 (urgent) for 3x augmentation + undersampling.

Classifier	Feature Engineering	Augmentation		Under	Acc	Non-urgent 0			Urgent 1		
						P	R	F1	P	R	F1
SVM	TF-IDF vectors (n gram word-level)	Approach #1	3X	√	0.91	0.95	0.95	0.95	0.34	0.33	0.34
			9X	√	0.89	0.95	0.93	0.94	0.27	0.35	0.31
		Approach #2	3X	√	0.90	0.95	0.95	0.95	0.31	0.30	0.30
			9X	√	0.88	0.95	0.93	0.94	0.25	0.32	0.28
		Approach #3	3X	√	0.92	0.94	0.97	0.96	0.38	0.20	0.26
			9X	√	0.91	0.94	0.97	0.95	0.35	0.24	0.28
Random Forest		Approach #1	3X	√	0.90	0.95	0.94	0.95	0.33	0.40	0.36
			9X	√	0.89	0.95	0.93	0.94	0.30	0.41	0.34
		Approach #3	3X	√	0.90	0.95	0.94	0.95	0.34	0.41	0.37
			9X	√	0.90	0.95	0.93	0.94	0.33	0.42	0.37
Boosting (XGBoost)		Approach #1	3X	√	0.85	0.95	0.88	0.92	0.20	0.38	0.26
			9X	√	0.70	0.95	0.71	0.81	0.12	0.53	0.20
		Approach #2	3X	√	0.77	0.95	0.80	0.87	0.14	0.42	0.21
			9X	√	0.66	0.95	0.67	0.79	0.11	0.55	0.19
		Approach #3	3X	√	0.85	0.95	0.89	0.92	0.20	0.38	0.27
			9X	√	0.87	0.95	0.91	0.93	0.25	0.40	0.31

8.4.1.2. BERT on the UNITE Dataset

When using BERT, Table 8.10 (below) shows the prediction performance for the different methods of manipulating the data. As mentioned, only augmentation was performed; no feature engineering was necessary. The performance of *R* for class 1 in BERT with the original data was not too low in comparison with the traditional ML results. Although it rose from (0.52) to 0.82 with the undersampling technique. However, for the negative class, recall decreased from

0.98 to 0.86. To achieve more balance between the two classes, 3x augmentation + undersampling was used (see Table 8.10 below).

Table 8.10: The performance results of the BERT model with original data, with three approaches to augmentation (see Table 8.3 above) using 3x and 9x (see Table 8.2 above) with and without undersampling and with undersampling without augmentation. Underlined: best performance of R for class 1 (urgent), **Bold:** best performance, balancing between class 1 (urgent) and class 0 (non-urgent) in the UNITE dataset.

Augmentation		Under	Acc	Non-urgent 0			Urgent 1		
				P	R	F1	P	R	F1
×		×	0.95	0.96	0.98	0.97	0.67	0.52	0.58
Approach #1	3X	×	0.95	0.97	0.97	0.97	0.62	0.63	0.63
	9X	×	0.94	0.97	0.96	0.97	0.54	0.59	0.57
	3X	√	0.92	0.98	**0.93**	0.96	0.46	**0.75**	0.57
	9X	√	0.93	0.97	0.95	0.96	0.50	0.63	0.56
Approach #2	3X	×	0.95	0.97	0.97	0.97	0.62	0.62	0.62
	9X	×	0.94	0.97	0.97	0.97	0.57	0.57	0.57
	3X	√	0.91	0.98	**0.92**	0.95	0.41	**0.77**	0.54
	9X	√	0.94	0.97	0.96	0.97	0.55	0.62	0.58
Approach #3	3X	×	0.94	0.97	0.97	0.97	0.60	0.59	0.59
	9X	×	0.94	0.97	0.97	0.97	0.61	0.59	0.60
	3X	√	0.89	0.98	**0.89**	0.94	0.36	**0.79**	0.50
	9X	√	0.94	0.97	0.97	0.97	0.57	0.58	0.58
×		√	0.86	0.98	0.86	0.92	0.32	0.82	0.46

Hence, the best classifier performance on the UNITE dataset was obtained with BERT using Approach 3 with 3x augmentation + undersampling.

To verify the effectiveness of the different data manipulation techniques to deal with the imbalanced data problem, the same methods were utilised on the Stanford dataset. In these experiments, augmentation was limited to 3x only, since 9x would have generated more instances in the minority class than in the majority class. Also, only Approach 3 was applied (see Table 8.3 above), which provided the best performance for the 3x augmentation + undersampling technique on the UNITE dataset.

8.4.1.3. BERT on the Stanford Dataset

Table 8.11 (below) shows the results of BERT on the Stanford dataset. Similar results were obtained for the UNITE dataset; the only difference being in the performance of the two techniques with 3x augmentation with and without undersampling. This is possibly because the distribution of non-urgent cases differs between the two datasets (see Figures 8.5 and 8.6

above). Whereas, as clarified in Figure 8.6 (above), the distribution of non-urgent cases for 3x achieved almost the same as the distribution of non-urgent cases for 3x + undersampling.

Table 8.11: The performance results of the BERT model with original data, with three approaches to augmentation (see Table 8.3 above) using 3x (see Table 8.2 above) with and without undersampling and with undersampling without augmentation. Underlined: best performance of R for class 1 (urgent), **Bold:** best performance of R, balancing between class 1 (urgent) and class 0 (non-urgent) for the Stanford dataset.

Augmentation		Under	Acc	Non-urgent 0			Urgent 1		
				P	R	F1	P	R	F1
×		×	0.91	0.94	0.96	0.95	0.80	0.73	0.76
Approach #3	3X	×	0.91	0.95	**0.94**	0.94	0.75	**0.78**	0.77
	3X	√	0.91	0.95	**0.94**	0.95	0.76	**0.78**	0.77
×		√	0.89	0.97	0.89	0.93	0.65	0.89	0.75

8.4.2. Adaptive Intervention Models

This section provides a discussion of the experimental results for the results related to the example adaptation intervention models.

8.4.2.1. Basic Adaptation Scenario

In this scenario, depending on urgent posts ranking (probability score given by the classification model), the aim was for the adaptive intervention model to automatically retrieve the most important urgent posts and reduce the number of posts that are read by an instructor. In this case, the naive Bayes with count vector was used with Approach #1 for 3X augmentation with the undersampling model (the best performance among different approaches in naive Bayes with count vector) as a case study. For example, if the time available for an instructor to read posts is limited to five posts, then the model will retrieve only five posts. Table 8.12 (below) presents the results of the comparison between the basic model (all posts) and the adaptive model that selects only the (five) most urgent posts for the urgent class (1), which clearly outperformed the basic model on all evaluation criteria.

Table 8.12: The performance results of the naive Bayes model with count vector as feature engineering with Approach #1 to augmentation (see Table 8.3 above) using 3x with undersampling. First row: basic model with all data; second row: filtering model with the top five most urgent posts for class 1 (urgent) in the UNITE dataset.

	1		
# Posts	P	R	F1
All	0.16	0.59	0.25
5	0.40	1.00	0.57

8.4.2.2. Expanded Adaptation Scenario

In the second scenario, an adaptation filtering model was proposed based on the number of learner posts. Pearson's correlation was used to calculate the correlation between the number of posts written per learner and the number of posts from those who require immediate attention. This process resulted in a strong correlation (0.65).

The results of the Fisher-Jenks algorithm to cluster learners are shown in Table 8.13 (below). To obtain the two groups (for urgent/ non-urgent learners), clusters 1 and 2 were merged to reflect learners with a high number of posts as these are significantly more communicative than learners in cluster 0.

Table 8.13: Clustering learners based on their number of posts.

Cluster	Count	Mean	Std	Minimum	Maximum
0	734	3.30	3.06	1	15
1	57	27.26	12.50	16	62
2	6	107.16	34.31	84	173

Posts from the low-number-of-posts group (non-urgent learners) were removed from each fold (using stratified four-fold cross-validation). The number of posts in the testing set is shown in Table 8.14 (below) for both: (i) the basic model, which contains all learners; and (ii) the filtering model, which only contains learners with a high number of posts (urgent learners). Hence, the number of posts in the filtering model was much lower than for the basic model. For example, in fold 1, the number of posts dropped from 1156 to 533 basic to filtering, reducing the number of posts the instructor needs to read. Thus, whilst the overall recall was somewhat reduced (by 11%), the load on the instructor was also significantly reduced ($p<<0.5$).

Table 8.14: Number of posts in the testing set. First row: basic models; second row: filtering models on the UNITE dataset.

Fold	Model	Number of posts in testing set
1	Basic	1156
	Filtering	533
2	Basic	1156
	Filtering	561
3	Basic	1155
	Filtering	551
4	Basic	1155
	Filtering	552

8.4.3. Discussion

Finding urgent messages is vital for instructors involved in running MOOC courses. However, this is a daunting prospect for MOOC instructors due to the sheer volume of posts that need to be read. Thus, classification models for automatically analysing posts and predicting their urgency are very much needed; some accurate models have been proposed in the past (Khodeir, 2021; Alrajhi and Cristea, 2023; Guo *et al.*, 2019). Whilst some of these models have obtained cutting-edge performance, the performance dips somewhat for the underrepresented class. Please note that it is to be expected that any such dataset would be unbalanced, with the non-urgent posts being the predominant class, in practically any online learning system. However, simply increasing performance may not be enough, and indeed, perhaps impossible for highly unbalanced data. Dealing with class imbalances is the best way to improve data and therefore improve performance by proposing different methods to tackle the issues of class imbalance and resampling data.

Here the model first considered fine-grained learner modelling that deems each post as a feature of a learner, which, if urgent, needs to be dealt with on its own. Next, as recent research has shown correlations between urgency and the number of posts made by learners, this indicates that learners posting urgent posts are likely to make many posts which enables such learners to be classified at the macro-scale as an 'urgent learner' (Alrajhi *et al.*, 2021). Such learners would need to be treated as a priority by instructors.

Modelling learners based on posts only is a simplification of the learner model; any model is a simplification of real-world conditions. However, the author believes that the posts of learners can provide insight into specific learner characteristics and needs. For instance, the language of the post can reveal anxiety or a certain level of background knowledge or impatience, thus covering various learner-model variables.

Learner models can contain several parameters and be simpler or richer. Indeed, learner models can reflect various aspects of a learner; they often include various parameters such as current level of confusion, motivation, understanding, etc. Interventions to reduce learner dropout from MOOCs could include changing the difficulty or type of problems, referring the learner to modules for missing prerequisite knowledge, peer referrals, encouraging communications, etc. In this research project, the author adds to this rich tapestry of user model dimensions by extracting urgency based directly on user posts, an approach that has been overlooked in previous user modelling. Importantly, the richness of data obtained by using post-based user modelling was considered, in the sense that posts may reflect various aspects

of a learner such as boredom, interest, knowledge, fluency, etc. This learner model can be used by itself or in conjunction with other user parameters (if known) to further enrich the user model. This, however, does not detract from the merit of the parameters introduced with this approach.

Finally, whilst the current results are very specific to MOOC posts analysis, the proposed techniques may serve as a template for other similar NLP classification tasks using ML with severely skewed datasets.

8.5. Error Analysis

To understand the reasons for the errors obtained in the testing set in every fold, an in-depth re-analysis of the model was conducted. To achieve this, the examples of mistakes that the best algorithm (BERT - Text Augmentation + Undersampling) made on UNITE data were manually inspected. Specifically, false negatives (FN), which the model categorised as non-urgent (although they are labelled as urgent), were considered to be the most critical errors, as the aim was to capture *all* urgent cases. To put the results and especially the errors in context, the miss-predictions of the classifier with human-level performance for the different folds (using stratified k-fold cross-validation, choosing a value of k = 4 (four folds)) were compared as explained in the methodology under Section 8.3.2.2.1. The results are shown in Table 8.15 (below).

Table 8.15: FN results for the best algorithm versus disagreement between human annotators.

Fold	FN (Total)	Human disagreement
1	23	19
2	14	11
3	19	16
4	14	13

The results revealed that most of the FN cases were also mirrored in the disagreement between annotators (i.e., for 19/23 false negatives misclassified by the classifier, the human annotators also disagreed for fold 1, etc., see Table 8.15 above). This further supports the notion that decision-making among annotators is difficult, as well as that the more difficult cases are both hard for humans and classifiers to categorise; examples of each fold are shown in Table 8.16 (below).

Table 8.16: Anonymised examples of FN results and disagreement between human annotators on UNITE data.

Fold	Example
1	I have some difficulties to understand diagrams. But it seems very important to give a meaning and a context to words used in analysis.
2	I had done this, the [programming-platform] is going on. But I need also the [other-platform]. I installed a old Version [other-platform], also the Newest. I couldn't found the [other-platform] for [setup].
3	Further to my comment on the previous "step" I am yet to be convinced!
4	I don't understand the reason of this message when I type [code-removed] Warning message:[error-message-removed]

Table 8.16 provides a better understanding of why humans and ML struggle in certain cases. For example, in fold 1, the learner does not understand the diagram, but s/he is happy about providing a meaning and context for the words used in the analysis. Some annotators believe that this post is non-urgent because the learner did not request assistance. However, another annotator may find that the learner has difficulty understanding the concept and so the post is urgent and requires an intervention. Such clashes may explain why the model was unable to detect the above-mentioned urgent cases.

8.6. Epilogue

On MOOC platforms, deciding the right moment for instructor intervention is an important challenge to be overcome to better support learners. Building an automated model to detect posts that require urgent intervention represents a promising solution to this problem. However, the available MOOC post datasets naturally contain only a few urgent cases, leading to imbalanced data, which explains the difficulty in creating models to detect such cases accurately. In this chapter, three techniques (text augmentation, text augmentation + undersampling, and undersampling) were analysed and compared to improve the quality of such data. Also, several new pipelines incorporating different text augmenters were provided. The results show that an increase in model performance can be obtained via undersampling, and a combination of text augmentation + undersampling achieves the best performance in balancing between the two classes.

These results help automatically retrieve the most-urgent posts for instructors to consider. To show how this can be applied, two adaptive models were used for illustration based on two types of user models: (i) personalised instructor intervention based on a fine-granularity learner model, and (ii) filtering results based on a higher granularity learner model.

It further inspected wrongly classified urgent instances and found that the problem does not simply lie with the classifier: it also stems from the data, which humans also find difficult to annotate. This indicates that the difficulties faced by human annotators in classifying such posts are also faced by these models.

The next chapter shows how (XAI) techniques can be applied to interpret a MOOC intervention model for urgent-post detection by analysing learner posts to help instructors determine when posts need immediate attention and to aid annotators in producing high-quality datasets.

CHAPTER 9: AN EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) APPROACH FOR URGENT INSTRUCTOR-INTERVENTION MODELS

9.1. Prologue

Deciding when instructor intervention is needed based on learners' posts and their urgency in the context of MOOC environments is a known challenge (Almatrafi, Johri and Rangwala, 2018). To solve this challenge as clarified in SLR in Chapter 2, prior approaches used automatic ML models to predict urgency with more accurate performance achieved when DL methods are applied. These models are characterised as 'black-box' approaches as the results are opaque to humans (Von Eschenbach, 2021) and it is not easy to explain such models' prediction results, especially those of DNNs (Lipton, 2018). Therefore, XAI is used in general to understand such results to enhance trust in AI-based decision-making. Although instructor intervention models need to be accurate in their decisions, it is difficult to achieve this as urgency decisions are hard to make, even for humans (Chandrasekaran *et al.*, 2015b). This difficulty concurred with the author's experience with the data labelling process for this task. Also, the large number of observations that need to be performed may increase the cognitive overload on annotators (Dong *et al.*, 2020) (including physical consequences such as eye strain/blurry vision), which causes them to struggle to make an appropriate decision. As the advancement of AI is allowing humans and machines to collaborate to solve complex issues, XAI can be increasingly used to support MOOC instructors and annotators.

This chapter deals with the intervention problem by showing how XAI techniques can be applied to interpret a MOOC intervention model for urgent-posts detection by analysing learner posts as posts were selected from a MOOC course and annotated using human experts. The

initial goal is establishing proof of concept using explainable AI for the task of urgent intervention as this had not been done before in this research area. In NLP, correct labelling and annotating text data correctly are critical issues which play an important role in supervised ML model prediction. Therefore, considering that making urgency decisions has recently been confirmed to be hard for humans (Chandrasekaran *et al.*, 2015b), the current research project evaluated and studied the confidence levels between human annotators' decisions (annotator agreement confidence) and compared these with the ML model's decisions on every instance of learner posts. To understand how and why the model decisions are made, intervention model prediction was explained and compared with human decision-making using Captum (Kokhlikyan *et al.*, 2020) as an interpretation tool, which is a state-of-the-art approach for interpreting Transformer models (Bennetot *et al.*, 2021).

This chapter showed how pairing a good predictor such as BERT (as a widely used language model in the field of NLP) with XAI results and especially colour-coded visualisation can be used to support instructors make decisions on urgent intervention. Also, it showed that XAI can be used not only to support instructors making decisions on urgent intervention but also further used to support annotators in creating high-quality, gold-standard datasets for urgent intervention. Thus, the RQs were formalised as follows:

- ***RQ6.1:*** *How can a transparent XAI model be constructed to further support instructors' decisions to intervene based on an urgent-posts-intervention-need-detection model?*
- ***RQ6.2:*** *How can XAI be employed to improve human annotators' decisions about the urgency of posts (i.e., deciding on which posts need intervention)?*

The following are the most important contributions of this chapter; to the best of the author's knowledge, this is the first time that: (i) text classification explainability has been applied to an instructor intervention model, (ii) the AI prediction error has been shown to be connected to human (lack of) confidence (i.e., appearing for the same instances, here, posts), and (iii) how explainable models can be used for annotator support to for creating high-standard corpora.

9.2. Related Work on Explainable Artificial Intelligence

A review of the literature on urgent instructor intervention in MOOCs shows that the area has recently gained great momentum. A variety of text classification models have been proposed to classify urgent posts ranging from traditional ML (Almatrafi, Johri and Rangwala, 2018) to other DL (Guo *et al.*, 2019) and Transformers as embedding (Khodeir, 2021) with different

levels of inputs (Alrajhi and Cristea, 2023). That said, none of these works sought to interpret and explain the model's decision-making. In general, even though some traditional ML approaches such as classification trees and naive Bayes algorithms are simple to understand and interpret, they are less accurate than other approaches (Kowsari *et al.*, 2017), especially for large data sets. Therefore, in intervention task complex models which achieve better performance have been proposed (Guo *et al.*, 2019; Sun *et al.*, 2019; Khodeir, 2021). These models are considered to offer 'black-box' approaches and are consequently difficult for end-users to understand, as depicted in Figure 9.1 (below) (Kumar, Dikshit and Albuquerque, 2021), although some large models are interpretable (Rudin, 2019).

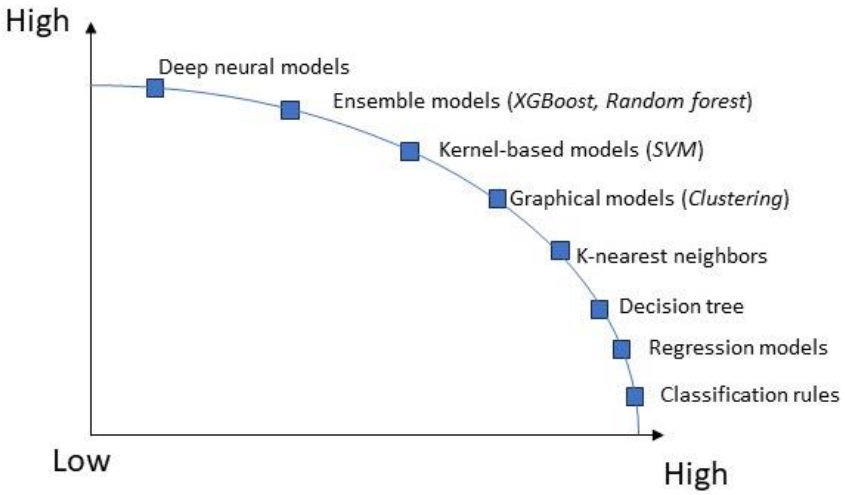

Figure 9.1: Predictive ability vs interpretability trade-off (Kumar, Dikshit and Albuquerque, 2021).

Recently, outside of our specific area, a new research direction has become very active: aiming to explain and interpret 'black-box' predictions and ML models in general for different sectors. Model interpretability is a field of XAI that attempts to explain model internals and results in human-understandable terms (Gilpin *et al.*, 2018; Adadi and Berrada, 2018). A wide range of powerful tools have been proposed; for example, Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh and Guestrin, 2016), the SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), InterpretML (Nori *et al.*, 2019), and Captum (Kokhlikyan *et al.*, 2020). Explainability in AI is essential to enable developers to understand and improve their models and for end-users to increase model-decision trust (Confalonieri *et al.*, 2021). Please note that while *interpretability* and *explainability* are often used interchangeably, some papers distinguish between them (Došilović, Brčić and Hlupić, 2018). Here, however, for simplicity, the current work does not make that distinction.

The BERT model has been extremely popular within the NLP domain, being applied in high-performance text classification models e.g., (Fonseca *et al.*, 2020; Pereira *et al.*, 2021). The architecture of BERT consists of a deep layer which can lead to the results being difficult to interpret. Importantly, in relation to the current research project, recent studies have proposed techniques for using XAI combined with BERT, which, as mentioned recently, is applied in text classification models to achieve high performance. (Kokalj *et al.*, 2021) proposed Transformer-SHAP (TransSHAP) by adapting and extending SHAP (Lundberg and Lee, 2017) to operate on BERT by building custom functions and visualising the results sequentially. They demonstrated that the visualisation approach used on TransSHAP was simpler than that of other tools (LIME and SHAP). However, this approach is considered limited in terms of only supporting random word sampling, which may result in unintelligible and grammatically incorrect sentences and wholly uninformative texts. Another study by (Szczepański *et al.*, 2021) proposed a new approach for explainable BERT-based fake news detectors using two XAI techniques (LIME and Anchors) on the Kaggle dataset. Their findings support the use of multiple methods to construct explanations. However, there is a problem with Anchor as it is not always able to find an explanation.

In contrast, Captum is an open-source multi-modal (image, text, audio, video) library for Transformer model interpretability (Bennetot *et al.*, 2021; Kokhlikyan *et al.*, 2020). Captum is an open-source library developed by Facebook AI and offers cutting-edge techniques such as *integrated gradients* that make it simple for researchers and developers to identify which features contribute to a model's decision and output (Captum, 2021). This package has been drawing great attention from researchers some of whom have used this package in their applications. For instance, (Levy *et al.*, 2022) utilised Captum to interpret a BERT model that was used as one of a set of different ML models to predict primary *current procedural terminology* (CPT) codes from pathology reports.

Hence, this research project built an explainable instructor intervention classifier model as a text classification task by deploying the Captum package as it is one of the most commonly used tools for use with Transformer models (Bennetot *et al.*, 2021).

The work that lies closest to the current project is that of (Hu, Mello and Gašević, 2021) which used XAI to analyse online discussions. However, no methods of XAI have been applied yet to urgent intervention to support instructors in MOOC environments. Moreover, none of

these works clearly explain how to connect AI prediction error to human (lack of) confidence or use explainable models for annotator support in creating high-standard corpora.

9.3. Methodology

This section summarises the methods used to generate the Gold-standard corpus with consideration of the measure of confidence between annotators, together with the tool and technique used to explain the ML model. Thus, this research project consists of four basic stages (see Figure 9.2 below) as follows:

- First, construct an 'urgent' gold-standard dataset via the use of human experts annotating posts and computing their label confidence levels (Section 9.3.1).
- Second, use BERT to build an automatic urgent intervention model (Section 9.3.2).
- Third, automatically explain the model as a local explanation and calculate the probability and word attribution of urgency (Section 9.3.3).
- Fourth, visualise the importance of words in post, compare the two approaches (machine and confidence), and discuss the results (Section 9.3.4 and Section 9.4).

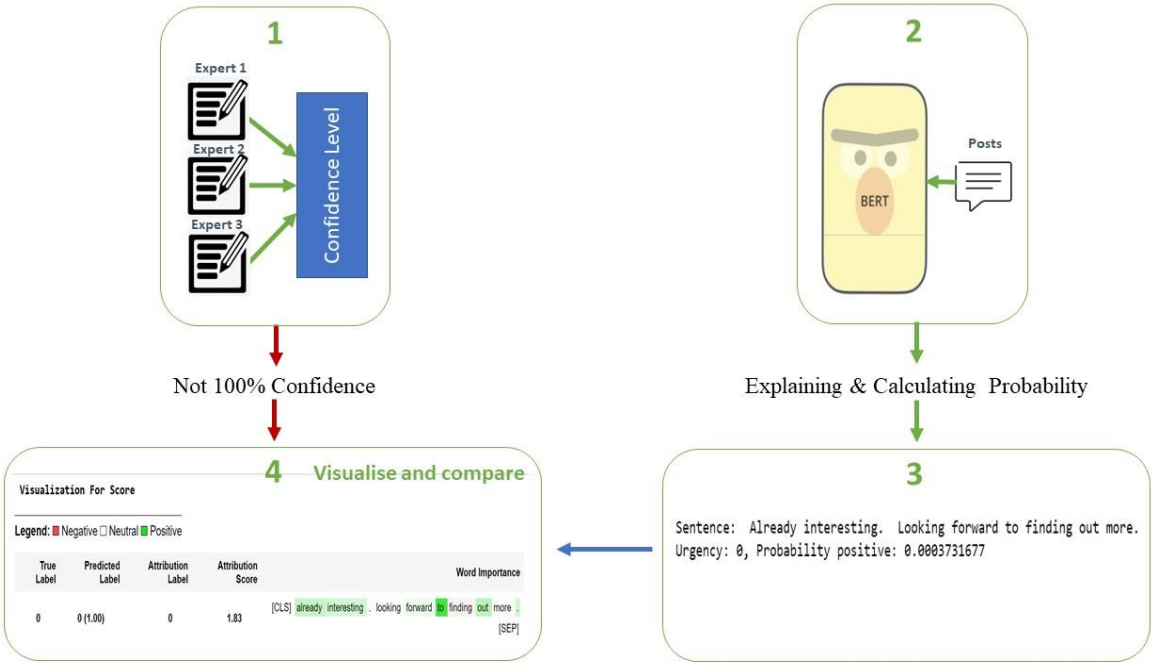

Figure 9.2: Human annotator vs machine pipeline: basic stages.

9.3.1. Adding Confidence Level to the Gold-Standard Dataset

In this research project, the Gold-standard corpus (as explained in Section 3.2.2.2) was used as a case study. To add the annotator agreement confidence level, the three annotators' decisions were considered after converting them to a binary value. Therefore, the annotator agreement confidence level was calculated thus (Figure 9.2, step 1):

- If the three annotators agreed → 100% agreement confidence.
- Otherwise → < 100% agreement confidence (i.e., ∼67% = 2/3 agreement).

9.3.2. Fine-tuning the BERT Model

As a preprocessing step, the data were split into training and testing sets using the stratify method (Farias, Ludermir and Bastos-Filho, 2020) to preserve the percentage of samples for each class (80% training and 20% testing). Thus, the distribution of the training set was 0: 3922, 1: 706 and that of the testing set was 0: 981, 1: 177. Then, the training set was then split again: 90% of the data were used for training and 10% for validation.

BERT was fine-tuned, as mentioned before, without any engineering features. The 'bert-base-uncased' version was used, which means there is no distinction between capital letters and lowercase letters. Next, the text input was prepared, and the model was trained by defining the batch size = 8, number of training epochs = 4. Finally, the prediction model performance on the testing set was evaluated, and the pre-trained model was saved for later use in interpretation (Figure 9.2, step 2).

9.3.3. Interpreting the BERT Model

After training the model, the prediction of the BERT model was automatically interpreted and explained by using the Captum package which supports classification models. The predictions were interpreted via the BertForSequenceClassification in Captum from Captum_BERT colab (Captum, 2022). This was achieved by creating the *layer-integrated gradients explainer* and *attribute* methods to generate feature importance and identify which words (tokens) have the highest attribution to the model's output. Based on the gradient of the model's output (prediction) with respect to the input, integrated gradients (Sundararajan, Taly and Yan, 2017) provide a way to calculate the attribution score of each input feature of a deep learning model (here, BERT). This attribution score can be used to determine which words are important to the

outcome that the model predicts. The final attribution score is calculated by the average value for each word (Figure 9.2, step 3).

In this experiment, three posts with < 100% confidence between annotators were inspected, which means the final label was set by majority voting (2/3, with one annotator disagreeing). These posts were selected to showcase three scenarios reflecting the differences in agreement between the human annotators: 1: large difference; 2: slight difference; 3: in-between. This will be further clarified in Section 9.4.

9.3.4. Visualising and Comparing

The final step was to visualise the explainability results with the attribution score and highlight the word importance as input for use by instructors and annotators and potential future decision-making support methods. The visualisation was performed by using VisualizationDataRecord method. Green highlights are used to indicate the tokens that contribute positively to the model's prediction (i.e., features that have a positive impact on pushing the prediction towards a specific class); red highlights indicate tokens that have a negative impact on the model's prediction (i.e., those which push the prediction towards a different class) (see Figure 9.2, step 4).

9.4.Results and Discussion

This section presents the results obtained from BERT to predict urgent posts and how learner-posts decisions can be evaluated and explained using Captum to further support instructors' decision-making and address ***RQ6.1*.** Then, the results on the agreement (confidence) between the annotators were illustrated using three scenarios based on agreement with the end goal to improve the human annotators' decisions and address ***RQ 6.2***.

The results obtained from BERT to predict urgent posts show that the average accuracy score was high (0.92). However, as explained before, as the data is extremely unbalanced, additional metrics were used to evaluate the classifier (precision, recall and F1-score) for every class to achieve a comprehensive understanding of the outcomes (see Table 9.1 below). Please note that here, whilst working with a decent classifier, the focus is not on the optimisation of the classifier but on the explanation of the obtained results. Thus, BERT has been selected here as it is one of the state-of-the-art classifiers; however, the method of explainable decisions

about urgent posts for instructors and comparing machine prediction to human classification is generalisable, and so can be used with other DL models.

Table 9.1: The results of the BERT classifier.

	Precision	Recall	F1-score
0	0.95	0.95	0.95
1	0.73	0.71	0.72

These measurements are based on the confusion matrix, as explained in Section 3.4, which is depicted as a table with four different combinations of predicted and true values: TN, FP, FN, and TP, as the results from the BERT classifier reported (see Figure 9.3 below).

Figure 9.3: Confusion matrix of the BERT classifier.

As previously mentioned, one of the goals was to analyse learner posts and explain the text classification decisions using Captum to understand the reasons behind the predictions and help instructors with their decision-making process in relation to offering intervention. Here a random post prediction from the test set was chosen; then, the explainability results were illustrated with highlighted text (see Figure 9.4 below). The attribution score = 1.45 and the different colours reflect the effect of word attribution towards the prediction; the level of highlighting depicts the importance of the feature for the classification. Specifically, the green highlight depicts a positive contribution (got, looking, understanding, be, …), whilst the red highlight contributes by decreasing the prediction score (forward, useful, …). In the case of the example below, it was found that the predicted label is non-urgent (0) and the true label is also non-urgent (0) with a confidence level of 100% between annotators. Such visualisation can further be used by an instructor to understand the decisions and recommendations of a classifier for urgency detection in learners' MOOC forum posts.

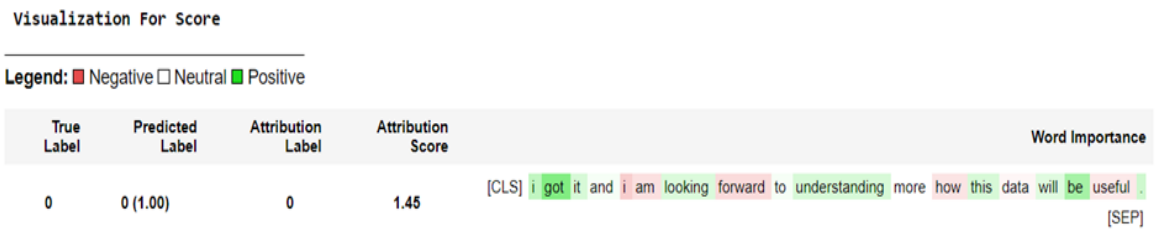

Figure 9.4: Screenshots of Captum explanations.

Then, the agreement between the three annotators was calculated, as previously explained; it was found that the number of posts that have 100% confidence between annotators was 4190; in contrast, the number of posts < 100% confidence was 1596 from the total data. From the testing data, the total number of posts among annotators with 100% confidence was 833. However, there were 325 posts with < 100% confidence.

The second goal of this research project was to help annotators with data labelling decision-making. Thus, the relationship between machine-produced (BERT model) results and the confidence agreement level between human annotators was also analysed. Using the confusion matrix, different cases can be studied (see Table 9.2. below).

Table 9.2. Machine prediction correctness (from the BERT confusion matrix), vs human annotator classification correctness, with (binary) confidence between (human) annotators and number of posts for each case, **Bold**/*Italics*: **cases that should**/*could* be explained to annotators.

Cases	True class (human annotators)	BERT confusion	BERT prediction	Confidence between human annotators	Number of posts
1	1	TP	1	100%	79
2	**1**		**1**	**< 100%**	**47**
3	0	FP	1	100%	17
4	**0**		**1**	**< 100%**	**30**
5	1	FN	0	100%	17
6	*1*		*0*	*< 100%*	*34*
7	0	TN	0	100%	741
8	0		0	< 100%	193

The aim here is to help annotators to find urgent cases that BERT can classify as urgent. Thus, the focus here is on true positives (TP), especially in case 2 where the confidence level between human annotators was < 100%. The reason for focusing on TP is that it was desirable for instances of TP to be found by both the algorithm and the annotators. There were 126 TP cases, as reported in Figure 9.3 (above). Interestingly, the experimental and investigation results show that for 79 out of the 126 instances, the classifier and the annotators agreed that the posts need urgent intervention with a confidence level = 100%. For the remaining 47 cases (case 2), it was found that the confidence level between annotators was < 100%. Therefore, 47 cases need explanation and visualisation to the annotator who disagreed with the other two annotators to potentially change their minds. In addition, FP, where a post is considered by the algorithm (BERT) as urgent, but not by the annotators, may be a potential issue if the label should be True but is not. The number of FPs was 47 with 17 cases with a confidence level = 100% and 30 cases with a confidence level of < 100% (case 4). That means that at least one of the annotators from 30 cases believes it is urgent, like BERT. Thus, these are the cases that should be explained and shown to annotators, especially with the highlights illustrating the reason for BERT's decision, to help the human annotators to refine their decisions. In general, however, any of the posts where annotators disagree could potentially be reinspected by the annotators to ensure that they increase their confidence levels.

Next, some of these posts were inspected from those deemed TPs to interpret the probability of predicting urgency by the classifier and to understand if and how this may be related to the disagreement between annotators.

To better understand in-depth the findings, three scenarios based on the agreement between human annotators were considered and studied as shown in Table 9.3 (below); these three cases were selected, according to the level of annotation (large difference, slight difference, or in between). Please note that in Table 9.3, the annotators' rating is shown before conversion to binary as urgent or non-urgent to understand their real decisions.

Table 9.3: Three scenarios based on TP with agreement between human annotators: 1: large difference; 2: slight difference; 3: in-between.

#	Text	First annotator	Second annotator	Third annotator
1	What's the shortcomings of the crowdsourced data? It's hard for me to understand.	1	6	7
2	I wonder if there is also a correlation between future orientation index and GDP per capita when the search terms are two years ahead and two years before (e.g. search terms "2009" and "2013" in the search year 2011).	3	4	4
3	I am seeking to build the following: 1-Multiple big data DB on VPS over the internet so they will be MySQL on CentOS. 2-Multiple DB manipulation engines that will Read or write on them from the BigData source. 3-Multi-agent simulations on a given basemap 4-The data model running on the simulation and DB manipulation engines will be an XML based model. Can anyone help me to build this environment?	2	6	4

9.4.1. Scenario 1 (Large Difference)

In this scenario, the observation was that some posts led to large differences between annotators. Therefore, the model was interpreted and the important words were highlighted, as shown in Figure 9.5 (below). The words 'hard for me' are the most important words that affect the decision. Thus, this may draw the attention of the annotator and lead them to making a correct decision.

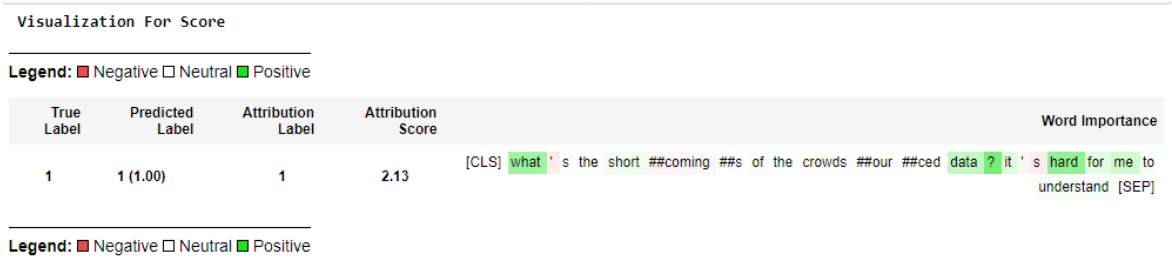

Figure 9.5: Screenshots of Captum explanations for scenario 1 (large difference and < 100% confidence between annotators).

9.4.2. Scenario 2 (Slight Difference)

In this scenario, the agreement is strong on being a threshold case (between urgent and non-urgent). When visualising (see Figure 9.6 below), it finds that the words 'I wonder if' are

important. The meaning of ‘wonder’ involves asking for help, but also thinking. Thus, this scenario can be used as a confirmatory analysis for annotators.

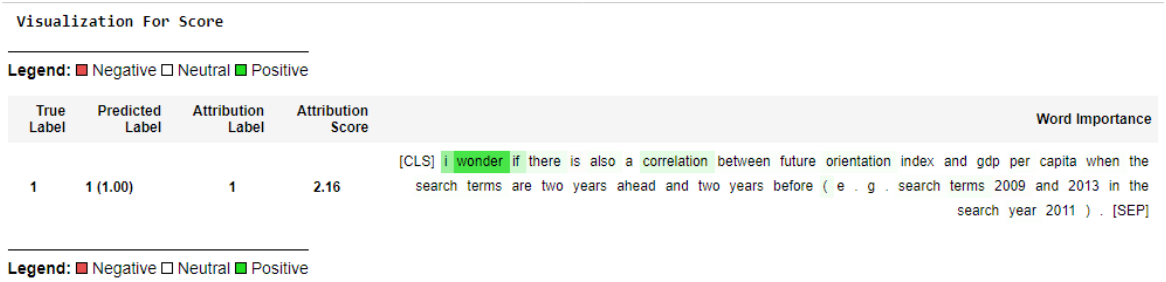

Figure 9.6: Screenshots of Captum explanations for scenario 2 (slight difference and < 100% confidence between annotators).

9.4.3. Scenario 3 (In-Between)

In this scenario, the score was incremental (2, 4, 6). To understand it, visualisation (see Figure 9.7 below) shows that the words ‘can anyone help’ and the punctuation mark ‘?’ are important words for the algorithmic decision. The annotator difference may be due to some annotators considering that by using the word ‘anyone’, the learner is asking for help from their peers, not the instructor.

Figure 9.7: Screenshots of Captum explanations for scenario 3 (in-between and < 100% confidence between annotators).

9.4.4. Discussion

There is a need for classification algorithms that can automatically analyse posts and determine how urgent they are, and some reliable models have previously been offered as shown before in the SLR. However, the story does not end here as with any black-box system, explainability is key and related to trust in the system. Moreover, correct labels are crucial. However, the type of posts appearing in a learning system are hard even for experts to reliably classify, as the experiments with annotators show. This further supports the addition of (automatic) explanations for the recommendation to better contextualise the information presented for

instructors. Indeed, highlighting the most important words would potentially simplify the intervention detection process for instructors, which is important in an educational context, such as for MOOCs, in which many posts are available for the instructors' evaluation. Similarly, this approach could facilitate the annotators' work on deciding whether a post is urgent or not. That is, this method can facilitate the work of instructors and annotators, leading to further improvements in instructor intervention and the dataset annotation process.

Here, thus, XAI was used in a novel way by turning around the approach to explain not the errors in the algorithm but the errors in human annotation (which may well lead to or explain errors in the algorithm[16]). Thus, three scenarios were analysed based on TPs. As explained, these are the cases it really wants to deal with and help the disagreeing annotator to check their decision to increase the quality of the dataset.

The results show how the colour-based highlighting functionality of XAI can provide an in-depth understanding of where the different decisions of annotators, as well as those of the algorithm, may stem from. Such explanations can be of use for instructors as well as for annotators. Thus, whilst searching for the best algorithms for urgent instructor-intervention and explainable models to support instructor decision-making, it also found serendipitous gains in the automatic explanations of annotators. Such systems could thus support annotators in facilitating/fast-tracking their work in detecting intervention points, bringing them to a common denominator, and helping them make informed decisions on a sample of pre-labelled data to then be able to confidently label new, unseen data rigorous and systematically.

9.5. Epilogue

Today, with the advent of DL, these models are showing remarkable performance in many tasks such as NLP. The problem with these models, however, is a lack of transparency and interpretability (Došilović, Brčić and Hlupić, 2018). Explainability in AI is crucial for end users to increase model-decision trust as well as for developers to understand and improve models (Confalonieri *et al.*, 2021).

The objective of this research project was to provide an explanation of the ML decision as a local explanation method for a specific text classification problem; namely, that of explaining

[16] Please note that it was not considered here that the algorithm is incapable of making errors. However, being able to compare their own decision(s) against the algorithm may give human annotators additional insight to revise (some of) their decisions.

individual predictions in the urgent intervention task in a MOOC environment to assist instructors with making appropriate interventions. Additionally, it aimed to evaluate the annotators' agreement confidence obtained from labelling an urgent instructor intervention task in a MOOC environment. In particular, the author would like to highlight the contribution of explaining individual predictions in the urgent intervention task and assessing annotators' decisions when labelling a MOOC post corpus. A BERT model was presented to classify urgent post cases. To better understand what causes labelling errors, an interesting discovery was made on the relationship between the ability of the classifier to find urgent cases and the confidence level between human annotators on making a data-labelling decision. Moreover, a new method for supporting annotators was offered. Here, the field of urgency prediction was advanced by proposing a method for potentially supporting instructor intervention as well as annotators' decision-making in data labelling tasks.

The next chapter discusses the general findings of the thesis along with its limitations and potential avenues for future work.

CHAPTER 10: DISCUSSION

10.1. Prologue

In this thesis, the need for instructor intervention in MOOC environments is explored by detecting the need for intervention based on text content posted to discussion forums by MOOC learners. It sought to detect the need for instructor intervention by using three research perspectives: (i) posts, topics, and learners using several ML approaches, (ii) detecting posts in new contexts to add priority to an intervention strategy based on learner behaviour, and (iii) adding user modelling to make intervention models more adaptable based on both learners and instructors. Furthermore, amongst other approaches, XAI was employed to make it easier for instructors and annotators to detect when interventions may be required. As a result, the models presented in this study can assist in the process of identifying when instructor intervention is required in MOOC environments based on textual content gathered from MOOC discussion forums.

First, this chapter covers the different fundamental aspects related to this thesis, such as demonstrating the importance and effect of instructor intervention in MOOC environments, discussing issues related to instructor intervention, and clarifying the normality of the data associated with MOOCs. Then, it moves on to discuss the findings presented in the literature. A summary of the thesis's outcomes with its overall findings and contributions is also provided. Furthermore, it considers the limitations related to intervention-prediction data and models. Finally, opportunities for additional future research are identified.

10.2. The Impact of Instructor Intervention in MOOCs

In MOOC discussion forums, instructors play a critical role: communicating and offering real-time intervention to help solve MOOC learners' problems by providing insightful responses. Such communication can have a significant learning impact (Ntourmas *et al.*, 2019) and is related to the retention of learners on MOOC courses.

Many researchers have investigated learner engagement, retention, and dropout in traditional online courses in general and MOOC environments specifically, focusing on different factors, one of which is instructors. In online courses, (Das, 2012) discussed the importance of the online presence of course instructors to facilitate learner engagement. Among MOOC researchers, (Hew, 2016) proposed investigating different features, one of which was instructor accessibility and passion in prompting learner engagement. Furthermore, as (Hone and El Said, 2016) clarified, one of the most significant predictors of MOOC learner retention was interaction with the course instructor. Therefore, there is an urgent need to better enable MOOC instructors to identify urgent posts and at-risk learners and provide help to them via discussion forums.

There is a call to cancel the role of manual instructor intervention and create an automatic intervention system. Although this solution may be useful in reducing the burden on instructors, the author believes that in certain intervention cases, there must be direct interaction with humans (i.e., instructor intervention).

10.3. The Issue of Instructor Intervention in MOOCs

As reviewed in the previous section, instructor intervention in MOOCs is a crucial area. However, instructor intervention in MOOCs in terms of identifying urgent posts and at-risk learners is a daunting prospect for instructors due to the sheer volume of posts and the fact that urgent cases requiring intervention are rare compared to non-urgent ones. The greater the number of participants in a MOOC forum, the harder it is for instructors to offer timely assistance (Ntourmas *et al.*, 2022). In addition, the decision to offer intervention is a subjective decision (Chandrasekaran *et al.*, 2015b) that is associated with instructors' personal preferences. This was highlighted by (Ntourmas *et al.*, 2018) who demonstrated that instructors' interventions in the discussion forums of the two different MOOC courses varied to some extent based on subjective factors. Also, usability issues related to discussion forum design are an issue; (Ntourmas *et al.*, 2019) revealed that in the OpenEdX discussion forum, course

designers must consider several issues when deciding on how instructor intervention is to be facilitated. Therefore, there is a need to develop new intervention systems for MOOCs that help instructors to identify posts/learners who require intervention and make these processes more adaptive to help both instructors and learners.

10.4. The Nature of MOOCs Data (Imbalanced Nature)

Generally, the nature of the data collected from MOOC courses is expected to have similar characteristics in that most aspects of such MOOC datasets would be imbalanced. For example, (i) the distribution of urgent posts on discussion forums vs non-urgent posts is unequal with the latter being the predominant class (Almatrafi, Johri and Rangwala, 2018); (ii) in the case of learners dropping out, many more MOOC learners drop out (around 90%) vs the 7–10% who complete their courses (Malliga, 2013); (iii) there is a very low level of course buy-in: less than 1% of all learners enrolled on an online course choose to purchase the final certificate (Alshehri *et al.*, 2021).

This thesis focuses on identifying posts on discussion forums that need intervention and learners at risk of dropping out using different ML approaches while considering the imbalanced nature of the data classifier which is biased toward predicting the dominant class. Thus, in this thesis, selecting an appropriate evaluation metric was considered when evaluating these models. The most common classifier measures were reported as average *Acc*, *P*, *R*, and *F1* scores for each class. Also, *BA* was used to identify posts; one of the target classes, *non-urgent,* appears much more frequently than other classes like *urgent*, and in identifying at-risk learner *dropout* which outnumbered *completers*.

In addition, the performance of the proposed model falls somewhat for the underrepresented class, hence, there is a need to improve the data. Thus, in this thesis, to alleviate the problem of imbalanced data in identifying urgent posts as discussed in Chapter 8, different data balancing techniques were proposed; namely, text augmentation, text augmentation with undersampling, and undersampling to improve the quality of text data as an input to the classifier that identifies posts requiring intervention.

10.5. Literature Findings

As shown in the literature review (Chapter 2), all the previous research focused on identifying posts that need instructor intervention by following a one-size-fits-all approach, without any

personalisation in an intervention based on learners, despite long-term personalisation research in education. In addition, the other issue due to the nature of MOOCs is that urgent instances for intervention are less frequent than non-urgent ones, which leads to unbalanced data. In the case of urgent-post detection, this aspect has been neglected. The studies that came closest to this (Almatrafi, Johri and Rangwala, 2018) and (Khodeir, 2021) took into account various common techniques such as data splitting and evaluation metrics but failed to address this issue by improving data imbalance.

In identifying topics, one previous study focused on assisting instructors to understand, find, and navigate the most important topic clusters by linking the topics discussed in the forum posts in relation to the relevant weekly lectures (Atapattu T, 2016). Also, a recent study (Yang, Ren and Wu, 2022) determined learners' topic attention by utilising a technique based on the TEAM model, and then visualised the topic attention of different learner groups. However, these studies did not analyse the text of posts to extract urgent language by correlating topics with urgent posts.

From a learner perspective, the literature identified dropout by using post features in addition to other features (Borrella, Caballero-Caballero and Ponce-Cueto, 2019; Xing *et al.*, 2016). However, these studies did not inspect the posts written by learners or their written post history to identify learners at risk of dropout.

10.6. Thesis Findings

This section describes the numerous ways in which the current thesis has advanced the extant literature. As shown in the literature review, most previous research has focused on identifying posts that need instructor intervention. Also, there have been some attempts to analyse topics and identify potential dropout learners based on discussion forums. However, this thesis has expanded the research approach by seeking to identify posts, topics, and learners, proposing a new direction for intervention in MOOCs by adding learner behaviours and user modelling to offer a more effective intervention model.

The umbrella research question in this thesis (How can urgent instructor intervention need be detected based on learner posts in MOOC environments?) was addressed by applying NLP techniques and different ML models using supervised and unsupervised algorithms. This includes supporting instructors in three main ways: (i) keeping track of discussions and building models that automatically identify urgent posts; (ii) analysing and visualising topics

about which learners posted on specific courses to identify urgent language; (iii) detecting at-risk learners from their temporal posts. In addition, among these models, the current thesis proposes models that can also be adapted to both instructors *and* learners. Moreover, it applies XAI to help both instructors and annotators. All these tools will help instructors to provide effective intervention to learners who may require it.

The first main research question tackles the classification of urgent posts that need instructor intervention based on different features about various dimensions of posts in addition to textual data and different textual inputs using several ML approaches. This research question was addressed by conducting two experiments: the first focused on the text-only content of posts and then added other numerical data dimensions (*sentiment, confusion, opinion, question and answer*) in addition to text. The second experiment moved to focus on the text-only content of posts and added the name and the domain of the course to the text input. The inputs are different types of text data e.g., word-only or word with character, used two different embedding approaches (BERT or word2vec) to represent words. The Stanford MOOCPosts dataset was used because it contains 11 courses that cover three different domains. Also, the proposed models can be compared with other available (state-of-the-art) models as most researchers used this data (as explained in the SLR). However, in the second experiment, some posts were removed because they contained an empty course name; as mentioned, the name and the domain of the course were added to the text input. ***RQ1*** is split into four sub-questions, ***RQ1.1*** and ***RQ1.2*** for the first experiment, and ***RQ1.3*** and ***RQ1.4*** for the second experiment, as follows:

- ***RQ1.1:*** *Is there a relationship between the various dimensions of the learners' posts and their need for urgent instructor intervention?*

This was addressed by analysing and visualising the relationship between the ratio number of urgent and non-urgent posts across the five dimensions (*sentiment, confusion, opinion, question, and answer*). The results emphasised that there is an association between the percentages of non-urgent/urgent posts and these dimensions. Interestingly, it focused on four (neutral) scales and compared the values of (4) and (4.5) for *sentiment* with the rates of non-urgent/urgent; also, for *confusion* with the rates of non-urgent/urgent. This showed that there *is* a relationship between specific values (4 and 4.5) for the *sentiment* and *confusion* scales with the proportion of urgent/urgent posts. As explained previously (Chapter 4, Section 4.2.3.1), this

is the reason for choosing >4 as the threshold to define urgent posts in the Stanford dataset. The results also indicate that there is a significant correlation between *urgency* and *confusion* as well as *question*.

- ***RQ1.2:*** *Does using several dimensions as features in addition to textual data increase the model's predictive power for identifying posts that require the need for urgent instructor intervention when using deep learning?*

This was addressed by developing and training different models as a basic model (text-only) and a multidimensional model that integrates different numerical features (which are the five dimensions: *sentiment, confusion, opinion, question, and answer*) in relation to posts in addition to text features. The findings are interesting and highlight that combining several dimensions as features in addition to textual data (multidimensional model) increases the DL model's ability to identify when urgent instructor intervention is required (see Section 4.2.3.2). This is because using different characteristics about posts facilitates the detection of urgent posts.

- ***RQ1.3:*** *What is the preferable combination between different deep learning models to construct the best predictor model amongst them to identify posts that need instructor intervention?*

This was addressed by applying different simple and hybrid deep neural networks (known as the 'plug & play' technique) for different input levels (word-based and word-character based), based on different embeddings (word2vec or BERT). Based on the results (Chapter 4, Section 4.3.3), models using BERT for word embedding outperformed all the word2vec-based models. That implies that using BERT to represent words is preferable. The best value from all the models for *R* for class (1) is **0.81** and *BA* is **0.872** in CNN + LSTM + Attention model based on BERT at the word-level. Here, it is preferred to detect all urgent cases that should be focused on *R*. Also, it focuses on *BA* which is a widely used metric for the binary classification of imbalanced datasets.

- ***RQ1.4:*** *Do word-character-based approaches outperform word-based approaches for the post urgency problem and is this different when using BERT for word embedding, compared to more traditional models (e.g., word2vec)?*

This was addressed by comparing two input levels (word-based versus word-character based) with two different embeddings (word2vec or BERT). The models that employed word-character-based input with word2vec for word embedding tend to outperform those using word-based input. In contrast, there is no improvement for models that use BERT for word embedding in different base inputs (word only and word-character) (see Table 4.5). In other words, BERT is capable of representing words on its own without any support.

Another main research question that this thesis attempted to answer is related to topics and extracting the language of urgency by analysing text posts across a specific course and providing a visual representation of these topics. One course (SciWrite) from The Stanford MOOCPosts dataset was used as a case study because it contains a large number of posts with a high proportion of urgent posts. The sub-questions for the main question ***RQ2*** are as follows:

- ***RQ2.1:*** *Can the language of urgency be detected from learners' posts?*

To extract urgent language and understand the topics, LDA was used to cluster words from forum posts into different topics. Then, these topic lists and trending terms were linked with urgent posts as it is a useful indicator for exploring urgent language. The results showed the top ten terms on six topics (the optimal number of LDA topics in this experiment). Then, under the assumption that this ensured that they were the most representative posts of that particular topic, the proportions of urgent and non-urgent posts for each topic with a dominant contribution of more than 80% were calculated. Furthermore, this shows why these non-urgent statements utilise wording that appears to be urgent (as explained further in Chapter 5, Section 5.4.2).

- ***RQ2.2:*** *Can the language of urgency be visualised simply and intuitively?*

This was addressed by using different simple aids: (i) displaying word cloud visualisations (top ten terms) for each topic represented by a different colour; (ii) adding pyLDAvis interactively to provide instructors with a summary and interpretation of topics. To further assist instructors (and perhaps learners), (iii) the post tokens for each post were coloured according to the specific topic.

The other major research question was concerned with predicting learners who might drop out and may need intervention based on their posting history. The dataset used to implement

this experiment was the Dropout dataset from FutureLearn (as clarified in Section 3.2.2.3). The sub-questions for this question ***RQ3*** are as follows:

- ***RQ3.1:*** *Which multi-input models (processing several recent posts) are useful for predicting learners who may drop out (thus may need instructor intervention)?*

This was addressed by establishing several intervention models that utilise two forms of supervised multi-input ML classification models (other deep learning architectures and Transformer). In the Transformer model, an approach for siamese and dual BERT was used to create multi-inputs with binary text classification which were termed *multi-siamese BERT* and *multiple BERT,* respectively. The results (see Table 6.2) showed that the intervention model represented by the Transformer models (*multi-siamese BERT* and *multiple BERT*) can more accurately identify at-risk learners, predict dropout learners, and determine the need for intervention.

- ***RQ3.2:*** *Does clustering learners based on their number of posts prior to the prediction step improve prediction outcomes?*

This was explored by clustering learners based on their number of posts before prediction. Then, the same experiments were examined for the best previous intervention models (Transformer-based) with specific groups of learners. According to the results and against this assumption, it provided negative values for predicted outcomes (as discussed in Chapter 6, Section 6.4, Table 6.3).

Another main research question was analysing the behaviour of learners who need urgent intervention and the possibility of designing an architecture for prioritising instructor intervention based on learner behaviour. The Gold-standard corpus from Futurelearn was used because this data enables the study of learner behaviour in terms of step access. ***RQ4*** was divided into four sub-questions, as follows:

- ***RQ4.1:*** *Is there a relationship between the number of posts written by learners who need urgent intervention and the average number of posts?*

The observation from inspecting learners' writing behaviour illustrated that learners often tend to write more posts overall if they write more posts that require intervention (as represented in

Figure 7.6, Chapter 7, Section 7.3.1); there tends to be a positive relationship between the average number of posts and the urgency of the posts.

- ***RQ4.2:*** *Is there a relationship between high-frequency (HF) commenter learners who require urgent intervention and their average number of step access instances?*

A study of the number of step(s) accessed by the *urgent* (HF commenters) group and the *non-urgent* group was achieved by calculating the average number of steps accessed. The urgent group's step access had a lower average count than the other group (see Figure 7.7 - left).

- ***RQ4.3:*** *Is there a relationship between the number of HF commenter learners and completion-rates?*

This was addressed by visualising the relationship between different learner groups and the completion rates. The results revealed that only 13% of HF commenters who need immediate assistance are expected to finish the course as opposed to 27% of the non-urgent group (see Figure 7.7 - right). This, in the researcher's opinion, is one of the causes of the high dropout rate.

- ***RQ4.4:*** *How can an intervention priority framework based on behaviour be designed?*

This was achieved by proposing a novel framework to provide an automated intervention priority model for MOOCs containing two phases: (i) a prediction phase by using BERT as a classifier model; (ii) an intervention priority phase that adds priority (high, mid or low) based on different risk-level groups. To further confirm the efficacy of the proposed model, the relationship between the different risk groups of learners identified (high, mid, and low) and their completion rates was computed. As a result, most completion rates for high-risk learners were quite low, while those for mid-risk learners were average; those for low-risk learners were very high (as depicted in Figure 7.8. Chapter 7, Section 7.3).

For the other main research question, to improve prediction models for instructor intervention based on posts and make these models more personalised and adaptive, user modelling (specifically learner and instructor modelling) was added to enable an instructor to decide when to intervene. These adaptive models are based on proposing a solution for unbalanced data which is one of the issues affecting MOOC discussion forums. The UNITE

dataset was used (which was derived from FutureLearn) because it contains very rare urgent cases (only 7%). ***RQ5*** was divided into two sub-questions, as follows:

- ***RQ5.1:*** *How can the data imbalance issue (urgent versus non-urgent) in learners' posts be addressed?*

This was addressed by applying comprehensive data balancing techniques comprising text augmentation, text augmentation with undersampling, and undersampling (see Section 8.3.2.2). Also, several new pipelines for combining different augmenters for text augmentation were proposed. Among these models, combining 3x augmentation + undersampling usually achieves the best performance (see Section 8.4.1).

- ***RQ5.2:*** *What would an adaptive intervention model to assist instructors in dealing with urgent posts look like?*

To improve the intervention task, the adaptive intervention models (interactive systems that can be adapted or adapt themselves to their current users) were constructed based on the instructor and learner models. Two scenarios were suggested (see Section 8.3.3): the basic one (semi-automatic instructor intervention) then an expanded scenario was used based on coarse granularity and expanded learner models (semi-adaptive instructor intervention). This personalises (by automatic adaptation) the identification process of urgent posts in MOOCs for instructors, the primary users who need to manage their workloads, as well as, indirectly, catering for the needs of learners as secondary users, to have their urgent messages identified (and ultimately, resolved). As the results show, this approach will improve the instructor intervention process, reduce the number of posts that instructors need to read, and enhance the quality of instructor interactions with learners.

Also, this research project attempted to answer the final main research question in this thesis which is related to creating a transparent XAI model to detect urgent intervention to support instructors' decisions to intervene in posts as well as annotators' decisions. The Gold-standard corpus from Futurelearn was applied because it is manually labelled which provides an opportunity to study the annotation process. The sub-questions of ***RQ 6*** are as follows:

- ***RQ6.1:*** *How can a transparent XAI model be constructed to further support instructors' decisions to intervene based on an urgent posts-intervention-need detection model?*

This was answered by developing a BERT-based automatic urgent intervention model. Next, the model was explained by visualising important words using the Captum tool (as shown in Figure 9.4). An instructor can use this visualisation to better comprehend the decision and suggestions made by a classifier to detect urgency in learner discussions on MOOCs.

- ***RQ6.2:*** *How can XAI be employed to improve human annotators' decisions about the urgency of posts (i.e., deciding on which posts need intervention)?*

This was addressed by connecting AI prediction error to human (lack of) confidence, focusing on (TP) and (FP) (as explained in Section 9.4). Thus, three scenarios with < 100% confidence between annotators were analysed, based on true positives (TP). Therefore, this shows how the colour-based highlighting functionality of XAI can provide an in-depth understanding of the algorithm's decision-making process. Emphasising the key phrases could make it easier for annotators to determine whether a post is urgent.

10.7. Limitations

In any academic research, the limitations must be clarified to improve them in future work. Thus, the main limitations of the current thesis are highlighted as follows:

First, although the instructor intervention task in MOOC discussion forums is very important, it is not easy as such decisions are very subjective (Chandrasekaran *et al.*, 2015b). Thus, creating data that serves this field of research is very challenging. It can be seen in the literature that researchers used different methods to create datasets: (i) labelling data as the instructor's decision to intervene in threads as guided by a data-driven approach in which 0 = if the instructor did not respond to the thread and 1 = an intervention occurred. The author believes that this method is inaccurate because there might have been some posts where the instructor decided that an intervention was required but did not intervene (because they missed the post or ran out of time, etc.); similarly, there may be unnecessary interventions to posts due to subjective issues; (ii) using crowdsourcing to label data as in the Standford MOOCpost dataset which hired three coders to label each domain, then they set the label of urgency as an unweighted average of the two scores from two coders; still, their agreement is not very high (as clarified in Section 3.2.1). In this thesis, the author followed the Stanford approach but used

four experts in the field; three of them are university computer science instructors in addition to the author of this thesis. That said, the presented data labelled by experts still had a low level of agreement. Thus, creating a gold standard data set is very challenging due to the subjectivity involved in making a decision to intervene as well as being time-consuming. However, a plan to create a gold standard data set with high agreement between annotators and rich data is a request for future researchers.

Second, some of the models proposed in this thesis such as automatic classification in general and the potential solution for unbalanced data or extracting urgent language may not be general enough for all online courses and MOOC platforms as it has been applied to only one specific course (i.e., from the FutureLearn dataset to solve unbalanced data or to predict dropout learners, and from the Stanford dataset to extract urgent language). However, for the unbalanced problem as shown in Chapter 8 for the course from FutureLearn, further validation of the best solution on the highly popular and well-used Stanford dataset was provided, thus strengthening the case for the generalisability of this approach and its applicability across other MOOC courses and domains.

10.8. Future Work

Although the current research project contributes to predicting instructor intervention need based on MOOC discussion forums, there is still room for future work to improve on the intervention predicting task, as described in the following:

- In general, utilising other datasets, as well as other courses and environments, to further generalise the findings and evaluate whether implementing other NN models or combining other different NN can increase performance in terms of classifying urgent posts or identifying dropout learners.
- Other general further work can link with the work on pedagogical interventions for automated guidance to instructors (Chandrasekaran *et al.*, 2015a).
- Another future direction is considering communication on posts in MOOC platforms in other languages (i.e., non-English); for example, Chinese, Hindi, and Arabic, etc. This approach seeks to make the findings of this thesis more generalisable to MOOCs operating in other countries and languages.
- In analysing topics, analysing other courses remains an avenue for future research. In addition, further research can consider using other tools for topic modelling.

- In identifying learners based on their temporal sequence of posts, plan to replicate the proposed models with other courses and different numbers of posts to further explore the generalisability of these findings. Moreover, add clickstream data as additional features.

10.9. Epilogue

The literature review highlighted the importance of instructor intervention in MOOC environments. Recently, researchers have paid attention to the problem of intervention and developed a set of computational models that help mitigate this problem. However, there is still much work to do in terms of improving models' performance, extracting urgent language, inspecting learners' posting history, studying learner behaviour, and making models better adapted to the needs of instructors and learners. This thesis fills this gap by considering different aspects of instructor interventions in MOOCs based on discussion forums beginning with basic features like posts, topics, and learners, and then expanding to study learner behaviour and adaptions. This research project also contributes to dealing with the imbalanced data issue which is one of the characteristics of MOOC environments. In addition, it has proposed how XAI can be used in addressing the instructor intervention problem. The outcomes of these contributions have been discussed in this chapter along with recommendations for future opportunities for development. The beneficiaries of the findings of the thesis in terms of its outputs are MOOC instructors (primary users), MOOC learners (secondary users), and MOOC providers. The following chapter concludes this thesis.

CHAPTER 11: CONCLUSION

Determining the need for instructor intervention in MOOC discussion forums has become an extremely important issue in distance education due to the commitment to openness and the need to cater for huge numbers of learners and vast numbers of posts. Such intervention is required to support learners and thus may reduce drop-out rates. However, the critical challenge here is the extremely high ratios of learners to instructors and the nature of MOOC discussion forums in terms of the vast number of posts of which only a low number thereof require urgent intervention. Thus, this thesis tackled the intervention problem from three main perspectives: (i) posts, (ii) topics, and (iii) learners to improve on the extant intervention models. Then, it sought to expand the identification of posts based on (iv) posts with learner behaviour and adding priority in intervention and (iv) posts with user modelling and solving one of the main issues of highly unbalanced post data. Finally, it applied XAI to improve not only the instructor intervention task but also the annotators' decision-making issue. The above was achieved by implementing different architectures, models, and experiments using the two different MOOC platforms (Stanford MOOCPosts and FutureLearn). These proposed models can be applied as intelligent systems in MOOC environments.

The Stanford corpus was the foundation for the majority of earlier studies on instructor intervention, as clarified in the SLR. In this thesis, in addition to employing the Stanford corpus, a new instructor intervention corpus was created based on the FutureLearn platform which was annotated by human experts similar to the previous Stanford MOOCPosts dataset in how it annotated intervention decisions. Different ways of constructing a gold standard corpus were proposed (as clarified in detail in Chapter 3).

Note that this study is the first to conduct a SLR in the field of instructor intervention in MOOC discussion forums to identify and analyse the extant studies in this field (Chapter 2). To overcome the limitations in the extant literature, the current research project inspected novel approaches. The initial approach to resolving the intervention issue involved considering posts (Chapter 4). Two experiments were conducted using the Stanford MOOCPosts dataset. In the first experiment, the purpose of the research was to predict automatic intervention based on learner post content incorporating NLP and other features captured from the posts (*sentiment, confusion, opinion, question, and answer*), the research looked at how these dimensions related to the rate of the number of urgent posts. The findings demonstrate that including these dimensions as features in addition to text features improves DL models' performance and makes intervention more accurate. To construct and train this multidimensional DL, a novel architecture based on sub-models was developed.

The second experiment in classifying posts used a 'plug & play' approach by proposing a classification model for identifying when a given post needs instructor intervention. This was based on various simple and hybrid neural networks with different types of inputs: (i) word-based level input; (ii) configuring what is referred to as word-character-based input by adding character-based input in addition to word-based input. The words were represented using different word embedding (word2vec or BERT). The end goal is to establish a good combination in terms of performance. The results show that the BERT-based models outperformed the models that used word2vec for word embedding with word input only. The best model is the CNN + LSTM + Attention model based on BERT at the word-level which achieves promising results (BA = 0.875). It is noteworthy to mention that the proposed model outperformed the cutting-edge model. Also, the results show that it is preferable to utilise BERT as a standalone tool for embedding without any additional input in the form of characters.

In relation to topics (Chapter 5), the findings highlight that learners express their need for urgent intervention via discussion forums using special language. Thus, it is useful to extract this language to (i) help instructors in their intervention and (ii) learners when writing such content. The instructor may be able to assist more successfully if visualisation is used. Using a course from the Stanford MOOCPosts dataset as a case study, learner posts were analysed to investigate the language signalling that urgent intervention is needed. The analysis revealed that some words are connected to one another and reflect a demand for quick action, particularly in posts at the thread level. It is significant here that this research project is the first to propose a *context-dependent urgency language;* that is, a language expressing the need for

urgent intervention in a MOOC, and that also it demonstrated some simple and easily reproducible methods for extracting and visualising the above.

In terms of learners (Chapter 6), the research project attempted to predict learner dropout and their need for intervention from their most recent posts. Various ML models were built using the FutureLearn dataset (Dropout) including other deep learning architectures and Transformer with multi-input to enable instructors to intervene more effectively. To add more than two inputs to the Transformer models, multi-siamese BERT and multiple BERT based on siamese and dual BERT were developed. The findings show that the intervention model may identify at-risk learners more accurately with the inclusion of the Transformer model.

To provide more valuable interventions (Chapter 7), priority of intervention was included using the FutureLearn dataset (Gold-standard corpus). Firstly, an analysis of learner posts for urgency was offered showing that learners with high step access rates require less intervention to their posts. This might be because if they have written several posts that require attention, learners may become less motivated to access the course materials. Also, it verified that the majority of course completers did not require significant intervention with regard to their posts. Based on these findings, a framework and algorithm were developed to prioritise instructor intervention, encouraging instructors to support and assist their learners by concentrating on high-risk learners first, thus improving the possible outcomes of such valuable interventions. The results demonstrate that most completion rates for high-risk learners are quite low, while those for mid-risk learners are average, and those for low-risk learners are very high.

The difficulty in developing models to effectively recognise urgent cases is explained by the fact that MOOC post datasets only include a small number of urgent cases, leading to imbalanced data (Chapter 8). In this study, the issue of imbalanced data was solved by applying different strategies. Also, the study makes instructor intervention more valuable by adding adaptation based on instructors' and learners' models using the UNITE dataset. To enhance the quality of such data, three strategies (data augmentation, data augmentation+undersampling, and undersampling) were employed and compared. Additionally, several new pipelines that included various data augmenters were offered. The results demonstrate that undersampling can improve model performance to detect urgent cases, and that combining data augmentation and undersampling yields the best results in achieving class balance. Adding adaptation with two different scenarios will improve instructor tasks. Finally, incorrectly classified urgent cases were investigated in more detail; it was discovered that the issue is not limited to the classifier;

it also originates from the intervention task, which is immensely challenging for humans to annotate.

Finally, the goal of the research on XAI was to explain the ML decisions made for a particular text classification problem by explaining individual prediction in the urgent intervention task in a MOOC environment that may help instructors with their interventions (Chapter 9). In addition, the field of urgency prediction was advanced by proposing a new method for supporting annotators.

APPENDIX A

Manually classify comments from online classes

You are asked to analyse **5790** comments which have been written on a FutureLearn **MOOC** (**M**assive **O**pen **O**nline **C**ourses) platform for several of their courses. For each such course, there are thousands of learners and their comments to each step of the course, and it is difficult for the instructor to answer to all of them. Other learners can see the comments and answer them as well, or express their 'like'-ing of a specific comment. However, the retention on such MOOC courses is low (in average, around 10%). You need to evaluate if the comments posted need instructor intervention or not – i.e., if the instructor needs to respond to that comment or question (How urgent is it that the instructor get involved in response to the post?). Urgency indicates the degree to which the instructor(s) should be concerned with the content of the post. If a post is very urgent, then the instructor should respond to the post as soon as possible. If a post is not urgent, then the instructor might not have to respond to the post at all. For this purpose, you will be labelling the comments/questions with ratings from 1 to 7. We ask you to use your own judgement. You will enter your results in a spreadsheet.

To help in the task, different degrees of urgency *to the instructor* are mapped to scores as follows:

- No reason to read the post → 1
- Not actionable; read if time → 2
- Not actionable; maybe interesting → 3
- Neutral: respond if spare time → 4
- Somewhat urgent: good idea to reply teaching assistant might suffice. → 5
- Very urgent: good idea for instructor to reply. → 6
- Extremely urgent: instructor definitely needs to reply → 7

Finally, we need you to pay attention to two important points:

• You should take enough time to read and understand each comment, make your own decision.

• If you find by chance any personal information in any comment, treat it in a confidential way and do not store or use it in any way.

Note:

The course name is: **Big Data.**

Instructions:

- Open your spreadsheet.
- You care only about the text column.

- The text cell in each row contains one forum posts (comment).
- In post_type column, Comment mean ‘main comment’ while subComment mean ‘reply to other comment’.
- The cell with column header 'Urgency(1-7)' need ratings from 1 to 7.

APPENDIX B

The results on naive Bayes with other feature engineering and the other traditional models (logistic regression, support vector machine, random forest and boosting model - extreme gradient boosting (XGBoost)) rendered similar results as those shown in the results section (naive Bayes model with count vector as a feature engineering) in the UNITE dataset.

Table B.1: The performance results of the naive Bayes model with various types of feature engineering with original data, with three approaches to augmentation (see Table 8.3 above) using 3x and 9x (see Table 8.2 above) with and without undersampling and with undersampling without augmentation in the UNITE dataset.

Feature Engineering	Augmentation		Under	Acc	Non urgent 0			Urgent 1		
					P	R	F1	P	R	F1
TF-IDF vectors (word level)	×		×	0.93	0.93	1.00	0.96	0.00	0.00	0.00
	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.62	0.05	0.10
		9X	×	0.89	0.94	0.94	0.94	0.26	0.26	0.26
		3X	√	0.77	0.96	0.79	0.86	0.17	0.57	0.26
		9X	√	0.84	0.95	0.87	0.91	0.20	0.43	0.28
	Approach #2	3X	×	0.93	0.93	1.00	0.96	0.69	0.05	0.10
		9X	×	0.91	0.94	0.97	0.95	0.31	0.20	0.25
		3X	√	0.79	0.96	0.81	0.88	0.18	0.53	0.27
		9X	√	0.88	0.95	0.92	0.93	0.24	0.32	0.27
	Approach #3	3X	×	0.93	0.93	1.00	0.96	0.72	0.05	0.10
		9X	×	0.91	0.94	0.96	0.95	0.30	0.21	0.24
		3X	√	0.80	0.96	0.82	0.89	0.19	0.52	0.27
		9X	√	0.87	0.95	0.92	0.93	0.24	0.35	0.28
	×		√	0.47	0.98	0.44	0.60	0.11	0.90	0.19
TF-IDF vectors (n gram word level)	×		×	0.93	0.93	1.00	0.96	1.00	0.00	0.01
	Approach #1	3X	×	0.93	0.94	0.99	0.96	0.52	0.13	0.21
		9X	×	0.90	0.95	0.94	0.95	0.30	0.31	0.31
		3X	√	0.86	0.95	0.89	0.92	0.24	0.45	0.31

		9X	√	0.87	0.95	0.91	0.93	0.24	0.39	0.30
	Approach #2	3X	×	0.93	0.94	0.99	0.96	0.52	0.12	0.20
		9X	×	0.90	0.95	0.95	0.95	0.31	0.29	0.30
		3X	√	0.85	0.95	0.89	0.92	0.23	0.45	0.30
		9X	√	0.88	0.95	0.92	0.93	0.26	0.36	0.30
	Approach #3	3X	×	0.93	0.94	0.99	0.96	0.47	0.13	0.20
		9X	×	0.90	0.95	0.95	0.95	0.30	0.30	0.30
		3X	√	0.86	0.95	0.89	0.92	0.24	0.42	0.30
		9X	√	0.87	0.95	0.91	0.93	0.25	0.38	0.30
	×		√	0.63	0.97	0.62	0.76	0.13	0.72	0.22
TF-IDF vectors (n gram character level)	×		×	0.93	0.93	1.00	0.96	0.31	0.02	0.03
	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.57	0.06	0.11
		9X	×	0.92	0.94	0.99	0.96	0.39	0.12	0.19
		3X	√	0.88	0.95	0.92	0.94	0.28	0.40	0.33
		9X	√	0.91	0.94	0.97	0.95	0.34	0.20	0.25
	Approach #2	3X	×	0.93	0.93	1.00	0.96	0.57	0.06	0.11
		9X	×	0.93	0.93	0.99	0.96	0.41	0.08	0.13
		3X	√	0.90	0.95	0.95	0.95	0.32	0.33	0.33
		9X	√	0.92	0.93	0.99	0.96	0.35	0.10	0.16
	Approach #3	3X	×	0.93	0.93	1.00	0.96	0.56	0.07	0.12
		9X	×	0.93	0.94	0.99	0.96	0.48	0.14	0.21
		3X	√	0.89	0.96	0.92	0.94	0.31	0.44	0.36
		9X	√	0.92	0.94	0.98	0.96	0.39	0.18	0.25
	×		√	0.56	0.98	0.53	0.69	0.13	0.87	0.22

Table B.2: The performance results of the logistic regression model with various types of feature engineering with original data, with three approaches to augmentation (see Table 8.3 above) using 3x and 9x (see Table 8.2 above) with and without undersampling and with undersampling without augmentation in the UNITE dataset.

Feature Engineering	Augmentation		Under	Acc	Non urgent 0			Urgent 1		
					P	R	F1	P	R	F1
Count vector	×		×	0.92	0.94	0.98	0.96	0.38	0.14	0.21
	Approach #1	3X	×	0.91	0.94	0.96	0.95	0.34	0.24	0.28
		9X	×	0.89	0.95	0.94	0.94	0.28	0.32	0.30
		3X	√	0.84	0.95	0.87	0.91	0.21	0.45	0.29
		9X	√	0.88	0.95	0.92	0.93	0.26	0.38	0.31
	Approach #2	3X	×	0.91	0.94	0.96	0.95	0.33	0.25	0.28
		9X	×	0.89	0.95	0.94	0.94	0.28	0.30	0.29
		3X	√	0.83	0.96	0.86	0.90	0.21	0.51	0.30
		9X	√	0.88	0.95	0.92	0.93	0.26	0.36	0.30
	Approach #3	3X	×	0.91	0.94	0.97	0.95	0.35	0.24	0.28
		9X	×	0.90	0.94	0.95	0.95	0.28	0.28	0.28
		3X	√	0.84	0.96	0.87	0.91	0.23	0.50	0.32
		9X	√	0.88	0.95	0.93	0.94	0.26	0.34	0.29
	×		√	0.71	0.96	0.72	0.82	0.15	0.64	0.24
TF-IDF vectors (word level)	×		×	0.93	0.93	1.00	0.96	0.00	0.00	0.00
	Approach #1	3X	×	0.93	0.94	0.99	0.96	0.49	0.15	0.24
		9X	×	0.89	0.95	0.93	0.94	0.29	0.35	0.32
		3X	√	0.83	0.96	0.86	0.91	0.22	0.52	0.31
		9X	√	0.87	0.95	0.90	0.93	0.25	0.44	0.32
	Approach #2	3X	×	0.93	0.94	0.99	0.96	0.56	0.16	0.25
		9X	×	0.91	0.95	0.96	0.95	0.34	0.29	0.31
		3X	√	0.83	0.96	0.85	0.90	0.22	0.55	0.31
		9X	√	0.88	0.95	0.92	0.94	0.26	0.35	0.30
	Approach #3	3X	×	0.93	0.94	0.99	0.96	0.52	0.16	0.24
		9X	×	0.90	0.95	0.95	0.95	0.31	0.30	0.30
		3X	√	0.85	0.96	0.88	0.91	0.23	0.49	0.31
		9X	√	0.88	0.95	0.92	0.93	0.27	0.40	0.32
	×		√	0.72	0.97	0.72	0.83	0.16	0.68	0.26
	×		×	0.93	0.93	1.00	0.96	1.00	0.00	0.01

TF-IDF vectors (n gram word level)	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.69	0.09	0.17
		9X	×	0.90	0.95	0.94	0.94	0.30	0.35	0.32
		3X	√	0.86	0.96	0.89	0.92	0.25	0.46	0.32
		9X	√	0.86	0.95	0.89	0.92	0.24	0.45	0.32
	Approach #2	3X	×	0.93	0.93	1.00	0.96	0.60	0.08	0.14
		9X	×	0.90	0.95	0.94	0.94	0.29	0.32	0.30
		3X	√	0.85	0.96	0.88	0.92	0.23	0.46	0.31
		9X	√	0.85	0.95	0.88	0.92	0.22	0.43	0.29
	Approach #3	3X	×	0.93	0.93	1.00	0.96	0.65	0.09	0.16
		9X	×	0.90	0.95	0.95	0.95	0.33	0.34	0.33
		3X	√	0.87	0.95	0.91	0.93	0.26	0.43	0.33
		9X	√	0.88	0.95	0.91	0.93	0.28	0.42	0.33
	×		√	0.74	0.96	0.75	0.84	0.16	0.62	0.25
TF-IDF vectors (n gram character level)	×		×	0.93	0.93	1.00	0.96	1.00	0.01	0.01
	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.65	0.04	0.07
		9X	×	0.93	0.93	0.99	0.96	0.58	0.09	0.16
		3X	√	0.93	0.95	0.97	0.96	0.46	0.28	0.35
		9X	√	0.93	0.94	0.99	0.96	0.53	0.12	0.19
	Approach #2	3X	×	0.93	0.93	1.00	0.96	0.72	0.05	0.10
		9X	×	0.93	0.93	1.00	0.96	0.64	0.09	0.16
		3X	√	0.93	0.95	0.97	0.96	0.48	0.32	0.39
		9X	√	0.93	0.94	0.99	0.96	0.58	0.14	0.22
	Approach #3	3X	×	0.93	0.94	1.00	0.97	0.71	0.13	0.22
		9X	×	0.93	0.94	0.99	0.97	0.63	0.16	0.26
		3X	√	0.91	0.95	0.94	0.95	0.36	0.41	0.38
		9X	√	0.93	0.94	0.99	0.96	0.57	0.25	0.34
	×		√	0.75	0.97	0.76	0.85	0.18	0.69	0.29

Table B.3: The performance results of the support vector machine model with various types of feature engineering with original data, with three approaches to augmentation (see Table 8.3 above) using 3x and 9x (see Table 8.2 above) with and without undersampling and with undersampling without augmentation in the UNITE dataset.

Feature Engineering	Augmentation		Under	Acc	Non urgent 0			Urgent 1		
					P	R	F1	P	R	F1
Count vector	×		×	0.93	0.93	1.00	0.96	0.00	0.00	0.00
	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.62	0.08	0.14
		9X	×	0.91	0.94	0.96	0.95	0.34	0.25	0.28
		3X	√	0.88	0.95	0.92	0.93	0.25	0.36	0.30
		9X	√	0.89	0.95	0.94	0.94	0.27	0.29	0.28
	Approach #2	3X	×	0.93	0.93	1.00	0.96	0.57	0.07	0.12
		9X	×	0.92	0.94	0.98	0.96	0.39	0.20	0.26
		3X	√	0.86	0.95	0.90	0.92	0.22	0.37	0.27
		9X	√	0.91	0.94	0.96	0.95	0.33	0.27	0.29
	Approach #3	3X	×	0.93	0.93	1.00	0.96	0.57	0.08	0.13
		9X	×	0.92	0.94	0.98	0.96	0.42	0.21	0.28
		3X	√	0.87	0.95	0.91	0.93	0.25	0.38	0.30
		9X	√	0.91	0.95	0.96	0.95	0.35	0.29	0.32
	×		√	0.66	0.95	0.66	0.78	0.12	0.58	0.19
TF-IDF vectors (word level)	×		×	0.93	0.93	1.00	0.96	0.00	0.00	0.00
	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.64	0.09	0.15
		9X	×	0.92	0.94	0.98	0.96	0.38	0.20	0.26
		3X	√	0.91	0.95	0.96	0.95	0.34	0.29	0.32
		9X	√	0.91	0.94	0.96	0.95	0.33	0.25	0.28
	Approach #2	3X	×	0.93	0.93	1.00	0.96	0.59	0.09	0.15
		9X	×	0.93	0.94	0.98	0.96	0.46	0.20	0.28
		3X	√	0.90	0.95	0.94	0.95	0.32	0.34	0.33
		9X	√	0.92	0.94	0.97	0.96	0.38	0.24	0.29
	Approach #3	3X	×	0.93	0.93	1.00	0.96	0.71	0.07	0.13
		9X	×	0.93	0.94	0.99	0.96	0.51	0.16	0.25
		3X	√	0.92	0.94	0.97	0.96	0.38	0.26	0.31
		9X	√	0.93	0.94	0.98	0.96	0.46	0.21	0.29
	×		√	0.72	0.96	0.73	0.83	0.16	0.65	0.25
	×		×	0.93	0.93	1.00	0.96	0.50	0.00	0.01

TF-IDF vectors (n gram word level)	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.56	0.07	0.12
		9X	×	0.91	0.95	0.96	0.95	0.34	0.28	0.30
		3X	√	0.91	0.95	0.95	0.95	0.34	0.33	0.34
		9X	√	0.89	0.95	0.93	0.94	0.27	0.35	0.31
	Approach #2	3X	×	0.93	0.93	0.99	0.96	0.55	0.08	0.14
		9X	×	0.91	0.94	0.96	0.95	0.31	0.23	0.26
		3X	√	0.90	0.95	0.95	0.95	0.31	0.30	0.30
		9X	√	0.88	0.95	0.93	0.94	0.25	0.32	0.28
	Approach #3	3X	×	0.93	0.93	1.00	0.96	0.55	0.05	0.09
		9X	×	0.92	0.94	0.98	0.96	0.41	0.18	0.25
		3X	√	0.92	0.94	0.97	0.96	0.38	0.20	0.26
		9X	√	0.91	0.94	0.97	0.95	0.35	0.24	0.28
		×		√	0.67	0.96	0.66	0.79	0.14	0.68	0.23
TF-IDF vectors (n gram character level)		×		×	0.93	0.93	1.00	0.96	1.00	0.00	0.01
	Approach #1	3X		×	0.93	0.93	1.00	0.96	0.67	0.02	0.05
		9X		×	0.93	0.93	1.00	0.96	0.74	0.06	0.11
		3X		√	0.93	0.94	0.99	0.97	0.61	0.23	0.33
		9X		√	0.93	0.93	1.00	0.96	0.73	0.08	0.15
	Approach #2	3X		×	0.93	0.93	1.00	0.96	0.73	0.05	0.09
		9X		×	0.93	0.93	1.00	0.96	0.68	0.07	0.13
		3X		√	0.94	0.95	0.99	0.97	0.60	0.27	0.37
		9X		√	0.93	0.94	1.00	0.96	0.67	0.11	0.19
	Approach #3	3X		×	0.94	0.94	1.00	0.97	0.81	0.13	0.22
		9X		×	0.93	0.94	1.00	0.97	0.70	0.14	0.24
		3X		√	0.93	0.95	0.97	0.96	0.51	0.34	0.41
		9X		√	0.94	0.94	0.99	0.97	0.69	0.18	0.29
		×		√	0.76	0.97	0.76	0.85	0.18	0.70	0.29

Table B.4: The performance results of the random forest model with various types of feature engineering with original data, with three approaches to augmentation (see Table 8.3 above) using 3x and 9x (see Table 8.2 above) with and without undersampling and with undersampling without augmentation in the UNITE dataset.

Feature Engineering	Augmentation		Under	Acc	Non urgent 0			Urgent 1		
					P	R	F1	P	R	F1
Count vector	×		×	0.93	0.93	1.00	0.96	0.80	0.01	0.02
	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.54	0.04	0.07
		9X	×	0.93	0.94	0.98	0.96	0.46	0.19	0.27
		3X	√	0.88	0.95	0.91	0.93	0.27	0.41	0.33
		9X	√	0.90	0.95	0.95	0.95	0.33	0.34	0.34
	Approach #2	3X	×	0.93	0.93	1.00	0.96	0.64	0.05	0.09
		9X	×	0.92	0.94	0.98	0.96	0.40	0.14	0.21
		3X	√	0.87	0.95	0.90	0.93	0.24	0.41	0.31
		9X	√	0.92	0.94	0.97	0.96	0.37	0.25	0.30
	Approach #3	3X	×	0.93	0.93	1.00	0.96	0.67	0.04	0.08
		9X	×	0.93	0.93	0.99	0.96	0.42	0.09	0.15
		3X	√	0.90	0.95	0.93	0.94	0.32	0.40	0.35
		9X	√	0.92	0.94	0.97	0.96	0.41	0.25	0.31
	×		√	0.68	0.96	0.68	0.80	0.14	0.68	0.23
TF-IDF vectors (word level)	×		×	0.93	0.93	1.00	0.96	0.67	0.01	0.01
	Approach #1	3X	×	0.93	0.93	1.00	0.96	0.62	0.04	0.07
		9X	×	0.92	0.94	0.98	0.96	0.41	0.20	0.27
		3X	√	0.88	0.95	0.92	0.94	0.29	0.43	0.35
		9X	√	0.89	0.95	0.93	0.94	0.28	0.34	0.31
	Approach #2	3X	×	0.93	0.93	1.00	0.96	0.58	0.05	0.08
		9X	×	0.92	0.94	0.98	0.96	0.41	0.20	0.27
		3X	√	0.87	0.95	0.91	0.93	0.25	0.41	0.31
		9X	√	0.90	0.95	0.95	0.95	0.32	0.33	0.32
	Approach #3	3X	×	0.93	0.93	1.00	0.96	0.67	0.04	0.08
		9X	×	0.93	0.94	0.99	0.96	0.50	0.18	0.27
		3X	√	0.89	0.95	0.93	0.94	0.31	0.41	0.36
		9X	√	0.92	0.95	0.97	0.96	0.42	0.32	0.36
	×		√	0.71	0.96	0.71	0.82	0.14	0.63	0.24
	×		×	0.93	0.93	1.00	0.96	0.59	0.05	0.09

TF-IDF vectors (n gram word level)	Approach #1	3X	×	0.92	0.94	0.98	0.96	0.45	0.23	0.31
		9X	×	0.91	0.95	0.95	0.95	0.36	0.39	0.38
		3X	√	0.90	0.95	0.94	0.95	0.33	0.40	0.36
		9X	√	0.89	0.95	0.93	0.94	0.30	0.41	0.34
	Approach #2	3X	×	0.93	0.95	0.98	0.96	0.47	0.26	0.34
		9X	×	0.91	0.95	0.95	0.95	0.33	0.31	0.32
		3X	√	0.89	0.95	0.93	0.94	0.30	0.39	0.34
		9X	√	0.88	0.95	0.93	0.94	0.27	0.36	0.31
	Approach r#3	3X	×	0.93	0.94	0.98	0.96	0.48	0.25	0.32
		9X	×	0.91	0.95	0.95	0.95	0.38	0.41	0.39
		3X	√	0.90	0.95	0.94	0.95	0.34	0.41	0.37
		9X	√	0.90	0.95	0.93	0.94	0.33	0.42	0.37
	×		√	0.86	0.95	0.90	0.92	0.21	0.36	0.26
TF-IDF vectors (n gram character level)	×		×	0.93	0.93	1.00	0.96	0.69	0.03	0.05
	Approach #1	3X	×	0.93	0.93	0.99	0.96	0.48	0.06	0.11
		9X	×	0.93	0.93	0.99	0.96	0.34	0.05	0.08
		3X	√	0.93	0.94	0.98	0.96	0.51	0.22	0.31
		9X	√	0.93	0.93	0.99	0.96	0.43	0.07	0.12
	Approach #2	3X	×	0.93	0.93	0.99	0.96	0.38	0.06	0.10
		9X	×	0.93	0.93	0.99	0.96	0.38	0.08	0.13
		3X	√	0.93	0.95	0.98	0.96	0.46	0.26	0.34
		9X	√	0.92	0.93	0.99	0.96	0.40	0.10	0.16
	Approach #3	3X	×	0.93	0.94	0.99	0.96	0.57	0.14	0.23
		9X	×	0.93	0.94	0.99	0.96	0.51	0.15	0.23
		3X	√	0.92	0.95	0.97	0.96	0.46	0.33	0.39
		9X	√	0.93	0.94	0.99	0.96	0.48	0.18	0.26
	×		√	0.67	0.97	0.66	0.79	0.15	0.75	0.24

Table B.5: The performance results of the boosting model (XGBoost) with various types of feature engineering with original data, with three approaches to augmentation (see Table 8.3 above) using 3x and 9x (see Table 8.2 above) with and without undersampling and with undersampling without augmentation in the UNITE dataset.

Feature Engineering	Augmentation		Under	Acc	Non urgent 0			Urgent 1		
					P	R	F1	P	R	F1
Count vector	×		×	0.93	0.93	1.00	0.96	0.74	0.04	0.08
	Approach #1	3X	×	0.92	0.94	0.98	0.96	0.43	0.17	0.24
		9X	×	0.88	0.95	0.92	0.93	0.24	0.34	0.28
		3X	√	0.81	0.96	0.83	0.89	0.19	0.51	0.27
		9X	√	0.83	0.95	0.86	0.91	0.20	0.43	0.27
	Approach #2	3X	×	0.93	0.94	0.99	0.96	0.48	0.17	0.25
		9X	×	0.91	0.94	0.96	0.95	0.34	0.25	0.29
		3X	√	0.81	0.96	0.83	0.89	0.19	0.53	0.28
		9X	√	0.89	0.95	0.93	0.94	0.27	0.32	0.29
	Approach #3	3X	×	0.92	0.94	0.98	0.96	0.41	0.14	0.21
		9X	×	0.91	0.94	0.96	0.95	0.31	0.25	0.28
		3X	√	0.82	0.95	0.84	0.89	0.19	0.48	0.27
		9X	√	0.88	0.95	0.92	0.93	0.26	0.36	0.30
	×		√	0.72	0.96	0.73	0.83	0.14	0.58	0.23
TF-IDF vectors (word level)	×		×	0.93	0.93	1.00	0.96	0.67	0.04	0.08
	Approach #1	3X	×	0.93	0.94	0.98	0.96	0.45	0.17	0.25
		9X	×	0.87	0.95	0.91	0.93	0.23	0.33	0.27
		3X	√	0.80	0.96	0.83	0.89	0.18	0.51	0.27
		9X	√	0.83	0.95	0.86	0.91	0.20	0.45	0.28
	Approach #2	3X	×	0.93	0.94	0.99	0.96	0.45	0.15	0.23
		9X	×	0.91	0.95	0.96	0.95	0.37	0.27	0.31
		3X	√	0.79	0.95	0.82	0.88	0.16	0.47	0.24
		9X	√	0.89	0.95	0.93	0.94	0.28	0.34	0.31
	Approach #3	3X	×	0.93	0.94	0.99	0.96	0.50	0.16	0.25
		9X	×	0.91	0.94	0.96	0.95	0.32	0.26	0.29
		3X	√	0.81	0.95	0.84	0.89	0.18	0.48	0.27
		9X	√	0.88	0.95	0.91	0.93	0.25	0.38	0.30
	×		√	0.69	0.96	0.70	0.81	0.13	0.60	0.22
	×		×	0.93	0.93	1.00	0.96	0.62	0.04	0.07

TF-IDF vectors (n gram word level)	Approach#1	3X	×	0.93	0.94	0.99	0.96	0.51	0.14	0.22
		9X	×	0.92	0.95	0.96	0.96	0.40	0.31	0.35
		3X	√	0.85	0.95	0.88	0.92	0.20	0.38	0.26
		9X	√	0.70	0.95	0.71	0.81	0.12	0.53	0.20
	Approach #2	3X	×	0.93	0.94	0.99	0.96	0.54	0.15	0.23
		9X	×	0.92	0.94	0.97	0.96	0.40	0.26	0.32
		3X	√	0.77	0.95	0.80	0.87	0.14	0.42	0.21
		9X	√	0.66	0.95	0.67	0.79	0.11	0.55	0.19
	Approach #3	3X	×	0.93	0.94	0.99	0.96	0.51	0.15	0.23
		9X	×	0.90	0.95	0.95	0.95	0.32	0.30	0.31
		3X	√	0.85	0.95	0.89	0.92	0.20	0.38	0.27
		9X	√	0.87	0.95	0.91	0.93	0.25	0.40	0.31
	×		√	0.74	0.95	0.76	0.84	0.14	0.49	0.21
TF-IDF vectors (n gram character level)	×		×	0.93	0.94	0.99	0.97	0.65	0.15	0.24
	Approach #1	3X	×	0.93	0.94	0.99	0.97	0.61	0.20	0.31
		9X	×	0.93	0.94	0.99	0.96	0.53	0.19	0.28
		3X	√	0.92	0.95	0.96	0.96	0.44	0.39	0.42
		9X	√	0.93	0.94	0.98	0.96	0.53	0.24	0.33
	Approach #2	3X	×	0.93	0.94	0.99	0.97	0.61	0.21	0.31
		9X	×	0.93	0.94	0.99	0.96	0.53	0.18	0.26
		3X	√	0.91	0.95	0.95	0.95	0.39	0.40	0.39
		9X	√	0.93	0.94	0.98	0.96	0.47	0.23	0.31
	Approach #3	3X	×	0.93	0.95	0.98	0.96	0.56	0.29	0.39
		9X	×	0.93	0.95	0.98	0.96	0.49	0.29	0.37
		3X	√	0.91	0.96	0.94	0.95	0.38	0.51	0.44
		9X	√	0.92	0.95	0.97	0.96	0.46	0.36	0.40
	×		√	0.77	0.97	0.78	0.86	0.18	0.65	0.29

BIBLIOGRAPHY

Abebe, R., Hill, S., Vaughan, J. W., Small, P. M. and Schwartz, H. A. 'Using search queries to understand health information needs in africa'. *Proceedings of the International AAAI Conference on Web and Social Media*, 3-14.

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

Adadi, A. and Berrada, M. (2018) 'Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)', *IEEE access,* 6, pp. 52138-52160.

Agarwal, B. and Mittal, N. 'Text classification using machine learning methods-a survey'. *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*: Springer, 701-709.

Agrawal, A. and Paepcke, A. (2019) *The Stanford MOOCPosts Data Set*. Available at: https://datastage.stanford.edu/StanfordMoocPosts/ (Accessed: 18/3/2019).

Agrawal, A., Venkatraman, J., Leonard, S. and Paepcke, A. 'YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips'. *the 8th Intl. Conference on Educational Data Mining*.

Ahmadaliev, D. K., Medatov, A. A., Jo'rayev, M. M. and O'rinov, N. T. (2019) 'Adaptive educational hypermedia systems: an overview of current trend of adaptive content representation and sequencing', *Theoretical & Applied Science,(3)*, pp. 58-61.

Alamri, A., Alshehri, M., Cristea, A., Pereira, F. D., Oliveira, E., Shi, L. and Stewart, C. 'Predicting MOOCs dropout using only two easily obtainable features from the first week's activities'. *International Conference on Intelligent Tutoring Systems*: Springer, 163-173.

Alamri, A., Sun, Z., Cristea, A. I., Senthilnathan, G., Shi, L. and Stewart, C. 'Is MOOC learning different for dropouts? A visually-driven, multi-granularity explanatory ML approach'. *Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Athens, Greece, June 8–12, 2020, Proceedings 16*: Springer, 353-363.

Alamri, A., Sun, Z., Cristea, A. I., Stewart, C. and Pereira, F. D. 'MOOC next week dropout prediction: weekly assessing time and learning patterns'. *International Conference on Intelligent Tutoring Systems*: Springer, 119-130.

Albawi, S., Mohammed, T. A. and Al-Zawi, S. 'Understanding of a convolutional neural network'. *2017 international conference on engineering and technology (ICET)*: Ieee, 1-6.

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. (2022) *Learner Profiling: Demographics Identification Based on NLP, Machine Learning, and MOOCs Metadata.* Durham University.

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.

Almatrafi, O. and Johri, A. (2022) 'Improving MOOCs Using Feedback/Information from Discussion Forums: An Opinion Summarization and Suggestion Mining Approach', *IEEE Access*.

Almatrafi, O., Johri, A. and Rangwala, H. (2018) 'Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums', *Computers & Education,* 118, pp. 1-9.

Alrajhi, L., Alamri, A. and Cristea, A. I. (2022) 'Intervention Prediction in MOOCs Based on Learners' Comments: A Temporal Multi-input Approach Using Deep Learning and Transformer Models'.

Alrajhi, L., Alamri, A., Pereira, F. D. and Cristea, A. I. (2021) 'Urgency Analysis of Learners' Comments: An Automated Intervention Priority Model for MOOC'.

Alrajhi, L., Alharbi, K. and Cristea, A. I. 'A Multidimensional Deep Learner Model of Urgent Instructor Intervention Need in MOOC Forum Posts'. *International Conference on Intelligent Tutoring Systems*: Springer, 226-236.

Alrajhi, L. and Cristea, A. I. 'Plug & Play with Deep Neural Networks: Classifying Posts that Need Urgent Intervention in MOOCs'. *International Conference on Intelligent Tutoring Systems*: Springer, 651-666.

Alrajhi, L., Pereira, F. D., Cristea, A. I. and Aljohani, T. (2022) 'A Good Classifier is Not Enough: A XAI Approach for Urgent Instructor-Intervention Models in MOOCs'.

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,* 31, pp. 215-233.

Alshehri, M. and Cristea, A. I. (2022) 'MOOCs Paid Certification Prediction Using Students Discussion Forums'.

Alzetta, C., Adorni, G., Celik, I., Koceva, F. and Torre, I. 'Toward a user-adapted question/answering educational approach'. *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, 173-177.

Amarasinghe, I. and Hernandez-Leo, D. (2019) 'Adaptive Orchestration of Scripted Collaborative Learning in MOOCs'.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N. and Zwerdling, N. 'Do not have enough data? Deep learning to the rescue!'. *Proceedings of the AAAI Conference on Artificial Intelligence*, 7383-7390.

Arguello, J. and Shaffer, K. 'Predicting speech acts in MOOC forum posts'. *Ninth International AAAI Conference on Web and Social Media*.

Asmussen, C. B. and Møller, C. (2019) 'Smart literature review: a practical topic modelling approach to exploratory literature review', *Journal of Big Data,* 6(1), pp. 1-18.

Atapattu, T. and Falkner, K. 'A framework for topic generation and labeling from MOOC discussions'. *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, 201-204.

Atapattu T, F. K. T. H. (2016) 'Topic-wise classification of MOOC discussions: A visual analytics approach'.

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.

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

Bashar, A. (2019) 'Survey on evolving deep learning neural network architectures', *Journal of Artificial Intelligence,* 1(02), pp. 73-82.

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

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003) 'A Neural Probabilistic Language Model', *Journal of Machine Learning Research,* 3, pp. 1137-1155.

Bennetot, A., Donadello, I., Qadi, A. E., Dragoni, M., Frossard, T., Wagner, B., Saranti, A., Tulli, S., Trocan, M. and Chatila, R. (2021) 'A Practical Tutorial on Explainable AI Techniques', *arXiv preprint arXiv:2111.14260*.

Berrar, D. (2019) 'Cross-validation', *Encyclopedia of bioinformatics and computational biology,* 1, pp. 542-545.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) 'Latent dirichlet allocation', *the Journal of machine Learning research,* 3, pp. 993-1022.

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) 'Enriching word vectors with subword information', *Transactions of the association for computational linguistics,* 5, pp. 135-146.

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.

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'. *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 6*: Springer, 121-130.

Borrella, I., Caballero-Caballero, S. and Ponce-Cueto, E. 'Predict and intervene: Addressing the dropout problem in a MOOC-based program'. *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, 1-9.

Borrella, I., Caballero-Caballero, S. and Ponce-Cueto, E. (2022) 'Taking action to reduce dropout in MOOCs: Tested interventions'.

Bozkurt, A., Akgün-Özbek, E. and Zawacki-Richter, O. (2017) 'Trends and patterns in massive open online courses: Review and content analysis of research on MOOCs (2008-2015)', *International Review of Research in Open and Distributed Learning: IRRODL,* 18(5), pp. 118-147.

Brahimi, T. and Sarirete, A. (2015) 'Learning outside the classroom through MOOCs', *Computers in Human Behavior,* 51, pp. 604-609.

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

Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z. and Wong, F. M. F. (2014) 'Learning about social learning in MOOCs: From statistical analysis to generative model', *IEEE transactions on Learning Technologies,* 7(4), pp. 346-359.

Brodersen, K. H., Ong, C. S., Stephan, K. E. and Buhmann, J. M. 'The balanced accuracy and its posterior distribution'. *2010 20th international conference on pattern recognition*: IEEE, 3121-3124.

Captum (2021) *Captum - Model Interpretability for PyTorch*. Available at: https://captum.ai/docs/introduction.html.

Captum (2022) *Captum_BERT*. Available at: https://colab.research.google.com/drive/1pgAbzUF2SzF0BdFtGpJbZPWUOhFxT2NZ.

Capuano, N. and Caballé, S. 'Multi-attribute categorization of MOOC forum posts and applications to conversational agents'. *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*: Springer, 505-514.

Capuano, N., Caballé, S., Conesa, J. and Greco, A. (2021) 'Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis', *Journal of Ambient Intelligence and Humanized Computing,* 12(11), pp. 9977-9989.

Chandrasekaran, M., Ragupathi, K., Kan, M.-Y. and Tan, B. (2015a) 'Towards feasible instructor intervention in MOOC discussion forums'.

Chandrasekaran, M. K., Epp, C. D., Kan, M.-Y. and Litman, D. J. 'Using discourse signals for robust instructor intervention prediction'. *Thirty-First AAAI Conference on Artificial Intelligence*.

Chandrasekaran, M. K. and Kan, M.-Y. 'Countering Position Bias in Instructor Interventions in MOOC Discussion Forums'. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 135-142.

Chandrasekaran, M. K., Kan, M.-Y., Tan, B. C. and Ragupathi, K. (2015b) 'Learning instructor intervention from mooc forums: Early results and issues', *arXiv preprint arXiv:1504.07206*.

Chaplot, D. S., Rhim, E. and Kim, J. 'Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks'. *AIED Workshops*, 54-57.

Chaturvedi, S., Goldwasser, D. and Daumé III, H. 'Predicting instructor's intervention in MOOC forums'. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1501-1511.

Chen, K. Z. and Yeh, H. H. (2021) 'Acting in secret: Interaction, knowledge construction and sequential discussion patterns of partial role-assignment in a MOOC'.

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.

Chen, Y., Gao, Q., Yuan, Q. and Tang, Y. (2019) 'Facilitating students' interaction in MOOCs through timeline-anchored discussion', *International Journal of Human–Computer Interaction,* 35(19), pp. 1781-1799.

Chitsaz, M., Vigentini, L. and Clayphan, A. (2016) 'Toward the development of a dynamic dashboard for FutureLearn MOOCs: insights and directions', *Show Me The Learning. Proceedings ASCILITE*, pp. 116-121.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) 'Learning phrase representations using RNN encoder-decoder for statistical machine translation', *arXiv preprint arXiv:1406.1078*.

Chua, S.-M., Tagg, C., Sharples, M. and Rienties, B. (2017) 'Discussion analytics: Identifying conversations and social learners in FutureLearn MOOCs', *MOOC analytics: live dashboards, post-hoc analytics and the long-term effects*, pp. 36-62.

Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P. and Scott, A. M. (2020) 'A full systematic review was completed in 2 weeks using automation tools: a case study', *Journal of clinical epidemiology,* 121, pp. 81-90.

Clark, K., Khandelwal, U., Levy, O. and Manning, C. D. (2019) 'What does bert look at? an analysis of bert's attention', *arXiv preprint arXiv:1906.04341*.

Cobos, R. and Ruiz-Garcia, J. C. (2021) 'Improving learner engagement in MOOCs using a learning intervention system: A research study in engineering education', *Computer Applications in Engineering Education,* 29(4), pp. 733-749.

Coffrin, C., Corrin, L., De Barba, P. and Kennedy, G. 'Visualizing patterns of student engagement and performance in MOOCs'. *Proceedings of the fourth international conference on learning analytics and knowledge*, 83-92.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y. and Cohen, I. (2009) 'Pearson correlation coefficient', *Noise reduction in speech processing*, pp. 1-4.

Cokluk, O. (2010) 'Logistic Regression: Concept and Application', *Educational Sciences: Theory and Practice,* 10(3), pp. 1397-1407.

Confalonieri, R., Coba, L., Wagner, B. and Besold, T. R. (2021) 'A historical perspective of explainable Artificial Intelligence', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 11(1), pp. e1391.

Conneau, A., Schwenk, H., Barrault, L. and Lecun, Y. (2016) 'Very deep convolutional networks for text classification', *arXiv preprint arXiv:1606.01781*.

Coulombe, C. (2018) 'Text data augmentation made simple by leveraging NLP cloud APIs', *arXiv preprint arXiv:1812.04718*.

Cristea, A. I., Alamri, A., Kayama, M., Stewart, C., Alsheri, M. and Shi, L. 'Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses'. Association for Information Systems.

Crossley, S., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T. and Bergner, Y. (2015) 'Language to Completion: Success in an Educational Data Mining Massive Open Online Class', *International Educational Data Mining Society*.

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S. and Baker, R. S. 'Combining click-stream data with NLP tools to better understand MOOC completion'. *Proceedings of the sixth international conference on learning analytics & knowledge*: ACM, 6-14.

Curiskis, S. A., Drake, B., Osborn, T. R. and Kennedy, P. J. (2020) 'An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit', *Information Processing & Management,* 57(2), pp. 102034.

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

Das, S. (2012) *Increasing instructor visibility in online courses through mini-videos and screencasting.*: Faculty Focus Special Report.

Dawei, W., Alfred, R., Obit, J. H. and On, C. K. (2021) 'A literature review on text classification and sentiment analysis approaches', *Computational Science and Technology: 7th ICCST 2020, Pattaya, Thailand, 29–30 August, 2020*, pp. 305-323.

De Notaris, D. 'Reskilling higher education professionals: Skills and workflow in the making of a MOOC'. *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 6*: Springer, 146-155.

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*.

Dong, H., Wang, W., Huang, K. and Coenen, F. (2020) 'Automated social text annotation with joint multilabel attention networks', *IEEE Transactions on Neural Networks and Learning Systems,* 32(5), pp. 2224-2238.

Došilović, F. K., Brčić, M. and Hlupić, N. 'Explainable artificial intelligence: A survey'. *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*: IEEE, 0210-0215.

Downes, S. (2008) 'Places to go: Connectivism & connective knowledge', *Innovate: Journal of Online Education,* 5(1), pp. 6.

Drobot, I.-A. (2023) 'FutureLearn and Coursera: Communication on Two MOOC Platforms', *Massive Open Online Courses-Current Practice and Future Trends*: IntechOpen.

Elman, J. L. (1990) 'Finding structure in time', *Cognitive science,* 14(2), pp. 179-211.

Ezen-Can, A., Boyer, K. E., Kellogg, S. and Booth, S. 'Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach'. *Proceedings of the fifth international conference on learning analytics and knowledge*, 146-150.

Farias, F., Ludermir, T. and Bastos-Filho, C. (2020) 'Similarity Based Stratified Splitting: an approach to train better classifiers', *arXiv preprint arXiv:2010.06099*.

Ferguson, R. (2012) 'Learning analytics: drivers, developments and challenges', *International Journal of Technology Enhanced Learning,* 4(5/6), pp. 304-317.

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.

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.

Fonseca, S. C., Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Carvalho, L. S. and Cristea, A. I. (2020) 'Automatic subject-based contextualisation of programming assignment lists', *International Educational Data Mining Society*.

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

Garousi, V., Bauer, S. and Felderer, M. (2019) 'NLP-assisted software testing: a systematic review'.

Geng, S., Niu, B., Feng, Y. and Huang, M. (2020) 'Understanding the focal points and sentiment of learners in MOOC reviews: A machine learning and SC-LIWC-based approach', *British Journal of Educational Technology,* 51(5), pp. 1785-1803.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. and Kagal, L. 'Explaining explanations: An overview of interpretability of machine learning'. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*: IEEE, 80-89.

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*.

Gong, M. (2021) 'A novel performance measure for machine learning classification', *International Journal of Managing Information Technology (IJMIT) Vol,* 13.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G. and Cai, J. (2018) 'Recent advances in convolutional neural networks', *Pattern recognition,* 77, pp. 354-377.

Guo, S. X., Sun, X., Wang, S. X., Gao, Y. and Feng, J. (2019) 'Attention-Based Character-Word Hybrid Neural Networks with semantic and structural information for identifying of urgent posts in MOOC discussion forums', *IEEE Access,* 7, pp. 120522-120532.

Gütl, C., Rizzardini, R. H., Chang, V. and Morales, M. 'Attrition in MOOC: Lessons learned from drop-out students'. *Learning Technology for Education in Cloud. MOOC and Big Data: Third International Workshop, LTEC 2014, Santiago, Chile, September 2-5, 2014. Proceedings 3*: Springer, 37-48.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. and Bing, G. (2017) 'Learning from class-imbalanced data: Review of methods and applications', *Expert Systems with Applications,* 73, pp. 220-239.

Haniya, S. (2019) 'Developing an Intervention to Advance Learning At Scale'.

Harris, J. D., Quatman, C. E., Manring, M., Siston, R. A. and Flanigan, D. C. (2014) 'How to write a systematic review', *The American journal of sports medicine,* 42(11), pp. 2761-2768.

Hayes, A. F. and Krippendorff, K. (2007) 'Answering the call for a standard reliability measure for coding data', *Communication methods and measures,* 1(1), pp. 77-89.

He, J. Z., Bailey, J., Rubinstein, B. I. P. and Zhang, R. (2015) 'Identifying At-Risk Students in Massive Open Online Courses'.

Herman, R. L. (2012) 'Letter from the Editor-in-Chief: the MOOCs are coming', *The Journal of Effective Teaching,* 12(2), pp. 1-3.

Hew, K. F. (2016) 'Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS', *British Journal of Educational Technology,* 47(2), pp. 320-341.

Hirschberg, J. and Manning, C. D. (2015) 'Advances in natural language processing', *Science,* 349(6245), pp. 261-266.

Hochreiter, S., urgen Schmidhuber, J. and Elvezia, C. (1997) 'LONG SHORT-TERM MEMORY', *Neural Computation,* 9(8), pp. 1735-1780.

Hodgson, R., Cristea, A., Shi, L. and Graham, J. 'Wide-scale automatic analysis of 20 years of ITS research'. *International Conference on Intelligent Tutoring Systems*: Springer, 8-21.

Hollands, F. M. and Tirthali, D. (2014) 'MOOCs: Expectations and reality', *Center for Benefit-Cost Studies of Education, Teachers College, Columbia University,* 138.

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

Hossin, M. and Sulaiman, M. N. (2015) 'A review on evaluation metrics for data classification evaluations', *International journal of data mining & knowledge management process,* 5(2), pp. 1.

Hu, Y., Mello, R. F. and Gašević, D. (2021) 'Automatic analysis of cognitive presence in online discussions: An approach using deep learning and explainable artificial intelligence', *Computers and Education: Artificial Intelligence,* 2, pp. 100037.

Huang, Y. and Wang, R. 'Analysis of Influencing Factors of MOOC Learners' Loyalty Based on Online Review Text Mining'. *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*: IEEE, 693-698.

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*.

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

Ipaye, B. and Ipaye, C. B. (2013) 'Opportunities and Challenges for Open Educational Resources and Massive Open Online Courses: The Case of Nigeria', *Commonwealth of Learning*.

Itani, A., Brisson, L. and Garlatti, S. (2018) 'Understanding Learner's Drop-Out in MOOCs'.

Jacobi, C., Van Atteveldt, W. and Welbers, K. (2016) 'Quantitative analysis of large amounts of journalistic texts using topic modelling', *Digital journalism,* 4(1), pp. 89-106.

Jansen, D. and Schuwer, R. (2015) 'Institutional MOOC strategies in Europe', *Status Report Based on a Mapping Survey Conducted in October-December 2014*, pp. 4.

Jarnac de Freitas, M. and Mira da Silva, M. (2020) 'Systematic literature review about gamification in MOOCs', *Open Learning: The Journal of Open, Distance and e-Learning*, pp. 1-23.

Jiang, L., Zhang, H. and Cai, Z. (2008) 'A novel Bayes model: Hidden naive Bayes', *IEEE Transactions on knowledge and data engineering,* 21(10), pp. 1361-1371.

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. and Goshtasbpour, F. (2022) 'JIME Virtual Special Collection–2012 to 2022: The Decade of the MOOC'.

Joseph, M. R. (2020) 'ROLE OF MOOCs IN MODERN EDUCATION', *Journal of Applied Science And Research,* 8(2), pp. 13-17.

Jungiewicz, M. and Smywiński-Pohl, A. 'Data Augmentation for Sentiment Analysis in English–The Online Approach'. *International Conference on Artificial Neural Networks*: Springer, 584-595.

Kalyanathaya, K. P., Akila, D. and Rajesh, P. (2019) 'Advances in natural language processing–a survey of current research trends, development tools and industry applications', *International Journal of Recent Technology and Engineering,* 7(5C), pp. 199-202.

Kamath, C. N., Bukhari, S. S. and Dengel, A. 'Comparative study between traditional machine learning and deep learning approaches for text classification'. *Proceedings of the ACM Symposium on Document Engineering 2018*, 1-11.

Kennedy, J. (2014) 'Characteristics of massive open online courses (MOOCs): A research review, 2009-2012', *Journal of Interactive Online Learning,* 13(1).

Kesim, M. and Altınpulluk, H. (2015) 'A theoretical analysis of MOOCs types from a perspective of learning theories', *Procedia-Social and Behavioral Sciences,* 186, pp. 15-19.

Khalil, M. and Ebner, M. (2017) 'Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories', *Journal of computing in higher education,* 29(1), pp. 114-132.

Khodeir, N. A. (2021) 'Bi-GRU urgent classification for MOOC discussion forums based on BERT', *IEEE Access,* 9, pp. 58243-58255.

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.

Kloft, M., Stiehler, F., Zheng, Z. and Pinkwart, N. 'Predicting MOOC dropout over weeks using machine learning methods'. *Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs*, 60-65.

Klusener, M. and Fortenbacher, A. (2015) 'Predicting Students' Success Based on Forum Activities in MOOCs'.

Kobayashi, S. (2018) 'Contextual augmentation: Data augmentation by words with paradigmatic relations', *arXiv preprint arXiv:1805.06201*.

Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S. and Robnik-Šikonja, M. 'BERT meets shapley: Extending SHAP explanations to transformer-based classifiers'. *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 16-21.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C. and Yan, S. (2020) 'Captum: A unified and generic model interpretability library for pytorch', *arXiv preprint arXiv:2009.07896*.

Koné, M., May, M., Iksal, S. and Oumtanaga, S. 'A Collective Dynamic Indicator for Discussion Forums in Learning Management Systems'. *Computer Supported Education: 11th International Conference, CSEDU 2019, Heraklion, Crete, Greece, May 2-4, 2019, Revised Selected Papers 11*: Springer, 88-110.

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S. and Barnes, L. E. 'Hdltex: Hierarchical deep learning for text classification'. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*: IEEE, 364-371.

Kraus, S., Breier, M. and Dasí-Rodríguez, S. (2020) 'The art of crafting a systematic literature review in entrepreneurship research', *International Entrepreneurship and Management Journal,* 16(3), pp. 1023-1042.

Kumar, A., Dikshit, S. and Albuquerque, V. H. C. (2021) 'Explainable artificial intelligence for sarcasm detection in dialogues', *Wireless Communications and Mobile Computing,* 2021, pp. 1-13.

Kurtz, G., Kopolovich, O., Segev, E., Sahar-Inbar, L., Gal, L. and Hammer, R. (2022) 'Impact of an Instructor's Personalized Email Intervention on Completion Rates in a Massive Open Online Course (MOOC)'.

Lai, S., Xu, L., Liu, K. and Zhao, J. 'Recurrent convolutional neural networks for text classification'. *Twenty-ninth AAAI conference on artificial intelligence*.

Lallé, S. and Conati, C. 'A data-driven student model to provide adaptive support during video watching across MOOCs'. *International Conference on Artificial Intelligence in Education*: Springer, 282-295.

Lambert, S. R. (2020) 'Do MOOCs contribute to student equity and social inclusion? A systematic review 2014–18', *Computers & Education,* 145, pp. 103693.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE,* 86(11), pp. 2278-2324.

Lee, D., Watson, S. L. and Watson, W. R. (2019) 'Systematic literature review on self-regulated learning in massive open online courses', *Australasian Journal of Educational Technology,* 35(1).

Levy, J., Vattikonda, N., Haudenschild, C., Christensen, B. and Vaickus, L. (2022) 'Comparison of machine-learning algorithms for the prediction of current procedural terminology (CPT) codes from pathology reports', *Journal of Pathology Informatics,* 13, pp. 100165.

Li, S., Ao, X., Pan, F. and He, Q. (2022) 'Learning policy scheduling for text augmentation', *Neural Networks,* 145, pp. 121-127.

Liang, D., Xu, W. and Zhao, Y. 'Combining word-level and character-level representations for relation classification of informal text'. *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 43-47.

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.

Lipton, Z. C. (2018) 'The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery', *Queue,* 16(3), pp. 31-57.

Litman, D. 'Natural language processing for enhancing teaching and learning'. *Thirtieth AAAI Conference on Artificial Intelligence*.

Liu, P., Wang, X., Xiang, C. and Meng, W. 'A survey of text data augmentation'. *2020 International Conference on Computer Communication and Network Security (CCNS)*: IEEE, 191-195.

Liu, Z., Liu, S., Liu, L., Sun, J., Peng, X. and Wang, T. (2016) 'Sentiment recognition of online course reviews using multi-swarm optimization-based selected features', *Neurocomputing,* 185, pp. 11-20.

Liyanagunawardena, T. R., Adams, A. A. and Williams, S. A. (2013) 'MOOCs: A systematic study of the published literature 2008-2012', *International Review of Research in Open and Distributed Learning,* 14(3), pp. 202-227.

Loria, S. (2018) 'textblob Documentation', *Release 0.15,* 2(8), pp. 269.

Lundberg, S. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', *arXiv preprint arXiv:1705.07874*.

Macina, J., Srba, I., Williams, J. J. and Bielikova, M. 'Educational question routing in online student communities'. *Proceedings of the Eleventh ACM Conference on Recommender Systems*: ACM, 47-55.

Madabushi, H. T., Kochkina, E. and Castelle, M. (2020) 'Cost-sensitive BERT for generalisable sentence classification with imbalanced data', *arXiv preprint arXiv:2003.11563*.

Mahesh, B. (2020) 'Machine learning algorithms-a review', *International Journal of Science and Research (IJSR).[Internet],* 9, pp. 381-386.

Makcedward (2020) *makcedward/nlpaug*. Available at: https://github.com/makcedward/nlpaug.

Malliga, P. (2013) 'A survey on MOOC providers for higher education', *International Journal of Management & Information Technology,* 7(1), pp. 962-967.

Massimiani, M., Lacko, L. A., Swanson, C. S. B., Salvi, S., Argueta, L. B., Moresi, S., Ferrazzani, S., Gelber, S. E., Baergen, R. N. and Toschi, N. (2019) 'Increased circulating levels of Epidermal Growth Factor-like Domain 7 in pregnant women affected by preeclampsia', *Translational Research,* 207, pp. 19-29.

Mazari, A. C., Boudoukhani, N. and Djeffal, A. (2023) 'BERT-based ensemble learning for multi-aspect hate speech detection', *Cluster Computing*, pp. 1-15.

Mazzolini, M. and Maddison, S. (2007) 'When to jump in: The role of the instructor in online discussion forums', *Computers & education,* 49(2), pp. 193-213.

McAuley, A., Stewart, B., Siemens, G. and Cormier, D. (2010) 'The MOOC model for digital practice'.

McNemar, Q. (1947) 'Note on the sampling error of the difference between correlated proportions or percentages', *Psychometrika,* 12(2), pp. 153-157.

Meet, R. K. and Kala, D. (2021) 'Trends and Future Prospects in MOOC Researches: A Systematic Literature Review 2013-2020', *Contemporary Educational Technology,* 13(3).

Mehrabi, M., Safarpour, A. R. and Keshtkar, A. (2022) 'Massive open online courses (MOOCs) dropout rate in the world: a protocol for systematic review and meta-analysis', *Interdisciplinary Journal of Virtual Learning in Medical Sciences,* 13(2), pp. 85-92.

Meier, Y., Xu, J., Atan, O. and van der Schaar, M. (2015) 'Personalized Grade Prediction: A Data Mining Approach'.

Mikolov, T., Le, Q. V. and Sutskever, I. (2013) 'Exploiting similarities among languages for machine translation', *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. 'Distributed representations of words and phrases and their compositionality'. *Advances in neural information processing systems*, 3111-3119.

Min, W. N. S. W. and Zulkarnain, N. Z. (2020) 'Comparative Evaluation of Lexicons in Performing Sentiment Analysis', *JACTA,* 2(1), pp. 14-20.

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.

Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L. and Ghazi Saeedi, M. (2021) 'Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods', *Interactive Learning Environments,* 29(1), pp. 142-163.

Mrhar, K., Douimi, O. and Abik, M. (2021) 'A Dropout Predictor System in MOOCs Based on Neural Networks', *Journal of Automation, Mobile Robotics and Intelligent Systems*, pp. 72-80.

Nadkarni, P. M., Ohno-Machado, L. and Chapman, W. W. (2011) 'Natural language processing: an introduction', *Journal of the American Medical Informatics Association,* 18(5), pp. 544-551.

Najmani, K., Benlahmar, E. H., Sael, N. and Zellou, A. (2022) 'A Systematic Literature Review on Recommender Systems for MOOCs', *Ingénierie des Systèmes d'Information,* 27(6).

Nanda, G., Douglas, K. A., Waller, D. R., Merzdorf, H. E. and Goldwasser, D. (2021) 'Analyzing Large Collections of Open-Ended Feedback from MOOC Learners Using LDA Topic Modeling and Qualitative Analysis', *IEEE Transactions on Learning Technologies*.

Naqvi, S. R., Ullah, Z., Taqvi, S. A. A., Khan, M. N. A., Farooq, W., Mehran, M. T., Juchelková, D. and Štěpanec, L. (2023) 'Applications of machine learning in thermochemical conversion of biomass-A review', *Fuel,* 332, pp. 126055.

Ni Ki, C., Hosseinian-Far, A., Daneshkhah, A. and Salari, N. (2021) 'Topic modelling in precision medicine with its applications in personalized diabetes management', *Expert Systems*, pp. e12774.

Nikolaev, I., Botov, D., Dmitrin, Y., Klenin, J. and Melnikov, A. 'Use of Topic Modelling for Improvement of Quality in the Task of Semantic Search of Educational Courses'. *21st International Workshop on Computer Science and Information Technologies (CSIT 2019)*: Atlantis Press, 104-111.

Nori, H., Jenkins, S., Koch, P. and Caruana, R. (2019) 'Interpretml: A unified framework for machine learning interpretability', *arXiv preprint arXiv:1909.09223*.

North, M. A. 'A method for implementing a statistically significant number of data classes in the Jenks algorithm'. *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*: IEEE, 35-38.

Ntourmas, A., Avouris, N., Daskalaki, S. and Dimitriadis, Y. (2018) 'Teaching assistants' interventions in online courses: a comparative study of two massive open online courses'.

Ntourmas, A., Avouris, N., Daskalaki, S. and Dimitriadis, Y. (2019) 'Evaluation of a Massive Online Course Forum: Design Issues and Their Impact on Learners' Support'.

Ntourmas, A., Daskalaki, S., Dimitriadis, Y. and Avouris, N. (2021) 'Classifying MOOC forum posts using corpora semantic similarities: a study on transferability across different courses', *Neural Computing and Applications*, pp. 1-15.

Ntourmas, A., Dimitriadis, Y., Daskalaki, S. and Avouris, N. (2022) 'Assessing Learner Facilitation in MOOC Forums: A Mixed-Methods Evaluation Study'.

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.

Onah, D. F. and Pang, E. L. 'MOOC design principles: topic modelling-PyLDavis visualization & summarisation of learners' engagement'. 13th annual International Conference on Education and New Learning Technologies.

Onah, D. F., Sinclair, J. and Boyatt, R. (2014a) 'Dropout rates of massive open online courses: behavioural patterns', *EDULEARN14 proceedings*, pp. 5825-5834.

Onah, D. F., Sinclair, J. E. and Boyatt, R. 'Exploring the use of MOOC discussion forums'. *Proceedings of London International Conference on Education*, 1-4.

Osman, A. I. A., Ahmed, A. N., Chow, M. F., Huang, Y. F. and El-Shafie, A. (2021) 'Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia', *Ain Shams Engineering Journal,* 12(2), pp. 1545-1556.

Otter, D. W., Medina, J. R. and Kalita, J. K. (2020) 'A survey of the usages of deep learning for natural language processing', *IEEE transactions on neural networks and learning systems,* 32(2), pp. 604-624.

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.

Pappano, L. (2012) 'The Year of the MOOC', *The New York Times,* 2(12), pp. 2012.

Park, J.-H. and Choi, H. J. (2009) 'Factors influencing adult learners' decision to drop out or persist in online learning', *Journal of Educational Technology & Society,* 12(4), pp. 207-217.

Pati, D. and Lorusso, L. N. (2018) 'How to write a systematic review of the literature', *HERD: Health Environments Research & Design Journal,* 11(1), pp. 15-30.

Paton, R. M., Fluck, A. E. and Scanlan, J. D. (2018) 'Engagement and retention in VET MOOCs and online courses: A systematic review of literature from 2013 to 2017', *Computers & Education,* 125, pp. 191-201.

Pennington, J., Socher, R. and Manning, C. 'Glove: Global vectors for word representation'. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Pereira, F. D., Pires, F., Fonseca, S. C., Oliveira, E. H., Carvalho, L. S., Oliveira, D. B. and Cristea, A. I. 'Towards a Human-AI Hybrid System for Categorising Programming Problems'. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 94-100.

Prabhakar Kaila, D. and Prasad, D. A. (2020) 'Informational flow on Twitter–Corona virus outbreak–topic modelling approach', *International Journal of Advanced Research in Engineering and Technology (IJARET),* 11(3).

Pranckutė, R. (2021) 'Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world', *Publications,* 9(1), pp. 12.

Prenkaj, B., Velardi, P., Stilo, G., Distante, D. and Faralli, S. (2020) 'A survey of machine learning approaches for student dropout prediction in online courses', *ACM Computing Surveys (CSUR),* 53(3), pp. 1-34.

Qiu, S., Xu, B., Zhang, J., Wang, Y., Shen, X., de Melo, G., Long, C. and Li, X. 'EasyAug: An automatic textual data augmentation platform for classification tasks'. *Companion Proceedings of the Web Conference 2020*, 249-252.

Raffaghelli, J. E., Cucchiara, S. and Persico, D. (2015) 'Methodological approaches in MOOC research: Retracing the myth of P roteus', *British Journal of Educational Technology,* 46(3), pp. 488-509.

Raghu, M. and Schmidt, E. (2020) 'A survey of deep learning for scientific discovery', *arXiv preprint arXiv:2003.11755*.

Ramesh, A., Goldwasser, D., Huang, B., Daume, H. and Getoor, L. (2020) 'Interpretable Engagement Models for MOOCs Using Hinge-Loss Markov Random Fields'.

Rani, S. and Kumar, P. (2019) 'Deep learning based sentiment analysis using convolution neural network', *Arabian Journal for Science and Engineering,* 44(4), pp. 3305-3314.

Reutemann, J. (2016) 'Differences and Commonalities–A comparative report of video styles and course descriptions on edX, Coursera, Futurelearn and Iversity', *European Stakeholders Summit on experiences and best practices in and around MOOCs*.

Ribeiro, M. T., Singh, S. and Guestrin, C. '" Why should i trust you?" Explaining the predictions of any classifier'. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135-1144.

Rivard, R. (2013) 'Measuring the MOOC dropout rate', *Inside Higher Ed,* 8, pp. 2013.

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.

Robinson, A. C. (2015) 'Exploring class discussions from a massive open online course (MOOC) on cartography', *Modern Trends in Cartography*: Springer, pp. 173-182.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C. and Gehlbach, H. (2016) *Forecasting student achievement in MOOCs with natural language processing.*

Rogers, A., Kovaleva, O. and Rumshisky, A. (2021) 'A primer in BERTology: What we know about how BERT works', *Transactions of the Association for Computational Linguistics,* 8, pp. 842-866.

Rose, C. and Siemens, G. 'Shared task on prediction of dropout over time in massively open online courses'. *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, 39-41.

Rossi, D., Stroele, V., Braga, R., Caballe, S., Capuano, N., Campos, F., Dantas, M., Lomasto, L. and Toti, D. (2022) 'CAERS: A Conversational Agent for Intervention in MOOCs' Learning Processes'.

Rossi, L. A. (2023) *Coursera Forums*. Available at: http://github.com/elleros/courseraforums (Accessed: 02/06 2023).

Rossi, L. A. and Gnawali, O. 'Language independent analysis and classification of discussion threads in Coursera MOOC forums'. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*: IEEE, 654-661.

Rousseeuw, P. J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics,* 20, pp. 53-65.

Rudin, C. (2019) 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence,* 1(5), pp. 206-215.

Sallam, M. H., Martín-Monje, E. and Li, Y. (2022) 'Research trends in language MOOC studies: a systematic review of the published literature (2012-2018)', *Computer Assisted Language Learning,* 35(4), pp. 764-791.

Sanchez-Gordon, S. and Luján-Mora, S. (2018) 'Research challenges in accessible MOOCs: A systematic literature review 2008–2016', *Universal Access in the Information Society,* 17(4), pp. 775-789.

Schuster, M. and Paliwal, K. K. (1997) 'Bidirectional recurrent neural networks', *IEEE Transactions on Signal Processing,* 45(11), pp. 2673-2681.

Schutzberg, A. (2019) 'Using Tough Love to Promote Active Learning'.

Sciarrone, F. and Temperini, M. (2019) 'Simulating Peer Assessment in Massive Open On-line Courses'.

Scpus: Tips and Tricks. Available at: https://blog.scopus.com/tips-and-tricks (Accessed: 24/11/2021 2021).

Shah, D. (2016) *Monetization Over Massiveness: Breaking Down MOOCs by the Numbers in 2016*. Available at: https://www.edsurge.com/news/2016-12-29-monetization-over-massiveness-breaking-down-moocs-by-the-numbers-in-2016#:~:text=In%202016%2C%202%2C600%20new%20courses,to%20MOOC%20companies%20in%202016. (Accessed: 18/07/2023).

Shah, D. (2021) *By The Numbers: MOOCs in 2021*. Available at: https://www.classcentral.com/report/mooc-stats-2021/ (Accessed: 24/7/2023).

Sharma, H. and Sharma, A. K. (2017) 'Study and analysis of topic modelling methods and tools–A survey', *American Journal of Mathematical and Computer Modelling,* 2(3), pp. 84-87.

Sharmin, S. and Zaman, Z. 'Spam detection in social media employing machine learning tool for text mining'. *2017 13th international conference on signal-image technology & internet-based systems (SITIS)*: IEEE, 137-142.

Shimabukuro, J. (2016) 'What's Wrong with MOOCs: One-Size-Fits-All Syndrome' (Accessed 29/06/2022.

Shorten, C., Khoshgoftaar, T. M. and Furht, B. (2021) 'Text data augmentation for deep learning', *Journal of big Data,* 8(1), pp. 1-34.

Sinha T, C. J. (2015) 'Connecting the dots: Predicting student grade sequences from bursty MOOC interactions over time'.

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.

Slade, S. and Prinsloo, P. (2013) 'Learning analytics: Ethical issues and dilemmas', *American Behavioral Scientist,* 57(10), pp. 1510-1529.

Smaili, E. M., Khoudda, C., Sraidi, S., Azzouzi, S. and Charaf, M. E. H. (2022) 'An Innovative Approach to Prevent Learners' Dropout from MOOCs using Optimal Personalized Learning Paths: An Online Learning Case Study', *Statistics, Optimization & Information Computing,* 10(1), pp. 45-58.

Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. Y. 'Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks'. *Proceedings of the conference on empirical methods in natural language processing*: Association for Computational Linguistics, 254-263.

Soni, V. D. (2020) 'Global Impact of E-learning during COVID 19', *Available at SSRN 3630073*.

Springer Link Search Tips. Available at: https://link.springer.com/searchhelp (Accessed: 24/11/2021 2021).

Stracke, C. M. and Bozkurt, A. 'Evolution of MOOC designs, providers and learners and the related MOOC research and publications from 2008 to 2018'. *Proceedings of International Open & Distance Learning Conference (IODL19)*, 13-20.

Stump, G. S., DeBoer, J., Whittinghill, J. and Breslow, L. 'Development of a framework to classify MOOC discussion forum posts: Methodology and challenges'. *NIPS Workshop on Data Driven Education*, 1-20.

Sun, D., Li, T., You, F., Hu, M. and Li, Z. 'Prediction of learning behavior characters of MOOC's data based on time series analysis'. *Journal of Physics: Conference Series*: IOP Publishing, 012009.

Sun, X., Guo, S., Gao, Y., Zhang, J., Xiao, X. and Feng, J. 'Identification of urgent posts in MOOC discussion forums using an improved RCNN'. *2019 IEEE World Conference on Engineering Education (EDUNINE)*: IEEE, 1-5.

Sunar, A. S., White, S., Abdullah, N. A. and Davis, H. C. (2016) 'How learners' interactions sustain engagement: a MOOC case study', *IEEE Transactions on Learning Technologies,* 10(4), pp. 475-487.

Sundararajan, M., Taly, A. and Yan, Q. 'Axiomatic attribution for deep networks'. *International conference on machine learning*: PMLR, 3319-3328.

Syed, S. and Spruit, M. 'Full-text or abstract? examining topic coherence scores using latent dirichlet allocation'. *2017 IEEE International conference on data science and advanced analytics (DSAA)*: IEEE, 165-174.

Szczepański, M., Pawlicki, M., Kozik, R. and Choraś, M. (2021) 'New explainability method for BERT-based model in fake news detection', *Scientific reports,* 11(1), pp. 1-13.

Troyano, J. A., Carrillo, V., Enríquez, F. and Galán, F. J. 'Named entity recognition through corpus transformation and system combination'. *International Conference on Natural Language Processing (in Spain)*: Springer, 255-266.

Tsironi, E., Barros, P., Weber, C. and Wermter, S. (2017) 'An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition', *Neurocomputing,* 268, pp. 76-86.

Tucker, C., Pursel, B. K. and Divinsky, A. (2014) 'Mining student-generated textual data in MOOCs and quantifying their effects on student performance and learning outcomes', *The ASEE Computers in Education (CoED) Journal,* 5(4), pp. 84.

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M. and Baik, S. W. (2017) 'Action recognition in video sequences using deep bi-directional LSTM with CNN features', *IEEE Access,* 6, pp. 1155-1166.

Using FutureLearn (2020). Available at: https://www.futurelearn.com/using-futurelearn.

Van de Oudeweetering, K. and Agirdag, O. (2018) 'MOOCS as accelerators of social mobility? A systematic review', *Journal of Educational Technology & Society,* 21(1), pp. 1-11.

Van der Maaten, L. and Hinton, G. (2008) 'Visualizing data using t-SNE', *Journal of machine learning research,* 9(11).

Van Houdt, G., Mosquera, C. and Nápoles, G. (2020) 'A review on the long short-term memory model', *Artificial Intelligence Review,* 53(8), pp. 5929-5955.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. 'Attention is all you need'. *Advances in neural information processing systems*, 5998-6008.

Veletsianos, G. and Shepherdson, P. (2015) 'Who studies MOOCs? Interdisciplinarity in MOOC research and its changes over time', *International Review of Research in Open and Distributed Learning,* 16(3), pp. 1-17.

Veletsianos, G. and Shepherdson, P. (2016) 'A systematic analysis and synthesis of the empirical MOOC literature published in 2013–2015', *International Review of Research in Open and Distributed Learning,* 17(2), pp. 198-221.

Vigentini, L., León Urrutia, M. and Fields, B. 'FutureLearn data: what we currently have, what we are learning and how it is demonstrating learning in MOOCs'. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 512-513.

Von Eschenbach, W. J. (2021) 'Transparency and the black box problem: Why we do not trust AI', *Philosophy & Technology,* 34(4), pp. 1607-1622.

Voudoukis, N. and Pagiatakis, G. (2022) 'Massive Open Online Courses (MOOCs): Practices, Trends, and Challenges for the Higher Education', *European Journal of Education and Pedagogy,* 3(3), pp. 288-295.

Wang, W. Y. and Yang, D. 'That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets'. *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2557-2563.

Wang, X., Jiang, W. and Luo, Z. 'Combination of convolutional and recurrent neural network for sentiment analysis of short texts'. *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 2428-2437.

Wang, X., Wen, M. and Rosé, C. P. 'Towards triggering higher-order thinking behaviors in MOOCs'. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 398-407.

Web of Science Core Collection: Search Tips. Available at: https://clarivate.libguides.com/woscc/searchtips (Accessed: 24/11/2021 2021).

Wei, J. and Zou, K. (2019) 'Eda: Easy data augmentation techniques for boosting performance on text classification tasks', *arXiv preprint arXiv:1901.11196*.

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.

Wen, M., Yang, D. and Rosé, C. P. 'Linguistic reflections of student engagement in massive open online courses'. *Eighth International AAAI Conference on Weblogs and Social Media*.

Whang, S. E., Roh, Y., Song, H. and Lee, J.-G. (2023) 'Data collection and quality challenges in deep learning: A data-centric ai perspective', *The VLDB Journal*, pp. 1-23.

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*.

Wise, A. F., Cui, Y., Jin, W. and Vytasek, J. (2017) 'Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling', *The Internet and Higher Education,* 32, pp. 11-28.

Wise, A. F., Cui, Y. and Vytasek, J. 'Bringing order to chaos in MOOC discussion forums with content-related thread identification'. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*: ACM, 188-197.

Wong, A. W., Wong, K. and Hindle, A. (2019) 'Tracing forum posts to MOOC content using topic analysis', *arXiv preprint arXiv:1904.07307*.

Wong, J.-S. (2018) 'MessageLens: a visual analytics system to support multifaceted exploration of MOOC forum discussions', *Visual Informatics,* 2(1), pp. 37-49.

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.

Wu, B. (2021) 'Influence of MOOC learners discussion forum social interactions on online reviews of MOOC', *Education and Information Technologies,* 26(3), pp. 3483-3496.

Wulf, J., Blohm, I., Leimeister, J. M. and Brenner, W. (2014) 'Massive open online courses', *Business & Information Systems Engineering,* 6(2), pp. 111-114.

Xiang, R., Chersoni, E., Long, Y., Lu, Q. and Huang, C.-R. 'Lexical Data Augmentation for Text Classification in Deep Learning'. *Canadian Conference on Artificial Intelligence*: Springer, 521-527.

Xing, W. and Du, D. (2018) 'Dropout prediction in MOOCs: Using deep learning for personalized intervention', *Journal of Educational Computing Research*, pp. 0735633118757015.

Xing, W. L., Chen, X., Stein, J. and Marcinkowski, M. (2016) 'Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization'.

Xiong, W. and Litman, D. 'Evaluating topic-word review analysis for understanding student peer review performance'. *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*: University of Pittsburgh.

Yan, W., Dowell, N., Holman, C., Welsh, S. S., Choi, H. and Brooks, C. 'Exploring Learner Engagement Patterns in Teach-Outs Using Topic, Sentiment and On-topicness to Reflect on Pedagogy'. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*: ACM, 180-184.

Yang, B. (2022) 'Analysis and Visualization of While-Learning Social Help Seeking Aligned with Learning Content to Facilitate Data-informed Support in MOOCs'.

Yang, C. Y., Ren, W. Y. and Wu, F. T. (2022) 'Diagnose Topic Attention: What are the Preference and Demand with Different Learner Groups in MOOCs Discussion Forums'.

Yang, D., Wen, M., Howley, I., Kraut, R. and Rose, C. 'Exploring the effect of confusion in discussion forums of massive open online courses'. *Proceedings of the second (2015) ACM conference on learning@ scale*: ACM, 121-130.

Yang, T.-Y., Brinton, C. G., Joe-Wong, C. and Chiang, M. (2017) 'Behavior-based grade prediction for MOOCs via time series neural networks', *IEEE Journal of Selected Topics in Signal Processing,* 11(5), pp. 716-728.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. 'Hierarchical attention networks for document classification'. *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480-1489.

Yenigalla, P., Kar, S., Singh, C., Nagar, A. and Mathur, G. 'Addressing unseen word problem in text classification'. *International Conference on Applications of Natural Language to Information Systems*: Springer, 339-351.

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*.

Young, T., Hazarika, D., Poria, S. and Cambria, E. (2018) 'Recent trends in deep learning based natural language processing', *ieee Computational intelligenCe magazine,* 13(3), pp. 55-75.

Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M. and Le, Q. V. (2018) 'Qanet: Combining local convolution with global self-attention for reading comprehension', *arXiv preprint arXiv:1804.09541*.

Yu, J. L., Alrajhi, L., Harit, A., Sun, Z. T., Cristea, A. I. and Shi, L. (2021) 'Exploring Bayesian Deep Learning for Urgent Instructor Intervention Need in MOOC Forums'.

Yu, Z., Haghighat, F., Fung, B. C. and Yoshino, H. (2010) 'A decision tree method for building energy demand modeling', *Energy and Buildings,* 42(10), pp. 1637-1646.

Zhang, W., He, Y., Wang, L., Liu, S. and Meng, X. (2023) 'Landslide Susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing', *Geological Journal*.

Zhang, X., Zhao, J. and LeCun, Y. 'Character-level convolutional networks for text classification'. *Advances in neural information processing systems*, 649-657.

Zhang, Z., Robinson, D. and Tepper, J. 'Detecting hate speech on twitter using a convolution-gru based deep neural network'. *European Semantic Web Conference*: Springer, 745-760.

Zhao, C., Han, J. G. and Xu, X. 'CNN and RNN Based Neural Networks for Action Recognition'. *Journal of Physics: Conference Series*: IOP Publishing, 062013.

Zheng, S., Wisniewski, P., Rosson, M. B. and Carroll, J. M. 'Ask the instructors: Motivations and challenges of teaching massive open online courses'. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 206-221.

Zhu, M., Sari, A. and Lee, M. M. (2018) 'A systematic review of research methods and topics of the empirical MOOC literature (2014–2016)', *The Internet and Higher Education,* 37, pp. 31-39.

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.