Learner Profiling: Demographics Identification Based on NLP, Machine Learning, and MOOCs Metadata

TAHANI MUSLIH M ALJOHANI

Learner Profiling: Demographics Identification Based on NLP, Machine Learning, and MOOCs Metadata

Tahani M. M. Aljohani

A thesis presented for the degree of

Doctor of Philosophy in Computer Science

Supervised by:

Prof. Alexandra I. Cristea

Dr. Chris G. Willcocks

Artificial Intelligence and Human Systems Research Group

Department of Computer Science

Durham University in the United Kingdom

2022

Dedication

This thesis is dedicated to the sake of Allah the Almighty

"all praise is due to him"

Learner Profiling: Demographics Identification Based on NLP, Machine Learning, and MOOCs Metadata

Tahani M. M. Aljohani

Submitted for the degree of Doctor of Philosophy in Computer Science

Abstract

Massive Open Online Courses (MOOCs) have become universal learning resources, and the COVID-19 pandemic is rendering these platforms even more necessary. Many types of research are ongoing to improve the learning resources provided to learners via MOOCs. These platforms also bring an incredible diversity of learners in terms of their demographics; thus, much MOOCs research relies on the learners' demographics data. Traditionally, these data are extracted from pre-course questionnaires that are filled-in by the learners themselves. However, besides introducing potential cognitive overhead (asking learners to fulfil tasks outside of the main purpose of learning), this leads to a clear bias in any research based on these questionnaires. The latter is because only about 10% of the MOOCs learners provide (a given type of) demographics data (with the intersection of all types of demographic data being significantly below 10%), while others do not provide any type of their demographic data. Thus, the population data obtained via questionnaires is not representative of the actual population in the MOOCs. To resolve this issue, a research area called Learner Profiling (LP) is investigated in this thesis. This area naturally extends from a research area called Author Profiling (AP), which aims at identifying traits about authors in different domains. Instead, LP aims to identify learners' demographics in the online educational domain. This research specifically focused on identifying the employment status, gender, and academic level of learners in MOOCs. Classifying the employment status of learners was based on the semantic representation of their comments, and *comparing the sequential with the parallel ensemble deep learning architecture* (Convolutional Neural Networks and Recurrent Neural Networks). This obtained an average high accuracy of 96.3% for the best proposed method; *using NLP based approach* for balancing the training samples. Additionally, the task of classifying the gender of learners was tackled based on the syntactic knowledge from the learners' comments. Different tree-structured Long-Short-Term Memory models were compared and, as a result, *the researcher proposed a novel version of a bi-directional composition function* for existing architectures. In addition, 18 different combinations of word-level encoding and sentence-level encoding functions for this task were compared and evaluated. Based on the results, the novel bi-directional model outperforms all other models and the highest accuracy result among the proposed models is the one based on the combination of Feedforward Neural Network and the Stack-augmented Parser-Interpreter Neural Network (82.60% classification accuracy). Next, the learner's academic level was identified based on training small size - rich data - i.e. not only textual content (data including learner activity data). The researcher argues here that to classify a learner trait from the sparse textual content, researchers need to use additionally other features stemming from the MOOC platform, such as derived from learners' actions on that platform. Accordingly, *time stamps, quizzes, and discussions* were examined, as learners' behavioural data sources for the classification problem. This novel approach for the task achieves a high accuracy (89% on average), even with a simple classifier, irrespective of data size. To conclude, such classification models as used in this thesis show that they can achieve highly accurate results and that pre-course questionnaires to extract the demographic information with a high cognitive overhead could become obsolete.

Declaration

The work in this thesis is 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 degree or qualification, and it is all the author's own work unless referenced to the contrary in the text.

List of Publications

The works presented in this thesis have been published in journals or conferences. The relevant publications are listed below:

- (Chapter 5): Published a paper at the 4th International Conference on Information and Education Innovations. Durham, UK. (**Best Presentation Award**).
  - Aljohani, T., Cristea, A.I.: Predicting learners' demographics characteristics: Deep learning ensemble architecture for learners' characteristics prediction in MOOCs. In: Proceedings of the 2019 4th International Conference on Information and Education Innovations. pp. 23–27. ACM (2019).

- (Chapter 6): Published a poster at the ACM Women UK Inspire 2019. Canterbury, UK.
  - Aljohani, T., Yu, J., Alrajhi, L: Author-Profiling of Learners' Gender. ACM Women UK Inspire (2019).

- (Chapter 5): Published a paper at the Intelligent Tutoring Systems Conference. Athens, Greece.
  -Aljohani, T., Pereira, F.D., Cristea, A.I., Oliveira, E.: Prediction of users' professional profile in MOOCs only by utilising learners' written texts. In: International Conference on Intelligent Tutoring Systems. pp. 163–173. Springer (2020).

- (Chapter 6): Published a paper at the International Conference on Advanced Data Mining and Applications. Sydney, Australia. (**Best Paper and Presentation Award**).
  -Aljohani, T., Yu, J., Cristea, A.I.: Author profiling: Prediction of learners' gender on a MOOC platform based on learners' comments. International Journal of Computer and Information Engineering14(1), 29–36 (2020).

- (Chapter 5 and 6): Published a paper at Frontiers in Research Metrics and Analytics: Advanced Analytics and Decision Making for Research Policy and Strategic Management.
  -Aljohani, T., Cristea, A.I. : Learners demographics classification on MOOCs during the COVID-19: author profiling via deep learning based on semantic and syntactic representations. Frontiers in Research Metrics and Analytics, 34 (2021).

- (Chapter 7): Published a paper at Intelligent Tutoring Systems Conference. Athens, Greece.
  -Aljohani, T., Cristea, A.I. : Training Temporal and NLP Features via Extremely Randomised Trees for Educational Level Classification. In: International Conference on Intelligent Tutoring Systems. Springer, Cham (2021).

Forthcoming Publications

- (Chapter 6): Published a paper at International Conference on Artificial Intelligence in Education (AIED 2022). Accepted.

Acknowledgements

My deepest thanks go to my father, my biggest inspiration for conducting this research. Your advice regarding research and publications; based on your own background of graduate studies was invaluable. I have learned from you the importance of hard work, impartiality, and objectivity in all life matters.

My heartfelt thanks go to my mother. The beliefs and values that you had raised me with them helped me in my life journeys, including the PhD journey. You taught me how to commit to my responsibilities and I learned from you how to be a good mother; no matter what the conditions. You taught me how to be dedicated and empathetic, and show respect to others.

My greatest thanks go to my husband (Dr Thamer Alrefai) for surrounding me with respect, trust, and care. Your strong belief in me and what I can achieve helped me believe in myself. Without the sacrifices you made to support my dreams, especially during my PhD, I may never have achieved them. You are a grace from God for me.

My Special thanks go to my brothers (Dr Ahmad, Eng Khaled, Eng Majed, and Basel) and my sisters (Amani, Dr Asmaa, and Eng Alaa) for always motivating and pushing me forward toward my goals. Thanks for your continuous advice and support, and Thanks for being in good communication whenever I need. You have been beside me the whole time. Thanks to all of you.

For my soul, my daughters (Nour and Dalia), I am really proud of you. Being clever kids at school and behaving well to others, are significant achievements of me as a mother. You are a source of happiness for me. May God help me in raising you with great values.

My sincerest thanks go to my supervisor, Professor Alexandra I. Cristea, for her unlimited support in conducting this research. I thank you for the invaluable advice, feedback, and scientific guidance you have provided throughout. Your honest academic support makes this research gain a lot of your excellent knowledge and expertise in the domain. I would also like to thank my second supervisor, Dr Chris G. Willcocks, for his advice during this research and I value the insights and guidance you provide. Thanks also to all members of our research group for all the helpful discussions and advice; especially my co-authors: Filipe Dwan Pereira, Jialin Yu, and Laila Alrajhi.

I want to show my gratitude to everyone who has helped and encouraged me during my PhD journey; especially my neighbours and friends in the UK, who have provided help to me and my kids. These people are exceptional, I admire their humanity and their support for my responsibilities of being a mother and a PhD student.

Last but not least, I would like to thank the Ministry of Education in Saudi Arabia and the Saudi Cultural Bureau in the UK for their full funding of this research and for helping in facilitating this project. I also want to thank the Department of Computer Science at Durham University for having me and allowing me the use of their facilities which are first-rate.

وزارة التعليم
Ministry of Education

Contents

List of Figures

List of Tables

List of Acronyms

MOOCs Massive Open Online Courses

AP Author Profiling

LP Learners Profiling

LA Learner Analytics

NLP Natural Language Processing

DL Deep Learning

ML Machine Learning

AI Artificial Intelligence

ANN Artificial Neural Network

FFNN Feedforward Neural Network

CNN Convolutional Neural Network

RNN Recurrent Neural Network

LSTM Long-Short Term Memory

BiLSTM Bi-directional Long-Short Term Memory

RecNN Recursive Neural Network

TreeLSTM Tree Long-Short Term Memory

SPINN Stack-augmented Parser-Interpreter Neural Network

SATA Structure-Aware Tag Augmented

GloVe Global Vectors for Word Representation

PCA Principal Component Analysis

SMOTE Synthetic Minority Over-Sampling Technique

KNN K-Nearest Neighbour

SVM Support Vector Machine

NB Naïve Bayes

LR Logistic Regression

DT Desion Tree

RF Random Forest

ET Extra Tree

SA Sentiment Analysis

POS Part of Speech

TF-IDF Term Frequency Inverse Document Frequency

Chapter 1

Introduction

1.1 Prologue

This chapter offers an introduction to the topic of this thesis. First, the research scope is defined in Section 1.2. This helps to further define the research motivation in Section 1.3 and the research problem in Section 1.3.1. In Section 1.4, the research questions are listed. Then, the aim and objectives of this research are discussed in Section 1.5. This is followed by an outline of the research contributions in Section 1.6. Finally, the thesis structure is presented in Section 1.7.

1.2 Research Scope

After Salman Khan introduced the Khan Academy in 2006, which focused on providing online courses mainly consisting of mathematics content, substantial efforts followed in Khan's footsteps, aiming to offer online education. This resulted in the emergence of so-called Massive Open Online Courses (MOOCs) [218]. MOOC platforms, such as the platform founded in 2012 by Coursera[†] [26], offer a chance for students to engage in online learning in a variety of subjects. The courses are provided by universities, institutions, and even by leading companies such as Microsoft. Many educational researchers are interested in improving online educational platforms, including MOOC environments, to promote education, and most of the research in this area is

[†] Coursera is a global online learning platform that was among the first to be created and deployed (htps://about.coursera.org/).

classified under the umbrella of the new area of *Learning Analytics (LA)* [4]. Up to now, several studies have also assessed demographic data from users of online education and MOOC platforms in an attempt to better understand their reach [222].

Along with the advancement of data analytics tools, the availability of data from MOOC platforms opens up significant opportunities to investigate learner behaviours and trends. Together, these data and the tools to analyse them may enable improvements in online educational platforms and their design, as well as learning experiences and outcomes. An important problem in the case of MOOCs is that data represent unknown learners for which, in most cases, completed demographic data are unavailable [11].

On the other hand, a relatively recent research direction was proposed and promoted since 2007, namely *Author Profiling (AP)* [21]. AP is notable in that it promotes the use of automatic tools – developed based on state-of-the-art advances in computer science in the area of Natural Language Processing (NLP) and Machine Learning (ML) – to analyse online data for the purpose of identifying authors' demographics and characteristics [168].

Based on the research area of MOOC platforms, the big data relating to learners that they generate, and recent advances in computing systems for AP, this motivated a noteworthy research question for this thesis concerning the possibility of inferring MOOC learners' traits based on their online writing in an online educational discussion forum. Therefore, in pursuing this research question, this thesis combines the LA direction in MOOCs research, with the AP research direction, thereby a research area called *Learner Profiling (LP) are examined.*

1.3 Research Motivation

The available courses on MOOC platforms are sponsored by top-ranked educational institutions that provide diverse certified programs. This creates an incredible opportunity for low-income learners who aspire to receive an elite education. However, researchers have questioned whether MOOCs can provide a successful education for students [63]. The distinct characteristic of MOOCs is their flexibility, which arises because the recorded courses are accessible online at any time and any place, which facilitates near-universal access [11]. However, while MOOCs have become an important source of education, they still have limitations and require substantial improvements, particularly before they can provide a high level of personalised education. One such type of personalisation, which is of special interest for this thesis, relates to the learner-specific education opportunities on MOOCs based on demographic characteristics [63].

The area of AP is an important source of inspiration for this research. AP is used widely nowadays because of its applications in many domains, including marketing, forensics, and other fields [76]. AP studies show that AP, as a solution, is feasible and reliable because its fundamental idea is logical and reflected in real-world data: namely, that similar users – in terms of their demographic characteristics – tend to have similar patterns in their writing styles [14]. According to Jonathan Schler and his team [188], bloggers have different writing patterns based on their gender. For example, they found that females write more pronouns and negation words in their blogs than males who write more articles and prepositions. Hence, female users can be distinguished and detected based on such similarities in their writing patterns.

In addition, in online contexts, users tend to share more personal information than in other settings [105], which frequently provides an indication of their demographic characteristics. These considerations motivated this research, drawing on the concepts of the domain of AP, to apply the available techniques to online learners using their comments from forum discussions, as well as other behaviours on MOOC platforms in order to classify their demographics.

1.3.1 Research Problem

Due to today's fast-paced lifestyles, enrolling at a university or another educational institution necessitates commitment and sacrifice. As a result, MOOC-style online learning is expected by some to become the standard [106]. MOOCs have grown in popularity over the last decade as digitisation has increased rapidly. Along with this, many face-to-face courses suddenly stopped during the ongoing COVID-19 pandemic [230], and so the majority of new MOOC users this year are learners who are trying to find replacements for their suspended classes [193]. This situation makes MOOCs an optimal alternative as they offer remotely-accessible classes from the world's leading institutions [181], [6].

It is generally accepted that the perfect type of online education is a form of personalised education [77]. Therefore, it is relevant that demographic information is one of the main candidate factors for providing such a personalised system and improving learning outcomes [102]. The current methodology for obtaining such data from learners who use MOOCs is via surveys that are completed by the learners themselves. Although MOOC providers ask users to provide demographic information during registration, most learners seem unaware of its value to their learning and only approximately 10% of them fill it in, which is low the proportion of actual population in MOOCs, according to Almatrafi's study [11].

The main concern with the low 10% response rate from learners to self-reported demographic surveys on MOOCs is that this heightens the risk of obtaining unrepresentative outcomes. MOOC data derived from questionnaires with a low response rate are likely to underestimate the target population. As a result, the misrepresentation could skew estimates of demographic effects on variables of interest, including course completion rates or other learning outcomes [222]. Many studies in the field of LA that have investigated MOOCs have relied on demographic data as one of the main research parameters for their experiments, but to the best of the author's knowledge, most of these studies have used pre-course, open responses to identify learners' characteristics, which have also been utilised later for different research aims [67].

Due to a confluence of factors, therefore, research into online education platforms, specifically regarding MOOCs, has become increasingly important. This research targets the introduction of new approaches to provide reliable data – equivalent to the self-report surveys with learners – but far more representative and reflective of the actual population. Without relying on these often-incomplete surveys or other expensive solutions, this thesis seeks to introduce automatic methods to extract learners' demographic data via less costly, AP-driven solutions. In addition, this research considering gaps in AP area, which is explained in Chapter 3, Section 3.4.3.

1.4 Research Questions

The research questions were developed based on the research problem and the particular gaps observed in the literature (further described in Chapter 3). The research questions are stated as follows:

Umbrella Research Question: How can Author Profiling (AP) and Learning Analytics (LA) be combined in relation to MOOC platforms to perform Learner Profiling (LP)?

To address this umbrella research question, the researcher determined a number of specific learner profile demographics: gender, employment status, and level of education. The reasons for considering these demographics for this research were based on their importance in personalisation systems in the domain (further details are explained in Sections 5.2, 6.2, and 7.2). These represent targets of the classification models in this research.

The source of AP data are normally text, and so textual data from learners, which is represented by their discussions in the learning environment, as presented in Chapters 5 and 6. Additionally, the researcher further explored how adding other metadata can be used as other candidate predictors, which is presented in Chapter 7.

Hence, the main research questions are:

- **RQ1:** Is it possible to classify employment status based on the comments that learners exchange on MOOCs discussion forums?
- **RQ2:** Can advanced textual features extracted from MOOCs discussion forums be used to classify a learner's gender?
- **RQ3:** Can a learner's level of education be classified based on a MOOC discussion forum data on a course-level classification?
- **RQ4:** Can the use of metadata in addition to the MOOC discussion forum data improve the classification of a learner's level of education?

1.5 Research Objectives

To achieve the identified research questions, the following objectives were addressed in this thesis:

O1: To construct a new and large dataset using MOOCs, intended for later use and application with NLP and Machine Learning (ML) models (explained in Chapter 4, Section 4.2.1). This is important in addressing the main research questions of this thesis.

O2: To investigate ensemble deep learning models and examine them in the area of LP. This addresses the RQ1 above, and further explanation is given in Chapter 5.

O3: To solve the issue of unbalanced data using an NLP approach. This addresses RQ1 above (further explained in Chapter 5).

O4: To examine deep learning algorithms for gender profiling and introduce new learning architectures. This centres on RQ2 above (further explained in Chapter 6).

O5: To examine NLP and non-NLP approaches based on available MOOC metadata, considering them for learners' educational level classification in course-level data. This is intended to answer RQ3 and RQ4, as further explained in Chapter 7.

1.6 Research Contributions

This thesis proposes the field of Learner Profiling (LP) in new context, which is an extension of the AP area and a combination of LA and AP.

This thesis also contributes to the field of online education research, by creating a novel approach for combating the bias that may result from using incomplete or unrepresentative surveys. In particular, the research findings are expected to assist decision-making regarding learner-related issues of personalisation and recommendation systems in MOOC platforms.

Further contributions are as follows:

- Collected *new educational data from MOOCs* for the research, which are rich in size in terms of labelled samples (see Section 4.2.1).

- Investigated three approaches with separate classification goals: *learners' gender*, *learners' employment status*, and *learners' educational level*. The decision of focusing on these three demographics is mainly based on the importance and the emerge of them in MOOC studies, see Sections 5.2, 6.2, and 7.2.

- Implemented *novel methods*, described in Section 4.2.6, as follows:
  - Comparing sequential ensemble DL architecture with parallel ensemble DL architecture (based on Convolutional Neural Networks [CNN] and Recurrent Neural Networks [RNN]) in Chapter 5, to classify learners' employment status.
  - Using a new NLP-based approach to the area of demographics profiling for balancing samples during training models in Chapter 5, which is used also in Chapter 6.
  - Introducing a novel version of a bi-directional composition function for existing treeLSTM architectures, and comparing 18 different combinations of word-level encoding and sentence-level encoding functions (Chapter 6), to classify learners' gender.
  - Examining available MOOC-related metadata (timestamps, quizzes, and discussions), which introduced new features – not only NLP features – to the area of demographics profiling, to classify learners' level of education (Chapter 7).

1.7 Thesis Outline

The thesis is organised into nine chapters as follows:

- **Chapter 1:** This chapter introduces the research scope, motivation, and problem. From this, the research questions are derived. Then, the objectives necessary for carrying out this research are outlined. Finally, the contributions of the thesis are listed.
- **Chapter 2:** It gives an overview of the main technical concepts relevant to the research. The chapter starts with an introduction to ML methods: DL and conventional ML. It also provides an introduction to the stylometry features used in the field of AP.
- **Chapter 3:** It reviews current models in the literature used for AP-related tasks. The chapter starts with a description of the available public data, as well as presents a discussion of

works performed for AP. The chapter concludes with an investigation of relevant literature on MOOC demographics.

- **Chapter 4:** It describes the research methodology used to answer the research questions. The study's data sources and data collection procedures are explained. This is followed by an analysis of learners' texts based on primary analytical methods. After that, a general mathematical definition of the research aim and approach is provided. Then, the ethical considerations of conducting this research are discussed. Finally, the conceptual framework of the research is presented.

- **Chapter 5:** It gives a description of the experimental setting used for employment status classification in this thesis. First, the study is introduced with an overview. The study's data source is then described. Following this, the processes of data collection, analysis, and preparation are detailed. Next, a description of the methodology is provided. Finally, evaluation processes, including the obtained results, are discussed.

- **Chapter 6:** It gives a description of the experimental setting used for gender classification in this thesis. First, an overview of the study is presented. Next, data collection, analysis, and pre-processing are described. This is followed by an explanation of the utilised approaches and learning algorithms. Finally, the evaluation processes are discussed along with the obtained results.

- **Chapter 7:** It gives a description of the experimental setting used for education level classification in this thesis. First, an overview is presented. Then, data preparation and feature engineering are described. This is followed by an explanation of the applied approaches and learning algorithms. Finally, the evaluation process, in terms of model performance and feature importance, is discussed.

- **Chapter 8:** It provides an extensive discussion of the thesis and explains how it represents a valuable contribution to – and extension of – the prior literature. It covers the importance of MOOCs and the significance of LP as a new research area; makes overall observations in terms of data preparation, pre-processing, bias, and other considerations; and discusses

ML approaches and performance. After that, certain limitations are identified and future research areas are suggested.

- **Chapter 9:** It concludes the thesis by providing a summary of its key contributions and findings.

The next chapter presents the main technical concepts relevant to the research, including classification ML approaches DL and conventional ML used in this research, as well as the stylometry features used in the field of AP.

Chapter 2

Classification Approaches

2.1 Prologue

Machine Learning (ML) is a sub-field of Artificial Intelligence (AI). It studies algorithms that have the ability to learn from samples (i.e., data) to address specific problems without using specific programmed instructions [82]. The learning behaviours differ between ML algorithms, which means that the choice of an algorithm depends on the nature of the problem addressed.

In this research, the targeted problem falls under the umbrella of supervised ML. Supervised learning methods use two types of algorithms: conventional ML algorithms and Deep Learning (DL) algorithms. DL models are currently the state of the art in Natural Language Processing (NLP), and they are used in Chapters 5 and 6 of this thesis. In addition, Transformers, which represent a new generation of DL models that are being applied in NLP tasks, are used in Chapter 7. This research also investigates conventional ML approaches in Chapter 7. The samples used for model training in the AP area are mainly based on textual features known as stylometry features [105]. However, in Chapter 7 this concept is expanded into the LP area by the addition of non-textual features for training the models.

The aim of this chapter is to provide a *theoretical background* of the methods used to conduct this research. This is achieved by providing an overview of the main technical concepts that are relevant to the research. In doing so, this chapter aims to establish the context that is necessary to accomplish the thesis' objectives (presented in Chapters 5, 6, and 7). This chapter starts by providing an introduction to Supervised Learning and further discusses specific concepts such as DL, conventional ML, stylometry features.

2.2 Supervised Learning

Typically, ML algorithms are classified based on three approaches: unsupervised, semi-supervised, and supervised. The choice of approach mainly depends on the nature of the problem and the number of labelled samples available in a dataset [13]. For instance, if no labelled samples are available, then there is no other option other than to use unsupervised learning (i.e., clustering). If the dataset contains a small portion of labelled samples, with a large proportion being unlabelled data, then semi-supervised learning becomes the method of choice. Also, if there is a large proportion of labelled data, then supervised learning is the preferred method. It is reported that supervised approaches provide more reliable results than unsupervised and semi-supervised approaches [13]. Hence, due to its competitive performance when significant quantities of labelled samples are available, supervised learning is a widely used approach [82].

In the following sections, the ML algorithms that are used in this thesis are discussed. Abstract descriptions of the models and how they learn from the data are given.

2.3 Deep Learning

2.3.1 Artificial Neural Networks

Before discussing the DL models used in this research, it is crucial to understand their architecture. Hence, this section provides an understanding of the architecture of DL models. DL models consist of what are known as Artificial Neural Networks (ANN). ANN are a subset of ML with a complex structure inspired by the human nervous system. The human brain is believed to contain around 100 billion neurons that are connected by electrochemical connections called ‘synapses’. As a result, an ANN is made up of a huge number of artificial neurons that are organized in layers. The neurons are ‘fully connected’, which means that each layer’s nodes are linked to those in the adjacent layers. Each connection has a weight parameter that represents the link in the biological synapse. Finding appropriate weight values that characterise a problem is referred to as training for a specific task [82].

The training step in a single neuron can be expressed mathematically as follows:

$$S = \sum_{i=1}^{n} w_i x_i \tag{2.1}$$

A neuron's input is a weighted sum S obtained by multiplying each output from the previous layer's neurons $x_1, x_2, ...x_n$ with its respective weight $w_1, w_2, ...w_n$ and summing the results. Next, the weighted sum S is calculated using a nonlinear function (activation function) to get the final output. Figure 2.1 presents a bird's-eye view of a single neuron's training process.

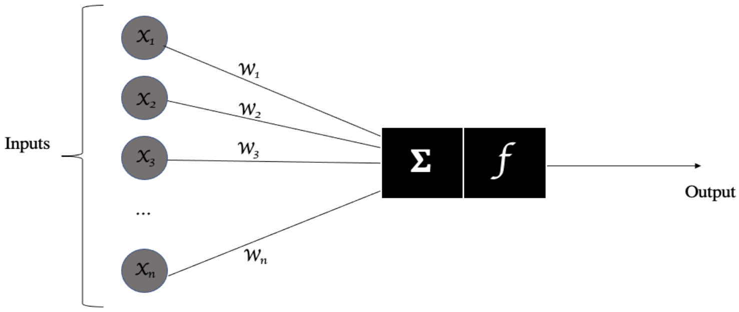


Figure 2.1: Training process for a single artificial neuron

In a classification task, the output (final) layer employs a softmax activation function, also known as multinomial logistic regression, which is used to generate a categorical probability distribution between the task's classes. When a neural network receives an input x from the training set, it generates an output $\widehat{y}$ that differs from the desired output y in most cases. A loss function is then used to determine the gradient error value between y and $\widehat{y}$. Also, cross-entropy is used because the network in the classification task needs to generate the output that is a distribution of probability values.

Each layer's neurons are fully connected, and each connection has its own weight. Thus, as the network develops in size, the number of parameters (such as the weights) also increases rapidly. Large numbers of parameters elevate the risk of the network becoming overly complex and losing its generalisation ability to new inputs; this is known as overfitting. However, different methods known as regularisation techniques are a viable way to overcome this difficulty. One simple technique, for example, is called dropout regularisation [205].

Training in ANN mainly consists of two phases: forward pass and backward pass. The aim of the forward pass is to generate the output $\widehat{y}$, which is completed through layer-by-layer propagation of the input from the input layer to the output layer. By contrast, the aim of the backward pass is to propagate the gradient values obtained by the loss function backwards from the output layer, providing each neuron with a gradient value that roughly corresponds to its contribution to the output. Upon building a neural network, a large number of variables (parameters) must be chosen, including the number of hidden neurons, learning rate, epoch, and batch size. Some of them have a high level of significance in terms of enhancing the training process [30].

In the next subsections, the DL techniques used in this thesis are introduced, namely Convolutional Neural Networks (CNN) and Long Short-term Memory (LSTM) networks. Both of these DL techniques are used in Chapter 5. In addition, the tree LSTM-based models used in this thesis are introduced in Chapter 6 and the transformers in Chapter 7.

2.3.2 Convolutional Neural Networks

CNNs is an ANN model that consists of a number of layers that include neurons with their biases and weights. However, the difference is that the neurons in CNNs take more consideration of the *spatial structure* of features [3]. CNNs have played a major role in making significant advances in the area of computer vision. In recent years, CNN have become the state-of-the-art area and are a method of choice to solve almost all detection or recognition problems [82].

The basic structure of a CNN consists of three layers: a convolutional layer, pooling layer, and fully connected layer. In the convolutional layer, the filters convert the volume of data into feature maps. These feature maps are then processed in the pooling layer, which reduces the parameters. The output features are then processed into the fully connected layer. Figure 2.2 presents these layers in a standard structure, which contains one layer from each of the three layers.

As a complement to image classification approaches where matrices containing pixels as inputs

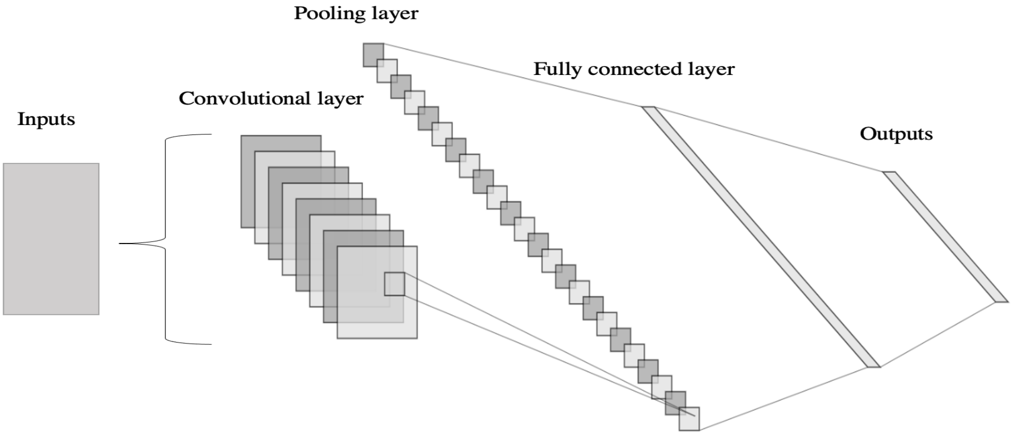


Figure 2.2: Structure of a standard CNN

are processed, the input matrix for text classification contains token vector representations (often as a word or character). Hence, the input vector for text classification is a one-dimensional representation (1D), which is unlike the case with image data, where they have a two or more dimensional representation for each input [235]. In 1D convolution, where input is sequential (e.g., text or audio data), a long vector (array) convolves into a shorter vector. This transformation is accomplished with the help of predetermined parameters in a CNN such as filters and strides [55].

A convolution ci applies a non-linear function f as follows:

$$ci = f(\Sigma_{j,k} w_{j,k} (x_{[i:i+h1]})_{j,k} + b) \tag{2.2}$$

where i is the current input vector, j is a position in the convolution kernel/filter k, and h is the number of words in spans (size of the convolution). In addition, b is a bias term, w is a weight, x is the current word embedding, and $[i : i]$ represents a sub-matrix of x ([3], [49]). The CNN architecture and parameters used in this thesis are further explained in Chapter 5.

2.3.3 Recurrent Neural Networks

Recurrent Neural Network (RNN) is an ANN model that differs from the standard ANN in the way of handling inputs. In an ANN, once an input has been processed, the state of the current ANN is lost. This means that the ability to process a sequence of inputs and extract information from this sequence is limited. By contrast, in an RNN, the values (outputs) of the current state in the ANN are processed and concatenated with the input of the next step of the sequence; this occurs in each step, enabling all states in the network to remain active throughout the sequence and, thus, acting as a *memory*. As languages employ a sequence of tokens (words or characters) that are used to build a sentence, this sort of learning behaviour is important to consider for solving many tasks in the area of NLP. RNN-based models have accomplished outstanding performance for sequential data [235].

Another key distinction between ANN and RNN is backpropagation. Backpropagation refers to how a backward pass updates a weight. Instead of calculating each weight's gradients once, they are calculated multiple times – once for each time step – and then summed. However, this introduces an issue known as the *vanishing gradient* problem; as the number of time steps in the sequence becomes sufficiently larger, the gradient value is propagated to an earlier state, which progressively vanishes (i.e., becomes smaller). Thus, a more advanced RNN architecture was designed to eliminate the vanishing gradient problem [197], explained in the next section.

2.3.3.1 Long Short-Term Memory

The principle of Long Short-Term Memory (LSTM) is attributed to RNN. In RNN, each time step uses only one recurrent output. However, in LSTM, a new component called a cell is introduced, which produces a second recurrent output. The cell includes four gates with various functions, allowing the network to control what information to send or forget to the next time step. This sophisticated structure leads to improved performance for long-term dependencies such as lengthy sentences [123].

The functions of the four gates in an LSTM cell are as follows: the forget gate specifies what information should be forgotten in the current cell state; the input gate decides what information from the input should be added to the new cell state, and this is done by two sub-gates: the first determines which values to update, while the second calculates a new cell state value that needs to be added to the current values. Finally, there is an output gate that defines which values to output [89].

The following formulas briefly describe the memories/gates that are inside the hidden unit of an

LSTM, which help the model to remember the term information:

$$
\begin{aligned}
f_t &= \sigma(w_f x_t + u_f h_{t-1} + b_f) \\
i_t &= \sigma(w_i x_t + u_i h_{t-1} + b_i) \\
o_t &= \sigma(w_o x_t + u_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ tanh(w_c x_t + u_c h_{t-1} + b_c) \\
h_t &= o_t \circ tanh(c_t)
\end{aligned}
\tag{2.3}
$$

where t is the time step, h is the hidden state, f_t is the 'forget gate', i_t is the input gate, c_t is the cell state, u is the weighted metric, b is the bias term, w is the weight term in these functions, σ is the sigmoid function, and o is the Hadamard product [239] , [49]. Figure 2.3 shows the LSTM gates in a single cell.

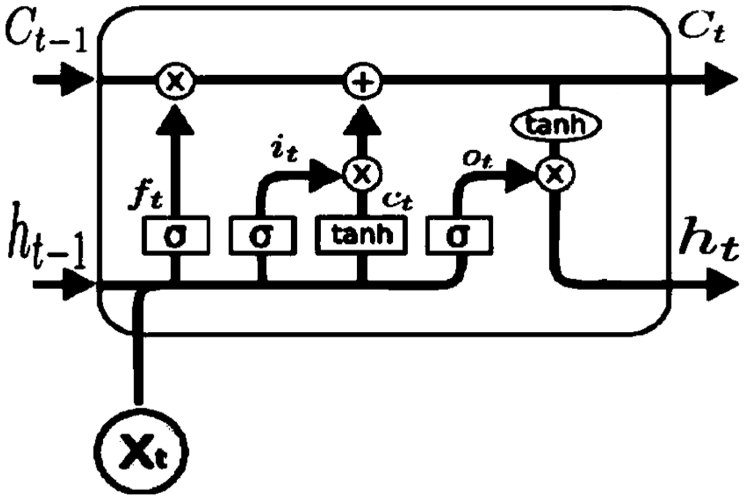


Figure 2.3: Illustration of a single LSTM cell

An extended version of LSTM, called Bidirectional Long Short-Term Memory (BiLSTM) [189], is used in this thesis in Chapters 5 and 6. The distinguishing aspect of BiLSTM is that it duplicates the input layer in a reverse direction. This is useful in NLP when seeking to understand sentences comprehensively in two directions: from left to right and vice versa.

More details on how LSTM and CNN are applied in this thesis to meet the research objectives, specifically to classify the employment statuses of MOOC learners, is presented in Chapter 5.

2.3.4 Recursive Neural Networks

Recursive Neural Networks (RecNN) are used in NLP to represent a sentence in a tree-like, recursive structure. This structure can be learned during training or given by a parse tree [111]. A RecNN model converts an input word to a vector, which is a leaf node. The node's pairs are then composed into phrase pairs by a composition function. This is called an intermediate representation of a tree. Lastly, the root node is considered to be the representation of the whole sentence.

Although pure RecNN, which only reach the closest constituent parts within a sentence, are more effective in terms of getting the meaning composition of the sentence, they still reach a limited space of information of the sentence and do not give the whole picture of the sentence's meaning. Thus, tree-sequential is useful to process human-wise when reading a sentence from left to right. This can provide the whole picture based on the current steps in the tracking vector of the tree-sequential. Also, adding a transition process during sentence encoding in a dynamical way helps to improve sentence understanding. During the sequential process of word sequences, the model will have the current status of a word that summarizes the whole left context, and through this summary, some of the information is lost that enables disambiguation before reaching the last word of the sentence. This is because these tree-structured models start with the constituents of a sentence that has its merged words [34].

To produce better tree-structured models, previous NLP researches have examined sequential models. These have been used to achieve state-of-the-art results. Researchers have extended sequential models and then compared their performance [213]. LSTM, the most powerful neural network architecture in NLP, due to its superiority in memorising long length sequences, has also been proven effective in the form of an advanced model that is a new, expanded version (namely, TreeLSTM).

2.3.4.1 Tree LSTM

Tree LSTM belongs to the family of RecNN and is inspired by the original LSTM. Vanilla Tree LSTM is the first neural network model with the ability to pass tree-structured information over sequences [213]. The hidden state of the original LSTM is composed of a current input at a current time step and a previous hidden state of an LSTM unit in the previous time step. However, the hidden state of the tree LSTM is composed of a current input vector and the hidden states of two child units (in the case of a binary tree). This is illustrated in Figure 2.4.

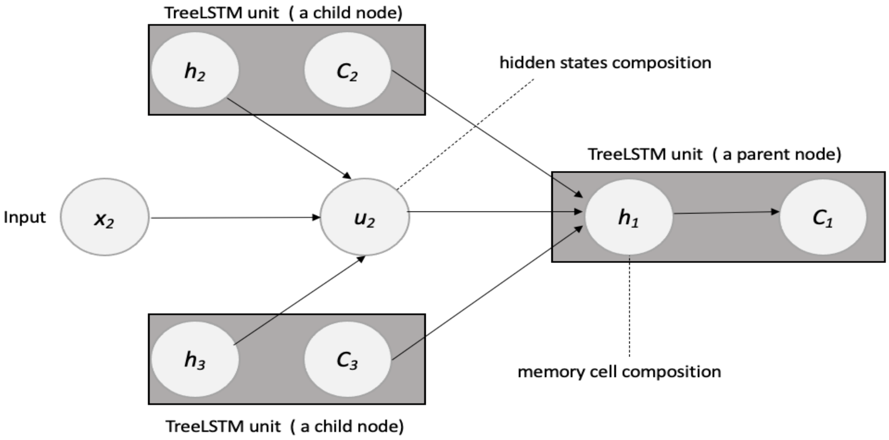

TreeLSTM unit (a child node)
h2
C2
hidden states composition
TreeLSTM unit (a parent node)
Input
x2
u2
h1
C1
h3
C3
memory cell composition
TreeLSTM unit (a child node)

Figure 2.4: Composition of memory cell and hidden state of a Tree LSTM unit with two child nodes

More details on how this type of algorithm is used to meet the objectives of the thesis's gender classification task based on learner data from MOOCs are provided in Chapter 6.

2.3.5 Transformers

More and more innovative DL models are now regularly introduced. As a result, it is becoming difficult to keep up with everything that's new in the area. However, a very recent neural network model in particular, known as a Transformer [225], [231], which was introduced in early 2019, has been very successful for common NLP tasks.

Most state-of-the-art NLP systems rely on RNN-based models such as LSTM, but this situation has changed with the advent of Transformers. Transformers are underpinned by an encoder-decoder structure [43] plus an attention mechanism [225]. The encoder is made up of encoding layers that process the input one by one, while the decoder is made up of decoding layers that handle the encoder's output in the same way. Each encoder/decoder layer uses an attention mechanism to weigh the importance of each item of the input data individually. Transformers handle input data sequentially, like RNN, but they do not always handle the input data in order, which is dissimilar to

RNN. This is because, in a Transformer, there is no need to process the beginning of the sentence before the end; instead, the Transformer recognizes the context that provides each word with its meaning in the text. The architecture of a Transformer is complex to understand, but Figure 2.5 presents a descriptive graph of its architecture and how its components perform.

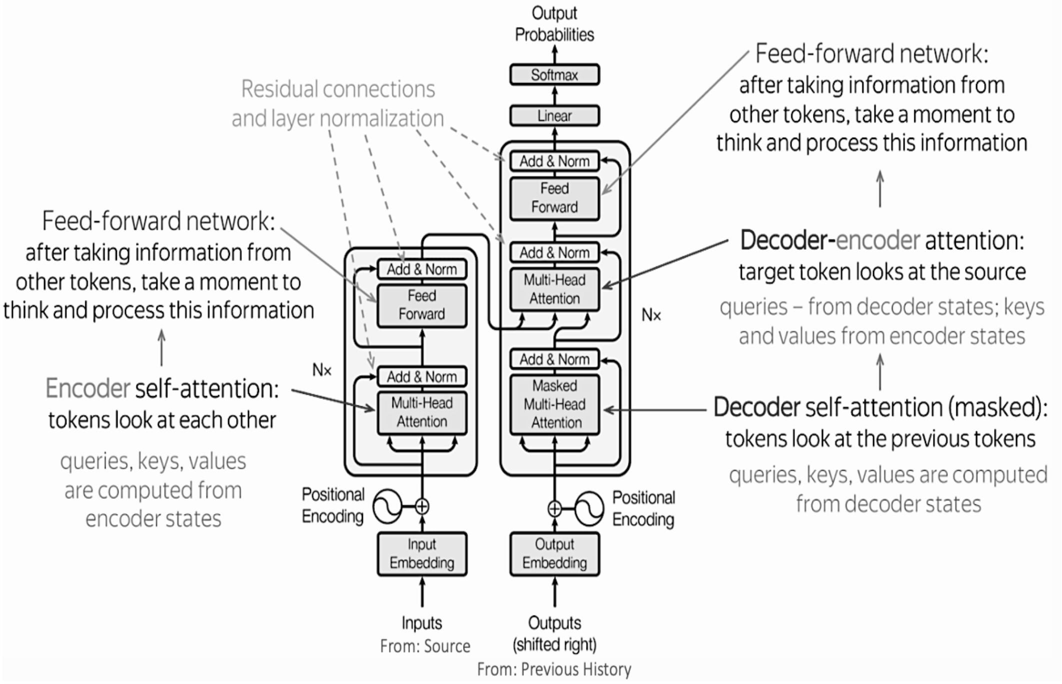

Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Multi-Head Attention
Masked Multi-Head Attention
Input Embedding
Output Embedding
Positional Encoding
Nx
Inputs
From: Source
Outputs (shifted right)
From: Previous History
Residual connections and layer normalization
Feed-forward network: after taking information from other tokens, take a moment to think and process this information
Encoder self-attention: tokens look at each other
queries, keys, values are computed from encoder states
Decoder-encoder attention: target token looks at the source
queries – from decoder states; keys and values from encoder states
Decoder self-attention (masked): tokens look at the previous tokens
queries, keys, values are computed from decoder states

Figure 2.5: A Transformer architecture; and it components' functions. cited from [226]

Furthermore, combining Transformer framework with Transfer Learning [220] generates what are known as Pre-trained Models, which are now at the cutting edge concepts of NLP research. The most powerful models of these models is Bidirectional Encoder Representations from Transformers (BERT) [56], which is mainly motivated to many other Pre-trained models.

2.3.5.1 Pre-trained Models

BERT:

Bidirectional Encoder Representations from Transformers (BERT) are context-dependent embeddings. BERT is based on the complex neural network architecture of Transformer, and its large version includes: 24 layers, 1024 hidden states, 16 heads, 340M parameters. It has been pre-trained on a large amount of unlabeled textual data in order to develop a language representation that can be fine-tuned for a particular NLP task [56]. Various approaches have recently been proposed to enhance BERT's prediction metrics or computing speed. XLNet and RoBERTa are two models that introduced improved prediction metric of BERT.

XLnet:

XLnet has shown to achieve 2% to 15% improvement over BERT performance [234] on different benchmark data. The large version of XLnet, which includes 24-layers, 1024-hidden, 16-heads, has been trained based on 133 GB of data (16 GB of the data which is the same training corpora of BERT, and 97 GB is additional data).

Transformers models basically are pre-trained on unlabeled data extracted from two different data sources. The first is the BooksCorpus [244], which includes 800M words, The aim of creating this corpus is to align the books to their movie releases. This provides rich descriptive explanations for the visual content of the movies, which go semantically far beyond the captions available in current datasets. The other source is English articles from the English Wikipedia (A description of this online encyclopedia[*], which includes 2,500M words. It is Wikipedia's first edition, founded on 15 January 2001 Both of these data are unlabeled data.

To make the training even better than BERT, XLNet algorithm predicts all tokens but in a random order, and this approach called Permutation Language Modeling. This is in contrary to masked language model in BERT. It only predicts 15% of tokens (the masked ones). This permutation language modeling also introduces a new approach for NLP area, which is the random order prediction of tokens instead of sequential order such as in the traditional language models. This makes the model more efficient in terms of processing word dependencies and relationships. XLnet is also inspired by Transformer XL models; as it addressed the issue of fixed-length contexts [53] and it demonstrated a high performance, even when the permutation-based training is absent [234].

RoBERTa:

RoBERTa was introduced by improving BERT's training methodology, which mainly involved retraining on more than 1,000% of BERT data (160 GB of text for pre-training[†] plus the 16 GB

[*] https://en.wikipedia.org/wiki/EnglishWikipedia

[†] The data were obtained from a web text corpus (38 GB), the Common Crawl News dataset (76 GB), and stories

of data used in pre-training BERT[*]) [127]. This means that RoBERTa uses more than 1,000% of BERT's computing power. RoBERTa also replaces the Next Sentence Prediction (NSP) task in BERT with dynamic masking, which changes the masked token dynamically while training (epochs). This leads to an improvement of 2% to 20% over BERT's performance on different benchmark data [127].

2.4 Conventional Machine learning

The approach of extracting features from data inputs is a key difference between conventional ML and DL. In conventional ML, features must be extracted manually by a data expert in a process known as feature engineering, while in DL, the algorithm itself extracts the features. Figures 2.6(a) and 2.6(b) provide a visual description of the two concepts of learning.

The conventional ML approaches used in this research are based on supervised learning (see Section 2.2). Unlike ANN models, which extract features and make reliable decisions independently, conventional ML models require human intervention in the form of feature engineering. This is essential in the early stages to enable an algorithm to make decisions based on what it has learned from the provided features.

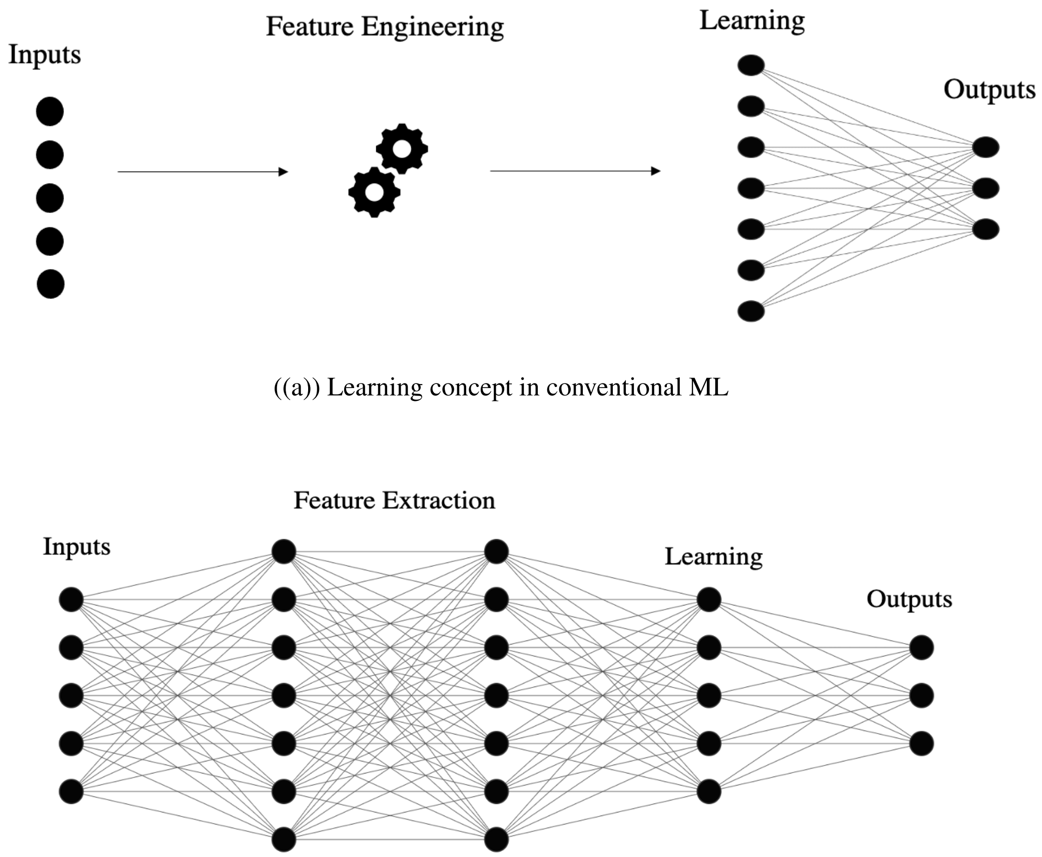
In this research, the conventional ML models that are used are Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), Logistic Regression (LR), and Extra Trees (ET).

2.4.1 Support Vector Machine

This algorithm is used widely for classification tasks. To perform training with SVM, the number of features *n* in a set of data is calculated for plotting each data item as a point in *n*-dimensional space, and each feature's value is the coordinate's value. This helps to define a line (separation boundaries or hyperplane) that divides the points into distinctive groups in regions, which are then classified differently. A hyperplane is defined by the distance between two points (called support

from Common Crawl (30 GB)

[*] The data consist of a corpus of books and English Wikipedia



((a)) Learning concept in conventional ML

((b)) Learning concept of DL

Figure 2.6: Learning concepts in conventional ML and DL

vectors). If the two closest points are at the furthest distance from a line, then it is the classifier line. The space between the line and each of these points is called the margin [39], [224].

2.4.2 Logistic Regression

Even though it is called regression, LR is not a regression algorithm. Rather, it is a classification algorithm commonly used for binary classification. It is used to predict discrete values, such as yes or no, from a set of independent variables. In practice, it forecasts an event's probability of occurrence by fitting data to a logit function. This also explains its name: logistic regression or, less commonly, logit regression. So, the output values are between 0 and 1 since it forecasts probability [194], [175].

2.4.3 Naive Bayes

This is a Bayesian classification method based on Bayes' theorem. The classifier assumes that the existence of one feature in a category has no relevance to the existence of other features. It is a simple model to design and is especially useful for huge data sets. NB is powerful and can outperform even highly sophisticated classification techniques. Text classification and issues with numerous classes are where NB is most commonly used [180].

2.4.4 K-Nearest Neighbours

KNN can be used to solve issues in both classification and regression. However, in the industry setting, it is more often used in classification tasks. KNN simply stores all available instances (data) and assigns new instances to them based on a majority vote of its K-nearest neighbours. The instance that has been allocated to the class is the most common among its K-closest neighbours, as determined using a distance function such as Euclidean or Manhattan distance. If $K = 1$, the instance is simply allocated to the nearest neighbour's class. Sometimes, selecting K might be challenging when performing KNN modelling [117].

2.4.5 Decision Tree

DT is usually used to handle classification tasks. It is a powerful method in ML as it works for both categorical and continuous dependent variables, which is remarkable. Using DT, the population is divided into two or more homogeneous sets. To generate as many distinctive sets as possible, the most meaningful attribute's variables are used [5]. RF and ET are extended versions of DT

2.4.6 Random Forest

RF is a DT approach but an ensemble method, which is because it is made up of a collection of DTs known as a forest. In this algorithm, each tree contributes a classification to a new object based on characteristics, which represents the 'votes' for that class. These trees grow based on the number of samples. If the training set has N instances, a random sample of N instances is taken with replacement. This sample set is going to be the training set that contributes to growing the tree. It is important that there is no pruning in RF, meaning that each tree is grown to the greatest degree feasible [36], [5].

2.4.7 Extra Tree

ET is a technique that combines the predictions from several decision trees (ensemble mood), and it is quite similar to RF [79]. It usually performs as well as or better than RF. Both choose a random collection of characteristics for each node's division. However, there are some differences between RF and ET. In the case of ET, it utilises the whole sample, whereas RF utilises only bootstrap replicas; this means it subsamples the inputs with replacement. The use of cut points for splitting nodes is another distinction between these two algorithms; in particular, RF selects the best split whereas ET chooses it at random. Once the split points are chosen, the two algorithms determine which of the subsets of attributes is the best. As a result, ET incorporates randomisation while maintaining optimization.

2.5 Computational Stylometry

AP is mainly concerned with analysing the writing style of a group of authors [188] in such a way as to automatically identify their anonymous demographics (e.g., age, gender, and educational level) [167]. This approach assumes that it is possible to infer traits of an author based on studying their writing style [16]. For example, studying how an author uses prepositions (e.g., using ‘4’ instead of ‘For’) and determiners (e.g., using ‘U’ instead of ‘You’) can indicate the age of the author, specifically in an informal context. Similarly, the frequent use of emoticons and non-dictionary idioms is a common writing style among young people, and so on [16].

Thus, computational stylometry is a text classification task within the area of NLP, in which labels need to be assigned by automated classifiers to objects (usually texts). For instance, the age attribute in AP has classes in the form of value ranges (10-20, 20-30, and so on); these ranges are the expected outputs in predicting the author's age [47]. AP is regarded as a subfield of text mining and also belongs to the so-called computational stylometry field [47]. Using ML approaches has led to state-of-the-art advances in this area [15].

2.5.1 Stylometry Features

The features it is possible to obtain based on an author's writing style are called *stylometry features*. This terminology covers a wide spectrum of features [167]. Many different vectors in the form of textual representation can be extracted and, as indicated in the literature, it can be found that they range from lexical and semantic to syntactic representations. These features are able to distinguish an author's traits based on their writing style, which improves a classifier's efficiency. They are typically characterized by researchers into five levels [147] as follows:

- Lexical features are the simplest form of data feature representation, which deal with the word and character levels. Both levels can capture differences in style and contextual information.
- Semantic features can capture the meanings of words, phrases, and sentences [147]. They are recognised to have an important role in identifying the traits of an author [73].
- Syntactic features such as word morphology deal with the internal structure of words within a sentence [21]. They are another important factor in analysing an author's writing style [66].
- Structural features represent the organisation of a document [147]. These features can be used in long documents such as emails, which sometimes contain signatures, but it is uncommon to find them in short sentences such as tweets.
- Domain/content-specific features refer to features that are only applicable to a specific domain, such as Twitter mentions [147].

Research has shown that semantic features and syntactic features are excellent factors to distinguish between classes of an author [21]. Thus, this thesis first focuses on a commonly used type of such stylometry features, namely, the semantic representation (see Chapter 5). Additionally, the more uncommon type – the syntactic type – is considered by discovering and fine-tuning this representation via DL models in Chapter 6, which is because such models have not been explored up to now for AP. Such features must be converted/represented effectively before they are processed by machines. Once the data are transformed into vector representations, they are passed to a classifier as the training input. These pre-processing steps, depending on experimental setting, are discussed in detail in Chapters 5, 6, and 7.

2.6 Epilogue

In rule-based systems, any unexpected input can cause the system to fail, which necessitates the implementation of additional rules or the modification of existing rules to resolve the issue. Significantly, therefore, the emergence of ML systems has overcome this limitation. Without having to write specific rules to address a problem, an ML system makes data-driven judgments. In this way, ML enables researchers to build computer programs that can learn from experience and make decisions automatically. The field of ML is concerned with the production of algorithms that learn from input data to discover underlying regular patterns. The extracted patterns can next be used to classify previously unseen instances in the data. In this thesis, the chosen models (DL and conventional ML) were used as comparative models. This is because both have been used in recent studies as state-of-the-art models and/or baseline models. More details are presented in the next chapter, which shows ML data and models used in literature for AP, as well as MOOC-related research in the area of extracting learner demographics.

Chapter 3

Related Works

3.1 Prologue

The ongoing wave of technology innovations in recent years has affected many aspects of life, including education systems. In education, one notable output of the digital age is the rise of so-called Massive Open Online Courses (MOOCs). MOOCs are educational information systems that provide a way to democratise knowledge, usually by offering free resources, and these platforms have experienced great success in attracting significant numbers of users. Owing to this phenomenon, users of MOOCs vary considerably in terms of their age, gender, location, employment status, and other factors.

However, in spite of this popularity, MOOC retention is low [51], [7]. A possible explanation is that the heterogeneity of users affects their diversity of needs and, thus, their involvement with MOOCs [76]. It has become more and more understood that a 'one size fits all' approach is not appropriate for MOOCs [11]; this is a critical avenue that 'no-barriers education' needs to support further. Importantly, while many MOOCs give their learners an opportunity to specify demographic data about themselves, the percentage of learners who complete these data is extremely low [11]. As such, adaptation based on such data would only be applicable to very few unless automatic methods for inferring user demographics are explored. Therefore, a research area called Author Profiling (AP) is applicable here as a way to infer these demographic characteristics. In this thesis, the author (the learner) specifically targets the understudied area of automatically extracting demographic data in MOOCs, which could serve as a means to design customised recommendations.

This chapter covers related works that have extracted the demographics of users based on user-generated texts. A definition of the research field of AP is provided, followed by a description of the data and models that have been employed for AP. Next, MOOC-related research in the area of extracting learners' demographics is discussed.

3.2 Author Profiling

On the World Wide Web, a huge amount of text is written daily by different kinds of authors, including blogs or reviews through many platforms [167]. This provides a considerable and noteworthy source of data. Thus, researchers have started to analyse these data from online sources ranging from emails and blogs to news articles. In addition, improvements in the field of computer science, especially in NLP and ML, have given rise to the discipline and practice of computational stylometry [47]. The purpose of computational stylometry is to assist the growth of the stylometric analysis area. The advanced innovations in technology that the use of Authorship Attribution (AA) in online content has facilitated are relevant for different applications, including cybercrime identification [107]. The language used on social media is characterised by abundant personal content, especially when compared to other forms of texts, including emails or classic letters [115]. This has shaped new research questions around the possibility of inferring author traits based on their writing style and content on online platforms [115].

Historically, the analysis of writing style was initiated many years ago by the sociolinguistics research community, and this type of analysis is referred to as AA [1]. It was applied for the first time in 1994 to the literary texts of Shakespeare, and it has since been applied to other literary and historic texts in order to determine the linguistic patterns of various authors. Until 2000, AA was restricted largely to this domain, but new resources in sociolinguistics have emerged with the growth of the Internet, which has the ability to complement these stylometric analyses [148]. Due to this change in the area, instead of analysing the texts of known authors, other tasks have risen in the area called Author Profiling (AP) [168]. AP is concerned with analysing an author(s) texts to detect some of their characteristics [188]. Research into AP started in our current century, the 21st century, and it has been applied to areas ranging from formal texts such as students' essays in work done by Argamon et al. in 2005 [19] to informal texts such as blogs in Schler et al. in 2006 [188].

In AP, it is possible to classify users into groups based on similar patterns (features) that are extracted from their writings and learned via ML algorithms. This is in demand due to the high number of users on online platforms [15]. Studies have found that age, gender, and other demographic characteristics of each user are directly or indirectly connected to their writing style; the relevance of extracting these characteristics from a given post or document is crucial [105]. As a result, AP is a widely used and researched technique that is being used to address problems across a variety of domains [105].

3.2.1 Applications of Author Profiling

In 2013, researchers in the domain of AP studies began to focus on social networking platforms. Initial research efforts in this area were related to the extensive growth of these platforms, particularly their generation of more text data compared to other resources on the Internet. Social media platforms are an enticing yet demanding target for AP because users are usually anonymous [135]. In particular, this leads to substantial ambiguity on these platforms.

AP is also an attractive research area due to its diverse applications in a wide range of areas [151], including marketing, forensics, security, and education. For this reason, AP is considered a critical technique in the current information era [168]. In the field of security and forensics, AP tasks try to disclose the identity of authors by predicting and classifying their profiles, which can ensure that users are protected from online harm or identity theft [28], or even identify a terrorist source [176]. In marketing, AP systems seek to improve marketing strategies by allowing companies to learn the characteristics of online consumers who wrote about them, or even to identify suitable candidates for advertising their products online [135]. Digital text forensics is another example that shows the benefits of AP [28]. Knowing an author's demographics can help to predict their identity in the case of a given crime.

Additionally, AP tasks are valuable for the education system, which is especially relevant for this thesis. For example, an important application of AP is its use to assess and identify the level of knowledge of students based on their writing [168]. However, the research in the area of AP for educational purposes is limited and no research has focused on online educational platforms. Nevertheless, this is an encouraging sign regarding the significance of investigating the issue of AP in relation to MOOCs. This research contributes to the field by extending AP to a less investigated domain (namely, education) and a new platform (namely, MOOCs), thereby targeting learners on MOOCs using so-called Learner Profiling (LP). Figure 3.1 illustrates the umbrella of the AP field, including the newly proposed area of LP.

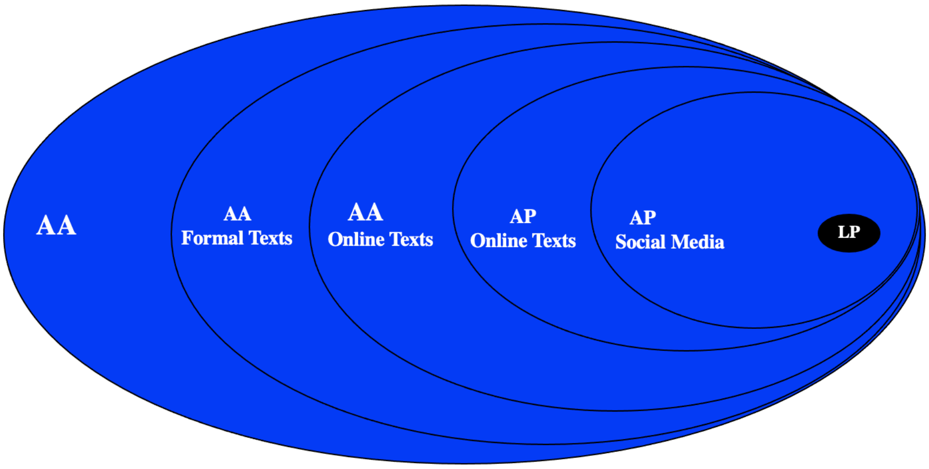


Figure 3.1: Authorship attribution umbrella, including the area of learner profiling

3.2.2 Demographics Classifications in Author Profiling

Classifying demographic and sociological characteristics such as gender, age, and educational level is an important target in the field of AP [21]. They are utilised to describe authors based on a specific demographic to group them into the classes in which they belong.

Up to now, only a small number of demographic types has been studied. The common characteristics of authors that have been examined in the area are gender, age, language variety, and personality type [29]. Each characteristic has different classes that need to be considered when building a classifier model. Actually, the main reason for the disparity in results between the characteristics is the diversity of the classes' essences [137]. Even when the same classifier is used, it produces different results for different characteristics, as is commonly seen in the AP literature [29].

Age and gender have received more attention compared to other demographics such as language variety and personality. Other author traits such as education level, region of origin, and mental health are still not widely used in AP tasks [152]. In particular, the demographic factor of employment status has not been investigated by any prior researcher, which highlights the originality of this thesis.

3.3 Author Profiling Corpora

AP has been studied in relation to many different data types, and so different approaches have been used [173]. The main characteristic among the available studies is the data genre used, which dominates the question of which features can be extracted. This is due to the tendencies that underlie the targeted data, in addition to the type of potential special characteristics that may exist in each specific data, such as domain symbols like use of hashtags in Twitter. Such issues in the data influence the approach. The availability of appropriate datasets for AP when using ML is of paramount importance. In this section, a number of the most common datasets used in the area are identified. A summary of these datasets, along with further data, is given in terms of the main characteristics in Table 3.1.

1. **Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) Dataset**: Due to the importance of the AP area, organisations have sought to provide resources or establish shared tasks[*] for AP research purposes. A well-known AP shared task is PAN, which is arranged annually as part of the Conference and Labs of the Evaluation Forum (CLEF) organisation. They provide corpora every year for AP, as well as other related tasks like Authorship Attribution (AA) [115]. PAN AP shared tasks are the most common shared tasks that provide datasets for the AP literature review. They started in 2013 and established goals for researchers to achieve in form of sub-tasks to classify the authors' traits, including

[*] These tasks include CS events for evaluation problems in digital text forensics and stylometry.

age, gender, personality, and language variety on social media [72]. PAN contains content from many social media platforms, and a large proportion of the content comes from Twitter [104], [206]. What makes PAN a well-known dataset in AP is not only the yearly competition that is associated with it, which encourages research efforts, but also the availability of different versions of the dataset with different sizes and classes. Table 3.1 provides further information about the PAN versions.

2. **The Blog Authorship Corpus**: This corpus has been available as a data resource since 2006[*]. It is used in a large spectrum of the AP literature review and it has four classes: gender, age, industry (sector), and zodiac sign of each author. The collection was created by Schler et al. [188] and came from a blogger.com website. The size is large as it contains 681,288 posts from 19,320 bloggers, amounting to over 140 million words.

3. **Fisher English Transcripts**: This dataset has two classes: gender and age. It was created by Cieri et al. [45] in 2003. It contains more than 160,000 transcripts of telephone conversations between random pairs of people. This collection of data includes a larger volume of transcribed telephone conversations compared to other similar datasets, and it can be used for any research initiative that aims to analyse telephone speech. The data were collected over approximately four months to ensure the collection of unique calls without repetition.

4. **The International Corpus of Learner English (ICLE)** : This dataset is a collection of 6,085 argumentative essays. The students who wrote the texts were at a higher-intermediate to advanced level of English learning. The labels of the data are the different mother tongues of learners to classify students based on their native language of students [83] (see Table 3.1).

   Also, the **TOEFL11 corpus** is another essay type of data used for native language profiling, which includes 12,100 essays (1,100 per language) written by TOEFL test-takers in 2006. These essays are evenly balanced from eight different topics [31].

5. **Internet Movie Database (IMDB) Corpus**: IMDB is a database for movies, video games, and television shows. The corpus consists of movie reviews only, and each of these reviews

[*] The Blog Authorship Corpus can be downloaded from http://u.cs.biu.ac.il/ koppel/BlogCorpus.htm

is associated with the gender of the author. The reviews are based on the top 250 movies of all time, according to the IMDB website rankings [153].

6. **The British National Corpus (BNC)**: This corpus was collected from different sources of English discourse texts in the late twentieth century in Britain, both written and spoken. The written texts were collected from many different sources, including popular stories, academic publications, journals, and newspapers. Labels of the data are based on gender, age, domicile in UK, and social class. [48].

7. **The Reuters Corpus**: This corpus includes formal texts used for AP, which were collected from news articles and all stories written in English and published in Reuters journal for 12 months between August 20, 1996, and August 19, 1997. Later these data was labelled manually based on each journalist's gender [124].

8. **Email Corpus**: This corpus was created by Estival et al. [65]. After collecting the emails, the dataset was been labelled by asking the email users to provide their demographic information by responding to a questionnaire. The data collected related to gender, age, first language, country of residence, level of education, and psychometric factors [65].

Table 3.1 includes more details of the data mentioned above and information abotu further AP data, which gives a clear view of what genres and classes there are in the available AP datasets.

Table 3.1: Common author profiling corpora

Data	Classes	Genres	Languages	Data Size
PAN2013 [104]	Female/Male, Ages: 10s/20s/30s	Blogs	English, Spanish	300,000
PAN2014 [206]	Female/Male, Ages: 18-24/25-34/35-49/50-64/65-xx	Blogs/Twitter/Hotel-Reviews	English, Spanish	24,000
PAN2015 [207]	Female/Male, Ages: 18-24/25-34/35-49/50-64/65-xx, Different-Personality-Types[*]	Twitter	Italian, English, Dutch, Spanish	33,000
PAN2016 [173]	Female/Male, Ages: 18-24/25-34/35-49/50-64/65-xx	Training: Twitter, Testing: Other-Genres	English, Spanish, Dutch	Sizes (Vary) [†]
PAN2017 [172]	Female/Male, Language-Variety-Types[‡]	Twitter	Spanish, English, Arabic, Portuguese	11,000
PAN2018 [171]	Female/Male	Twitter: Text/Images	Arabic, English, Spanish	12,600 [§]
PAN2019 [170]	Female/Male, Bot/Human	Twitter	English, Spanish	4,800
PAN2020 [169]	Fake News Spreaders (Yes/No)	Twitter	English, Spanish	100,000
PAN2021 [146]	Hate Speech Spreaders (Yes/No)	Twitter	English, Spanish	100,000

[*] Openness, Conscientiousness, Extroversion, Agreeableness, and Stableness

[†] Sizes are vary based on genre (cross-domain data). For the English dataset, the data size is as follows: Twitter: 428,000, blogs: 78,000, other social media: 348,000

[‡] Australia, Canada, Britain, Ireland, New Zealand, America, Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela, Portugal, Brazil, Gulf-Countries, Levantine, Maghrebi, Egypt

[§] (The size of each author is : 100 texts and 10 images, per author)

Table 3.1: Common author profiling corpora

Data	Classes	Genres	Languages	Data Size
Blogs-2006 [188]	Female/Male, Ages: 10s(13-17)/20s(23-27)/30s(33-47), Industry: (Education, Investment-Banking, Non-Profit), Zodiac-types[*].	Blogs	English	681,288
Fisher-2004 [45]	Female/Male, Ages: 16-29/30-49/Over-50	Telephone Scripts	English	16,000
IMDB-2010 [153]	Female/Male	IMDB Reviews	English	31,300
BNC-1993 [48]	Female/Male, Ages:15-24/25-34/35-44/45-59/60+, Domicile-in-UK: North/Midlands/South, Social-Class: (A/B)/C1/C2/(D/E)	Ancient British Texts	English	4,049
Essays-2003 [83]	11 Mother tongue groups[†]	English Learners' Essays	English	3,640
TOEFL-2013 [31]	11 Mother tongue groups[‡]	English Test Takers' Essays (2006–2007)	English	87,502
Reuters-2004 [124]	Female/Male	News Articles	English	800,000
Email-2007[65]	Male/Female, Ages: <25, 25–35, and >35, Education-level: No-tertiary-education/Some-tertiary-education, Native language[§], Psychometric[¶]	Emails	English, Arabic, Spanish	9,836
Blogs-2010 [145]	Female/Male	Blogs (Popular-Sites)	English	3,100

[*] Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio, Sagittarius, Capricorn, Aquarius and Pisces.

[†] Japanese, Chinese, Russian, Turkish, Italian, French, German, Norwegian, Bulgarian, Czech, Dutch, Polish, Swedish, Spanish, Finnish, and Tswana.

[‡] Japanese, Chinese, Korean, Hindi, Turkish, Arabic, Italian, German, French, Spanish, and Telugu.

[§] English, Arabic, and Spanish. Country: United States, United Kingdom, Australia, New Zealand, Egypt.

[¶] Agreeableness, conscientiousness, extroversion, neuroticism, and openness.

Table 3.1: Common author profiling corpora

Data	**Classes**	**Genres**	**Languages**	**Data Size**
Blogs-2011[184]	Ages of USA living only	Blogs (Journal[*])	English	24,500
Medical-Forum [149]	Ages: 10s/20s/30s	Breast Cancer Forum	English	1,997
Blogs-2011 [184]	Ages: 10s/20s/30s	Blogs	English	24,500
News Corpora [219]	Native language: English/Persian/Turkish /German	News Articles	English, Persian, Turkish, German	150 [†]
Comments-2018 [81]	Personality (MBTI)[‡]	Reddit	English	23,503
Facebook2013 [190]	Female/Male, Ages:10s/20s/30s, Personality[§]	Facebook	English	136,000

[*] http://www.livejournal.com/

[†] 150 articles per language.

[‡] MBTI categories: Extroversion-Introversion (EI), Sensing-Intuition (SI), Thinking-Feeling (TF), Judging-Perceiving (JP).

[§] Big Five Personality Traits.

3.3.1 Critical Evaluation

Based on the current review, it is clear that the genre of data used in AP is diverse. These data are collected from many domains such as blogs, social media, movie reviews, emails, essays, medical documents, and historical texts, and this explains the need of classifying the demographics of users for many domains.

Blog posts are widely used for AP due to the huge amounts of data created by bloggers on a daily basis. This makes them a good venue for collecting big data, in addition to the fact that these posts usually include varied different topics. One of the common sources of blog data used in the area of AP is "The Blog Authorship Corpus".

However, PAN dataset is regarded as the most common dataset that is widely used for AP studies, and a range of different tasks applied in them. It is not the only available data for AP, but the PAN dataset appears to be dominant in AP research, as it can be seen in next section, in Table 3.2. In addition to its characteristics such as size and genres variety, it also organizes well-known competitions for the AP task yearly, which has motivated researchers to participate.

On the other hand, Twitter data is easy to fetch, which is one of the reasons for its widespread use in the research community. Since Twitter provides Application Programming Interfaces (APIs)[*], this gives trustworthy access with public permission for researchers, companies, and other entities to fetch Twitter data that they need [118]. The data genre used, especially in PAN, is mostly based on Twitter. Even though a Twitter post is limited to only a few characters (250 characters), it is a major avenue for online users to communicate and express their ideas and it is a fundamental example of big data due to its growing number of users and their diversity[†] This can explain why stakeholders are interested in gaining a better understanding of people’s tweets [33].

Discovering other author profiles is important as well and should not be only limited on few number of traits/ demographics, neither limited to few domains or genres. Unfortunately, most of these public data are restricted to only a few demographics of authors in terms of gender, age, and native language or personality. Other important traits need to be discovered but the obvious

[*] https://help.twitter.com/en/rules-and-policies/twitter-api

[†] This is also the case with MOOC platforms; as they have huge numbers of diverse users.

dilemma that researchers may face will be availability of annotated data for these new traits. In some of the AP studies, data are crawled in order to get customised data for their own research. Crawling data is not a challenging work, but it is time-consuming and may need additional costly resources to annotate the data. In spite of that, a particular data have collected in this thesis to achieve its objectives and, which is attempted to fill up such gap in the AP literature.

3.4 Related Studies in Author Profiling

In this section, a comprehensive overview of works done for AP is presented. First, the emergence of AP research is discussed. Next, these works and the recent works are summarised in Table 3.2. A discussion of the models that have been examined in AP is given, and these are classified into two groups based on the type of ML they use: either used conventional ML approaches that require a feature engineering process or Deep DL approaches.

3.4.1 Conventional ML Approaches

- Raghunadha et al. [167] aimed to detect the gender of reviewers based on their reviews in a hotel website by applying four types of features: stylistic features, structural features, readability features (complexity of vocabulary and syntax in a text), and topic features (frequency of words occurring together). They used the ReliefF algorithm from WEKA[*] for feature selection. Based on these types of features, the authors established a final set of features that consisted of 43 features in total. These were fed to a RF classifier, and the model achieved 93.35% overall accuracy.

- The researchers in Swathi et al. [212] also experimented on the same data that used in [167], for users gender detection. However, they only used one type of data feature, namely, stylistic features. the authors used two classifiers: RF and Multinomial Naive Bayes (MNB). To improve the classifiers performance, a novel term weight measure was introduced by the authors that is used for selecting the features; it involved considering the class label in the data. By using this weighting schema, the highest accuracy was recorded by the MNB classifier at 91.45% [212].

[*] A java-based open-source tools used for ML research, which stands for Waikato Environment for Knowledge Analysis, available at: https://www.cs.waikato.ac.nz/ml/weka/)

- In Kocher et al. research [114], the authors used the PAN 2016 datasets (cross-genre) to classify the age and gender of authors. The research team compared the performance of 24 distance measures for feature selection in order to have the best set of features for their study. However, the large number of features involved in their study caused similarity and confusion in the measurement results, which led to very low accuracy results. In addition, the authors used a different genre training set than the test set (cross-genre), which was another cause for the lower accuracy. KNN was used as a classifier, and the results for English tweets were as follows : (i) On same genre datasets: 62.96% accuracy for gender and 45.99% for age; and (ii) On different genre datasets: 56.96% accuracy for gender and 34.61% for age [114].

- In other work, Ortega-Mendoza et al. [152] specifically looked at the significance of the first-person pronoun as a feature that can be used for age and gender identification. The first-person pronoun – namely, "I, me, mine, myself, my", as well as "I'm" – are commonly used words among social media users. Based on a previous study from the same research team [134], the authors hypothesised that personal phrases (first person pronoun) in a sentence can serve as a reflection of an author's attributes, which indicates the role of these terms in providing higher performance in the task of AP. They also introduced a term weighting scheme. They used three different datasets from six different social media collections in English (cross-genre), and used LinerSVM as a classifier for the task. Their proposed work shows that personal phrases are a useful feature to apply for age and gender identification. Accuracy results were varied based on the genre of text. The highest accuracy result for the gender classification was 84.25%, while the highest accuracy for age classification was 77.68% [152].

- Ashraf et al. [23] have trained stylistic features for gender and age identification using PAN 2016. They applied four categories of stylistic features: lexical, syntactic, character-based, and vocabulary richness (56 features in total). They explored many classifiers using the WEKA tool. RF, J48, and Logical Analysis of Data Tree (LADTree) were used. Since they used cross-genre data, results on the training set were higher compared to those on the test set. On the test set, the authors reported 57.6% accuracy for gender identification and 37.1 % accuracy for age identification [23].

- The aim of Markov et al. [136] was to classify both the gender and language variety of an author using a Twitter dataset consisting of multiple languages (PAN 2017). To do so, the authors used character n-grams, word n-grams, lemmas, and domain names. They also used various NLP feature representations, where text data were converted into a numerical vector representation. Various ML algorithms were also applied. The liblinear classifer (SVM) outperformed the other classifiers, and in English tweets they reported an accuracy of 81.3% for gender classification and 87.1% for language variety [136].

- In Markov et al.'s research [135], the authors used the doc2vec algorithm [122] to learn neural network-based document embeddings. In turn, these representations were used as the input to train a LR classifier. Both age and gender traits that are extracted from social media data (PAN 2014, PAN 2015, and PAN 2016) were examined in their study in the single genre mood and cross-genre mode. Their work contribution is the study's comparison of the traditional features with the document-embedding features in the same ML algorithm (LR classifier), which indicates that the introduced features yielded better performance in a single-genre setting only.

The above-mentioned studies are considered to represent different approaches in the AP literature regarding works using conventional ML, and they are also summarised, along with other related works, in Table 3.2.

3.4.2 Deep Learning Approaches

- A study performed by Surendran et al. [211] hypothesised that in order to increase the accuracy of AP, more relevant features need to be considered rather than centred around stylometric features only. They proposed the use of character-level and word-level features, with relevant features including emotional words, vocabulary richness, and readability metrics on a Twitter dataset. Notably, these features are applicable on other social media platforms for age and gender identification. The authors also compared the performance of conventional ML and DL approaches in their study. For the DL approaches, CNN achieved the highest accuracy for both traits: 90.1% accuracy for age identification and 97.7% accuracy for gender identification. It is reasonable to infer from these results that non-stylometric features may also be useful for an AP task.

- The unique aspect of the study undertaken by [18] is that they dealt with the AP task based on unsupervised learning. This is notable because AP solutions typically utilise supervised learning

[212]. The authors mentioned that the main issue when working with a DL model is that it requires a vast amount of annotated data. Thus, Ardehaly's team [18] proposed an idea called Learning from Label Proportion (LLP) [236]. To explain, group "bags" of unlabelled instances included in the training data and associated with label distributions for each of these bags. In addition, they used images (portraits of the users) and texts (users' names and counties), which were collected from Twitter profiles. The proposed DL model achieved good accuracy compared to the other supervised models for the AP task. The accuracy achieved for gender was 96%, for the white race 96%, and for the black race 86%. Their data is helpful because algorithms can easily learn and correctly identify these classes, particularly since face images and names have features that are clearly distinguishable among people based on their gender and race. As such, the study's results highlight the importance of using other non-textual features for AP.

- The authors in [72] introduced a new data representation to the field of AP, which is based on the subword technique [32]. The proposed system also mainly utilises the character n-grams embeddings technique. They used an algorithm called Deep Averaging Networks (DAN) to classify two author classes: language variety and gender in the PAN 2017 dataset. According to [96], DAN is a fast and competitive DL approach to use in many text classification tasks because it can magnify the most discriminant dimensions included in the embedding average. The model is also flexible when used with noisy data (e.g., data containing typos and abbreviations). The results of the model in the English version of the data were 76.6 % accuracy in terms of language variety and 76.5% accuracy for gender [72]. It is notable that this research group's work has opened a significant door for researchers in terms of advancing the area by trying new NLP embedding methods, as well as novel DL models.

- In a study performed by [115], a Bi-directional RNN (bi-RNN) with Gated Recurrent Units (GRU) [78] was used with an attention mechanism [225]. The team was inspired by a method applied in Neural Machine Translation (NMT) research. In fact, the greatest challenge in NLP tasks is the question of how to present input sequences in the most simple way but still possess the most important information. This challenge has been addressed in the field of NMT, and this is why the authors of the work reused the same algorithm used in NMT for AP (i.e., as both tasks are NLP tasks). The bi-RNN was able to weigh the most related information automatically, and the authors used the attention mechanism to further simplify the sequences. The accuracy of the proposed model was compared with previous models proposed for AP using CNN on the same dataset (PAN 2017). The model's accuracy outperformed the baseline CNN model by 1.45% for gender identification and 2.69% for language variety identification. The results on the English version of the dataset were 78.88% accuracy for gender and 79.08% for language variety [115]. One of the most notable features of this study is that it compared two important approaches in DL: the sequential approach (RNN) and the spatial approach (CNN). The results indicated that the sequential approach slightly outperformed the spatial approach. However, further investigation is needed to answer the question of how these two approaches perform when different data or different classes are applied to the sequential model (RNN).

- In Schaetti et al. [186], the study team tried to identify the gender and language variety of authors in a Twitter dataset using two proposed models: a CNN model, which used character feature only, and a TF-IDF-based model [86], which used character and word features. For both models, the same prepossessing techniques were applied. The accuracy of the TF-IDF-based model in detecting the gender of authors is 68% accuracy, and 83% accuracy was achieved for language variety in English tweets. The CNN model was revealed to have a better performance in detecting author gender, which was 78% accuracy, but it achieved a lower accuracy for the language variety class (65%). The TF-IDF-based model, in general, was identified as an effective approach for user modelling based on text [86]. However, the effectiveness of TF-IDF-based models must be investigated in data dominated by a specific kind of content [105], such as MOOC data.

- One of the AP works to consider the personality trait of an author was a study conducted by Liu's team [125]. They developed a personality detection model based on Bi-RNN with GRU. Twitter texts from PAN 2015 were used as a dataset for the model, which has gold standard personality labels: openness (OPN), conscientiousness (CON), extroversion (EXT), agreeableness (AGR), and stableness (STA). Their modelling capacity is strengthened by gaining more data over time in the study. The accuracy results of their model were as follows: EXT: 14.2%, STA: 18.8%, AGR: 14.7%, CON: 13.6%, OPN: 12.7%. In fact, personality trait were even more complex than other traits. It is well-known that the personality of an individual is temporal and not stable [137]. For this reason, Liu et al. created their modelling by gaining more data over time [125]. Their research indicates that author's personality is extremely difficult to identify by text only, even with DL approaches.

- A study done by Jiang et al. [98] introduced a model that processed different types of features in their work – characters, words, and topics – to identify different types of traits (gender, age, and industry) using The Blog Authorship Corpus. The novelty in their study was to introduce an ensemble DL model for the AP task for the three traits. The model included an RNN algorithm for characters input, CNN for words input, and the Latent Dirichlet Allocation (LDA) algorithm for the document/ topic level. Then, they were integrated using an ensemble DL structure model. The following accuracy results were recorded for the model: Gender: 79.2%, Age: 79.6%, and Industry: 37.6% [98]. Many AP studies have used The Blog Authorship Corpus, but Jiang et al.'s ensemble model provided a robust performance compared to other competitors for all the three classes on the same dataset.

Table 3.2 offers a comprehensive summary of the above-mentioned works and additional works that have been undertaken in the AP area, focusing in particular on studies involving English texts data from 2005 to 2022.

Table 3.2: Author profiling related works (2005-2022)

Year/Cation	Data	Classes	Features	Classifiers	Results
2005 [19]	Essays	Personality	text cohesion, function words, assessment and appraisal measures	SVM	(neuroticism 58.2%), (extraversion 58%)
2007 [65]	Emails	Gender, Age, Native Lang., Education, Country, Personality	structural features, named entities, word length, punctuation, function words, POS	RF	(69.26% Gender), (56.46% Age), (84.22% NL), (79.92% Education), (81.13% Country), (53.16% – 56.73% Personality)
2009 [159]	Blogs	Gender, Age, Location, Industry	word/char tokens, function words, paragraph and structural features	RF	(77.27% Age), (83.34% Gender), (78.01% Location), (82.12% Industry)
2010 [145]	Blogs	Gender	domain-terms, n-grams, words-endings, POS	SVM	88.56%
2011 [153]	Reviews (IMDB)	Gender	movie-metadata, word richness/ complexity, pronouns	Statistical Model	73.71%
2012 [216]	Essays	NL	word/ char n-grams, function words, P0S bigrams, spelling errors	Ensemble LR	(90.1% ICLE), (70.9%- 80.9% TOEFL subsets), (84.6% TOEFL full)
2013 [154]	Facebook	Gender, Emotion	words tokens, punctuation, P0S, emoticons	SVM	(59% Gender), (59.6% Joy), (32.3% Anger), (36.1% Disgust), (50.4% Surprise), (20% Sadness)

Table 3.2: Author profiling related works (2005-2022)

Year/Cation	Data	Classes	Features	Classifiers	Results
2013 [131]	Twitter	Political Alignment	twitter metadata/ network, n-grams	SVM	90.8%
2014 [185]	Facebook, Blogs, Twitter	Gender, Age	word unigrams	SVM	(83% Age), (91% Gender)
2015 [107]	Political papers	Gender, Age, Political Alignment	style markers, word/ char n-grams, lemmas, P0S n-grams	SVM	(74.6% Gender), (44.6% Age), (58.7% Political Alignment)
2015 [52]	Twitter	Ethnicity	friendship	LR	61% Ethnicity
2016 [23]	PAN 2016	Gender, Age	stylistic features: lexical, syntactic, character-based, and vocabulary richness	WEKA: RF,J48 LADTree	(57.6% Gender), (37.1 % Age)
2016 [199]	Health forum	Gender, Age	textual, forum related features	LR	(65.59% Age), (88.41% Gender)
2017 [114]	PAN 2014–2016	Gender, Age	comparing 24 distance measures	KNN	(62.96%Gender), (45.99%Age)
2017 [136]	Twitter	Gender, Language Variety	n-grams char/ word and domain names	SVM	(81.3% Gender), (87.1% Language Variety)

Table 3.2: Author profiling related works (2005-2022)

Year/Cation	Data	Classes	Features	Classifiers	Results
2017 [200]	Facebook, Twitter	English Nationality	formal linguistic, POS, lexicon/ metadata features	MNB*	77.32%
2017 [141]	PAN 2017	Language Variety	word embed for user tweets, linked users tweets	RNN/ attention	91.39% (for English accents)
2017 [186]	PAN 2017	Gender, Language Variety	character token (CNN), character/word token(TF-IDF)	CNN, RF (TF-IDF)	highest results: (78% gender (CNN)), (83% Language variety (TF-IDF))
2017 [142]	PAN 2017	Gender, Language Variety	char/word tokens	GRU/attention and CNN	(80.46% Gender), (87.17% Language Variety)
2017 [115]	PAN 2016	Gender, Language Variety	word tokens	Bi-GRU/attention	(78,88% Gender), (79,08% Language Variety)
2017 [72]	PAN 2017	Gender, Language Variety	subword/char, n-gram embeddings	Deep Averaging Networks	(79.6% Gender), (75.9% Language Variety)
2017 [17]	PAN 2017-2015	Gender, Age, Language Variety	domain terms, lemmas, POS, tweets characteristics, subjectivity and opinion mining	similarity based classification	(45%Age), (65%Gender), (23% Language Variety)
2017 [211]	Twitter	Gender, Age	sentence tokens	CNN	(90.1% Age), (97.7% Gender)

*Multinomial Naive Bayes

Table 3.2: Author profiling related works (2005-2022)

Year/Cation	Data	Classes	Features	Classifiers	Results
2017 [18]	Twitter	Gender, Race	image/ text: face image, user name, and county	Unsupervised Learning.	(96% Gender) ,(96% White race), (86% Black race)
2018 [152]	Multiple genre	Gender, Age	first person pronoun features	SVM	(84.25%Gender), (77.68% Age)
2018 [69]	Facebook	Gender, Age, Personality*	textual, visual, and relational features	ANN	(90% Age), (96% Gender), (65% Opn), (62% Con), (59% Ext), (56% Agr), (58% Neu)
2018 [187]	PAN 2017	Gender	Glove vectors, POS, function words	RNN	80.6%
2018 [162]	Texts/ Key-strokes	Gender, Age	keystrokes (press /release), char/word n-grams, CBOW vector	SVM	(63.50% Gender), (73.25% Age)
2018 [95]	SMS	Gender, Age	vocabulary, emoticon	RF, SVM, NB	(78.57% Gender), (55.42% Age)
2018 [223]	Twitter	Gender	bleaching text	SVM	59.8%
2018 [212]	Reviews	Gender	supervised term weight	MNB	91.45%
2018 [167]	Reviews	Gender	comprehensive textual features	RF	93.35%
2018 [98]	Blogs	Gender, Age, Industry	char, word, and topic features	CNN, LSTM, LDA	(79.2% Gender), (79.6% Age), (37.6% Industry)

*Extroversion (Ext), Agreeableness (Agr), Conscientiousness (Con), Neuroticism (Neu), Openness (Opn)

Table 3.2: Author profiling related works (2005-2022)

Year/Cation	Data	Classes	Features	Classifiers	Results
2018 [214]	PAN 2018	Gender	textual and image features	Fusion (GRU+VGG16)	86%
2019 [35]	Review	Gender, Age	images and texts	KNN	(73% Gender), (40% Age)
2019 [99]	Twitter	Bots, Human (Gender)	stylistic features	RF	95.95%
2019 [161]	Twitter	Gender	words n-grams	LR	81.72%
2019 [68]	PAN 2014/ 2015	Gender, Age	psycholinguistic dictionaries	SVM	Gender-Age: Blogs (70.5% - 38.4%) Reviews (70.8% - 31.6%) Tweet14 (68.1% - 40.2%) Posts (52.9% - 33.5%) Tweet17 (Gender only: 76.7%)
2019 [57]	PAN 2019	Bot Human (Gender)	char/word n-gram	CNN+RNN ensemble	(84% Bot), (58% Gender)
2019 [221]	Reviews	Nativity Language	document weight, frequent terms	RF	88.53%
2020 [163]	PAN 2019	Gender, Age, Fame/(Job)	-	BERT+ANN	(89% Gender), (83% Age) , (78% Fame), (73% Job)
2020 [38]	Twitter	Fake News Spreaders	words n-grams, frequencies of lexical features	ensemble LR	75%

Table 3.2: Author profiling related works (2005-2022)

Year/Cation	Data	Classes	Features	Classifiers	Results
2021 [208]	Twitter	gender	text and image	multimodal NNs	89.53%
2021 [174]	Kaggle Competition	Fake News Spreaders	writing style, SA, and co-authorship patterns	KNN	83%
2022 [113]	WhatsApp	Gender, Age	n-gram (words and phrase), LDA, emoji, contact card	RF	(75% Gender), (81% Age)

3.4.3 Critical Evaluation

A reasonable amount of research has been performed for AP, focusing on the extraction of an optimal set of features from the text. This can be used to infer the characteristics of an author based on these features and then feed them to a ML model.

Most prior studies have used conventional ML for the task of AP. Extracting features is one of the most complex aspects of AP and, as the available studies indicate, there is no standardisation yet. There is no coherent vision of how features are selected when using conventional ML for the task of AP, and studies performed under this category have no uniformity in their approaches. Taken together, therefore, the AP literature presented in this section suggests that feature selection in the area is only based on experimental settings. According to the comprehensive review of the AP literature that presented in Table 3.2, SVM appears to be the most frequently used classifier.

DL models reduce the effort needed to develop feature extraction methods. This is because DL models learn from high-level features of data, which is in contrast to conventional ML. In the latter, extensive effort is applied to extract features from the data to make the patterns within it more distinguishable and visible to learn with a classifier [217]. In AP, the use of DL is less investigated compared to the body of work that has used conventional ML. However, DL models have achieved the highest accuracy in the area. DL models are currently the state-of-the-art in the NLP field, especially the new generation of DL such as attention and BiLSTM.

The review of the AP literature indicates that the features used to feed classifiers in these works heavily extract from the content of the texts. This has clearly given rise to a limitation in terms of models generalising to become cross-domain models. Furthermore, content-based features utilise hundreds to thousands of features, ranging from lexical to syntactical features, and even this is a highly effective approach to solve AP tasks; however, it mainly depends on a deeper analysis of writing style.

3.5 Related Works on MOOCs

3.5.1 MOOC as an Educational Source

The first internationally recognised Massive Open Online Course (MOOC) was launched in 2008 following the growth of the Open Education movement. North America was MOOC place of birth [26], and the first MOOC was a course on *Connectivism and Connective Knowledge* provided by Dave Cormier from the University of Prince Edward Island [100]. Since then, MOOC have become a catch-all term for recent online courses [218], raising hopes for new opportunities in the world's higher education systems.

By the end of 2017, the landscape of MOOCs had grown to include 57 MOOCs platforms, 9,400 courses, more than 500 MOOC-based credentials, and approximately 100 million learners worldwide, following the launch of the first three major platforms, Coursera, Udacity, and edX, in 2012 [198]. The intention of democratising education and providing free access (open access) to as many people as possible (scalability) in a university-level education [26], and these the two key features of MOOCs. The values of MOOCs are rooted by the openness idea in education; the belief is that information should be freely shared regardless of economic or geographic barriers.

MOOCs have been driven from two different pedagogical directions by contrasting ideologies: connectivist MOOCs (cMOOCs) and content-based MOOCs (xMOOCs). cMOOCs are based on a connectivism philosophy of learning with informally generated networks, while xMOOCs take a more behaviourist approach [183]. Many MOOCs are influenced by the two ideologies, including Coursera, edX, and FutureLearn (see Figure 3.2).

MOOCs have successfully attracted a significant number of learners. Given that MOOCs are more affordable, less restricted, and more time- and location-flexible than standard higher education, and owing to this phenomenon, users in MOOCs are varied in terms of their demographic attributes such as gender, employment status, level of education, etc. [11]. In addition, when the educational industry was surviving on online teaching tools around the world at the start of 2020 due to the COVID-19 outbreak and lockdown, online learning environments such as MOOCs received widespread attention [22].

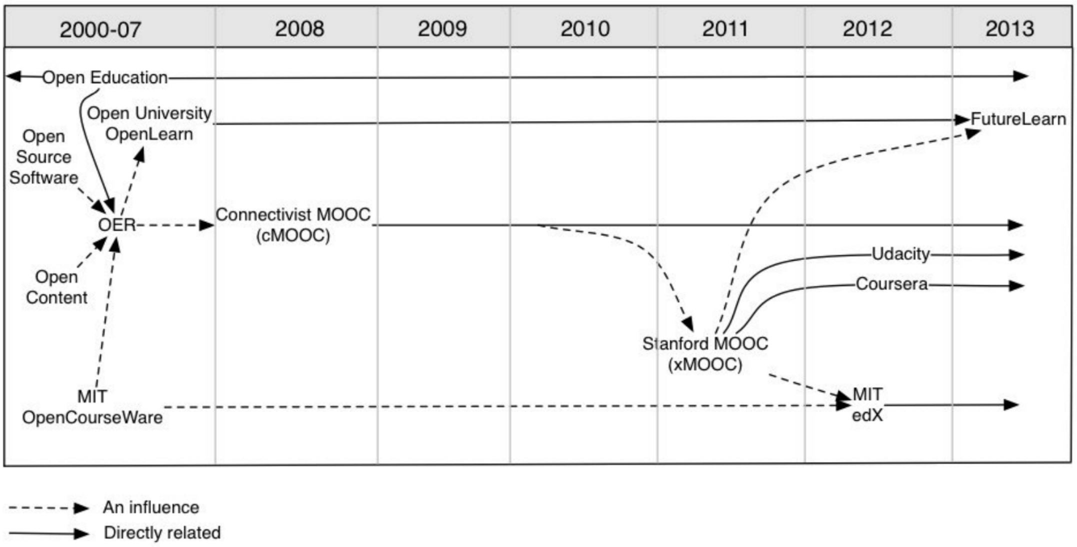


Figure 3.2: Open Education and MOOCs timeline, cited from [237]

Since MOOCs platforms bring an incredible demographic diversity of learners in one place, this diversity makes MOOC environments difficult to navigate; subsequently, this impacts the learning experience. To improve this critical avenue of a no-barriers education, it is important to build information systems that provide personalised recommendations for learners based on their personal needs.

In traditional education research, demographic characteristics have been effectively used as determinants (variables) of student achievement. This is because they are critical inputs into personalised systems. To gain an in-depth understanding of MOOC learners, MOOCs providers have begun to actively survey learners to obtain demographic profiles. Many studies of MOOCs have considered learner demographics for MOOC personalisation systems, as indicated by systematic reviews of the MOOCs literature in [156] and [120]. The problem is that when registering on MOOCs, users are always asked to take a survey to provide demographic information, but response rates as a percentage of all registrations are mostly low [37].

3.5.2 Demographics Data in MOOCs

MOOCs are educational tools that offer learning experiences to all people. MOOCs satisfy common demands and provide solutions to problems in the learning domain. They provide education as a societal service that can improve people's living standards in the long run. A primary problem in MOOCs research is that they utilise 'biased data' that does not include all demographic information of all learners. This is an obstacle to developing fair MOOC personalisation systems or delivering courses with equality and considering the diversity of MOOCs learners. This research aims to employ an automatic approach to extract learner demographic information, thereby providing unbiased data, which is vital for MOOCs personalisation purposes. This will be achieved by using and extending the most advanced approaches in the area of computer science.

All prior MOOCs studies have relied on questionnaire surveys (web-based surveys) to collect demographic data from users. However, the main issue with this approach is that the response rate for these surveys is low, which is known as the response error; this occurs in studies when the samples do not reflect the actual population for conducting a study or research relating to MOOCs. The researchers in [222] observed that response rates for web-based surveys are typically lower than surveys delivered by other modes.

This low rate response has been identified only by a few MOOC reserchers, and they developed designs to improve the responsiveness and representativeness of MOOC surveys. For example, the authors in [222] investigated whether a survey integrated into a MOOC environment could increase the response rates. Learners from six MOOCs of the University of Amsterdam were randomly allocated a demographic survey only via email or an embedded survey. In the embedded method, during the second week of the course, learners received an email with a short invitation to participate in the survey. By clicking a link in the email, the learners could then answer 17 questions about their demographics, prior knowledge, learning objectives, and motivation to use the MOOCs. If a learner has not filled the survey, 14 days later (i.e., after the first email), an email reminder was sent to the learners with a similar short text and the same link to the survey in order to increase participation [222]. The embedded survey increased the response rate from 6.9% to 61.5%, according to the findings. However, email-based surveys is time-consuming and there is no guarantee that all learners will provide their demographics.

In another study undertaken by [85], the authors demonstrated how third-party census information may be used to supplement the restricted survey data gathered by a MOOCs platform to offer a more complete picture of student background characteristics. They identified student communities based on self-reported student addresses. Then, as a predictor for socioeconomic status, they use statistics on a neighbourhood's average household income. This approach only helpful for a particular demographic characteristic, not all of them.

Other studies have solved the low response problem in web-based surveys using cash prizes in massive samples [74], [119]. However, using financial rewards can be extremely expensive, especially to obtain big samples. As a result, emphasising the reduction of such expenditure may be a more effective policy to guide MOOCs researches. So, optional quizzes (ungraded) can be included in the course environment. For example, learners may receive a questionnaire while they are actively participating in the course; they do not need to open other tabs, and content of the questionnaire is apparent to them right away [222].

To the best of the researcher's knowledge, no research has attempted to provide these data in a new strategy rather than mainly using those questionnaire surveys.

3.6 Critical Evaluation

It is important to mention at the outset that the technologies outlined in the above-mentioned studies can improve traditional MOOCs' surveys, but they are still not representative and suffer from data bias [222]. The fact is notable that almost all research on MOOCs with respect to learners' characteristics relies on personal information from learners, provided by the learners themselves in these surveys, for further analysis to know their personal needs. Whilst very useful, nevertheless, bias may arise in using such surveys, such as in the case of non-response bias, which is prevailing in MOOCs. In this research, therefore, the motivation is to classify user characteristics automatically.

One explanation that has been frequently emphasised by researchers is that there is a link between demographic factors and survey responses [164]. According to recent investigations, sociologically-relevant characteristics such as gender may influence how people respond to web surveys and MOOC surveys in particular. As a result, existing findings regarding the demographic distribution of MOOC participants may be incorrect [222], which can lead to biased results in the research.

There is evidence to show that these statistics will be skewed. Several scientific papers, for example, have reported that the number of survey responses tends to be greater among individuals who completed a course than among those who have not [46], [119]. As a consequence, people who perform poorly on assignments and quizzes tend to have lower response rates [222].

In this research, it was important to understand which demographic characteristics have a greater impact on MOOCs, as research parameters. It was found that gender, employment status, and educational level are the most critical learner demographics in relation to MOOCs (see Sections 5.2, 6.2, and 7.2). The literature indicates that MOOCs are mostly used by well-educated males who are seeking to improve in their careers. According to a study [44], students who have enrolled in MOOCs so far tend to be mostly well educated and employed, and men are more likely to use the courses than women. 79.4% of students hold a bachelor's degree or above, and 44.2% have completed post-secondary education. MOOC courses are taken by more men (56.9%) than women. More than half of the respondents (62.4%) say they are employed full-time or self-employed, while just 13.4% report they are unemployed or retired [44]. Further discussion about these particular demographics is given in Sections 5.2, 6.2, and 7.2 regarding employment status, gender, and educational level, respectively.

3.7 Epilogue

AP is a computational stylometry task that researchers have sought to solve using methods from both NLP and ML. Although the stylometric analysis area, in general, has achieved significant progress, the field of online stylometry still has some limitations due to the nature of online data, which contains substantial noise and short-length texts [1]. For AP, a deeper linguistic analysis is needed, typically involving many training samples, because the hypothesis of AP is to explore similar linguistic patterns among authors who share the same demographics [21]. Moreover, works that have achieved state-of-the-art results in AP usually utilise a large number of linguistic features [105]. This complicates the AP task in practice, particularly in the case of feature engineering. Also, online AP research in prior studies has mainly focused on social networks and targeted few characteristics such as gender, age, or native language [170]. However, other demographics such as education level, as well as important domains like education, have received less attention from the online AP community [14], [64].

AP is a research area that supports different domains, and it is thus a problem of special interest. In the domain of education, MOOCs offer a distinctive educational environment for learning compared to the traditional education system. MOOCs are more affordable, less restrictive, and more customisable in terms of location and time. In MOOCs, learners receive considerable freedom to choose an independent learning path. However, other researchers have reported that many MOOCs learners lack the self-regulation skills and abilities required to complete a programme on a MOOCs platform [12]. One way to remedy this problem is to guide the learner on the best suggestions for completing the course successfully. As a result, a MOOC recommendation system is critical to the learner's performance and helps to reduce cognitive overload [94].

Although MOOCs use open surveys for learners to specify their demographic data during registration, the actual percentage of learners who complete them is low. This creates what is called data bias in using such pre-course surveys, which arises due to non-response [37]. To fill the gaps identified in the literature, new directions of AP are investigated in this thesis. In particular, new domains are considered, along with new demographics, new NLP algorithms and tools, and a variety of non-NLP methods.

In the next chapter, the methodology is presented of how this thesis contributes and expands the area of AP to the educational domain.

Chapter 4

Methodology

4.1 Prologue

After highlighting the study's questions and objectives in Chapter 1, providing a technical background in Chapter 2, and discussing relevant literature in Chapter 3, this chapter provides an explanation of the research methodologies used to address the research questions and meet the respective objectives. This includes a description is given of the research data and the subsets used in this thesis in Section 4.2.1, including a primary prescriptive analysis of data, based on linguistic patterns of learners' comments in Section 4.2.2. A mathematical definition of the research purpose and its outcomes is then provided in Section 4.2.3. Additionally, an important aspect of the research, that of ethical considerations is discussed in Section 4.2.5. Finally, an illustration of the study's conceptual framework is presented in Section 4.2.6.

4.2 Research Methodology

This thesis investigates the use of machine learning approaches – both deep and conventional learning models – for the coined problem of learner profiling (LP). The machine learning models are fed by samples collected from an educational platform (MOOCs). In this research, a large-scale dataset is collected, which includes courses delivered by the University of Warwick via its FutureLearn platform (see Section 4.2.1). The dataset is used to identify learners' demographics, in order to help adaptive educational services, based on demographic data; this would only be applicable to very few unless more representative data are provided and learners demographics are identified.

4.2.1 Data Source

To address the purpose of this thesis[*], a large dataset was collected from courses available on FutureLearn[†], which is one of the relatively new MOOCs that has been developed since 2012. Futurelearn is a European online learning information system that facilitates remote and online learning; it is similar to the American platform; Coursera [179]. FutureLearn is a collaboration between many British universities, the British Library, and the BBC, which began in 2012. Since then, thousands of courses have been delivered on the platform by international institutions, businesses, and NGOs, which have contributed to the even greater expansion of the platform. Just before the COVID-19 pandemic in 2020, FutureLearn offered 327 courses produced by 83 different partners with more than 8 million learners.

For this thesis, permission was received[‡] to collect data from courses delivered by the University of Warwick on FutureLearn between 2013 and 2017. The authorised courses ranged across different domains. The specific courses were: *The Mind is Flat, Babies in Mind, Supply Chains, Big Data, Leadership for Healthcare, Literature and Mental Health*, and *Shakespeare and His World*. Taken together, these courses cover and synthesise different topics in social sciences, computer science, psychology, and literature. Notably, the courses cover both STEM and non-STEM fields[§], and so they represent different types of domains and, potentially, different types of learners. All of the courses were delivered repeatedly over consecutive years (called ‘runs’), resulting in a total of 27 runs [7].

It is critical to understand the nature of the data since it affects the research design. Delivering the courses over several years has allowed for a comprehensive and deep overview of learners’

[*] This research is based on ML; which needs an immense amount of data for analysing and making data-driven decisions based on input data.

[†] https://www.futurelearn.com/

[‡] Permission obtained by: Ray Irving (Ray.Irving@wbs.ac.uk) and Smith Nigel (nigel.smith@futurelearn.com).

[§] STEM is an acronym that stands for Science, Technology, Engineering, and Mathematics behaviour. Runs contain weekly learning units, and each week consists of several learning units and steps; these units or steps can be articles, discussions, videos, images, or quizzes (Pedagogical Resources). In each weekly learning unit and for any given step, learners can interact in various ways, by commenting, replying to, and liking comments from other users enrolled on the course.

When learners create an account at FutureLearn, they have the option to complete a survey about their demographic characteristics, or they can complete it later; in the second case, the survey step is skipped during registration. The demographic classes in the data are, however, scarce, since some learners do not complete all the information requested in the survey.

In addition, the system generates logs to record the learners' activities (e.g., steps, visit times, steps completed, or comments) that are correlated with their unique IDs. This is the original data that used for the LP in this research. Nevertheless, different sub-data are fetched, in order to address the different objectives of this thesis.

The resulting datasets are still large enough for the experiments undertaken in this thesis. This is because this thesis uses the largest classes in terms of the size, employment status class (Chapter 5), gender class (Chapter 6), and level of education class (Chapter 7). For the experiments in this thesis, they involved processing collected comments or other metadata from these learner IDs, which are associated with the learners' labels. Metadata was fetched only from learners whose characteristics were known, which means that the dataset is labelled, but this also means that the dataset is reduced significantly compared to its original size.

Table 4.1 provides an overview of the sizes of the datasets.

Data Subset	**Metadata**	**Runs**	**Courses**	**Users**	**Samples**	**Chapter**
Employment Status Profiles	Comments	27	7	9,538	381,298	Chapter 5
Gender Profiles	Comments	27	7	7,524	322,310	Chapter 6
Education Level Profiles	Comments, Time-Stamps, and Quizzes	21	4	12,984	Samples vary *	Chapter 7

Table 4.1: Overview of data subsets used in this research

*The samples in this dataset vary based on the features and courses. See Section 7.3.1, Chapter 7 for more details.

4.2.2 Preliminary Prescriptive Analytics

4.2.2.1 Employment Status

Identifying users' demographic characteristics based on textual features is pursued in this section by aiming to capture different patterns among user-generated texts, simply by measuring basic linguistic features. This step can usefully serve as an initial step [108], and to disclose any linguistic patterns among categories of employment status (i.e., working, not working, retired, as these categories will be discussed in Section 5.3.1), which is because it can provide an overall understanding of textual features. As shown in Table 4.2, this analysis covers the demographic categories to present their text distributions on the reported numbers of characters, words, and sentences, respectively.

	Not working	Retired	Working
Character-level			
Mean number	299	290	318
Minimum	1	1	1
Maximum	1,299	1,319	1,311
Word-level			
Mean number	53	51	56
Minimum	1	1	1
Maximum	258	263	247
Sentence-level			
Mean number	3	3	3
Minimum	1	1	1
Maximum	32	37	36

Table 4.2: Numerical representation – Basic textual pattern based on employment status categories

Unfortunately, this step did not provide any opportunity for interpretation, which is because it did not show any difference among the three categories, nor did it show an noticeable difference between the parameter 'number' concerning the writing style of the different groups (see Table 4.2). The three textual levels' minimum values are identical; maximum values show only minor variances, but still indicate more variation than mean values.

4.2.2.2 Gender

For the gender data, basic linguistic features were also measured, such as the number of tokens, including POS tags, in order to disclose any linguistic pattern in males and females. This analysis also contributed to answering one of the research questions mentioned in Section 1.4 regarding the question of whether a learner's gender can be inferred from their comments only. Here, the aim was to apply simple approaches, i.e., to confirm if ML approaches are more appropriate for solving the problem.

Specifically, in this thesis, the minimum, maximum, and mean values of different levels of tokens in comments were calculated for each gender category, as presented in Table 4.3. The table shows the normal distributions of each token level in the comments per category. According to the table, there are limited differences between males and females suggested by this basic investigation. It appears that they both have the same minimum numbers for the three levels. In the mean number, it seems that women write slightly more than men. However, the minimum values indicate that males and females are identical across the three levels.

Counts	Females			Males		
	Characters	Words	Sentences	Characters	Words	Sentences
Minimum	1	1	1	1	1	1
Maximum	1,319	263	90	1,311	254	83
Mean	293	52	4	344	60	5

Table 4.3: Numerical representation – Basic textual pattern based on gender categories

4.2.2.3 Educational Level

For the total data collected for the educational level experiment, which is presented in Table 4.1, the purpose of this step was to establish a general understanding of the different patterns among the users' texts. For this reason, all the data obtained from different courses were grouped, to facilitate the measurement of overall and basic linguistic features among the various categories of educational level (i.e., bachelor's, master's, and doctorate) (see Section 7.3.1). In Table 4.4, the only observable difference is apparent in the doctorate category. This may indicate that this group of users often writes shorter comments in length, on all three token levels. For Bachelor's and

Master's students, it seems that they both have a similar pattern of writing for the three levels of tokens, and no noticeable difference is indicated.

	Bachelor's	Master's	Doctorate
Character-level			
Mean number	302	298	186
Minimum	1	1	1
Maximum	1,315	1,320	1,105
Word-level			
Mean number	57	55	12
Minimum	1	1	1
Maximum	262	266	199
Sentence-level			
Mean number	2	2	2
Minimum	1	1	1
Maximum	36	41	23

Table 4.4: Numerical representation – Basic textual pattern based on educational level categories

To summarise, when considering the above-mentioned findings, they show the task's complexity and the lack of evidence of variance, according to this basic analysis point of view. This highlights the need to apply more advanced approaches, such as conventional ML or deep ML, to solve the main research questions, as well as the sub-questions (see Sections 5.3, 6.3, and 7.3), what this thesis is addressing.

4.2.3 Mathematical Definition

LP, which is a focus of this thesis, is a identification problem; one way to approach it is supervised learning. Supervised learning models may either be conventional ML or DL models, which – in conceptual terms – gain knowledge from a set of labelled samples in a dataset; this enables them to classify samples with their correct labels.

Technically, the process involves extracting common patterns from feature vectors that are used to represent these samples and match the patterns with new samples. In this way, when an unseen sample is provided, the model can identify which class the sample belongs to [13]. Thus, the classification models used in this thesis are fed via inputs, which are a combination of Samples (i) and their extracted Features (j), and which are processed within a matrix X that is represented in a domain R as follows:

$$X \in R^{i \times j}. \tag{4.1}$$

The classification process itself can be defined as a mapping function (f) as follows:

$$f = X \to Y, \tag{4.2}$$

where X is a collection of inputs (i,j) and Y is a fixed set of classes $y_1, y_2, .., y_n$, which correspond to the demographic categories in this research (these are presented later in Chapters 5, 6, and 7). An input $x \in X$ is a sample with a vector representation of features that belongs to a MOOC learner, and an output is a identification of a learner category $y \in Y$, where $f(x) = y$. The target is the correct category $y' \in Y$. If $y = y'$, this confirms that the identification is correct.

4.2.4 Performance Evaluation

The task of collecting labelled data is usually separated into three processes: training, validation, and testing. As the name indicates, the training data are used to train the model, while the validation set is used to fine-tune the model parameters and select most representative features to solve the problem. After model training, feature selection, and hyperparameter tuning, the test set is used to conduct performance measurements for the model [13].

In addition, when the dataset is small and the classifier is simple (such as in Chapter 7), K-fold cross-validation is a useful process for evaluating a model. K-fold cross-validation involves dividing the training data into K parts, or 'folds,' and training the model K times, each time leaving a different fold out. After each repetition, the model is used to identify the labels of the data that were left out [177].

After training, a performance metric is used for final evaluation. Typically, AP approaches are evaluated by computing their performance metrics based on *overall accuracy* [192], [166], but additional performance metrics are considered in this thesis, including *precision, recall, and F1*, as discussed in later chapters. The equations below explain the evaluation methods mathematically:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.5}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{4.6}$$

Where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives. **Accuracy** is ratio of the correctly predicted samples to the total samples in the used data. **Precision**, is the ratio of the number TP divided by the sum of the number of TP and the number of FP. **Recall** represents the ratio of the TP divided by the sum of the TP and FN. **F1-score** is a weighted mean of both precision and recall.

4.2.5 Ethical Considerations

Profiling users' demographics has many positive and potentially beneficial applications for society [105]. Unfortunately, however, as in so much of the research literature, new technologies can lead to negative outcomes, such as undermining privacy, identity theft, racial profiling.

The opportunity to apply LP in education systems can lead to many positive and beneficial outcomes. In the current research, the author is explicitly working on positive outcomes, as this research seeks to improve AP in the targeted application of estimating the demographic characteristics of learners using MOOC platforms. For instance, learners profiling on MOOCs may provide beneficial tools for stakeholders, and learners can be directed towards not missing out on opportunities or offers. Stakeholders can also be assisted in the domain of providing the best recommendations for learners based on their associated categories, as identified from the LP models applied in this research.

It is worth noting that the data collection procedure adopted in this research was based on general ethical standards, such as avoiding any potential harm to the research participants. The data were labelled via a self-report survey collected by learners themselves, and the researcher received all necessary permissions of use*, which allowed the use of this information for this research purposes. As such, demographic data were already declared by the learners by themselves (a trusted source of labelling), and they were aware that the information would be used for research purposes via Standard FutureLearn Practice. It was also made clear to the participants that there would be no undue influences, such as awards, and their participation was completely optional.

The author of this thesis is aware of the ethical issues raised in using profiling technologies as they could support undesirable or even illegal practices, including infringing upon privacy. Hence, before using profiling technologies, de-anonymisation techniques were applied, in order to achieve a balance between utility and intent. Personal information such as initials, email addresses, and phone numbers was not collected or handled by the researcher in this study. The researcher did not have access to such information and was only able to fetch demographic information. Therefore, high-level personal information from the participants was still anonymous and could not be identified.

4.2.6 Conceptual Framework

The research aim of this thesis is to identify demographic information about the MOOCs platform users (i.e., learners) using automatic approaches (i.e., author-profiling). This process is applied in order to generate representative data about learners that can be exploited to improve the outcomes of MOOC personalisation systems.

Selecting an appropriate research design is critical, and it should be based on the nature of the problem to be solved. The research methodology was designed based on the knowledge gained from the literature discussed in Chapter 3 related to this research's scope and research's questions (Section 1.4, Chapter 1). Accordingly, data were collected and prepared to meet the research

*(Via Standard FutureLearn Practice (https://www.futurelearn.com/info/terms/research-ethics-for-futurelearn), and from the Warwick University side. Permission granted by: Ray Irving (Ray.Irving@wbs.ac.uk) and Nigel Smith (nigel.smith@futurelearn.com)

questions and objectives stated in Chapter 1. This also guided the research in terms of what demographic information to identify (Sections 5.2, 6.2, and 7.2), what features can be extracted, and how to achieve this – as well as what models to employ and how (Sections 5.3, 6.3, and 7.3). This also assisted in covering different data-driven approaches, based on what was available in the research data. Figure 4.1 illustrates the thesis' overall conceptual framework.

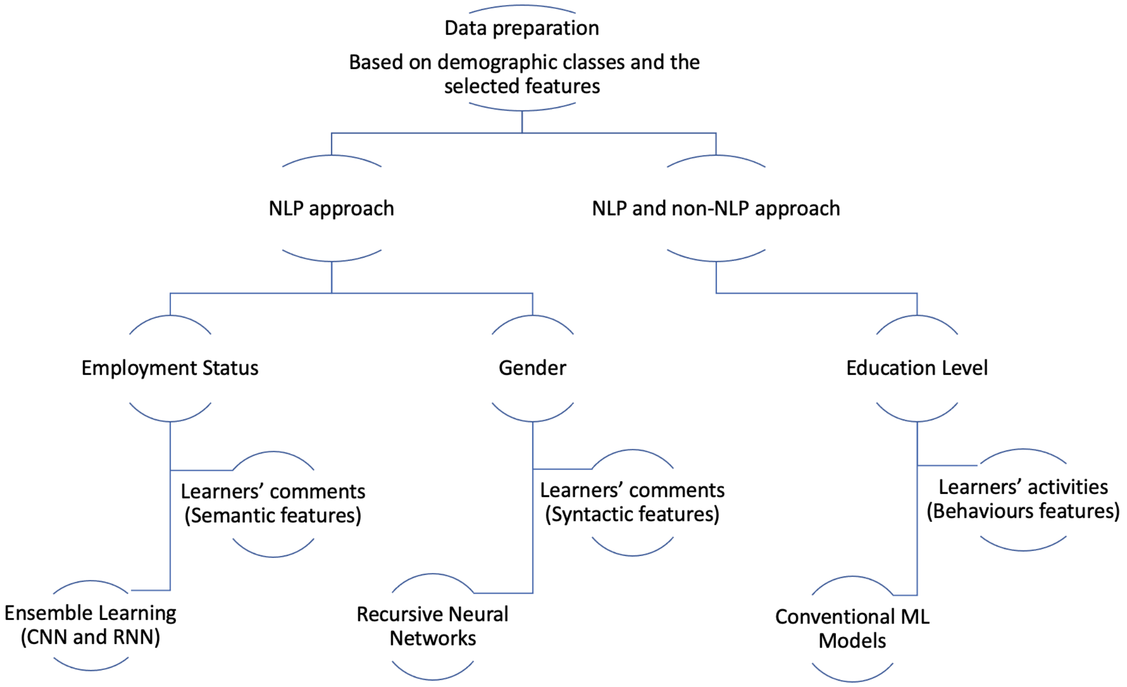

Data preparation
Based on demographic classes and the selected features
NLP approach
NLP and non-NLP approach
Employment Status
Gender
Education Level
Learners' comments (Semantic features)
Learners' comments (Syntactic features)
Learners' activities (Behaviours features)
Ensemble Learning (CNN and RNN)
Recursive Neural Networks
Conventional ML Models

Figure 4.1: Overall conceptual framework for the research

As shown in Figure 4.1, this research examines different types of textual features (specifically, learners' stylometry features) extracted from forum discussions on MOOCs (Chapters 5 and 6). This is in order to answer the two research questions RQ1: Is it possible to classify employment status based on learners' comments exchange in MOOC discussion forums?, and RQ2: Can advanced textual features extracted from MOOC discussion forums be used to classify a learner's gender?.

In addition, instead of exclusively focusing on only one type of metadata concerning MOOCs (To solve research RQ3: Can a learner's level of education be classified based on a MOOC discussion forum data on a course-level classification?), this research also investigates additional metadata, thereby reducing the heavy reliance on learners' comments by also leveraging other learner activities for the LP. This facilitates answering research question RQ4: Can the use of metadata in addition to the MOOC discussion forum data improve the classification of learners' level of education? (Chapter 7).

Different approaches are considered in this research for models' evaluation, including the use of evaluation metrics, comparison of different models performances, and assessment of features importance when appropriate. Full details of the methodological decisions towards answering each research question approaches are presented in the implementation chapters (Chapters 5, 6, and 7). An overall discussion of this thesis, including its limitations and recommendations for future work, is presented in Chapter 8.

4.3 Epilogue

Many studies that have investigated MOOCs have attempted to analyse learners based on their comments or other activities [116]. However, prior studies have mainly aimed to identify user behaviours, often using survey data to infer the users' demographic characteristics. In this thesis, learners' posts and activities are investigated from a different angle compared to the approaches used in earlier works on MOOCs. For this research, a large-scale dataset was constructed, where the original data source had different demographics (gender, employment status, and educational level were considered for this study), in order to achieve this study's objectives. The employment and gender classification models in this thesis are methodologically similar; using textual features, the educational classification is radically distinct because it includes non-textual features. The first two models use ANN algorithms to learn textual features implicitly, while the latter employs a feature engineering approach, in which features are specifically selected based on the nature of the dataset, before being fed into a conventional ML model. Despite the fact that AP models, usually, are based on textual forms of data (see Section 2.5), the author propose in this work to also consider other metadata, especially if such data is available, to study their effect on LP. In the next three chapters, more explanation is presented of the applied models and approaches for this thesis, as well as the results of the analysis, starting from Chapter 5, which presents employment status classification models.

Chapter 5

Deep Sequential Learning and Paraphrasing Approach for Employment Status Identification

5.1 Prologue

According to [109] and [178], the most common types of learners who are attracted to MOOC platforms are those who are seeking to enhance their *professional skills*. The employment status of a learner can inform MOOC personalisation systems in terms of what type of professional skills a learner might target. For this reason, it is important to have correct information about a learner's employment status, but a critical challenge is that this information is not available for all learners (see Section 1.3.1). To identify a learner's employment status in MOOCs, the researcher applied state-of-the-art models in the NLP area – namely, DL approaches (CNN and RNN). The researcher compared two basic styles of ensemble learning (parallel and sequential architectures) to show their respective performances in classifying learners employment status. This work only uses simple word tokens in semantic representation from learners' comments. However, in such a case, it is an essential requirement for a DL model to have a large data size (samples) to learn from; notably, the data used in this chapter met this requirement. One of the limitations identified relating to these comments is the imbalanced distribution of class categories. Thus, a new data augmentation technique for the area of demographic profiling is studied in this thesis.

In this chapter, the employment status classification experiments are described. First, the importance of the study is introduced. The study's methodology is then described. Following that, the processes of data collection, analysis, and preparation are detailed. In a turn, a description of the models architectures are provided. Finally, evaluation processes, including the obtained results, are discussed.

5.2 Importance of Employment Status Profiling for Learners

MOOCs are viewed as part of people's lifelong education and training. A majority of learners who take part in MOOCs are seeking *knowledge* or expertise [109], [178]. As a result, 60% of individuals who enrol in a MOOC are eager to develop their skills for their professional career [238], as well as to support their curriculum vitae [144]. This aligns with one of the objectives of MOOCs, namely, to offer a democratised education for all, especially for those learners who are economically unable to afford a high-quality education – which expands their career opportunities [44]. This could range from basic language learning to specialised technical IT skills. MOOCs are a useful way to stay current with industry and market developments [160]. In addition, it is clear that the COVID-19 pandemic has had a negative impact on unemployment rates and young graduates' career prospects. New employment and business models that include new employee skills are likely to emerge after the pandemic [22].

Thus, *employment status* is a considerable factor for completing courses in MOOCs [144]. Employment status impacts whether a user is likely to be completing a course to assist with job-seeking, research, brain-training, or entertainment. Similarly, learners who may have more free time are likely to have very different goals stemming from contrasting needs. As a result, there is an opportunity to personalise MOOC platforms in general depending on variances in employment status. MOOC courses, content, tools, steps, and other course elements can be modified to satisfy the needs of the population of MOOC learners who register for job-related goals, depending on their differences and preferences.

In the AP literature, occupation identification in general has received limited attention, as discussed in Section 3.2.2. Thus, this chapter contributes to the AP literature by investigating less examined user traits. In Section 3.4, Chapter 3, it was also shown that DL models are very popular in other NLP applications [241], but they have received less attention for solving AP applications. Conventional ML, especially SVM, has been the state-of-the-art for occupation classification in previous AP studies [188], [163]. However, there are insufficient studies in the AP area that have examined DL models. For example, in PAN 2019 [170], only three studies out of 55 considered DL models, which is an extremely low percentage. In other areas, DL models nowadays are widespread solutions, including for NLP tasks [54], [42].

To fill the identified gap in the literature, this chapter considers DL models for learner employment status profiling. This is feasible in this chapter due to the fact that the researcher has access to large samples from the target domain (namely, MOOCs). Specifically, to increase the likelihood of achieving high accuracy using DL approaches, models were selected for examination that were previously proven to be successful in NLP classification tasks [235]. It is an ensemble modelling of CNN and RNN (BiLSTM). It also has been proposed relatively recently for different traits for AP, but not for employment status classification.

5.3 Identification Methodology

A recent study [209] demonstrated that CNNs models are more effective in an embedding space represented by tokens compared to other models. This is due to the fact that it is not required in CNNs to have any knowledge of the structure of the language. Additionally, CNNs perform well in online data texts as they are reasonable at handling independent features, such as new words in a language [240]. On the other hand, RNNs models are considered effective for sequence modelling, such as analysing a sequence of words; handling semantics tokens; and handling sequential aspects of the data [235], based on the position/time index in a sentence [30]. However, they are still not effective enough to handle small parts of texts (e.g., characters), especially compared to CNNs [240]. As a result, a combination of the CNNs and RNNs in an ensemble technique could provide complementary information about the writing style features of a learner; and modelling semantic information of a text globally and locally [41]. Furthermore, RNNs have been improved through LSTMs [89], which has proven its superiority in memorising long length sequences. It is also one of the most powerful neural network architectures in NLP [55]. Thus, combining CNNs and LSTMs in an ensemble mode has shown higher rates of accuracy on many NLP tasks and many benchmark datasets, including IMDB[*]:(91.8%), Subj[†]:(94.0%), and TREC[‡]:(97.0%) [54].

As presented in Section1.4, the umbrella question in this chapter is:
RQ1: Is it possible to classify employment status based on the comments that learners exchange on MOOC discussion forums?

The sub-questions for this research question were:

- Is it possible to classify a learner's employment status from only the comments they exchange on the MOOC system based on a DL approach?
- Can the data imbalance issue among learners' comments based on their categories be solved using an NLP approach?
- Which DL architectures are viable to use for classifying learners' employment status?

In terms of this chapter's contributions, it combines the NLP and DL models to propose a new area, namely, the learners' employment status, based on available data; and only using comments. This is achieved using the sequential architecture of a CNN and an RNN in an ensemble model, which enables learning the semantic representation of comments, presented by an NLP algorithm, called Global Vectors for Word Representation (GloVe) [158]. Also, an NLP-based strategy is used to balance data for the first time in the area of AP in general, thereby promoting the high accuracy of this study.

5.3.1 Dataset for Employment Status Profiling

In this chapter, the difficult problem of classifying the employment status of learners was addressed based only on their comments. Comments are available metadata and ubiquitous across

[*] https://ai.stanford.edu/
[†] https://www.cs.cornell.edu/people/pabo/movie-review-data/
[‡] https://emilhvitfeldt.github.io/textdata/reference/

MOOCs [11], [9], and also discussion forums are one of the most common sources of metadata used to analyse MOOCs. Information from these forums is regularly used in learning communities for education and social interaction, and it simultaneously provides rich metadata for researchers to study learners and their needs [11]. Furthermore, the size of the labelled comments for the employment status categories used in this research is huge (see Section 4.2.1), which is appropriate for DL models.

Only comments from learners with IDs associated with their labels were collected. However, categories in this data vary in size (i.e., they are imbalanced data). This also meant that the dataset reduced significantly from the original size, yet was still large in size. The researcher gathered these samples from 27 runs of 7 courses, totalling 381,298 comments from 9,538 users. The data were labelled by the learners themselves based on an open survey at the beginning of each course. There were several types of work statuses (eight categories): Retired, Working Part-Time, Working Full-Time, Not Working, Self-Employed, Looking for Work, Unemployed, and Full-Time Student, as defined by the FutureLearn platform for the options available. For simplicity, the researcher further grouped these into more general types for the professional profile, as follows: Retired, Working (working part-time, working full-time, and self-employed) and Not Working (not working, looking for work, unemployed, and full-time student). This was done due to the fact that some of the original, fine-grained FutureLearn statuses were hard to differentiate and slightly ambiguous – such as ‘looking for work’ versus ‘unemployed’, and so on. Also, it is due to aiming at reducing the complexity and running time in the classification because computing a classifier is computationally expensive when the number of training instances is substantial [103]. In brief this research seeks to learn an embedding function that estimates whether a given comment originates from employed, retired, or unemployed learners, using a dataset of comments only.

5.3.2 Problem Definition

It is possible to define this problem mathematically as a mapping function (f):

$$f = S \to C \tag{5.1}$$

where S is a collection of samples (i.e., a vector representing the textual features of a learner's comments) and C is a fixed set of classes: $C = c1(Working), c2(Not - Working), c3(Retire)$, which are the class categories that this chapter's study aims to identify. The target is the correct class $c' \in C$. If $c = c'$, it follows that the classification is correct. The accuracy of the identified class is measured by a simple distance function such as softmax, which is shown as follows:

$$Distance = \begin{cases} 1 & if\ c = c' \\ 0 & if\ c \neq c' \end{cases} \tag{5.2}$$

5.3.3 Dealing with Bias

As the data for this thesis were labelled based on pre-course questionnaires filled in by the learners themselves, this represents traditionally higher human accuracy in terms of annotation [132]. This approach to obtaining labelled data is used widely in ML researches [182]. In addition, the researcher collected comments from each learner within six months of the date of registration, whenever possible, in case the learner's employment status changed. Duplicated comments were excluded and only the first comment was retained (i.e., the original comment written by a learner). This was because some learners, were found, copied and pasted other learners' comments. This meant that these copied comments were not written in their own personal style of writing.

Next, the researcher started with the relatively basic separation of the data into training and test sets. To further minimise any bias (e.g., by learning about the learners instead of the type of class), it was ensured that no comments written by the same learners were included in both the training and testing set at the same time. This warranted the use of independent samples in both the training and test sets to evaluate the model's generalisability and its achievement of unbiased results. Therefore, the researcher collected comments from only one run from each course for the testing dataset. This is because, mainly in each run, there is a different group of learners. Also, this provides enough samples for the test set. Data were used from the remaining runs for the training set.

After balancing the training data (as explained in Section 5.3.3.1), and to obtain the same class proportion, shuffling and stratification were applied to improve learning performance [110]. Stratification sampling separates the observations into homogeneous groups (by label), which balance the number of categories' samples in each batch. The samples are shuffled to mix up the order of the samples based on labels, which warrants that a sample has a chance to occur at any position in the data.

As a result, for the professional profile dataset, there were 320,483 comments in total from 6,969 users used for training (retired: 154,527, workers: 117,138, and non-workers: 48,818). After balancing this dataset, it was further divided into actual training (80%) and validation (20%). For the test dataset, there were 60,815 comments from 2569 users, of which retirees accounted for 17,302 samples, workers for 25,706, and non-workers for 17,798).

5.3.3.1 Text Augmentation

To improve performance for supervised learning problems [70], samples must be unbiased toward one class in order to reduce a tendency toward predicting the majority class. However, the classes in this study's training dataset were imbalanced. Therefore, to avoid expensive options in terms of time and money, more samples were produced using a text augmentation technique for oversampling. Specifically, sentences were paraphrased from the smaller size categories. To do so, large comments were tokenised using '.' for tokenisation from those minority groups and paraphrasing each sentence, and this took place until the same number of instances was achieved as in the majority class. Words were replaced using their synonyms and expressions by their paraphrases to generate new comments. To assist with the completion of this task, the paraphrase database PPDB was used [75], with over a billion paraphrase-pairs in total covering several languages. The idea behind this database is that if two strings $S1$ and $S2$, written in a language A, have the same translation f in another language B, then the pair $< S1, S2 >$ has the same meaning. As such, $< S1, S2 >$ can be extracted as a pair of paraphrases. In other words, words were replaced using their synonyms to generate new comments.

Also, this NLP technique (namely, the paraphrasing technique) was compared with the popular oversampling technique, which is known as the Synthetic Minority Oversampling Technique

(SMOTE) [40]. Hence, a training set was obtained with balanced samples to create a fair model during the training phase and reduce bias.

5.3.4 Data Pre-processing and Normalisation

Experiments were undertaken with different common AP normalisation steps on the MOOC data. However, based on the experiment in this study, certain steps were identified that had a significantly greater effect on classifier performance compared to others. It is important to remember that these normalisation processes should not harm the learners' writing style and, on the other hand, they should be supportive when using NLP libraries. For instance, some NLP libraries utilise white space to split tokens in Python, while others may need the correct form of words to extract the meaning or semantic content of the words.

Therefore, a pipeline of text normalisation was created to be used by other models presented in this thesis, and all comments were pre-processed. Pre-processing steps were applied that are commonly used for NLP tasks [171], [170]. More specifically, the pipeline steps are as follows:

- **Step 1**: As contractions often exist in English texts, the researcher expand these shortened versions of words in order to standardise the comments [130]. To illustrate, a phrase such as 'I'll be happy!' becomes 'I will be happy!'

- **Step 2**: All occurrences of URLs and hyperlinks were replaced with the string "URL" [130].

- **Step 3**: Special characters and punctuations can lead to noise in texts; therefore, the researcher separated all non-alphanumeric characters from words [50]. For example, 'Shakespeare course is interesting!' becomes 'Shakespeare course is interesting !'

- **Step 4**: The researcher used an adaptation of Peter Norvig's spell checker [*] to correct all typos in the comments.

- **Step 5**: The NLTK Tokenizer was used to tokenise words, after which the zero-padding strategy was applied based on the work of [42]. This created identical vectors lengths for all

[*] https://pypi.org/project/pyspellchecker/

comments. Using the length of the mean words sequences (here, 70 tokens), padding was applied to all sequences to ensure a uniform vector size for all vectors in the dataset, which is an important step in ML.

- **Step 6**: Before training classical models (such as SVM), weighting schemes were used to generate textual features for the models. More details are provided in Section 5.3.5.
- **Step 7**: In this semantic representation step, word representations (in this case, GloVe(300d)) were used for word input embeddings for DL models [158]. As recommended by [111], it generates a matrix of words based on co-occurrence statistics (see Figure 5.1).

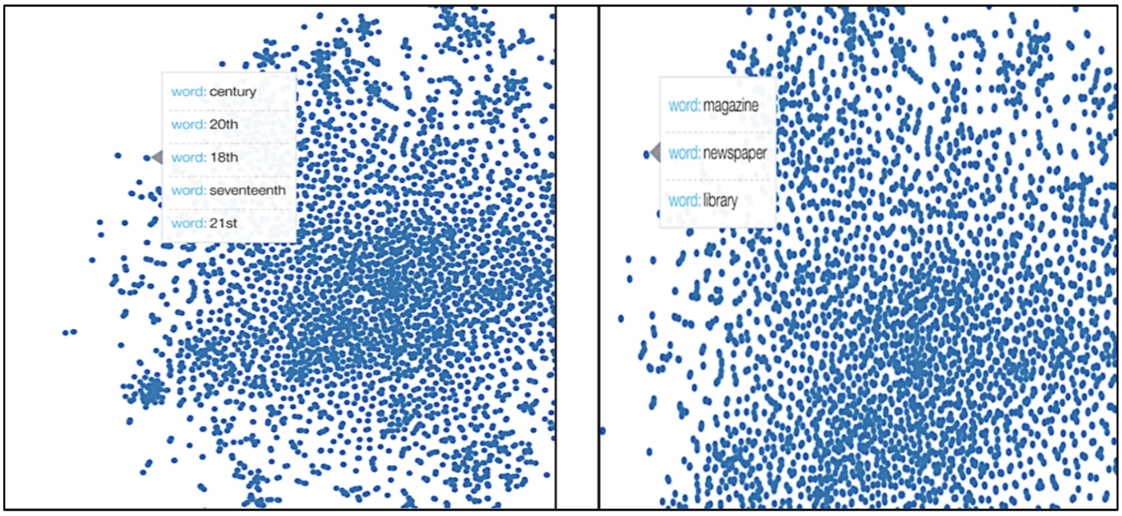


Figure 5.1: Word embedding using GloVe – Sample from the used MOOC dataset

The pre-trained GloVe algorithm was used, which gives pre-trained weights for the inputs (i.e., transfer learning) instead of starting from random weights (i.e., learning from scratch). These initial inputs are fed to the neural network models to provide semantic information, here regarding for word-level. This also converts text to numbers, which are numerical vector representations with numeric indexes to the tokens. In pre-trained models, any unknown word is treated as "unk". The UNK token here is initialised as the average of all embeddings of token vectors. Figure 5.1 shows an example of GloVe embedding.

It is important to mention that other NLP pre-processing steps such as stemming and lemmatisation have been examined. However, they were not used in this study because they were found to affect the models negatively.

5.3.5 Weighting Schemes

Different weighting schemes were applied in this chapter that are commonly used for feature extraction for traditional classifiers (conventional ML) [170]. Here, vector-based models are used to extract features from texts and convert them into matrices, which enables the representation of texts in vector space models. After that, the extracted features can be used as inputs for a classifier. Term Frequency-Inverse Document Frequency (TF-IDF) [86] was used, which is a baseline weighting schema in NLP and AP area [14]. Three forms of TF-IDF were applied, along with a simple weighting schema (namely, Word Count). The difference is that the word count generates vectors based on a word's occurrence in a corpus, while TF-IDF generates vectors based on a token's frequency. TF-IDF calculates two scores, tf and idf, which are combined into the TF-IDF formula, as follows:

$$tf(t,d) = \log(1 + freq(t,d)) \tag{5.3}$$

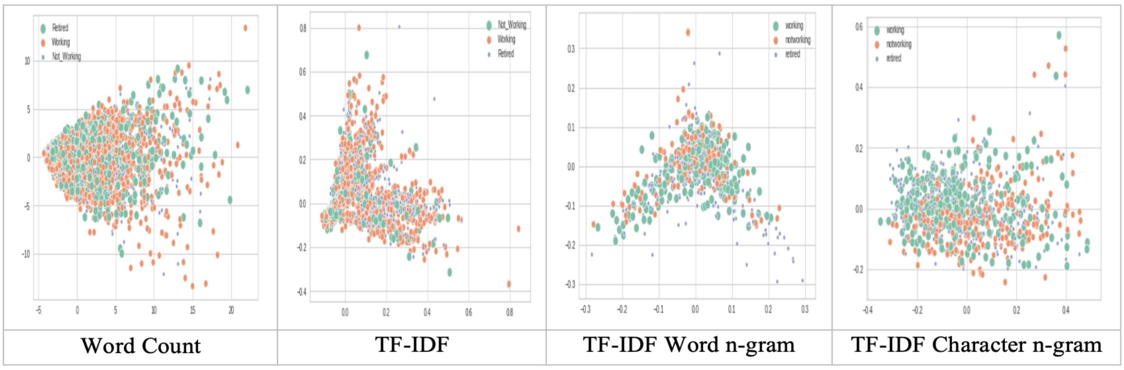
$$idf(t,D) = \log(\frac{N}{count(d \in D : t \in d)}) \tag{5.4}$$

$$tfidf(t,d,D) = tf(t,d).idf(t,D) \tag{5.5}$$

where t is a term in document (a sentence) d and D is the entire text corpus or *Dimension*.

TF-IDF is based on n-grams and can be based at either the character or word level. This generates TF-IDF vectors for each gram. It is well known in the AP domain that n-gram models increase classification accuracy, which is because they take into account sequences of words [161], [28]. For deeper representation of texts, this chapter examined n-gram TF-IDF based on $n = 2,3$ for characters and on $n = 3,4,5$ for words, as recommended by [161] and [28].

The issue of the above-mentioned schemas is that they generate numerous features. While this is typically a positive thing for training a classifier, the large number of features leads to more variance (noise). Therefore, the use of Principal Component Analysis (PCA) was considered, which is a popular dimension reduction technique. Mathematically speaking, PCA transforms each data point represented by a vector x and features n into a vector z with fewer dimensions m. This can be done through a linear transformation. Figure 5.2 is a simple visualisation of PCA-based vectors obtained for the three classes (working, not working, retired). These final feature representations are fed into the conventional ML classifiers.



Word Count | TF-IDF | TF-IDF Word n-gram | TF-IDF Character n-gram

Figure 5.2: Dimension reduction via PCA in the space of weighting schemes

5.3.6 Model Architecture

Since the initial step in Section 4.2.2.1, Chapter 4 did not explain any differentiation of textual features among learners, applying more advanced solutions is required. In this research, ML classifiers were applied to solve the problem. Although conventional ML involving learning by probabilistic representations of samples of text has proven to be effective in AP tasks in general [170], DL models have been less examined as AP solutions.

In this chapter, the performance of parallel ensembling (see Figure 5.3) was compared to the sequential ensembling of CNN and LSTM (see Figure 5.4). The sequential ensemble model was inspired by work done by [49], wherein both CNN and LSTM were utilised by means of an ensemble DL architecture in a sequential manner for a text classification task. The sequential ensembling architecture is presented in Figure 5.4. Different settings were tried and tested to tune the parameters for the CNN and RNN models, but only a description of the final settings selected for these models is given. Each of the CNN and RNN models in this study had identical settings when reused in different architectures. The sequential model architecture is described in the following sections.

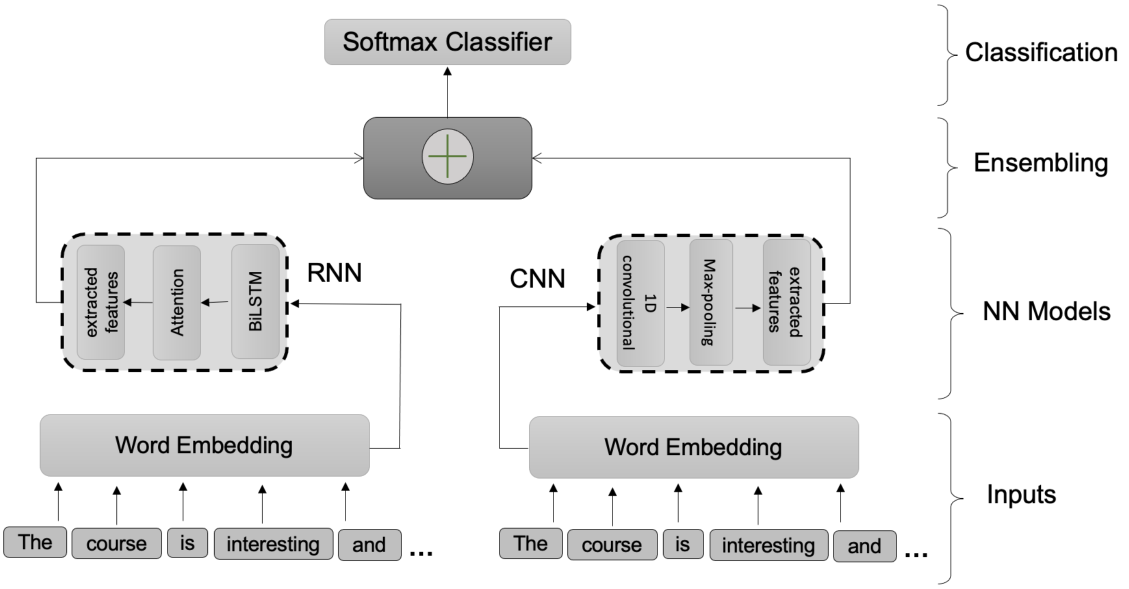


Figure 5.3: Parallel architecture of an ensemble learning: RNN and CNN

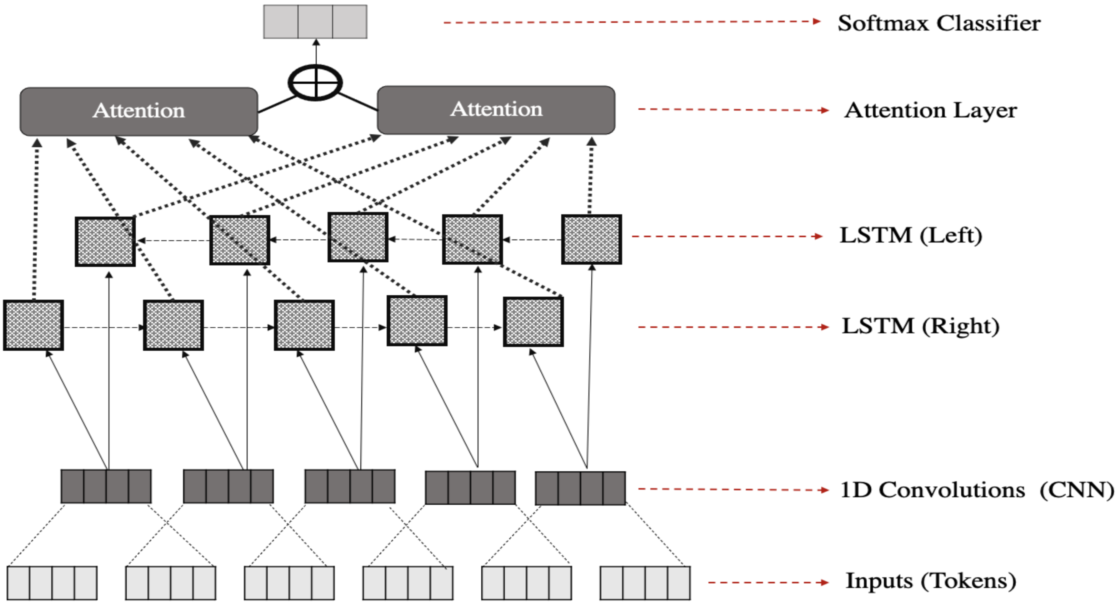


Figure 5.4: The sequential ensemble learning architecture

5.3.6.1 Embedding Layer

The first layer in the model is the embedding layer. In this layer, an embedding matrix is created and serves as a lookup table to search for words by indexes. It maps each comment sequence onto a real vector domain. Thus, an entire comment representation (X) is mapped to a matrix of size $s \times d : X \in R^{s \times d}$, where s is the maximum number of words in the longest comment (s = 70) and d is the embedding space dimension. It is common in text classification to have 256, 512, or 1,024-dimensional word embedding [9] when the data size is large. However, GloVe is utilised in this study, and the embedding size should be exact as the GloVe vectors size: d = 300. Using semantic pre-trained embedding yields better results and accuracy compared to randomly initialized embedding as it holds semantic embedding instead of random embedding [158].

5.3.6.2 Convolutional Neural Network

The second layer is a hidden layer containing the convolutional model. A 1D CNN was used with the strides set as equal, and valid padding was applied. For pooling, to extract the most important n-grams within the embedding space, the widely applied max-pooling process was used. The max-pooling operation also provides a combination of all pooling in each filter into one vector.

The final vector obtained was fed to a fully connected layer (FC). Experiments were undertaken with different numbers of neurons (i.e., gradually increasing from 10 to 50 neurons in the FC), and 50 neurons perform best, followed by ReLU as an activation function. To reduce overfitting, a dropout hyperparameter was added, which randomly removes words in sentences and forces the classification not to rely on any individual words. The dropout value was set as 0.5, which is considered a suitable dropout setting for many neural networks and ML tasks [205]. The final merged output matrix that is the output of the CNN model was fed as input to the RNN model as part of the sequence in the ensemble learning architecture.

5.3.6.3 Recurrent Neural Network

The RNN model consists of a Bidirectional Long Short-Term Memory (BiLSTM), plus an attention mechanism [225].‘Vanilla’ RNNs are known to suffer from the vanishing gradient problem. Thus, LSTMs were chosen for this research as they can solve this problem due to their complex internal structure and their ability to remember either long-term or short-term information [89]. To further enhance the LSTM structure and enable it to consider past word information, the bidirectional strategy was applied; this involved deploying two LSTMs to feed the data inputs in two different directions (one to read sequences forwards, the other to read sequences backwards). In other words, the system reads from the past to the future and from the future to the past, plus has an attention mechanism. Inputs of the two-LSTMs are then stacked together to better understand the token sequences. The BiLSTM in the model has 100 hidden units in total (50 neurons in each direction), and the likelihood of the dropout rate is set to 0.5 to regularise learning; this is followed by an FC layer of 30 units and activation functions (in this case, ReLU). It is known that the ReLU activation function is able to learn more quickly than other activation functions such as tanh and sigmoid [30]. The FC layer is also followed by a dropout layer (0.5).

5.3.6.4 Attention Mechanism

The combination of BiLSTM and the attention mechanism [54] is another way to improve classification accuracy. By assigning different weights to different pieces of contextual information, the attention mechanism distinguishes the important from the irrelevant. The attention technique is applied to provide separate emphasis to information derived from the forward hidden layer and the backward hidden layer in the BiLSTM.

For both CNN and RNN models, the models were integrated with the use of the Adaptive Moment Estimation (Adam) optimiser. As opposed to other stochastic optimisers, Adam's learning abilities are quicker and more reliable. It is built on gradient descent and maintains an adaptive learning level for each parameter [30]. To calculate classification error, two loss functions were used: Kullback–Leibler Divergence and Categorical Cross-Entropy. As there are three target categories in the present chapter's study, both loss functions were a viable option; both were found to produce similar results in this study.

5.3.6.5 Classification Layer

This layer has the probability distribution that describes the likelihood of each category based on the output features. In this chapter, the flattening layer was used for the representation of the output data generated from the previous layer to be fed into a softmax classifier (final classification layer). The softmax function is best used with the last layer of classification [61], which is because it uses the probability distribution of categories as a set of numbers between 0 and 1, whose sum is 1.

In summary, the GloVe representations, in the best performing model (ensemble sequential model), are fed into the CNNs to extract the most important embedded tokens. Next, the CNN layer outputs become the inputs for the Bi-LSTM, which is important for handling the sequencing of the data. This transfer is simply done by sharing the internal weights of neurons through the input sequence [239], [49]. This is followed by the mechanism technique, which retrieves the most important final information representation. After this, a simple classification output layer (softmax) is used to classify learners' employment status (see Figure 5.4). This approach is called *sequential ensemble learning*, which is because what has been learned through one approach to learning is subsequently fed into another one for further learning. This means that what is learned by the RNN depends on the question of what is learned by the CNN. However, in the parallel approach, there is no such degree of dependency as each model learns according to its own approach; after this, an average of the final outputs of each model is used for final calculation at the classification layer (see Figure 5.3).

In this chapter, the researcher sought to compare the performance of the two basic categories of ensemble learning. Parallel (bagging) ensemble learning involves base-learners, which work in parallel, whereas sequential (boosting) ensemble learning are base-learners that work in a chain [59]. Stacking is another option for sequential learning; however, boosting was preferred over stacking for sequential learning for fairness comparison because both bagging and boosting have similar types of deterministic strategies when combining results. The key difference between these two styles is the dependency between the base-learners. In the parallel ensemble, these base-learners are independent and this can reduce errors considerably by averaging or voting [126]. In the sequential fashion, base-learners are dependent, which means that any mislabelled sample in a previous step will influence the chain, which affects the overall performance [126].

5.4 Results and Discussion

Overall, the model for employment status classification was designed with an awareness of the following computational issues. Although the proposed model, as an ensemble model, can introduce a level of complexity, an attempt has been made to reduce this by only considering simple inputs to represent the data (tokens of words, instead of complex stylometric features), which can reduce computation time.

The classification performance of all models investigated in this chapter is summarised in Table 5.1. This is based on a comparison of the two balancing methods used during training, as discussed in Section 5.3.3.1, regarding SMOTE and Paraphrasing. Performance was measured by considering the ratio of the correctly identified samples to the total samples in our data (i.e., accuracy):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.6}$$

SVM was applied because it won the PAN competition, and it is reported that SVM with TF-IDF is the most effective solution for the identification of occupational traits [165]. In addition, Naïve Bayes (NB) and Logistic Regression (LR), which are also commonly used in the AP literature, were applied based on the two balancing methods: SMOTE and Paraphrasing. The experiment involved 40 iterations of training for each DL model in these experiments. The implementation of all models was done via Python libraries.

Generally speaking, DL classifiers achieved higher levels of accuracy for the task of employment status classification compared to traditional ML classifiers. In addition, all models recorded higher accuracy with the paraphrasing balancing strategy, with the exception of SVM (with TF-IDF) and CNN alone. Experiments were performed with a CNN model and RNN model to confirm the researcher's intuition that an ensemble method is more appropriate for the task than a 'single' deep model. For fairness of comparison, all DL models used identical parameters, as explained in Section 5.3.6. Also, all DL models were trained on an identical embedding layer. The performance of the single CNN model was superior to a single RNN, while sequential ensemble learning achieved better results than parallel ensemble learning. It was also found that CNN outperformed parallel ensemble learning. In the sequential ensemble model, 96.4% overall accuracy was obtained.

The other important point to acknowledge is that all models, with the exception of SVM (with TF-IDF) and CNN alone, achieved higher results when the Paraphrasing strategy was applied. This confirms that this strategy of balancing the data is highly effective. In general, the power of the data size during training is recognised, which stems from the remarkable results of almost all DL models that have been used in this study. The worst performer was RNN, but it still achieved a high accuracy of 76.3%, as shown in Table 5.1. In this study, conventional ML achieved the lowest performance, highlighting that this methodology may not be appropriate for this research unless more features are retrieved for learning.

These results are based on the test dataset. The test dataset was extracted separately from the training data (comments from a different group of learners), as explained in Section 5.3.1. The purpose of this was to avoid bias during learning. 10-fold cross-validation was applied in this study when a classifier had a limited number of parameters, such as in the case of SVM. It is also required when the availability of training samples is limited [111], but it is not required with DL models that have many parameters such as CNN models. This is because it increases the complexity during the training [82].

To present the ensemble model's results that used the paraphrasing balancing strategy in a comprehensive and realistic way, further results are presented using well-known performance measurements: F1-score, precision, and recall. In Table 5.2, the first column shows the precision, which is the ratio of the number of true positives (TP) divided by the sum of the number of false positives (TP) and the number of false positives (FP). Recall in the next column, represents the ratio of the TP divided by the sum of the TP and FN (false negative). Finally, the F1-score, which is a weighted mean of both precision and recall.

These performance measurements provide a full picture of the range of performance for the model. The left side of Table 5.2 shows details about the parallel model's performance for each category of the employment status, while the right part of the table shows the same information for the sequential model. As can be seen in the table, even at the detailed category level, the model

Model	Weighting Scheme	SMOTE	Paraphrasing
SVM	Word Vectors	75.0	75.9
	TF-IDF	64.2	63.5
	N-Gram TF-IDF	59.3	63.1
	N-Gram Character TF-IDF	60.5	63.3
Logistic Regression	Word Vectors	85.2	88.8
	TF-IDF	68.8	71.5
	N-Gram TF-IDF	64.3	69.2
	N-Gram Character TF-IDF	71.8	72.4
Naïve Bayes	Word Vectors	52.3	59.4
	TF-IDF	53.3	59.9
	N-Gram TF-IDF	49.5	59.8
	N-Gram Character TF-IDF	60.3	65.3
CNN	-	93.4	92.2
RNN	-	76.4	78.5
Parallel Ensemble	-	87.2	90.3
Sequential Ensemble	-	81.3	96.4

Table 5.1: Models' results: based on two balancing approaches

Employment Status	Parallel Model			Sequential Model		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Working	90	91	90	98	95	97
Not working	83	92	87	98	97	97
Retired	93	89	91	93	97	95

Table 5.2: Parallel and sequential models based on paraphrasing approach: classification of each class

performs exceptionally well. Examining the performance of the balancing technique utilised in this study is crucial. Thus, this section reports more than just the average results; specifically, detailed and non-equivocal results for the class categories (target: employment status) and the model results for each category are presented.

The 'not working' learners category includes a variety of types of people with their own motivations to learn and different goals. Thus, in identification studies, a more detailed separation of stakeholders is necessary. This chapter excluded the 'retired' from the 'not working' group and they were not considered under the not-workers category. This provided much greater accuracy in the identification results by differentiating these two groups; as they obviously have different learning goals on MOOCs. They are clearly learners who have the most free time to complete courses, learn, and improve their job-seeking efforts. Thus, they will have different forms of personalisation or recommendations.

The researcher found that the NLP-based balancing technique used in this chapter, which replaces words with synonyms, is very effective. It is based on a back-translation approach; it allows the substitution of terms with their synonyms while maintaining the same context. Also found that models were not affected negatively or biased as a result of this since the context has not changed, also it does not change any style of writing, such as the addition or removal of commas; it is only applied to individual words.

It is confidently to conclude from these experimental results that using this sequential architecture in ensemble models for learning data representations in this study, particularly associated with the paraphrasing strategy to balance data, can lead to excellent performance for learners' employment status classification in MOOCs.

5.5 Epilogue

Most MOOCs users are individuals who are looking for a job or who need to improve their professional skills [238]. Furthermore, the COVID-19 pandemic has resulted in mass unemployment, which will make MOOCs a great source for building a competitive workforce [196]. This chapter sought to identify different classes of employment status for the purpose of facilitating and enabling personalisation on MOOCs. For For example, non-working learners can be guided to take courses that are currently trendy in the job market to improve their chance of finding jobs, while learners who are currently working can be guided to take advanced courses to improve their skills.

In this chapter, NLP and DL models were combined to identify a learner's current job situation using limited, easily available data – namely, by using their comments. Instead of considering multiple stylometry features, the work in this chapter only uses simple word-tokens from learners' comments. The purpose of this choice was to reduce computational expense, especially after applying the relatively 'heavy' ensemble DL model.

It is an essential requirement for a DL model to have a large data size (samples) to learn from, However, one limitation in the data is the imbalanced distribution of class categories. Great attention was paid to this limitation. For this reason, a data augmentation technique was also explored in this chapter. It is possible to conclude from the experimental results that using the sequential architecture in an ensemble model of CNN and RNN for learning the data representation (associated with the paraphrasing strategy to balance data) can perform with high accuracy for this study, establishing a new state-of-the-art.

In the next chapter, another important demographic characteristic – gender – is further classified in order to identify a further critical piece of information about MOOC learners. A novel approach is used in the next chapter, which is new even to the area of demographic profiling.

Chapter 6

Bi-directional Mechanism Within Recursion Algorithms for Gender Identification

6.1 Prologue

Researchers have acknowledged the association of different behaviours with gender in the context of learning on MOOC platforms [27]. Course content can be personalised based on gender since the gender parameter has already been shown to influence the success of the learning process on MOOCs [144]. Hence, profiling learner gender, which serves as this chapter's focus, provides valuable information that is needed to strengthen the MOOCs literature. On this basis, the researcher applied a novel approach based on Recursive Neural Networks (RecNN) to learn syntactic knowledge extracted from learners comments themselves. In addition to proposing a novel version based on a bi-directional composition function, this chapter evaluates 18 different combinations of word-level encoding and sentence-level encoding functions in state-of-the-art candidates models, exploring their performances, particularly for learner gender profiling.

This chapter describes the experimental setting for gender profiling on MOOCs. At the outset, the importance of this research area and, in particular, the study in this chapter is introduced.

Next, the data collection, analysis, and pre-processing steps are described. This is followed by an explanation of approaches and learning algorithms that are applied. Finally, evaluation processes, including the obtained results are discussed.

6.2 Importance of Gender Profiling for Learners

In Section 3.4, a comprehensive review was presented of scientific research in the field of AP published between 2005 and 2022. As the review demonstrated, most prior research in demographic AP has focused on gender profiling, which clearly reflects the importance of this particular demographic characteristic in many different domains. Within the educational domain, particularly in relation to MOOCs, a great proportion of researchers have relied on gender information as a research parameter [195], [62]. For instance, MOOC studies have been keen to understand why males outnumber females in certain courses by a significant margin, and why men have recorded higher rates of receiving certificates than women, resulting in a greater rate of completion of MOOCs than females, which account for only 32% of the total [238]. The gender-based activity patterns on MOOCs are also different [11]. It has been found that females are generally more active in courses than males. Males and females are even distinct in terms of the types of courses they take [11]. Such differences are inherited from traditional education, where males have been shown to prefer Science, Technology, Engineering, and Math (STEM) courses to a higher degree than females [27]. In addition, studies have found that females had superior time management and study environment skills, while males had superior critical thinking skills [12]. Clearly, gender is one of the factors for MOOC success [144], and the examples just mentioned are only a few examples of gender differences in MOOCs; it is important to determine if MOOC platforms, environments, or designs are suitable for them, and they could be personalised based on gender differences. The main issue currently in MOOCs is the lack of complete gender information about learners (more details are given in Section 1.3.1). Providing information about the gender of learners in MOOCs is the main goal of this chapter.

According to the literature review (Section 3.4), techniques for gender profiling in informal text involve the application of content-related terms [20], applied dictionary-based analysis [233], or applied word functions or Part-of-Speech (POS) tag elements [188]. Only basic types of syntactic representations of text, such as POS, have been considered in previous studies, and prior researchers have simply examined it either in order techniques such as bag-of-words models or in sequential techniques such as RNN models [187]. These models, however, are not fully sufficient because they do not consider the ambiguity of natural language. For example, a sentence (I saw the child with the telescope) can have two meanings: (I saw the (child(with the telescope), which means I saw the child who had a telescope, or I ((saw the child) (with the telescope)), which means I used the telescope to see the child. This is because it is normally in languages that sentences can have different types of structures, which lead to differences in meaning. Such differences can be captured using more advanced syntactic representations. In addition, linguistics researchers have found that people with similar demographics are likely to express themselves using similar syntactic features naturally [204]. Therefore, in this chapter's research, it was considered to use advanced syntactic features for learner gender profiling.

6.3 Identification Methodology

Many complex NLP models have recently been developed to facilitate the analysis of syntactic features and the compositionality of human language [96]. The majority of these algorithms are based on the tree structure of texts and RecNN models [203]. The tree structure is a principal option for representing the syntactic representations as language naturally constructed in tree form [60]. RecNN models were designed to handle such a textual structure and reflect its syntactic representation. These models have achieved remarkable results on numerous text classification problems, including natural language inference [34], sentiment analysis [202], and discourse relation classification [228]. However, RecNN models have only been marginally explored for the area of NLP in general, and they have not yet been applied for AP. The cutting edge of recursive learning models consists of TreeLSTM-based models. These models operate at a higher semantic – and syntactic – level in terms of sentence processing based on linguistics, which enables a richer representation.

The umbrella research question in this chapter, also presented in Section 1.4, is given as follows:

- **RQ1:** Can advanced textual features extracted from MOOC discussion forums be used to classify a learner's gender?

  The following sub-questions were also established:

  - Which NLP representation can be applied to extract advanced syntactic-level representations from learners' MOOC discussion forum comments?
  - Which DL-based algorithms are available to handle advanced syntactic-level representations?
  - How can algorithms based on RecNN be designed so as to classify MOOC learners' gender?

The main contributions of this chapter are as follows: examining advanced syntactic features for the learner gender profiling; exploring state-of-the-art recursive models (tree-structured LSTM, SATA, and SPINN models), and applying them, for the first time to author-profiling (here, for learners gender profiling), and then improving the current recursive models by introducing a novel bi-directional strategy.

6.3.1 POS Tags Patterns

Since the gender classification task pursued in this chapter involves designing models focusing on syntactic attributes, basic linguistic POS tag patterns were also assessed in learners' writing to identify any distinctions at the level of gender. The frequency of each tag[*], as presented in Figures 6.1 and 6.2, was calculated based on gender. Also, the calculation was performed based on the mean "average" of these POS (see Figures 6.3 and 6.4). No noticeable differences are observable between either the frequency or average pattern, and it appears that men and women used the same amount of POS in their comments.

Heatmap approaches were also used to understand the correlation between the POS and discover a statistical measure linearly. Since there are many POS variables, the aim was to examine how dependent they are on each other, which may be shown in a 2D matrix called a correlation matrix.

[*] Description for each POS Tag is shown in Appendix A

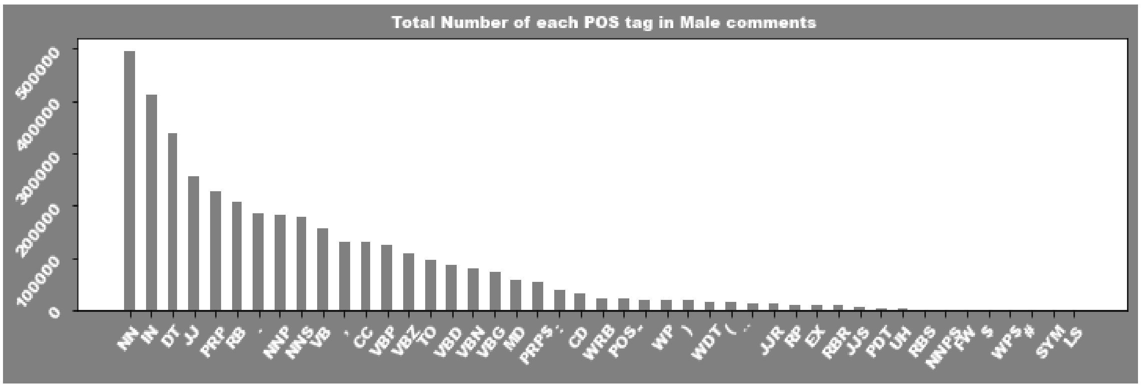


Figure 6.1: Part of speech frequency distribution in male comments

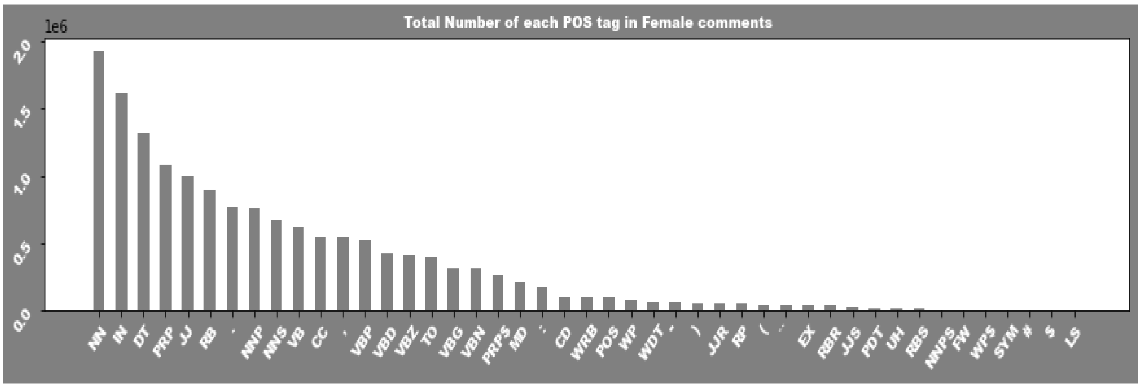


Figure 6.2: Part of speech frequency distribution in female comments

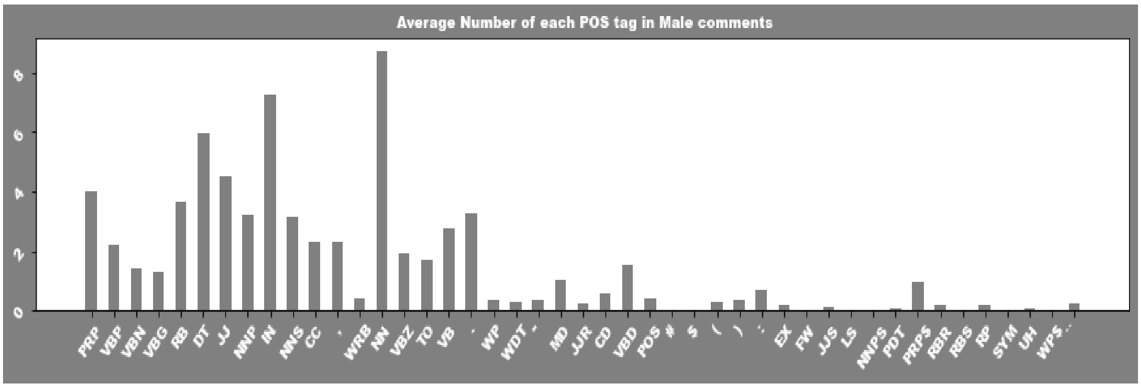


Figure 6.3: Part of speech average distribution in male comments

In the Figures 6.5 and 6.6, the lighter the colour between two variables, the stronger the correlation (and vice versa). For instance, and based on Figure 6.5, women's comments can be distinguished more if there are more sentences including NN and DT or IN, since they have a strong and positive correlation (correlation greater than 0.8). For males' comments, it seems that the predictor would perform better with comments that include MD and VB or JJ and IN.

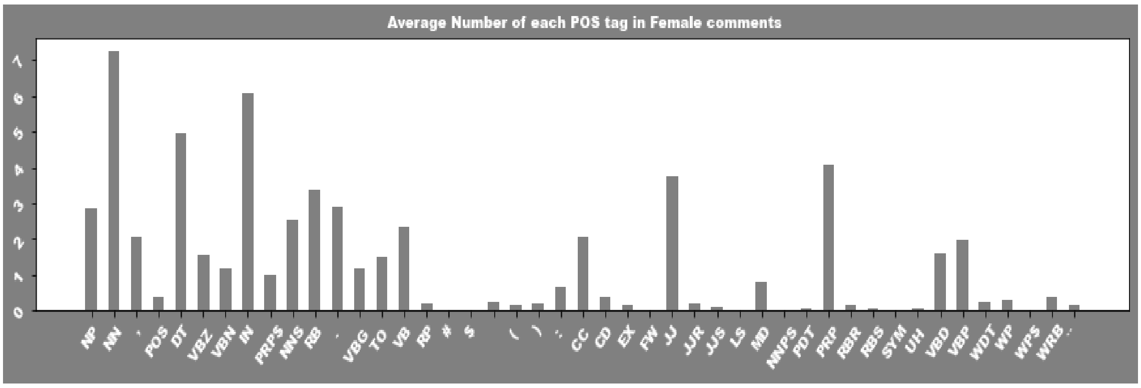


Figure 6.4: Part of speech average distribution in female comments

However, the distribution of POS patterns based on the mean, as shown in Figures 6.5 and 6.6, does not reveal differences in the writing styles of the two groups. This means that the chosen approach also failed to capture the differences in syntactic patterns, due to its simplicity.

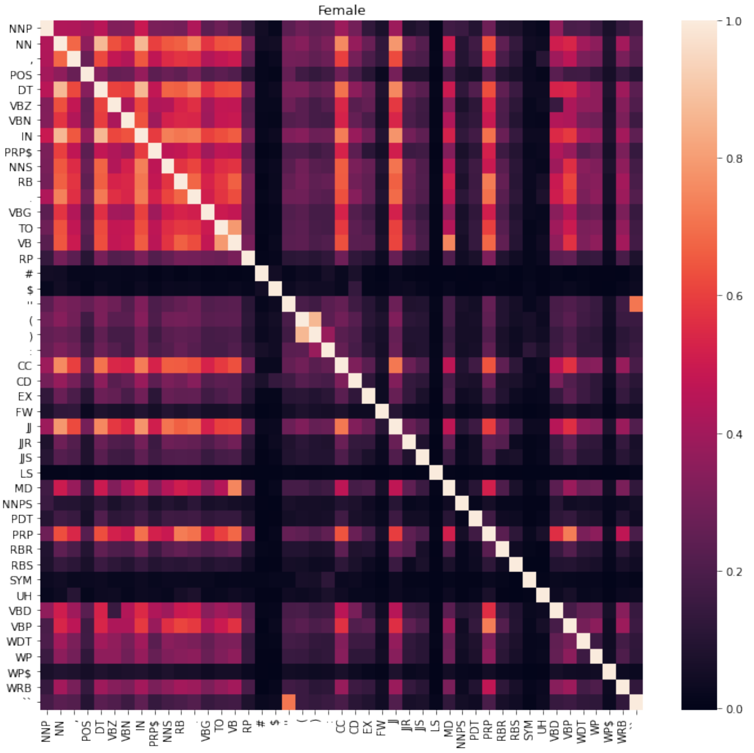


Figure 6.5: Correlation between POS in female comments

Although minor variances were observed among the variables in the heatmaps, there were far greater similarities. In general, the findings in this section may be affected by data imbalance issues, and their reliability may also suffer due to the risk of bias. In addition, the analysis is

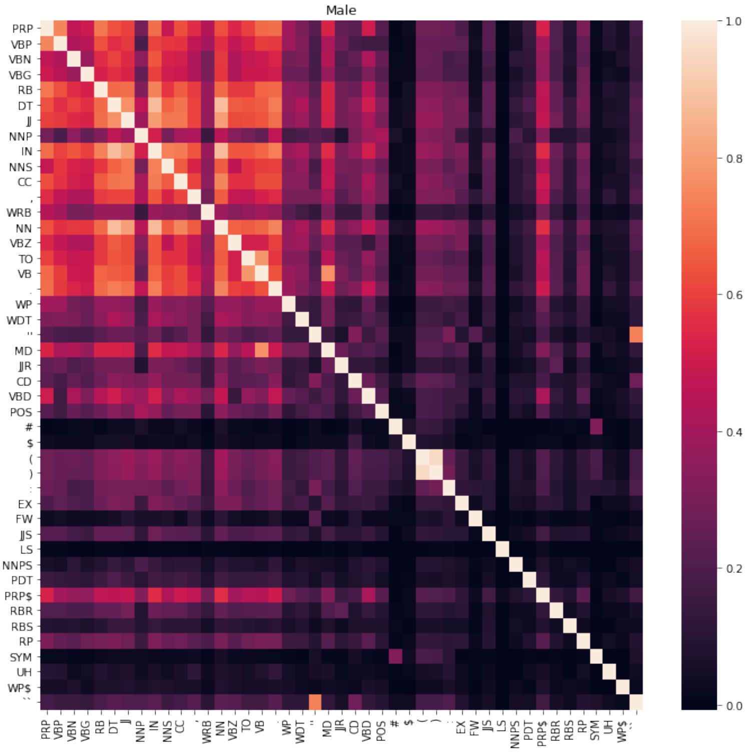


Figure 6.6: Correlation between POS in male comments

based on aggregated comments from all courses together, but learners may comment differently according to the topics. For instance, men may comment more on TECH courses than women and vice versa. Thus, DL approaches, which are more powerful in such noisy data and complex tasks [13], [30], have been examined for gender profiling this chapter.

6.3.2 Dataset for Gender Profiling

Comments as a form of user-generated content is often available across MOOCs [11], [10]. Discussion forums are one of the most widespread sources of metadata used to analyse MOOCs. These forums provide valuable insights into learning and social interactions on MOOCs, which can offer rich metadata to study learners and their needs [11]. Hence, the researcher seeks to tackle the difficult problem of identifying the gender of learners based only on their comments. It is also worth mentioning that the size of the labelled comments dataset used in this research in terms of the gender categories is huge, which is appropriate for DL models.

The data source used in this chapter was explained in Section 4.2.1. For the purpose of gender experimenting, the researcher only used comments labelled with the learners' gender. Approximately 322,310 samples were obtained (265,582 for Females and 56,728 for Males). These profiles were used as targets for gender clssification models. To handle bias, the pre-processing strategy explained in Section 5.3.3 was applied except that no effort was made to extract comments in this study based on the six months time window; the reason for this is that gender is not a changeable demographic characteristic. As a result, there were 61,157 comments in total from 2,568 users for validation and 183,258 comments from 4,956 users for training and testing. The two dimensions of this collected data were unbalanced in that females were 149,904 samples and males were 33,354; however, this was balanced using an effective balancing technique (namely, Paraphrasing) that – in this thesis – is applied to demographic profiling for the first time (see Section 5.3.3.1). The use of the data balancing technique was motivated by a verification of the approach's effectiveness for the study's domain based on the results obtained in Chapter 5. For text normalization, the same steps were applied as explained in Section 5.3.4. For the purpose of gender experiments, which require the syntactic representation of texts, these steps were expanded as follows:

- To obtain fixed vectors, the sentence tokens were padded to contain a maximum of 60 words (mean number of words in samples, in the gender data).
- As in step 7 in Section 5.3.4, word tokens were represented by GloVe, which is a low-dimensional word embedding (semantic vectors for each word for the leaf nodes or the initial inputs).
- Sentence tokens were applied (pythonic NLTK Tokenizer was used) for each comment, which is because the sentence level is the main text representation in this chapter. This is an effective technique that is concerned with the phrase level by cropping samples into sentences. This helps in two main points: first, it reduces the complexity during training time since samples become shorter; and second, it creates data (samples), which boosts the performance of DL (i.e., because the more samples, the better the learning and performance in DL) [82].
- **Syntactic Representation**: After applying the normalization steps, a parser was used based on an expert-designed grammar to handle the sentences/phrases (syntactic) level of the text.

The use of a constituency parser has proven effective in many related studies [111]. This chapter's study specifically used the Stanford Probabilistic Context-Free Grammar (PCFG) parser [112] because it is more accurate and provides the constituents of text, at the phrase level (e.g., NP, VP, and ADJP). See Appendix A[*], for further explanation of these tags. Also, Figure 6.7 shows an example of this text representation.

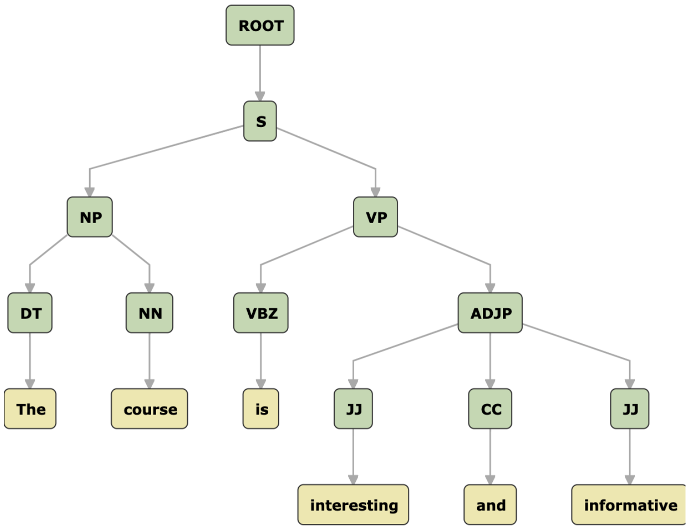


Figure 6.7: A constituency tree example based on PCFG parser

In python, the tree presented in Figure 6.7 is expressed in readable style; by square brackets (for sentence start and end or ROOT of sentence (S)) and parentheses (for each word and its tag)), as follows:

[('The', 'DT'), ('course', 'NN'), ('is', 'VBZ'), ('interesting', 'JJ'), ('and', 'CC'), ('informative', 'JJ')]

This is called the tree structure of a text. The tree structure can be used as an input for learning models. Although this study used a binary mode to establish a binary tree, it is possible to have a node with only one leaf; therefore, the decision was made to delete this node to improve performance and get a padded-like matrix.

[*] http://www.surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html

6.3.3 Problem Definition

The problem in this chapter is a supervised learning problem. Therefore, its mathematical definition takes the form of a mapping function (f):

$$f = S \rightarrow C \tag{6.1}$$

where S is a collection of samples (a vector representing the textual features of a learner's comments), and C a fixed set of classes: $C = c1(Female), c2(Male)$, which are the class categories in this chapter that the research aims to identify. The target is the correct class $c' \in C$. If $c = c'$, then the Identification is correct.

The accuracy of the identified class is measured using a simple distance function as follows:

$$Distance = \begin{cases} 1 & if\ c = c' \\ 0 & if\ c \neq c' \end{cases} \tag{6.2}$$

6.3.4 Recursion Algorithms

Gender profiling using MOOCs data is challenging. One of the principal reasons for this is the similarities between the writing styles of males and females, as shown in the analysis of the gender dataset (see Section 4.2.2.2 in Chapter 4). Thus, DL algorithms are used to solve the issue due to their ability to extract hidden and complex patterns in data [30]. Taking advantage of the high level of language structure (Grammar), three types of syntactic learning models were applied in this study: TreeLSTM, SPINN, and SATA. These models were chosen as they are state-of-the-art DL models for such text representations. Also, the researcher introduced new versions of these models based on a bi-directional composition function with different combinations. These models can learn by supervised learning.

6.3.4.1 Syntactic Textual Representation

The basic approach of many NLP models is to represent text as a sequence of words [55]. However, languages have different information structures. The tree structure is examined in this chapter, which is usually based on a language grammar. It is called a syntactic representation, which blends words into phrases naturally. This representation of text provides a comprehensive interpretation of a sentence's meaning. According to a linguistic principle, a sentence in natural language can be presented as a set of components that are nested constituently in a tree structure [155]. The tree structure may be provided as an input, learned from labelled samples (texts associated with their parse trees), or built implicitly by a neural network with no supervision (latent trees) [24]. The first type is commonly used due to its reported effectiveness [111], which extracts the structure using a pre-trained parser. Thus, this study gets the pre-determined structure from the PCFG parser model; the model is trained on identified treebanks, providing constituency trees that are handy in many sentence-level tasks (see Figure 6.7). In particular, these provide estimates of the optimal tree structures in any custom data. The parser was run using a careless probabilistic context-free grammar model, which outperformed standard PCFG modelling on less strictly grammatical raw data [87], such as comments in MOOCs.

6.3.4.2 Recursive Neural Networks

A RecNN model converts an input word to a vector, which is a leaf node; and the node's pairs are then composed into phrase pairs using a composition function. This is called an intermediate representation of a tree. In turn, the root node is considered the representation of the whole sentence.

6.3.4.3 TreeLSTM

TreeLSTM is inspired by the original LSTM, which processes tokens in a linear chain. Vanilla TreeLSTM is a type of RecNN models types that supports passing information recursively over sequences. TreeLSTM was introduced in Section 2.3.4.1, and it allows information to pass through trees [243], [213]. According to [24], "The recursive neural representations enables combining representations of more granular linguistic units into larger linguistic units (e.g., from characters to sentences, or from words to documents). The merging process is repeated recursively until the root node is reached".

For the same reasons that LSTM beats RNNs, TreeLSTM outperforms RecNNs [24]. The hidden state of the original LSTM is composed of a current input at a current time step and a previous hidden state of an LSTM unit in the previous time step. However, the hidden state of the tree LSTM is composed of a current input vector and the hidden states of two child units (in the case of a binary tree), as shown in Figure 6.8.

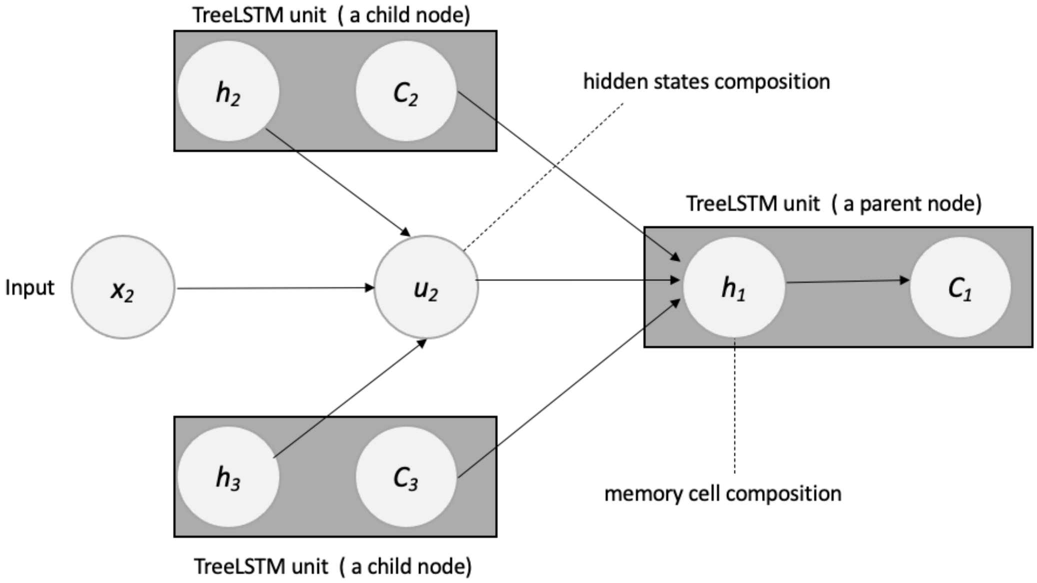

TreeLSTM unit (a child node)
h_2
C_2
hidden states composition
TreeLSTM unit (a parent node)
Input
x_2
u_2
h_1
C_1
h_3
C_3
memory cell composition
TreeLSTM unit (a child node)

Figure 6.8: Composition of memory cell and hidden state of a TreeLSTM unit: Two child nodes (2 and 3)

In a standard tree-structured LSTM cell, the composition functions are as follows:

$$\begin{bmatrix} i \\ f_l \\ f_r \\ o \\ g \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} \left(w \begin{bmatrix} h_1 \\ h_r \end{bmatrix} + b \right) \tag{6.3}$$

$$C = f_l \bigodot c_l + f_r \bigodot c_r + i \bigodot g \tag{6.4}$$

$$h = o \bigodot tanh(c) \tag{6.5}$$

where $h, c \in Rd$ refers to the hidden state and cell state, respectively, in the current cell. Also, in tree-structured LSTM, $h_l, h_r, c_l, c_r \in Rd$ represent the hidden states and cell states of a pair of child nodes (left and right); $g \in Rd$ refers to the composed inputs from both children; so each node or parent computed from two direct children, and $i, f_l, f_r, o \in Rd$ represent the input gate, two forget gates, and an output gate, respectively. These two separate forget gates from two children allow the network to choose to forget different information in each child node, which captures a more complex representation of the information from the same sentence.

In the above composition functions, $w \in R5d \times 2d$ and $b \in R5d$ are trainable parameters in the model, σ and tanh refer to the sigmoid and hyperbolic tangent functions, which apply non-linear transformations before the gate information is updated, and $\odot$ is the element-wise multiplication symbol, which is used because the dimensionality of elements on both sides is the same. The equations refer to a binary tree, but tree-structured LSTM is not limited to two-children cases; it can easily be extended to multiple children cases due to the flexible nature of the recursive neural network.

In this research, a binary tree setting was adopted, which is the most common type used in the related literature [203], [128]. The main problems are that TreeLSTM is well-known for having a long training time and experiencing difficulties in exploiting the advantages of batch computation, which is attributable to the diverse and complex structure of sentences. Notably, therefore, the SPINN model [34] has handled batch computation by flattening the tree structure during learning using a shift-reduce parser.

6.3.4.4 Stack-Augmented Parser-Interpreter Neural Network

The Stack-Augmented Parser-Interpreter Neural Network (SPINN) [34] enables efficient TreeLSTM training through the adoption of the idea of a shift-reduce parser from the compiler [2]. For this reason, it has remained the state-of-the-art model since 2015. A different composition function to construct the tree was introduced that increased the training accuracy and testing accuracy by 5.3% and 2.6%, respectively, on the Natural Language Inference (NLI) dataset compared to the baseline model (i.e., LSTM). The shift-reduce parser in SPINN increased the speed of learning for tree-structured models, allowing the handling of large-scale data. This is notable because previous models could not support batch computation. SPINN introduced a solution called tracker, which aims to summarise sentence information during training. Hence, a new composition function has an extra input for information, which is generated in real-time during the encoding of the sentence. This information provides higher accuracy, but it can only summarise limited information in a sentence.

SPINN provides a way to reconstruct the complex syntactic structure of the language by reading it from left to right with the help of a shift-reduce parsing algorithm [2]. The shift-reduce algorithm takes a sequence of inputs with length N and converts it to $2N$-1 length transitions; the sequence of transitions is either shifted or reduced. Then, the sequences of words from the sentence and related transitions are fed into the SPINN model. To encode this complex structure of the tree, two data structures are used, both of size N: stack and buffer. In the beginning, the sequence of inputs is fed into the buffer in order; when the transition is SHIFT, the top word in the buffer is pushed to the bottom of the stack, and when the transition is REDUCE, the bottom two words in the stack are popped out and combined into one word. Following this, the new word is pushed to the bottom of the stack. Having an example form MOOC data, this can be done by linearising as follows:
x: Test, results, were, encouraging
((Test results) (were encouraging))
a: shift, shift, reduce, shift, shift, reduce, reduce

The composition function in SPINN introduced a component called tracking LSTM, which is denoted e. This piece of extra input information is generated in real-time through the sentence-encoding process, and it consists of three components: two word-level embeddings from the two bottom positions of the stack and one word-level embedding from the top position of the buffer. This extra information e provides a representation of the current status of the sentence encoding process, as well as the current status of the buffer and stack. In addition, it supplies more information to the composition function. To generate e from the three components from the stack and buffer, a simple linear mapping is used. This information provides a global datum in each current cell, so it can expend the information in each step. It works as an indicator of the progress of sentence encoding. The composition function for SPINN is shown below.

$$\begin{bmatrix} i \\ f_l \\ f_r \\ o \\ g \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} \left(w \begin{bmatrix} h_1 \\ h_r \\ e \end{bmatrix} + b \right) \tag{6.6}$$

$$C = f_l \bigodot c_l + f_r \bigodot c_r + i \bigodot g \tag{6.7}$$

$$h = o \bigodot tanh(c) \tag{6.8}$$

6.3.4.5 Structure-Aware Tag Augmented

In 2019, the Structure-Aware Tag Augmented (SATA) model, a TreeLSTM-based model and principally an extended version of SPINN, was proposed as a model with additional information using a separate LSTM tree to model the sentence, as well as the extra information only contributing to the gate information, which empirically reached a better optimum over the tree [111]. SATA has achieved state-of-the-art accuracy in 4 out of 5 public datasets [111].

The extra information in SATA that adds to the model comes from a tag representation, which is generated as a by-product of the parser and creates an extra LSTM network to learn a higher representation of the tag at each node. This information from the new LSTM model is equivalent to the tracker LSTM part in the SPINN model. This new piece of information shares the same idea of a tracker LSTM, which is a representation of the current state for the encoding process (the level of the tree structure) and adds more information to the TreeLSTM encoding function. In addition, this provides more information on the syntactic structure of the sentence; however, this time, the extra information only contributes to the gate-information in the LSTM cell and does not influence the actual input information in the composition function, as shown below:

$$\begin{bmatrix} i_1 \\ f_{l1} \\ f_{r1} \\ o_1 \\ g \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} \left(w \begin{bmatrix} h_l \\ h_r \end{bmatrix} + b \right) \tag{6.9}$$

$$\begin{bmatrix} i_2 \\ f_{l2} \\ f_{r2} \\ o_2 \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} \left(w \begin{bmatrix} e \end{bmatrix} + b \right) \tag{6.10}$$

$$c = (f_{l1} + f_{2l}) \bigodot c_1 + (f_{r1} + f_{r2}) \bigodot c_r + (i_1 + i_2) \bigodot g \tag{6.11}$$

$$h = o_1 + o_2 \bigodot tanh(c) \tag{6.12}$$

The SATA model proves that the use of more linguistic information (tag information) helps in understanding sentences. Also, an advantage is that the SATA model allows for dynamic composition of the language tree, which can use all the information from each single sentence without losing any information. Thus, SATA is this study's final candidate for gender profiling using learner data from MOOCs.

6.3.4.6 Experiments Based on Bi-directional Mechanism

Both SPINN and SATA are powerful tree-based models. This study found that no study had examined the performance of these models by adding the bi-directional learning. Bi-directional learning has already shown its effectiveness in improving the sequential LSTM model. It is well-known that bi-directional LSTM outperforms vanilla LSTM for many NLP tasks [215], and bi-directional TreeLSTM alone shows one of best findings for sentiment classification [215]. Based on the information provided, a hypothesis was made that adding bi-directional TreeLSTM would improve the performance of SPINN and SATA. The researcher investigated propagating the top-down direction of information and the bottom-up direction using bi-directional TreeLSTM. In fact, the uni-directional TreeLSTM by default processes inputs from the bottom-up direction in a bottom-up manner through the tree (bottom-up manner). So, the researcher included the additional set of hidden state vectors in the top-down direction (from root to comment inputs), which then alters the model to the bi-directional paradigm. This is technically another TreeLSTM model, where the final hidden state is the final state vectors of the two LSTMs.

The syntactic learning in a TreeLSTM-based architecture in general consists of the following two steps: word-level encoding with a feedforward neural network or LSTM neural network; and sentence-level encoding with a tree-structured LSTM composition function. While previous literature has recommended using LSTM for word-level encoding, there is no such work to introduce bi-directional LSTM for the word-level encoding. Thus, this research also contributes to fill this gap, by adding the bi-directional LSTM at the word level as well. The motivation for this supplementary bi-directional technique is to increase the high-level representation of tree nodes during the recursive propagation across many branches.

To summarise, the gender profiling experiments were designed by comparing several versions of three tree-structured LSTM models: TreeLSTM, SPINN, and SATA. Also, novel versions of the bi-directional composition function were added to existing architectures, as well as to the word level. These versions of the bi-directional function for existing models on 18 different architectures (combinations) of word-level and sentence-level encoding on the research dataset (see Figure 6.9).

The first step is word-level encoding, and the encoding models evaluated in this research were one hidden layer feed-forward neural network, basic vanilla LSTM neural network with one hidden layer, and basic bi-directional LSTM neural network with one hidden layer. The second step is sentence-level encoding, for which the research constructed the TreeLSTM with a different composition function. Six versions of the composition function were evaluated: LSTM tree, composition function taken from the SPINN model, from the SATA model; and their bi-directional composition functions (the proposed version), as shown in Figure 6.9. The binary tree parser was applied in this study as it is used by the three models. For fair comparison purposes, all experimental settings were kept constant. All 18 models were trained on 30 epochs using a softmax as

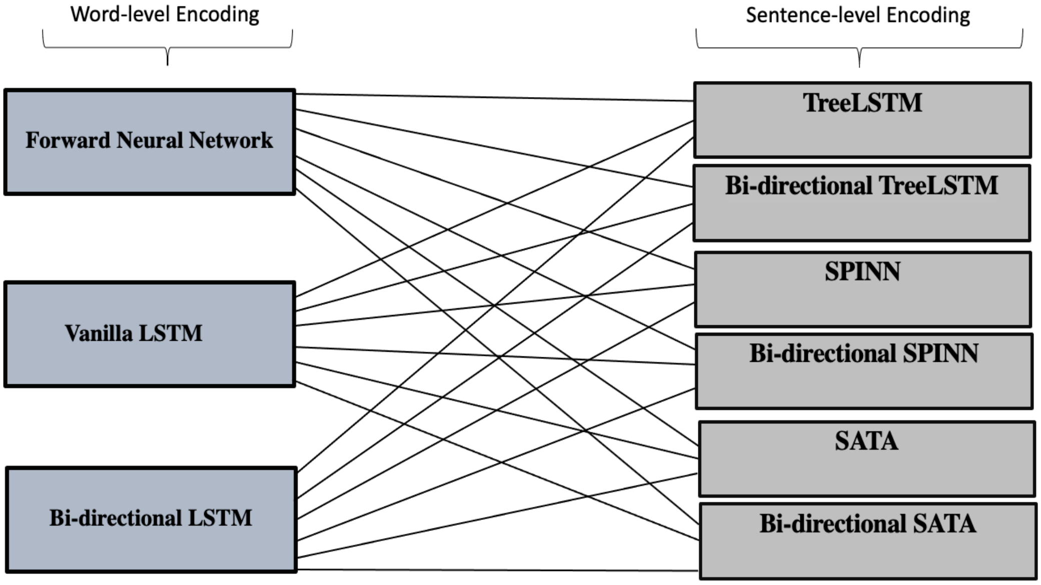


Figure 6.9: 18 architectures and combinations of models and versions of bi-directional TreeLSTM, SPINN, and STATA

the classifier at the last layer. Random search was used to tune the hyperparameters, and then the hyperparameters were validated on the validation set. During training, each model's performance was tracked based on the validation set, and parameters were saved when performance reached a new peak. The researcher used early stopping to stop training when there was no performance improvement after epochs 5. Generally, the optimal results were achieved on a batch size of 10, 0.1 learning rate, dropout of 0.3, and 100 hidden units.

6.4 Results and Discussion

The results of the models on the gender data were evaluated and the test accuracy results are reported in Table 6.1. Based on the experimental results, the models achieved competitive performance in relation to each other. In general, all models achieved high performance in identifying the gender class (80.5% or above), and all were more effective than the baseline models.

Model performance was also evaluated with classical ML, namely, SVM, RF, and LR. LR (introduced in Section 2.4.2), as a baseline. This is because LR won the PAN competition for developing the most successful solution in identifying gender traits [170]. LR has also been widely applied

Model	Classification Accuracy
SVM (n-gram TF-IDF)	75.8%
Random Forest (n-gram TF-IDF)	73.0%
Logistic Regression (n-gram TF-IDF)	68.7%
Bi-directional LSTM (Traditional Sequence)	78.80%
TreeLSTM	79.75%
SPINN	80.02%
SATA	79.20%
Forward Neural Network + TreeLSTM	81.49%
+With Bi-Directional Composition Functions	81.67%
Forward Neural Network + SATA	81.90%
+With Bi-Directional Composition Functions	82.20%
Forward Neural Network + SPINN	81.60%
+With Bi-Directional Composition Functions	**82.60%**
Vanilla LSTM + TreeLSTM	81.60%
+With Bi-Directional Composition Functions	80.70%
Vanilla LSTM + SATA	80.60%
+With Bi-Directional Composition Functions	81.49%
Vanilla LSTM + SPINN	80.60%
+With Bi-Directional Composition Functions	80.99%
Bi-directional LSTM + TreeLSTM	81.86%
+With Bi-Directional Composition Functions	82.49%
Bi-directional LSTM + SATA	82.10%
+With Bi-Directional Composition Functions	82.17%
Bi-directional LSTM + SPINN	81.47%
+With Bi-Directional Composition Functions	82.55%

Table 6.1: Accuracy of all models for gender identification: baseline and syntactic models

for gender classification. In this research, a 3-5 gram TF-IDF word-level model was utilised [165]. Bi-LSTM was also used as a baseline, which is usually used as a DL baseline in NLP experiments [139]. The TreeLSTM model was considered, which does not include tags during learning. This was possible in this study because the data are large; notably, some prior studies have indicated that large data size can help neural network models to learn syntactic rules even without including tags during learning, which means there is no need for external morphological information when the data size is large enough [111]. SPINN and SATA in their original structure are also considered as well.

The performance of the conventional ML baseline models, based on TF-IDF features, were the lowest. The sequential learning model (Bi-LSTM) fared somewhat better than conventional ML models, but not as well as all recursively learning algorithms. The proposed versions of the bi-directional strategy that were applied for all 18 models in Table 6.4 achieved higher results compared to every corresponding model. Every two versions of each model are very similar, but the bi-directional composition function models achieved slightly better results; nevertheless, they are competitive to each other and accuracies tend to be identical. This could promote the idea that the use of phrase-level representation is robust for learner gender classification.

The highest observed outcome in this chapter's experiments was 82.62%. This was achieved by the newly proposed model based on the simple Forward Neural Network combined with the SPINN model. It indicates that bi-directional learning is promising in terms of improving classification accuracy. This also shows the importance of the extra information that the model obtains during the training, which does not have to be limited to tags of constituents included in the SATA model. As the tracker LSTM in SPINN provides less information compared to SATA, this information may not make a significant contribution when the task is complex, as in learner gender profiling; also, by including more linguistic information, the accuracy was not substantially affected. Furthermore, it is evident that using only a simple model with fewer parameters for word encoding by the Forward Neural Network (which used linear mapping) still achieved high results. This might be attributable to the fact that using linear mapping better preserves word-level semantics, while the LSTM encoding alters the semantic meaning at the word level, thereby making it harder to structure the sentence from a syntactic perspective. This might also be related to task complexity.

The results in Table 6.1 are based on data test accuracy. The test dataset was extracted separately from the training data based on comments from different groups of learners. The purpose of this was to avoid bias during learning. Therefore, 10-fold cross-validation (CV) was not needed [111] as such methods are recommended only when training samples are small. However, it is usually applied when a simple classifier has been used (e.g., LR). It is also required when the size of the training samples is small [111], but it is not required with deep learning models that have many parameters such as LSTM models. The reason for this is that it increases complexity during the training [82].

Prior studies indicate that updating GloVe vectors for word-level encoding offers a minor gain on binary classification tasks with TreeLSTM models [111]. However, the size of the data used in this chapter's study was substantial, and updating the initialized GloVe vectors during training yields a boost in performance in capturing more accurate semantics of the contact. This is due to the fact that the meaning of certain words in the domain of education is different compared to other domains (e.g., words such as program, course, and degree). These word vectors were updated, but it created different embedding dimensions that may be higher; nevertheless, this is considered a viable way to boost classification performance. The researcher also updated GloVe at all models mentioned above.

Based on this chapter's findings, it can be claimed confidently that this study's results are robust. In large part, this is owing to the enormous size of the data used in the research. However, the proposed models are complex and need satisfactory computational resources. For example, I have used for this study the so-called Graphics Unit Processor (GPU), which is an expensive hardware, but is necessary for running any DL model. In spite of that, DL models are emerging as state-of-art techniques in NLP, and they do not demand heavily feature engineering, as is required for traditional ML. In AP research, using traditional models, which principally operate by experimenting with thousands of textual features, could be a viable way to analyse authors' writing styles. In this chapter, gender classification was handled based on differences in syntax among males and females, as indicated in the literature [232]. For example, males have been found to be more direct in their inquiries, while females are typically more polite (e.g., women may say: "I was wondering if you can help me?", while men say: "Please give me a hand?") [232]. Thus, a complex and advanced syntax learning approach has been considered to solve the gender profiling problem. Relatively few studies have been undertaken in NLP that have used syntactic representation based on tree-structured DL to explore information that is associated with syntactic parsing [128]. The main issue with these models is that parsing sentences takes time; for the data used in this chapter, for example, it would take days for parsing to complete due to the huge size. Also, using bi-directional tree LSTM increases model complexity compared to the use of the uni-directional tree LSTM. These models have considerable complexity and they require long training periods – often lasting days – due to the complexity of the models and the huge size of the used data (see Section 6.3.2).

In addition, it is known that the chance of having higher accuracy in DL models increases when more correct encoded information is fed into the model. In DL research in general, the more features/information that are fed into the model, the better the performance of the model, which is because it reduces the uncertainty of the model by providing extra information [55]. This could also be provided by bi-directional training. Nevertheless, the bi-directional versions were able to achieve comparable results only based on basic NLP normalisation tools (see Section 6.3.2) and using an innovative text-argumentation strategy, which was used earlier in this thesis in Section 5.3.3.1.

6.5 Epilogue

In this chapter, the researcher applied gender profiling to the critical domain of education. Various stakeholders in computer-based education, such as administrators, researchers, practitioners, educators, teachers, and ultimately learners, could benefit from personalised learning environments tailored to their needs. For example, females can be guided to take courses that include more feminine issues such as maternity, while males can be guided to take courses that include more interesting topics to them such as car or football.

DL models are widespread solutions in NLP tasks nowadays and hence these were considered for learner gender profiling. This is feasible in this chapter because a sufficiently large number of samples was obtained from a specific domain (in this case, MOOCs). In this chapter, the researcher investigated how new DL methods can be designed to identify the gender of learners in MOOCs based only on the comments exchanged. Therefore, cutting edge syntactic models were utilised, which have previously been used for other text classification tasks, and hence suggests that they represent viable candidates for AP as well. The models used in this chapter are complex and need sufficient computational resources. Tree-based models, in addition to the bi-directional version, have increased this complexity even further. However, the high accuracy, especially in terms of making classification over MOOC data, is particularly promising.

In the next chapter, another important demographic characteristic – level of education – is further classified in order to identify a further critical piece of information about MOOC learners. A novel approach based on further MOOCs metadada is used, which is new even to the area of demographic profiling.

Chapter 7

MOOC Metadata for Education Level Identification

7.1 Prologue

MOOCs are universal learning resources that are currently attracting tremendous numbers of users; particularly education seekers. An estimated 40% of learners enrol in MOOCs for educational reasons [238]. At the same time, the ongoing COVID-19 pandemic is rendering these platforms even more necessary [193]. Many face-to-face courses suddenly stopped during the pandemic [230], and so most new MOOC users this year have been individuals who are seeking to replace their suspended classes [193]. This positions MOOCs as an optimal alternative because they offer remotely (digitally) accessible classes from the world's leading institutions [181]. The ongoing pandemic is also expected to promote the demand for online education in the future, which is because it breaks any spatial or temporal limitations (see also Section 8.2.1). Therefore, in this research, a critical aim was to consider educational level of MOOC users.

In this chapter, the importance of education level profiling is discussed. Then, a description is given of the data preparation and feature engineering processes. This is followed by an explanation of the approaches and learning algorithms that are applied. Finally, the evaluation process, in terms of model performance and the importance of features, is discussed.

7.2 Importance of Education Level Profiling for Learners

A substantial percentage of the number of enrollments are education seekers, and education level is a well-known factor that influences the learning process in any education system [242]. In MOOCs, *education level* is a significant factor, among others, that has been found to affect the likelihood of completing a MOOC. According to [144], there is a positive correlation between the two; namely, the higher education level, the more likely a user is to complete a MOOC. Hence, it is one of the critical pieces of demographic information that is needed in order to personalise the systems in MOOCs. The ongoing COVID-19 pandemic has increased the numbers of this particular type of learner even more. According to a recent statistical report [193], enrollments at Coursera, a USA MOOC provider, increased by 640% just between mid-March to mid-April 2020 (10.3 million in 30 days) compared with the same interval in 2019. In the UK, the FutureLearn now has 13.5 million learners [71].

One of the advantages of MOOCs is their provision of college credits, via a certificate. The first attempts started in October 2013, when a contract was created between Antioch University and Coursera to license several of the university's courses on the Coursera platform, which would serve as credits for part of a bachelor's degree program. However, MOOC platforms will encounter many obvious challenges. Checking for plagiarism (cheating) or authorship is one of these points that could increase the credibility of MOOCs and may lead to more accreditation in online education [150]. Knowing the educational level of students is critical and it has many applications, including plagiarism detection. Therefore, the work presented in this chapter is a step toward achieving such creditability on MOOCs, specifically by extracting a learner's level of education automatically. This will further enable the enrichment of personalisation or the decision-making processes on MOOCs.

In the area of AP, educational profiling has received little attention in the related literature, as discussed in Section 3.2.2. Although AP has been implemented in educational texts before, it was only limited to identifying specific user traits, and it only worked with particular types of text (e.g., identifying student native languages based on their essays in English exams such as TOEFL) [31], [216].

One popular study was undertaken by [65] to classify educational traits – as either no tertiary education or some tertiary education – based on the content of the students' emails. They applied only NLP (textual) features (namely, structural features such as a signature, along with named entities, word length, punctuation, function words, and part-of-speech tagging). The best result the researcher achieved was based on bagging ensemble models such as Random Forest (RF), which achieved 79.92% accuracy. Importantly, this thesis uses the same textual features as [65] in this chapter's study, apart from the structural features that do not exist in MOOC comments (i.e., the research dataset). In addition, no named entities are used in this thesis because most of the available annotators were developed on a different domain (i.e., on news corpora), and they also have reported that it is not effective for their study.

Nearly one decade after [65], Wang et al. [229] in 2016 performed a study to identify education status based on videos that users were discussing. In particular, the researcher collected large numbers of microblogs. They set up a binary classification task in terms of identifying the users' status as having either a university or non-university background. Instead of focusing on writing style alone, their study went deeper by analysing the relationship between users and video-related information that they were talking about. For example, it was found that the group of Chinese users who watched English dramas tended to be well educated. This supported their assumption of the relatedness of such information to demographic classification. They extracted these words using Term Frequency-Inverse Document Frequency (TF-IDF) that describe the videos. The Decision Tree (DT) classifier outperformed the other models in achieving 81.2% accuracy for educational background classification. Their work can be regarded as the state-of-the-art in terms of high accuracy for applying AP to infer educational status. Their study has a common strand compared to this thesis, which investigates under-utilised features for education level profiling.

Another study was undertaken by [227] to examine the ability of Sentiment Analysis (SA) to infer user demographic characteristics. The researcher found that users with higher educational levels tended to express positive sentiment more frequently in their tweets compared to other users. Hence, SA is also considered in the present thesis as one of the educational level predictors.

7.3 Identification Methodology

This chapter seeks to identify the educational level of MOOC learners based on *course level*. This is different to the work done in Chapters 5 and 6. Instead of training classifiers with data collected from all courses in the MOOCs platform, this chapter aims to study classification models based on course level; this means that the classifiers are fed by data collected from each course separately. The intent is to examine how to identify a learner's demographic characteristics – in this case, their educational level – in MOOCs only based on data collected on a small scale (i.e., only one course). This leads to the need to consider other MOOC metadata, not just learners' comments because in course level numbers of comments are less. Such an experiment may provide an understanding of learner behaviour on MOOCs, particularly when metadata are used as effective predictors of level of education. Furthermore, this experiment can provide an early identification of the educational demographic, which is valuable considering the fast changes that are happening to these platforms during the ongoing COVID-19 pandemic. This can help, for instance, in providing information about a learner's current knowledge and how to tailor content; or even used this information as one of the factors that help in detect cheating on MOOCs (in case of a certificate is given by end of the course). Studying cheating detection systems or other systems in MOOCs is out of the scope of this thesis, but this chapter's study contributes to providing an early identification of learners' educational level, which can be used for further research into MOOCs.

This chapter uses both textual and behavioural data. The chapter specifically examines *time stamps, quizzes, and discussions*.

The research questions pursued in this chapter are given as follows, which were also presented in Section 1.4:

- RQ1: Can a learner's level of education be classified based on MOOC discussion forum data on a course-level classification?
- RQ2: Can the use of metadata in addition to MOOC discussion forum data improve the classification of a learner's level of education?

  The sub-questions for these research questions are:

- RQ3: What MOOC metadata is available for extraction, based on NLP and non-NLP features, for use as predictors for a learner's educational level?
- RQ4: How do the classifiers achieve high accuracy despite the simplicity of the applied features?
- RQ5: Which classifier can identify a learner's educational level regardless of data size?

The main contributions of this study are the following: first, this is the first attempt in the literature to identify the educational status of learners on MOOCs; second, this chapter investigates available MOOC metadata comprehensively for the task; third, this is the first time the AP approach has been linked with MOOC domain-related data, not only based on NLP features; and fourth, despite the simplicity of the applied features, high accuracy is achieved regardless of the data size, and even with inexpensive classifiers because of the careful selection of features and examine different classifiers for the problem.

7.3.1 Dataset for Education Level Profiling

The source of data for this thesis was explained in Section 4.2.1. For the purpose of education level experimenting, data were extracted from four courses delivered by the University of Warwick between 2013 and 2015. These courses bring together different topic domains, including Computer Science, Psychology, and Literature. These are both STEM courses[*] and non-STEM courses, and so they represent different types of domains and, potentially, different types of learners. Each course has been offered multiple times (known as 'runs') with 21 runs in total. The runs are of different durations as follows: Big Data (BG): three runs and nine weeks duration each. Babies in Mind (BM): six runs and four weeks duration each. The Mind is Flat (MF): seven runs and six weeks duration each. Shakespeare (SH): four runs and ten weeks duration each.

These courses are included in this study because they have enough samples for this chapter's experiments. In each week, learners complete a 'learning unit' that consists of several tasks (called 'steps'). These steps can involve videos, articles, quizzes, or discussions. The system generates

[*] STEM stands for Science, Technology, Engineering, and Mathematics.

a unique ID for each learner and also timestamps, which are the time of enrollment, time of submitting an answer, the time of accessing a step, the first time visiting a step, and when learners press the "Mark as Completed" button. The system also stores numerical and Boolean data related to learners' responses to different questions during a course. Learners in this dataset accessed 2,794,578 steps in total. For the chapter's experiments, there were 12,984 learners who declared their level of education out of the total learners in the dataset (245,255 learners), categorised as Bachelor (B), Master (M), and Doctorate (D).

In this chapter, the education level below a bachelor's degree was not considered. This is because learners in secondary education or lower were found to achieve a higher rate of course completion compared to others [222], even doctoral students. Therefore, the number of completed steps was higher, which corresponded to substantial missing data. There is a risk of the classification model becoming biased and generating false positives (even after balancing) towards learning more about one of the groups [191].

7.3.2 Problem Definition

The task of educational level classification is expressed mathematically as a mapping function (f) as follows:

$$f = S \rightarrow C \tag{7.1}$$

where S is a collection of samples (a vector representing the metadata features of a learner's ([Enrollment, Quiz, Time Spent, and Comment], as explained in Section 7.3.3). These features are examined separately and jointly (see Section 7.4).

C a fixed set of classes: $C = c1(Bachelor), c2(Master), c3(Doctoral)$, which are the class categories in this chapter that the classification is aimed at.

The target is the correct class $c' \in C$. If $c = c'$, it is possible to conclude that the classification is correct.

7.3.3 Feature Extraction

In this chapter, a comprehensive study was undertaken of the available features in the rich data available for the research. The aim was to extract potential features from the data based on the course level to serve as predictors for learners' level of education. The feature extraction process was based on three conditions:

1. ***Existence of Labels***: Features should belong to learners who have declared their education level. This is essential because this research is essentially based on supervised learning techniques.

2. ***Size of Samples for Features***: Some metadata are available in this chapter's dataset, but there are not enough samples. For example, the information of which a comment was moderated for inappropriate or offensive content is available in this chapter's data; however, when the researcher sought to extract them, it was noticed that they belong to only three learners, which is inadequate for the training process.

3. ***Relatedness***: Some available metadata in this chapter's dataset have not been extracted such as the question number or comment ID. This is because such metadata are obviously not predictors for education level, and also because they are automatically generated by the platforms, not by the learners.

The final set of extracted features belonged to four categories:

- **Enrollment Features**: These features belong to the date of enrollment for each learner (enrolled-at [timestamp] – when the learner enrolled).

- **Quiz Features**: These features belong to the date of submitting answers (submitted-at [timestamp]), responses data (i.e., the answer number selected, reflecting their ordered position [numerical]), and correctness data (for the correctness of the responses [Boolean]).

- **Time Spent Features**: These features belong to two types of dates related to steps: first visited-at (when the step was first viewed by the user [timestamp]), and last-completed-at (when the step was last marked as complete by the user [timestamp]).

-**Comment Features**: These features are comments written by a learner [text], date of post comments [timestamp], and the number of likes attributed to each comment [numerical].

7.3.4 Feature Engineering

Metadata were obtained from enrollments, quizzes, steps, and comments. Therefore, varied data sizes were obtained as there were different case scenarios for learners' activities. For example, certain learners watched videos but did not answer quizzes, some wrote comments while others did not, and so on (see Table 7.1). However, this issue was resolved by filling in missing data, as explained in the next sections.

	Enrollments			**Quizzes**			**Time Spent**			**Comments**		
Course	B	M	D	B	M	D	B	M	D	B	M	D
BD	870	737	160	5250	4860	1576	544	458	117	2326	2052	526
BM	1561	932	156	10065	6522	971	980	653	98	4650	2300	298
MF	2237	1424	269	48761	31015	6668	1249	836	187	9232	5844	2717
SH	2503	1747	388	136919	93311	22547	1802	1328	312	21363	14997	5887

Table 7.1: Samples size per course/ level

All the extracted features are in a raw format, and so they were normalised before feeding them into the ML models. For example, URLs were removed and any duplicated comments were excluded, keeping only the original comments written by learners. The importance of excluding duplicated comments stems from the finding that some learners copy and paste other learners' comments, which means that the copied comments do not reflect their personal way of writing. In addition, the researcher applied simple and advanced NLP techniques to the comments to convert them into textual representations, as they reflect an author's writing style [65]. All features were converted into numerical forms as follows:

1. **Temporal Features (5 Feature Sets)**:

   Any [timestamp] feature in the data (found in enrollment, quiz, and comment files) is normalised to:

   i *Hour*: Value of time hour within a day (values between 0 to 23).

   ii *Month*: Value of that month within a year (values between 1 to 12).

iii *Week Day*: Value of that day within a week (values between 1 to 7).

iv *Month Day*: Value of that day within a month (values between 1 to 31).

v *Year Day*: Value of that day within a year (values between 1 to 365).

See Table 7.2 for a description of the temporal features' symbols that were normalised for each file category in the dataset.

2. **Simple Textual Features (9 Feature Sets)**: the researcher converted comments [text] by simple NLP tools to textual representation based on character, word, and sentence level, as well as other special text levels. In turn, the researcher counted these textual representations as follows:

i *Character Count*: Total number of characters in a comment.

ii *Word Count*: Total number of words in a comment.

iii *Word Density*: Average length of words in a comment.

iv *Sentence Count*: Total number of sentences in a comment.

v *Sentence Density*: Average length of a sentences in a comment.

vi *Punctuation Count*: Total number of punctuation marks in a comment.

vii *Upper Case Count*: Total number of upper count words in a comment.

viii *Title Word Count*: Total number of proper case words in a comment.

ix *Stop Word Count*: Total number of stop words in a comment.

3. **NLP Features (2 Feature Sets)**:

The NLP features are extracted by pythonic implementation:

- *Part-of-Speech (POS)*: This is performed to extract the POS tags. The tags are explained at this website[*], also explained in Appendix A.

Then, the total number of nouns, verbs, adjectives, adverbs, and pronouns was calculated in each comment (see Table 7.2).

-*Sentiment Analysis (SA)*: A pythonic experiment was implemented to apply SA to thr used

[*] http://www.surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html

MOOCs comments, which assigns three polarities: positive (1), negative (-1), and neutral (0).

Feature Name	Symbol
Hour	enrolment[e_hour], quiz[q_hour], comment[c_hour]
Month	enrolment[e_month], quiz[q_month], comment[c_month]
Week Day	enrolment[e_week_day], quiz[q_week_day], comment[c_week_day]
Month Day	enrolment[e_month_day], quiz[q_month_day], comment[c_month_day]
Year Day	enrolment[e_year_day], quiz[q_year_day], comment[c_year_day]
Noun	['NN','NNS','NNP','NNPS']
Verb	['VB','VBD','VBG','VBN','VBP','VBZ']
Adjective	['JJ','JJR','JJS']
Adverb	['RB','RBR','RBS','WRB']
Pronoun	['PRP','PRP$','WP','WP$']

Table 7.2: Temporal features and POS symbols

4. **Time Spent Feature**:

   This represents the time spent on each learner activity (n) and is represented in seconds. It is computed as the difference between the time when the learner has fully completed a step (C) and the first time the same learner visited the step (V):

$$TimeSpent_n = C_n - V_n \tag{7.2}$$

7.3.5 Data Preparation for ML Models

Some missing values were noticed in the data, and these missing values were due to the fact that not all learners had completed all the activities in each step. These missing values were filled in by adding the average value of each feature. This step is important for creating vectors with fixed lengths for ML classifiers [157]. Also, the data were not in balance, and so this was resolved using the popular Synthetic Minority Oversampling Technique (SMOTE) [40]. Next, the dataset was split into training (80%) and test (20%) sets. In turn, the researcher further shuffled and stratified the dataset for better learning performance [110].

7.3.6 Computational Classifiers

One of the objectives in this chapter is to consider less expensive computational classifiers rather than expensive and complex models for education level profiling, wherever possible. Thus, the adopted approach has included multiple feature engineering steps, as explained in Section 7.3.4. The study examples were trained on many different supervised conventional learning algorithms. In particular, SVM, NB, RF, LR, and KNN, were used. Decision Tree (DT) was also applied as this was found to render a high performance for identifying the level of education in a previous study [229]. Also, this chapter investigated the Extra Trees (ET) classifier, which is a DT-based classifier that learns in an ensemble way.

7.3.6.1 NLP Baseline Models

For comparison purposes, the researcher used three baseline models that are commonly applied in text classification tasks. A description of these NLP baseline models is given; below.

TF-IDF: Term Frequency-Inverse Document Frequency (TF-IDF) a simple and old-fashioned, but it is still considered the state-of-the-art for many NLP applications. It is a lexically-dependent but semantically independent technique. For the study on education level classification in this chapter, both character n-grams ($n = 3,6$) and word n-grams ($n = 1,2$) were used, which were the best performing n-gram settings employed for AP in the recent PAN [25]. The following equation explains a standard TF-IDF technique mathematically:

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D) \tag{7.3}$$

where TF computes the term t frequency in a comment d and IDF computes term t's inverse frequency in the collection of comments D.

Word2Vec: Word2Vec is the first neural network-based modelling approach in NLP [140], word2vec is a semantics-dependent but context-independent embedding. This study used the skip-gram-600 model (one of the word2vec algorithms), which has two layers of shallow neural networks. It consists of average word vectors, which are built based on training on a corpus of 50 million tweets [90]. The following equation provides a mathematical explanation of the model:

$$P(w_o \mid w_c) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_c)}{\sum_{i \in \mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)} \tag{7.4}$$

In the skip-gram model, the conditional probability P is calculated for context words w_o and for a central (target) word w_c by a softmax operation on the vector v inner product. There are two dimensions for each word, where i is a word index in the dictionary and its context word vector is represented as u_i. Also, its central target word vector is represented as v_i. In the second dimension, the central target word and context word are indexed as o and c, respectively.

The features extracted from this model are fed to a simple ANN algorithm. In addition, three Transformers models are used as baseline models in this chapter, namely **(BERT, XLnet, and RoBERTa)**. An explanation of the Transformers models was explained in Section 2.3.5, Chapter 2.

7.3.6.2 Extra Trees Classifier

The researcher found that no study in the AP literature, particularly for educational traits identification, used the Extremely Randomised Trees (the Extra Trees (ET)), as can be seen in Section 3.4, Chapter 3. The ET model [79] is fundamentally an ensemble of DT, similar to other DT-based models such as RF. However, ET classifier is built using more non-pruned decision trees than RF, and the prediction is based on majority voting if the task is a classification task (see definition in Section 2.4.7, Chapter 2) [79].

Figure 7.1 illustrates the general workflow of the experiments visually.

7.4 Results and Discussion

For the educational level experiment in this chapter, a challenge associated with the use of DL models was the need for large amounts of data for training. This was especially challenging because, in this chapter's experiment, course level classification was the focus, which meant that

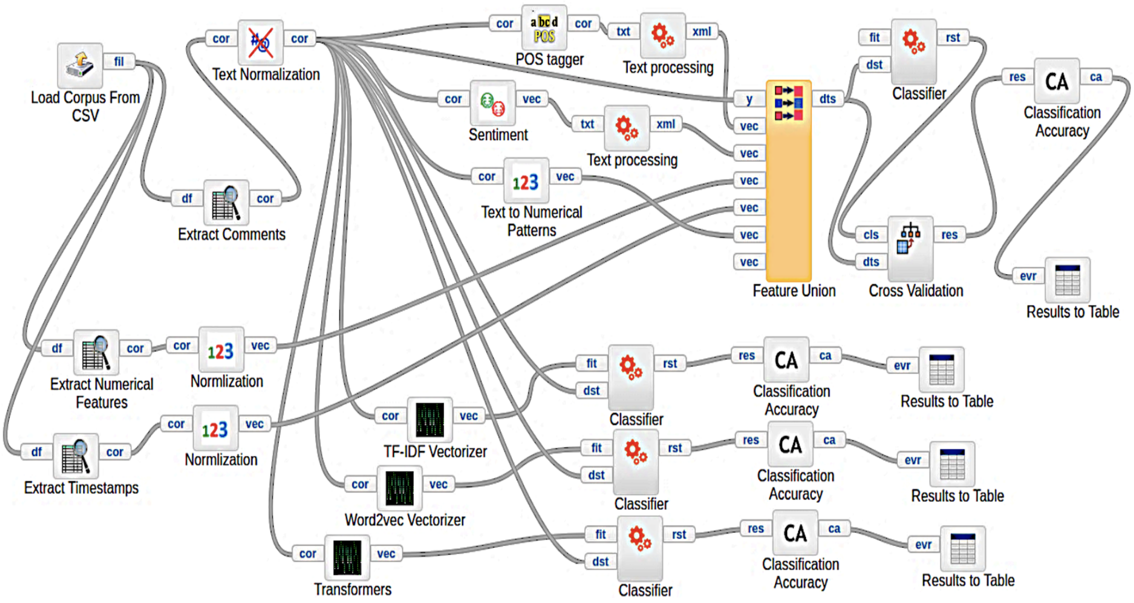


Figure 7.1: General workflow of the proposed level of education classification approach

fewer data were available for training. Thus, conventional ML models were applied for the experiment, which are optimal for experiments with fewer data (see results presented in Table 7.3). Thus, this chapter contributes to this issue by providing a course-level classification model adjusted to the learner educational profiling.

The fact that AP mainly relies on texts only could represent a limitation, especially when fewer textual samples are used. In addition, there are other metadata that have not yet been explored in the area. Furthermore, content-based features are usually used in the number of hundreds or even thousands of features for classification; these range from lexical and semantic to syntactical features, and they are based on grammars, n-grams, frequencies, and token levels, among others. This should be very effective when two objectives are met: enormous text samples from a specific domain (here, a course). When this is not the case, the researcher proposes that the task can be also solved using other approaches or by a deep examination of other potential features and metadata available in the MOOC domain.

In this chapter, the education level below a bachelor's degree was not considered. There is a large gap between course content that is appropriate for secondary or lower educational learners and content for those who have a bachelor's degree or higher. A larger data size with varied course content is needed in order to include learners in secondary education, or lower; this highlights a key area for further investigation.

Baseline models were applied, firstly, in this study to identify learners' level of education based only on text (i.e., a traditional way of solving NLP tasks, in general). The use of comments alone, based on these models performance, did not provide satisfactory results, see Table 7.3. This could be because comments in the dataset were dominated by a course context; models learned more about a course content rather than learners writing style. This may have affected the performance of these state-of-the-art NLP algorithms in this study, according to results presented in Table 7.3. Despite the simplicity of textual representations of TF-IDF or word2vec, they performed competitively compared to text representations via transformer models. BERT and XLnet outperformed word2vec and TF-IDF both at the character-level and word-level, but their results were not satisfactory for state-of-the-art models, especially RoBERTa, which recorded lower results than word2vec and TF-IDF. BERT performance was the highest among these baseline models.

This study also revealed that MOOC metadata outperformed baseline models except for enrollment features, that is, based on ET models. Quiz features, Time-spent features, and Comment features (which is even simpler than TF-IDF and word2vec) all achieved the highest accuracy. The Extra Trees (ET) classifier achieved the highest performance across all experimental settings, as shown in Table 7.3 along with the overall accuracy levels per course and feature category. These results were validated using 10-fold Cross-Validation (CV), which is well known for avoiding overfitting [97], especially when the data size is small. In each iteration (k), a single accuracy is estimated, after which all accuracies are averaged to obtain the final accuracy (A). The 10-fold CV accuracy is computed via the following formula (k-fold CV accuracy, where $k = 10$):

$$Accuracy = \frac{1}{k}\Sigma_{i=1}^{k} A_i \tag{7.5}$$

It is important to mention that since the pre-trained models (transformers) used in this chapter did not outperform the other simple models considered in this chapter, a further fine-tuning procedure was deemed necessary. The rationale for undertaking this fine-tuning procedure, according to [101], is that transformers need perfect fine-tuning to serve as successful tools for AP studies. This also could support the researcher's assumption that using simple and basic textual features is

Approach	**BD**	**BM**	**MF**	**SH**	**Average**
TF-IDF (char)	0.75	0.78	0.62	0.68	0.7075
TF-IDF (word)	0.80	0.84	0.75	0.66	0.7625
Word2vec	0.76	0.83	0.75	0.68	0.755
BERT	0.76	0.87	0.80	0.81	**0.81**
RoBERTa	0.58	0.79	0.70	0.73	0.70
XLnet	0.80	0.83	0.74	0.76	0.78
Enrollment + SVM	0.33	0.33	0.34	0.34	0.335
Enrollment + NB	0.37	0.39	0.35	0.34	0.3625
Enrollment + LR	0.35	0.38	0.35	0.36	0.36
Enrollment + KNN	0.57	0.67	0.64	0.57	0.6125
Enrollment + DT	0.65	0.76	0.72	0.67	0.7
Enrollment + RF	0.67	0.77	0.74	0.68	0.715
Enrollment + ET	0.67	0.78	0.74	0.68	**0.7175**
Quiz + SVM	0.34	0.34	0.34	0.34	0.34
Quiz + NB	0.38	0.39	0.36	0.36	0.3725
Quiz + LR	0.38	0.37	0.35	0.36	0.365
Quiz + KNN	0.81	0.88	0.82	0.69	0.8
Quiz + DT	0.87	0.91	0.85	0.72	**0.8375**
Quiz + RF	0.86	0.91	0.84	0.72	0.8325
Quiz + ET	0.85	0.90	0.84	0.72	0.8275
Time Spent + SVM	0.38	0.37	0.37	0.34	0.365
Time Spent + NB	0.41	0.41	0.35	0.38	0.3875
Time Spent + LR	0.48	0.46	0.41	0.38	0.4325
Time Spent + KNN	0.57	0.68	0.66	0.81	0.68
Time Spent + DT	0.75	0.80	0.81	0.87	0.8075
Time Spent + RF	0.80	0.85	0.86	0.86	0.8425
Time Spent + ET	0.82	0.85	0.87	0.84	**0.845**
Comment + SVM	0.33	0.36	0.39	0.36	0.36
Comment + NB	0.40	0.41	0.37	0.40	0.395
Comment + LR	0.43	0.39	0.46	0.42	0.425
Comment + KNN	0.60	0.74	0.63	0.60	0.6425
Comment + DT	0.79	0.86	0.81	0.78	0.81
Comment + RF	0.84	0.93	0.88	0.86	0.8775
Comment + ET	0.85	0.92	0.90	0.88	**0.8875**

Table 7.3: Overall accuracy per feature category and course with baseline models (Courses: Big Data (BG), Babies in Mind (BM), The Mind is Flat (MF), and Shakespeare (SH))

a potentially valuable direction to solve the chapter's issue.

The simple comment features outperformed all other features in this chapter's experiments. Therefore, the comment features were combined with each group of features. It was found that the time spent and comment combination of features achieved the highest accuracy compared to all approaches and settings in the experiments (see Table 7.4).

Approach	**BD**	**BM**	**MF**	**SH**	**Average**
Enrollment + Comment + DT	0.75	0.83	0.80	0.80	0.795
Enrollment + Comment + RF	0.80	0.89	0.85	0.88	0.855
Enrollment + Comment + ET	0.82	0.90	0.87	0.90	0.8725
Quiz + Comment + DT	0.81	0.90	0.84	0.73	0.82
Quiz + Comment + RF	0.84	0.91	0.85	0.74	0.835
Quiz + Comment + ET	0.85	0.91	0.85	0.74	0.8375
Time Spent + Comment + DT	0.79	0.85	0.81	0.78	0.8075
Time Spent + Comment + RF	0.84	0.91	0.88	0.86	0.8725
Time Spent + Comment + ET	0.85	0.94	0.90	0.88	**0.8925**

Table 7.4: Overall accuracy of features combination per feature category and course (Courses: Big Data (BG), Babies in Mind (BM), The Mind is Flat (MF), and Shakespeare (SH))

The results demonstrate that the selected features are so representative that they work well, even with extremely unbalanced data. The research has also demonstrated that using state-of-the-art NLP models is not supportive enough for what is supposed to be mainly a text classification task, which is attributable to the domain conditions. However, the results presented in this chapter are promising, which may reinforce the use of this approach.

The chapter's results also indicate that simple and computationally efficient algorithms compare with more complex deep models, with a lot of semantic information, can be effective solution for the problem. In practice, the most suitable target for the proposed classification task is users with some interaction with the system; this is because the feature extraction process relies on their interaction, time, comments, and other information. This could explain the reason behind achieving different results in each course, individually, for each category of features. Obtaining data in similar distribution both in course level and feature category level is one of the limitations in this chapter. Hence, the greater the learner interaction, the more information that is available for the models. In this chapter, the learners generated the required features for the chosen category goal.

7.4.1 Performance Significance

As shown in Table 7.3, three classifiers – namely, DT, RF, and ET – outperformed the other conventional ML. However, between themselves, DT, RF, and ET achieved similar results. The next question of interest, therefore, is whether the results of the three classifiers are significantly different or not. Therefore, the models were further compared using 5×2-fold cross-validation (CV) [58], meaning a two-fold CV with five iterations.This technique creates repetitive comparisons between supervised classification learning algorithms to enable a more efficient comparison. Therefore, the researcher evaluated the obtained results by pairwise comparison of the winning model to the others: ET versus DT, and ET versus RF. The idea was to determine whether the ET model achieved consistently higher performance (i.e., its high performance was not just a product of chance). Therefore, the function compared two classifiers (x and y) with labelled data that were split repeatedly five times as 50% training and 50% test data (see Figure 7.2).

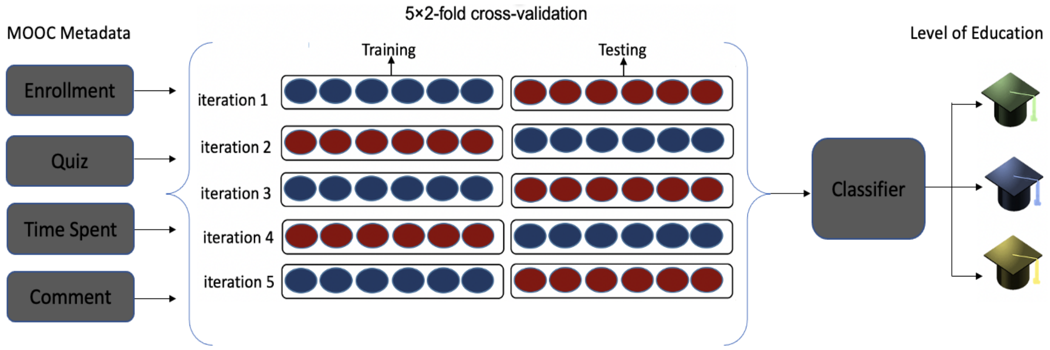


Figure 7.2: Training MOOCs features and the three conventional ML (DT, RF, and ET) based on 5×2-fold cross-validation

In each single iteration of the five, x and y are applied to the same split training set and then they are evaluated based on their relative performance P: (P^x and P^y) on the split testing set; then this process is repeated. This means that for each iteration (i), there are two different measurements of the classifier performance: $P_i^{(1)} = P_{1i}^x - P_{1i}^y$ and $P_i^{(2)} = P_{2i}^x - P_{2i}^y$. From these two measurements, the mean $(\overline{P})$ and variance (s^2) are estimated in the iteration:

$$mean = \overline{P_i} = \frac{P_i^{(1)} + P_i^{(2)}}{2} \tag{7.6}$$

$$variance = s_i^2 = (P_i^{(1)} - \overline{P}_i)^2 + (P_i^{(2)} - \overline{P}_i)^2 \tag{7.7}$$

The final difference variance, computed from the five iterations, is used to obtain the t-statistic (t), based on the following formula:

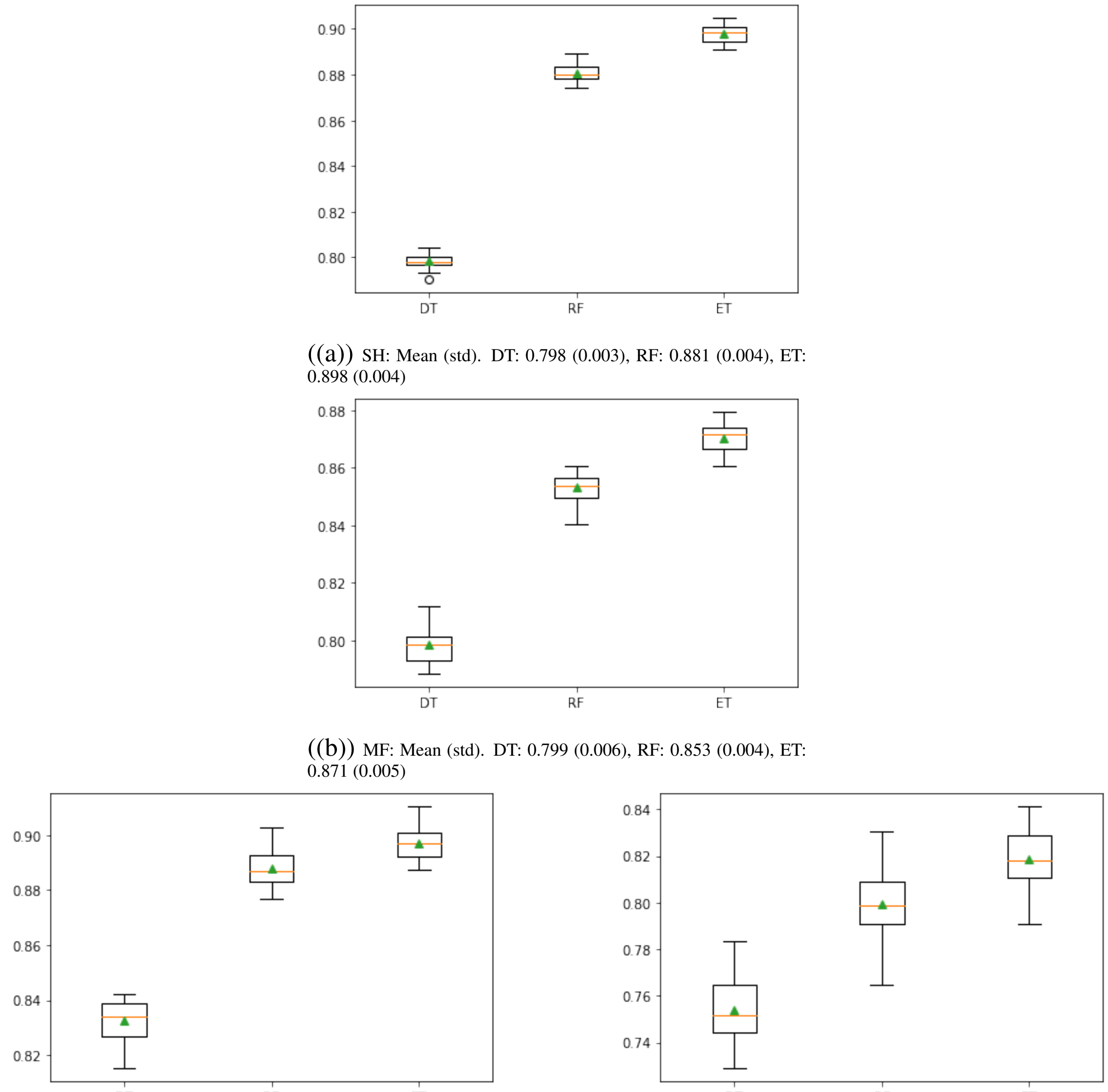
$$t - statistic = t = \frac{P_1^{(1)}}{\sqrt{(1/5)\sum_{i=1}^{5} s_i^2}}, \tag{7.8}$$

where $P_1^{(1)}$ is the score difference of the classifier in the first fold of the first iteration, and s_i^2 is the score difference estimated variance for the i^{th} iteration.

	p-value	t-statistic	significance
Enrollment + Comment :			
BD(ET Vs.DT)	0.002	-6.057	Yes
BD(ET Vs.RF)	0.053	-2.524	No
BM(ET Vs.DT)	0.000	-13.601	Yes
BM(ET Vs.RF)	0.004	-5.166	Yes
MF(ET Vs.DT)	0.000	-11.029	Yes
MF(ET Vs.RF)	0.006	-4.569	Yes
SH(ET Vs.DT)	0.000	-22.106	Yes
SH(ET Vs.RF)	0.001	-7.524	Yes
Quiz + Comment :			
BD(ET Vs.DT)	0.002	-6.064	Yes
BD(ET Vs.RF)	0.003	5.218	Yes
BM(ET Vs.DT)	0.178	1.567	No
BM(ET Vs.RF)	0.278	1.216	No
MF(ET Vs.DT)	0.000	10.492	Yes
MF(ET Vs.RF)	0.007	4.429	Yes
SH(ET Vs.DT)	0.008	-4.215	Yes
SH(ET Vs.RF)	0.674	-0.447	No
Time Spent + Comment :			
BD(ET Vs.DT)	0.001	-6.421	Yes
BD(ET Vs.RF)	0.265	-1.255	No
BM(ET Vs.DT)	0.000	-9.712	Yes
BM(ET Vs.RF)	0.168	-1.609	No
MF(ET Vs.DT)	0.000	-15.204	Yes
MF(ET Vs.RF)	0.019	-3.390	Yes
SH(ET Vs.DT)	0.000	-26.721	Yes
SH(ET Vs.RF)	0.001	-6.928	Yes

Table 7.5: Significance measurements for the three models comparison (ET versus DT and ET versus RF), for each features combination in each course

Also calculated was the probability *p-value*, compared to 0.05 (i.e., a significant difference between the two models, rejecting the null hypothesis, appears if *p-value* < 0.05). According to the results presented in Table 7.5, one can see that the *p-value* in most cases is less than 0.05. These results are also presented visually, which confirms the finding that ET outperforms both DT and RF, in Figures 7.3, 7.4, and 7.5.



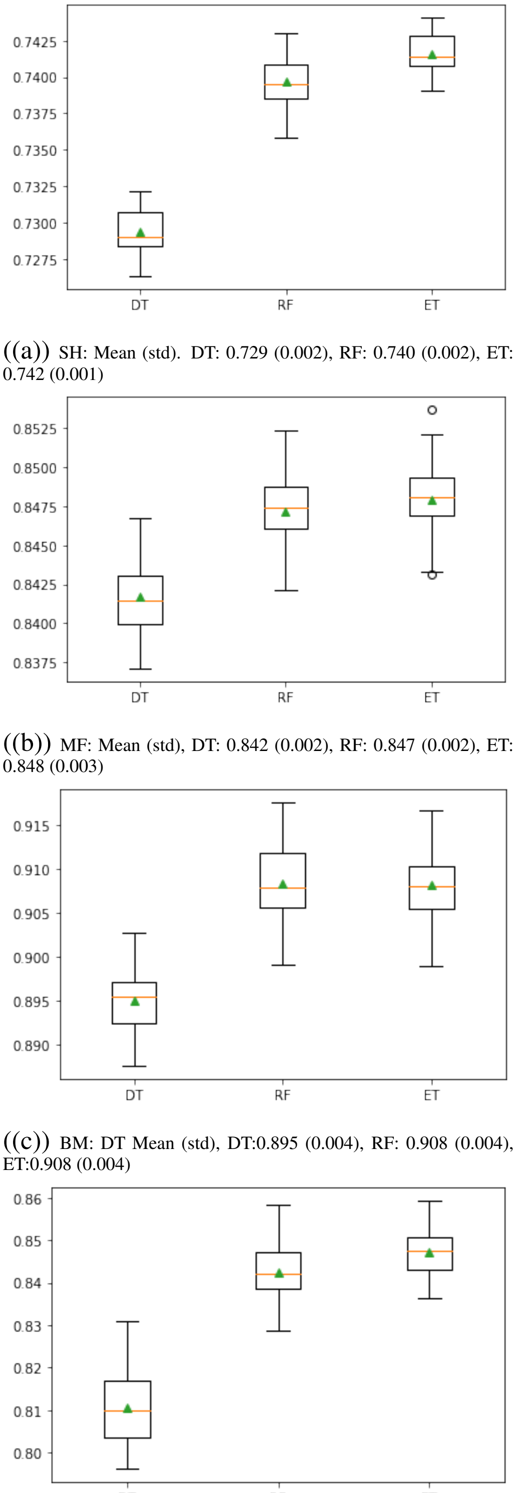
((a)) SH: Mean (std). DT: 0.798 (0.003), RF: 0.881 (0.004), ET: 0.898 (0.004)

((b)) MF: Mean (std). DT: 0.799 (0.006), RF: 0.853 (0.004), ET: 0.871 (0.005)

((c)) BM: Mean (std). : 0.833 (0.007), RF: 0.888 (0.006), ET: 0.897 (0.006)

((d)) BD: Mean (std). DT: 0.754 (0.013), RF: 0.799 (0.015), ET: 0.819 (0.013)

Figure 7.3: Models performance: enrollments and comments



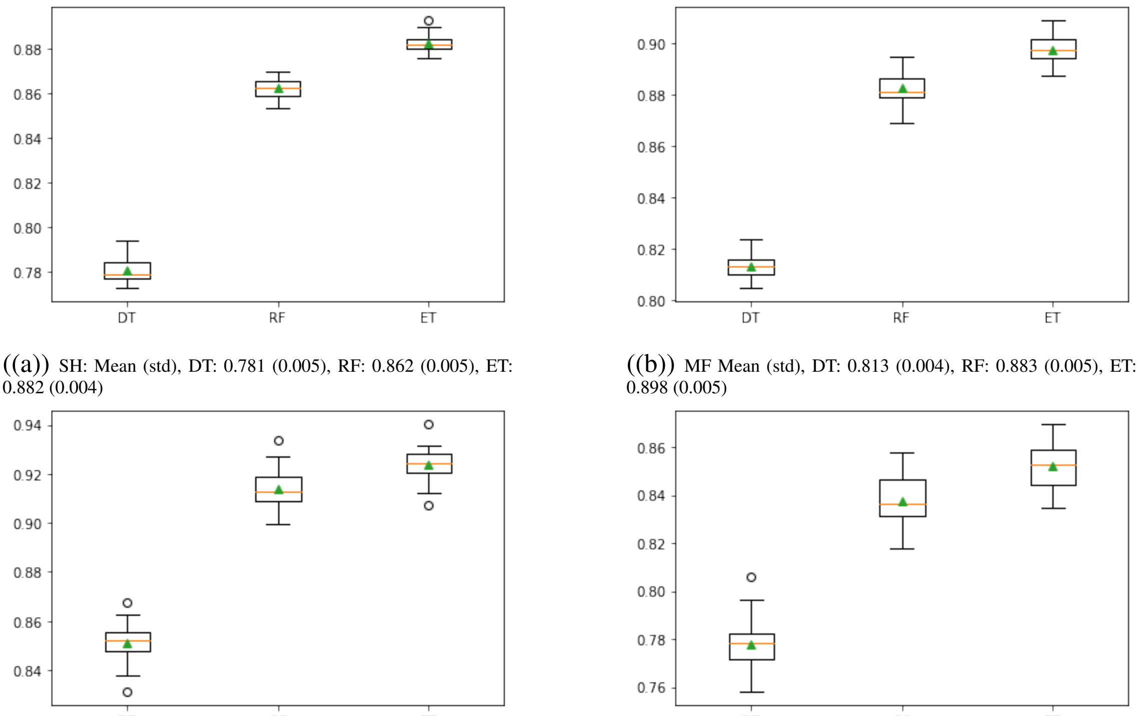
((a)) SH: Mean (std). DT: 0.729 (0.002), RF: 0.740 (0.002), ET: 0.742 (0.001)

((b)) MF: Mean (std), DT: 0.842 (0.002), RF: 0.847 (0.002), ET: 0.848 (0.003)

((c)) BM: DT Mean (std), DT:0.895 (0.004), RF: 0.908 (0.004), ET:0.908 (0.004)

((d)) BD: Mean (std), DT:0.811 (0.008), RF: 0.842 (0.006), ET:0.847 (0.005)

Figure 7.4: Models performance: quizzes and comments



((a)) SH: Mean (std), DT: 0.781 (0.005), RF: 0.862 (0.005), ET: 0.882 (0.004)

((b)) MF Mean (std), DT: 0.813 (0.004), RF: 0.883 (0.005), ET: 0.898 (0.005)

((c)) BM: Mean (std), DT: 0.851 (0.007), RF: 0.914 (0.007), ET: 0.924 (0.007)

((d)) BD: Mean (std), DT: 0.778 (0.011), RF: 0.838 (0.010), ET: 0.852 (0.009)

Figure 7.5: Models performance: time spent values and comments

7.4.2 Feature Importance

The above results are useful and applicable, but under their umbrella, many features are potentially hidden. In this section, a further examination in-depth of each feature set is given in terms of its individual importance. The idea is to obtain a general understanding of the degree of contribution of each to the classifier performance; this is intended to show which are the most influencing factors. This step also helps to decide which feature set to select, and then it eliminates irrelevant or redundant feature sets. This is further critical to minimise training time by decreasing the dimensions of the feature matrix. The researcher examined the Enrollment, Quiz, and Comment feature sets, respectively. Since the Time Spent feature only has one set, it was not included in this step (see Figure 7.8, which provides a visual description of the importance of each feature set). From the figure, it is clear that, for example, *e_hour* (enrollment hour) has a strong impact on the learner profiling for all examined courses; similarly, *q_hour* (quiz hour) is also a good predictor, unlike correct response, or *q_month* (quiz date in month).

The punctuation count feature seems the most predictor among comments features. Part-of-Speech (POS) features, including *pron_count*, *adv_count*, *verb_count*, have a similar impact on the classification of a learner's profile. However, sentiment analysis (SA) seems to have a noticeably limited influence on the outcome, which is much less than that of other features. Further investigation about theses feature (Features Correlation) is presented in Appendix B.

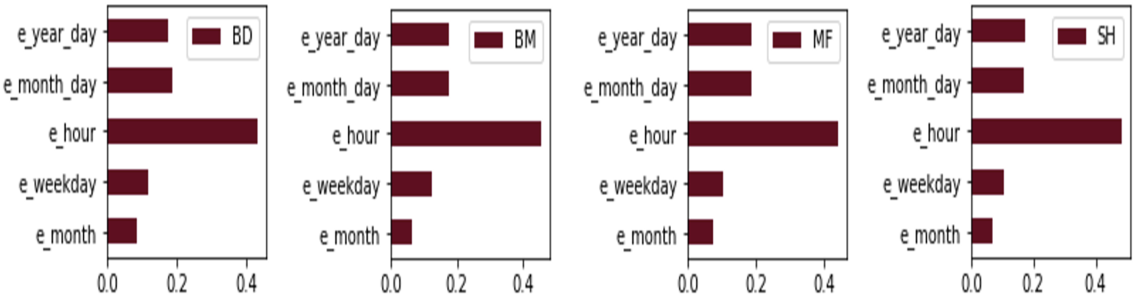


Figure 7.6: Importance of enrollment features

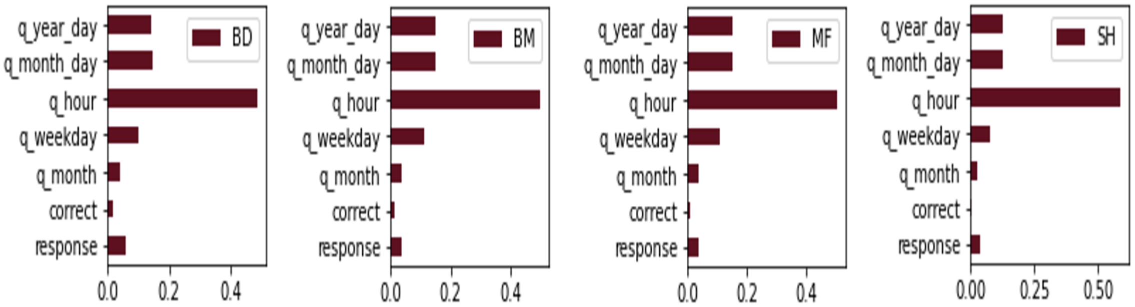


Figure 7.7: Importance of quiz features

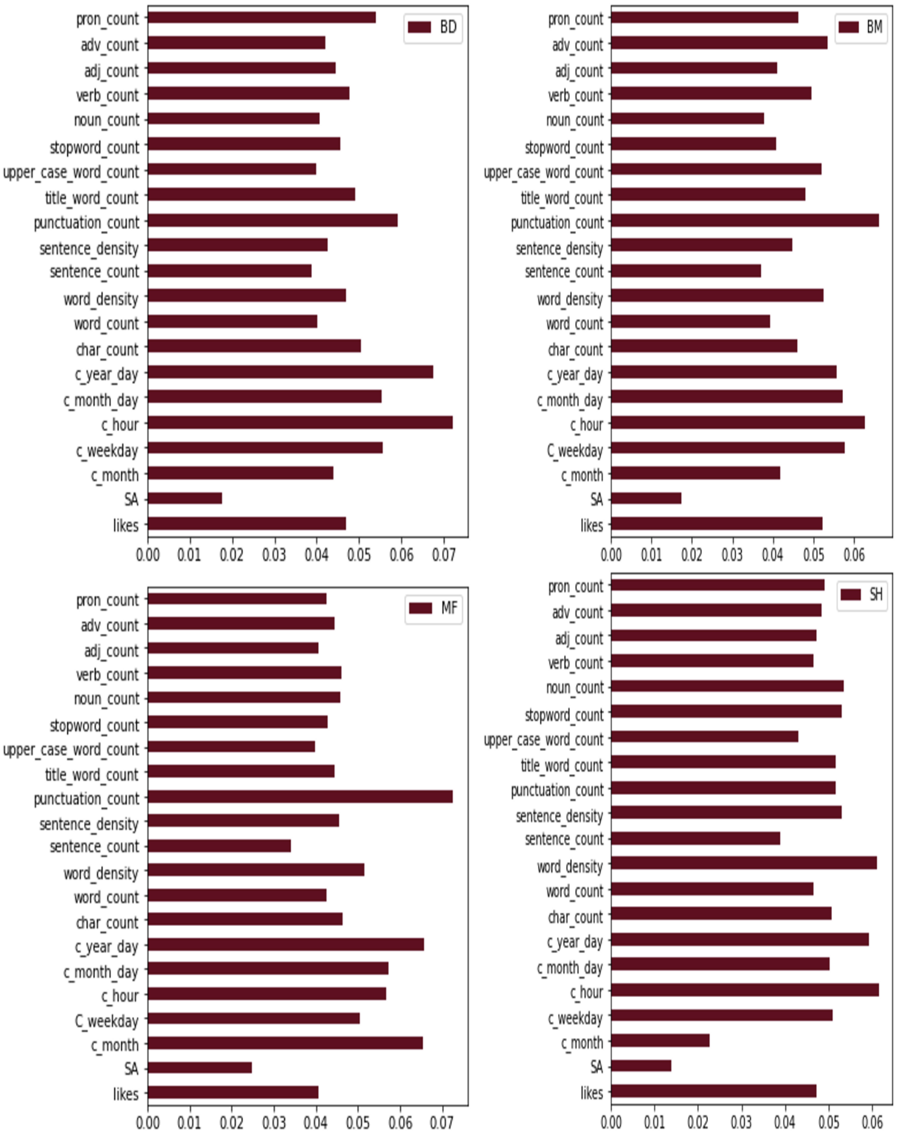


Figure 7.8: Importance of comment features

7.5 Epilogue

This chapter presented a solution to solve an educational level classification problem. The problem solved is noteworthy because it applies to the domain of MOOCs, in which a key challenge is the domain's simple textual representations about learners, as well as the very simple metadata available in the domain. Therefore, in this chapter, both textual and behavioural data were used (i.e., the time taken to complete a course's steps). The proposed approach not only achieved a high performance but also demonstrated that this task can be performed via inexpensive computational algorithms regardless of the data size.

The extracted features could help to classify these learners based on thier educational level to help personalising systems in MOOCs. For example, it is possible to help them to avoid taking an irrelevant course or to recommend a suitable course for their current level of education. This is because bachelors' students take basic courses compared to master's students who take advanced courses, for example. This information is also helpful for other systems in MOOCs, such as cheating detection.

The next chapter offers a discussion of the general findings of the experiments in this thesis and potential avenues for future research works.

Chapter 8

Discussion

8.1 Prologue

In this thesis, a research referred to as learner profiling (LP) is examined in new context, which focuses on identifying learner demographic characteristics on MOOCs. This is a strategy for, amongst other aims, overcoming biases that may arise from the current way of extracting learners' demographic data, which is based on traditional pre-questionnaires (see Section 1.3.1). Hence, the models provided in this research can serve as a means to design customised recommendation systems in MOOCs.

In this chapter, the importance of learners' demographics in MOOCs research is presented. Next, the significance of this research based on its findings and its overall contributions are described, as well as how the research approaches presented in this thesis have surpassed those described in prior literature. Furthermore, limitations of the approaches described in the previous chapters are discussed, and further research opportunities are identified.

8.2 Impact of Learner Demographics on MOOCs

Generally, learning must be adapted based on demographic factors because this enhances learning and improves outcomes. For example, teaching a language course is customisable based on age, and it is widely recognised that the use of age-appropriate materials is preferable when teaching children language courses compared to adults [133]. In the case of children, the consideration of keeping learners interested in the course, for instance, is a key factor that is not usually a concern in adult education.

At present, the current MOOCs are sufficient only for users who may be considered "advanced learners" and who do not require support to navigate courses and materials. As a result, it has been reported that only 7% of MOOCs learners complete their courses [133]. Thus, despite all the progress in the research on MOOCs, there are still many opportunities for development [80]. In particular, personalised recommendations are required [84] based on demographic information, which is in demand for many types of MOOCs studies, as discussed in Section 3.5.2, Chapter 3.

8.2.1 MOOCs After COVID-19

In the course of the pandemic, MOOCs have emerged as an effective solution for crisis management in education systems [201], [93]. The pandemic has boosted a global demand for MOOCs and promoted the demand for online education in the future, especially as it breaks any spatial or temporal limitations (see Section 1.3.1, Chapter 1) [71]. This pattern is expected to persist even after the pandemic, shaping the future of MOOCs and emphasising the importance of improving these platforms, as well as being prepared as an emerging system of education during the ongoing pandemic and beyond [196] [80]. This clearly explains the importance of further research into MOOCs in the future.

8.2.2 Personalisation Systems in MOOCs

To gain an in-depth understanding of MOOC learners, MOOCs providers have begun to actively survey learners to obtain demographic profiles. For instance, all prior MOOCs studies have relied on questionnaire surveys (web-based surveys) to collect demographic data from users. However, the main issue with this approach is the response error; this occurs in studies when the samples do not reflect the actual population for conducting a study or research relating to MOOCs [222]. According to [222], many studies in MOOCs have attempted to use different practices to encourage learners to respond to these surveys, see Chapter 3.However, as these methods still do not provide enough responses, these studies also call for replacing these surveys with alternatives methods [222]. Thus, my research instead aims to employ an automatic approach to extract learner demographic information, thereby, arguably, providing unbiased data, which is vital for MOOC personalisation purposes.

Regardless of the way they are obtained, demographic characteristics have been effectively used as determinants (variables) of student achievement, in traditional education research [198]. This is because they are critical inputs into personalised systems. Many studies of MOOCs have considered learner demographics for MOOC personalisation systems, as indicated by systematic reviews of the MOOCs literature in [156] and [120]. Developing a personalisation system in MOOCs, adapted to learners' gender, jobs, and educational level, is considered essential for increasing the engagement among learners in MOOCs [77].

For example, non-working learners can be guided to take courses that are currently trendy in the job market, to improve their chance of finding jobs, while learners who are currently working can be guided to take advanced courses to improve their skills. Also, females can be guided to take courses that include more feminine issues such as motherhood, while males can be guided to take courses that include examples based on more interesting topics to them, such as cars or football (note: adaptation could also determine the learners that are not conforming to these simple stereotypes, and create much more refined learner models). In addition, learners at a particular level of education can be helped to avoid taking a course irrelevant to them, instead being recommended a suitable course for their current level of education. For instance, Bachelor's students would be taking more basic courses, compared to the Master's students, who would be guided to more advanced courses.

To the best of my knowledge, the research presented in this thesis has attempted to extract this demographic data via a new strategy, based on automation, and thus arguably requiring less cognitive overhead for the learners (who can concentrate on the learning only and avoid answering questions if they so wish to). However, this approach introduces some limitations, such as the computational complexity of the approach. This is further discussed in section 8.5.

8.3 Impact of Learner Profiling

The term "Learners Profiling" has been mentioned in the literature review for different purposes; such as "identifying groups of students based on their similar academic behaviour" [143], "comparing similar learning behaviour of users of a mathematics training application" [88], or "gathering learner profiles to be used as a guide in an attrition study" [121]. However, no study of them has targeted MOOCs learners or aims to solve the problem of extracting the learners' demographic profiles based on computer science techniques. The term LP that is used for the purpose of this research has been defined clearly in the title of this thesis.

8.3.1 Research Findings

To address the research umbrella research question (Section 1.4, Chapter 1), the researcher determined a number of specific learner profile demographics for the classification problems: employment status, gender, and level of education; utilising MOOCs metadata that are processed by NLP and ML approaches. This is for combining LA and AP to perform LP. As explained in previous sections (Sections 5.2, 6.2, and 7.2) reasons for considering these demographics for this research were based on their importance in personalisation systems in the domain.

Thus, one of the main research questions was about the possibility of classifying employment status based on the comments that learners exchange on MOOC discussion forums. The subquestions for this question are as follow:

- RQ1: Is it possible to classify learners' employment status from only the comments they exchange on the MOOC system based on a DL approach?
  This is solved by collecting a huge data size of only learners' comments from all courses in the research data for training the proposed DL models for this particular classification problem (See Section 5.3.1). DL models work effectively when a large number of samples is available to learn from them. This can be clearly seen via the high-accuracy of classification that models achieved, according to the results presented in Table 5.1, Chapter 5.

- RQ2: Can the data imbalance issue among learners' comments based on their categories be solved using an NLP approach?

  This was solved by an effective paraphrasing technique, which did not change the meaning and structure of the sentence, which built upon a translation approach (see Section 5.3.3.1). The paraphrasing technique performance compared against a popular balancing method (SMOTE), and the classifiers recorded higher accuracy with the paraphrasing balancing strategy, as can bee seen in Table 5.1, Chapter 5.

- RQ3: Which DL architectures are viable to use for classifying learners' employment status?

  This was solved by using DL algorithms that already show effectiveness in the fields of NLP and AP, which is an ensemble learning of CNN and RNN, comparing two types of ensembles architectures (see Sections 5.3 and 5.3.6). Based on Table 5.2 in Chapter 5, which shows details about the parallel and the sequential model's performance for the employment status classification, it can be seen in the table, even at the detailed category level, these two types of ensemble architectures perform exceptionally well. However, sequential ensemble learning achieved better results than parallel ensemble learning, and 96.4% overall accuracy was obtained.

Another main research question in this thesis was concerned with examining advanced textual features that are extracted from MOOCs discussion forums to classify a learner's gender. This question is divided into three sub-questions, as follows:

- RQ1: Which NLP representation can be applied to extract advanced syntactic-level representations from learners' MOOC discussion forum comments?

  This is solved by examining an advanced NLP tool, to provide a syntactic representation of texts, called Parser. A constituency parser (PCFG) was used, due to its effectiveness in NLP literature, as discussed in Section 6.3.2, Chapter 6. It is based on an expert-designed grammar to handle the sentences/phrases (syntactic) level of the text. This representation of text was used because it provides a comprehensive interpretation of a sentence's meaning [155], as explained further in Section 6.3.4.1.

- RQ2: Which DL-based algorithms are available to handle advanced syntactic-level representations?

  The used parser presents a text in a tree structure, therefore, state-of-the-art recursive models able to handle such a structure are explored (tree-structured LSTM, SATA, and SPINN models), and they are applied for this particular classification problem. The results of these models were promising, as presented in Table 6.1 in Chapter 6.

- RQ3: How can algorithms based on RecNN be designed so as to classify MOOC learners' gender?

  To improve the performance of the utilised recursive models, a bi-directional strategy was added to them. This is because bi-directional learning has already shown its effectiveness in improving the sequential LSTM model in the NLP literature [215]. As a result, 18 different versions of the bi-directional function for existing model architectures were examined in (combinations) of word-level and sentence-level encoding. These bi-directional composition function models achieved slightly better results; nevertheless, they are similar to each other, as and accuracies tend to be identical, as can be seen in Table 6.1 in Chapter 6.

Also, this research attempted to answer two more main research questions. One of them was about classification of a learner's level of education based on a MOOC discussion forum data, specifically, on a course-level; while the other was about the use of metadata in addition to the MOOC discussion forum data to improve this classification problem. These research questions were divided into three sub-questions, as follows:

- RQ1: What MOOC metadata is available for extraction, based on NLP and non-NLP features, for use as predictors for a learner's educational level?

  After a comprehensive investigation of the research data, features that were extracted belong to three MOOCs metadata categories: time-stamps, quizzes, and discussions, to identify the educational status of learners. This was solved by a careful inspection that has been done in this research, during the feature extraction process (Section 7.3.3), feature engineering (Section 7.3.4), as well as during data preparation for ML models (Section 7.3.5).

- RQ2: How do the classifiers achieve high accuracy despite the simplicity of the applied features?

  As mentioned above, a careful inspection has been done in this research, thus, despite the simplicity of the applied features, this careful inspection was effective in term of reducing noises and biases in the data samples before feeding them to different classifies. This is also because the used features include different interactions of learners with the MOOCs environment.

- RQ3: Which classifier can identify a learner's educational level regardless of data size?

  Inexpensive classifiers achieved good results for this particular classification problem. This is due to the quality of the provided samples (features). This was confirmed by applying many different classifies that cover different types of models: NLP baseline models (based on TF-IDF and Word2Vec), Transformers (BERT, XLnet, and RoBERTa), and different types of conventional ML models (in particular, SVM, NB, LR, KNN, DT, RF, and ET). ET was found to render the highest performance for identifying the level of education based on a combination of feature categories; time spent and comments, as can be seen in Table 7.4, Chapter 7.

8.3.2 Research Data

Compared to other learning systems, MOOCs create massive amounts of data that analysts cannot easily or feasibly process manually. However, the availability of MOOC data on this considerable scale offers a unique opportunity to research a large human learning system, providing great data-driven solutions, in particular, to handle learners' needs based on their differences [4]. This is a significant benefit, as shown in this thesis, where the issue of dataset size is discussed in Section 4.2.1, Chapter 4). Notably, creating similarly-sized datasets may be difficult in any other type of educational system. This makes the data obtained in this research, which were collected from MOOCs platform, effective for the purpose of achieving the research objectives.

As shown in Chapters 5 and 6, deep learning models are always 'thirsty' for additional data to perform more effectively [30]. The effect of dataset size is obvious in all of the results presented in this thesis. In addition, the data is rich in metadata, which helps even further when the dataset size has been reduced to meet course level based experiment in Chapter 7. Furthermore, MOOCs are characterised by their involvement of a variety of disciplines and course subjects, which influences the diversity of content in the data [138]. More than one type of content is often used to reduce the bias in ML, where the tendency has been to learn using only one [138]. This should have reduced the level of bias in the experiments in this research. The data collected from different disciplines were explained in Section 4.2.1, Chapter 5.

The data used in this research are more appropriate than other public MOOCs data, especially considering the particular objectives of this thesis. Currently, there are six public data in MOOCs, according to [210]. For example, the Stanford MOOCPosts Data Set[*] is a public MOOCs dataset, yet it includes 29,604 posts; it is considerably smaller than the dataset used in this thesis. It contains the comments of learners labelled with their gender, but it does not include labels for learners' comments based on their occupation status. Additionally, although it includes educational information about learners, it does not include all the types of metadata used in Chapter 7 for the educational level experiments. Another available dataset is the Coursera Forum Discussion[†] (with 739,093 comments); however, this dataset does not contain any information about learner demographics. The Open University Learning Analytics dataset[‡] (OULAD), is another public dataset, but it contains no comments or the other types of metadata used in this thesis. The remaining public data cannot be compared to this thesis's data. HarvardX Person[§] includes only the activities of students from a single edX course, while KDD Cup 2015 [¶] includes learners' logs for only 30 days. The Act-Mooc[‖] dataset only includes the students' reaction networks (social network).

With these considerations in mind, the richness of the data used in this thesis offers a substantial advantage for the research. The final results that this thesis has yielded are very promising. In Chapter 5, an average accuracy of 96.3% was achieved for the best method for employment status

[*] https://datastage.stanford.edu/StanfordMoocPosts/
[†] https://github.com/elleros/courseraforums
[‡] https://analyse.kmi.open.ac.uk/open_dataset#description
[§] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147
[¶] http://moocdata.cn/challenges/kdd-cup-2015
[‖] https://snap.stanford.edu/data/act-mooc.html

classification, while in Chapter 6, 82.60% accuracy was obtained for gender classification. Lastly, the approach proposed in this thesis classified learners' academic level achieved high accuracy (89% on average).

8.3.3 LP Research

This section discusses how this research has improved on the prior literature in various ways. Evidently, most AP research in recent decades has been applied to social media platforms such as Twitter [15]. By contrast, this thesis expands the AP literature by proposing a new direction and a new domain, namely, the educational domain of MOOCs. MOOCs are a novel application domain for the AP area. To the best of the author's knowledge, this thesis is the only piece of research that has considered an online educational platform and examined its metadata for demographic profiling.

It is critical to analyse any new domain individually when dealing with AP methods. This is because AP methods typically do not perform the same across venues when applied to online platforms. The reason for this is due to the nature of the texts that are published on these venues, which are normally have different types of attributes, content, or symbols depending on the platform (e.g., hashtags on Twitter) [91]. Therefore, this research focuses on a specific domain that naturally has characteristic data patterns, which has different types of users, different goals, and different texts lengths and contents. An important implication of this is that effective models proposed for AP are generally unsuitable for direct transfer and use in exactly the same way for LP. As emphasized by Kaati et al. [105], AP models that were trained on texts from one particular domain significantly underperformed when applied to another domain, which means that AP models primarily rely on training data to achieve high performance. This observation has served as a key motivation for this research, and it prompted the researcher to discover, develop, and evaluate novel approaches that are different from what has been proposed in the literature.

The target of this research is the heterogeneous nature of MOOC environments in terms of their learner demographics, particularly based on *employment status*, *gender*, and *level of education*. In MOOCs, 60% of learners enrol to gain useful skills for their professional occupation [238].

According to the authors in [144], individuals frequently change employment during their lives, which means they will benefit from learning new work skills continuously in the future. In this context, MOOCs serve as a means for doing so [106]. In addition, 40% of learners have been shown to enrol in MOOCs for educational reasons [238], which highlights the importance of considering this demographic in this research. Both employment classification and educational classification have limited studies in the AP literature in general, as discussed in Section 3.2.2, Chapter 3. Thus, this research contributes to the AP area by investigating the less-examined user traits for profiling. Gender classification is also considered in this research; although it is heavily investigated in the AP area, a learner's gender is a critical factor for learning in personalised systems, as demonstrated in Section 3.2.2, Chapter 3 and Section 6.2, Chapter 6.

The main concern of this research is to provide representative data of different learner demographics in MOOCs. This focus is motivated by the limitations of the current methods for obtaining such information in MOOCs. This research responds to the problem by combining AP and LA in the domain of MOOCs to perform LP. This enables establishing different research questions, as well as finding solutions for them. Part of these research questions concentrates on the possibility of identifying learners' employment status and gender based on the comments they exchange. As noted previously, the size of the available data is a supportive factor; it can offer a solution to the problem via deep learning models that capture hidden features and outperform humans on many tasks [82]. However, there was concern about the balancing issue of these comments, among the different categories. The variable of data distribution has a strong effect on the performance of ML models, in general [13]. Thus, this issue has received sufficient attention in this thesis, particularly the investigation of a new NLP approach to generate more samples of the minority classes based on the paraphrasing technique (see Section 5.3.3.1, Chapter 5).

Also, two different approaches are examined in terms of DL models in this research: one is based on semantic representation and ensemble learning, which is a common approach in NLP (see Chapter 5); the other is based on syntactic representation and recursive learning, which is a less examined approach in NLP (see Chapter 6).

This research is also concerned with identifying a learner's educational level based on small-scale data (i.e., the course level), which is a task with the potential to help various decision-making issues in a MOOC at the end of the course. This also leads to considering the possibility of utilising metadata from MOOCs based on leveraging NLP and non-NLP features as predictors for the models. This problem is solved by considering learners' behaviours on MOOCs, which are extracted from different time stamps, quizzes, and comments. Since a small dataset was used (per course) in Chapter 7, simple classifiers were also considered. In spite of this simple methodology, high levels of accuracy were obtained due to careful procedures in data preparation and feature engineering.

As stated previously, the availability of personalised recommendations when delivering courses via MOOCs to learners, which are tailored based on their demographic characteristics, has become vital for MOOCs. The above-mentioned contributions to this aim are also outlined in Section 1.6 in Chapter 1, and details of the implementations are mentioned in Chapters 5, 6, and 7.

8.4 Overall Observations

8.4.1 Data Preparation

Learners in the data enrolled in different courses at different times. The experiments in this thesis were undertaken based only on data from learners that were extracted from the learners' earliest courses. This was achieved using the learners' dates of enrollment (enrollment information) and collecting their comments or other behaviours only within six months after enrollment (i.e., data were collected in a window of time). This step was important based on the assumption that a demographic label of a learner may has changed, this needs to be handled because that can leads to bias in providing models with inaccurate information (i.e., in terms of the label). This includes, for example, their level of education or job status, which potentially changes after enrollment. However, This does not apply to gender characteristic as this is a constant demographics of a person.

One of the main issues in AP studies is that there is no standardisation for text preprocessing that a researcher can apply to all studies for demographic profiling. In addition, working with texts extracted from discussion forums is challenging because the texts are informal and noisier than traditional or formal texts. They can include short sentences, incorrect spellings, slang terms, and abbreviations. In this thesis, similar pre-processing settings were applied, and these settings should be as basic as possible in order to avoid losing too many details that might reveal the author's identity of writing style. This is also common in the AP area – namely, having similar pre-processing steps in a study even for different demographics tasks [169]. For example, in this thesis, words or characters were removed that were not related to a user's writing style, such as web links, because their structure is not part of the comment's content. This is also a step recommended by [169], which helped to prevent the models from becoming reliant on terms that provide no essential information for classifiers.

Sequence (vector) length is also an important factor that heavily influences any ML model's performance. An initial finding in this thesis was that most comments left by users on MOOCs are long. Also, they vary in length; some contain only five words while others contain up to 200 words. To cover all words in all comments, the longest-sized comment should be set as the maximum; however, it should be noted this will result in sparsity due to zero padding, which is undesirable during the learning phase given that it generates computational expenses for the ML models. Due to this, the average length of the comments was used in each experiment in this thesis, and measures were taken to ensure that all samples – not only textual samples – had fixed lengths in each experiment.

This research also involved an investigation of a new balancing technique (Chapters 5 and 6), which is an NLP-based approach. The approach has been introduced for the first time in this thesis for demographic profiling, as discussed in Section 5.3.3.1, Chapter 5. However, this technique is not valid to use for educational level classification models, as shown in Chapter 7. This is because some features are numerical features, and even those that are extracted from comments are in numerical form; as a result, only traditional SMOTE technique was applied for the experiment in Chapter 7.

The experimental results in this research indicate that the developed methods and strategies, including the approach to handling the data and potential biases, have significantly influenced classifier performance.

8.4.2 Stylometry Features

This research also carefully investigated other important factors that significantly affect ML performance. In particular, since this research involved text classification problems, special attention was paid to textual representations that are appropriate for a specific classification problem.

A key finding of this thesis relating to stylometry features is that simple features (primitive analysis) alone yield unreliable results. This finding was documented in three sections of this thesis, namely, Sections 4.2.2.1, 4.2.2.2, and 4.2.2.3 in Chapter 4. In contrast, semantic features alone, as used in Chapter 5, and syntactic features alone, as in Chapter 6, resulted in more reliable performance levels.

Advanced syntactic knowledge assisted in classifying learners' gender in more advanced structures based on DL algorithms in Chapter 6, which also evidences its effectiveness in classifying learners' gender in this thesis.

8.4.3 Features of MOOCs

Identifying a learner's traits in a MOOC based on course level is notoriously challenging if the data available are very small or a learner has left no comments. However, it is possible to achieve this task if features other than textual features alone are considered. In this thesis, the suggestion is made that to identify a learner's academic level, it is worthwhile to use a multimodal approach in which varied features stemming from MOOC platforms are applied, such as features derived from a given learner's actions on the platform (see Section 7.3.3, Chapter 7). For example, temporal metadata concerning learners' reactions to course materials (e.g., videos or quizzes) can also be used, as discussed in Chapter 7. Such stemmed features indicated that the educational level of learners is correlated with their behaviour on MOOC platforms.

8.4.4 NLP and ML Algorithms

Many machine learning algorithms, including deep learning (e.g., BiLSTM and CNN) and conventional learning (e.g. SVM, DT, and LR) are investigated in this research. This thesis initially used DL models, as presented in Chapters 5 and 6, because they were a less investigated type of ML for AP [163]. However, DL systems do not require feature engineering, but they do tend to perform more complex jobs for learning (e.g., learning using fewer features, as seen in Chapters 5 and 6).

Deep ensemble learning has been examined in the AP literature, while, to the best of the author's knowledge, no previous AP study has developed a recursive approach based on textual features to identify user gender or other user traits. RecNN have only been applied to AP based on user relationships on Twitter [129], but not based on textual features (see Sections 6.2 and 6.3, Chapter 6). Approaches based on syntactic representations of texts and simple syntactics, for example based on POS tagging, have been studied in this research, but the deep syntactic representation based on the tree structure of texts, has not been explored in AP.

The results of the ML models in this thesis show that the use of ML (both DL and conventional ML), as well as with the help of NLP and MOOCs metadata, has significant potential to assist MOOC development. This also brings experimental evidence to similar opinions expressed in the literature regarding using ML and NLP approaches to develop MOOCs [94].

8.5 Limitations and Future Works

Although the research in this thesis contributes new knowledge about the identification of learners' demographic characteristics on MOOCs, it comes within some limitations, as is true for any research. Highlighting these limitations may assist in framing the research directions in the future, as discussed in this section.

- One of the restrictions in the AP area, in general, is that there is no standard for cleaning or preparing data, as can be seen in the literature presented in Chapter 3. This is because users use different writing styles, even those who belong to the same class (e.g., younger females have different writing styles than older females)[154], [153]; moreover, this occurs to an even greater degree when they are from a different target class (e.g., men versus women) [92]. In this thesis, a basic data cleaning process was performed in order to avoid impacting model performance. However, a plan for investigating the standardisation of the cleaning process for LP in the domain of MOOCs studies will be considered in the future.

- Another limitation of this research is that only English-speaking learners/data are considered. This generates a problem, because NLP solutions vary, depending on the language. For example, the work in Chapter 6 cannot be applied to languages other than English, as it supports English grammar only; this is because language-specific NLP tools have been used [204]. In addition, only one English MOOCs platform (FutureLearn) is used in this research. Even though the dataset is huge in size, analysing other platforms can be beneficial to further strengthen the findinfs. Therefore, examining other MOOCs, including platforms with different languages, is another future research direction.

- It is important to note that critical metadata used in the research in this thesis is based on learners' comments. This indicates that the models used in this study would be unable to identify the demographic characteristics of a learner who has left no comments in the discussion forum. Thus, the task of finding an alternative approach that is not reliant on comments or textual features can be considered in future research.

- A user's degree of experience and familiarity with MOOC environments could affect learner behaviour and their activities on MOOCs. In particular, there is a chance that users who are already familiar with learning on MOOCs may behave differently than other users (e.g., those who are learning via a MOOC for the first time). Therefore, learners can be categorised into two categories: those who have only participated in one run/course (short-term learners); and those who have participated in many runs/courses (long-term learners). To investigate such issues, this could be considered in future work, especially based on longitudinal data collections.

- The main challenge that arises when using textual data in the AP area is that these data vary significantly across user demographics (predictor variables) [92]. This presents a constraint when considering a single classification model (one framework), as revealed in the literature review in Chapter 3, which is due to the variations of writing style among users based on their demographics. Also, due to the different number of categories within a class in this research, each demographic characteristic was handled with a separate model, focusing on approaches that are appropriate to each one. This highlights a gap in the area of demographic profiling, namely, that there is no single model that can identify multiple demographic characteristics at the same level of performance. Hence, developing a solution to fill this gap is an important future direction for this research.

- This thesis examined three types of demographic information. However, more demographic information may be needed for a comprehensive personalisation system. The basic approach in this research was to examine each demographic factor separately. This approach was based on the previous point regarding demographic variability, which arises due to natural variations in writing styles among user demographics. However, future plans includes for this research to be continued and expanded, and the task of studying more types of demographic information, as one of the future areas of active LP research.

- Although the size of the data used in this thesis is huge, and while it is substantial enough to have enabled a satisfying analysis, one of its drawbacks is that it is only based on a single MOOC platform: namely, FutureLearn, the platform sponsored by the University of Warwick. The reason for this is that the author only received permission to use these data, but it is important to emphasise that it is optimal to have data that are not only large in size but also from different universities or even different countries. This is because it is still important to discover how the proposed models could perform based on different sets of learner comments or activities (e.g., those differing in terms of their location or culture), and this would also enable cross-platform research. Thus, in future work, this research can be expanded to examine other forms of online education.

- Another direction of follow up research relates to the cost of training for demographic data. The question arises if it is more economical to devise a convincing means to encourage or even enforce MOOC users to complete a survey on basic demographic data, for instance, within one minute before they can proceed with the course content, rather than consuming potentially significant computational resources to derive these characteristics from the other indirect inputs (as done in this thesis). As explained in section 1.3.1, MOOCs already make available such surveys, however, their uptake is low, only approximately 10% of the learners fill it in [11]. Nevertheless, prompt information and other motivational techniques might be able to be (for instance, adaptively) applied, to convince users to provide this data themselves, when the potential benefits of doing so are clarified to them. This would involve research into motivational theories and adaptation, amongst others.

- Other solutions could be preferred "economically", but they still need to be examined under MOOCs settings, where learners take courses optionally, without any obligation, compared to what they usually do within their universities settings. Thus, the question here, is whether to choose between machine learning-based solutions or other systems, such as rules-based solutions. In fact, such systems need an in-house developer (Human Knowledge Encoded) to enhance the inference of these online questionnaires. This could be a potential future work.

8.6 Epilogue

AP experiments have revealed that computational approaches are useful for investigating users' demographic data in a range of applications [105]. However, compared to other fields such as forensics and marketing, there is still a long way to progress in terms of adopting the benefits of using AP in an educational context [15]. This thesis includes groundbreaking investigations that have considered the use of AP approaches to identify learner demographics, thereby suggesting a new context of learner profiling (LP). To the best of the researcher's knowledge, this is the first study to explore demographic profiling in MOOC data. The outcomes of this thesis are highly significant: namely, using advanced computer science-based models to identify learner demographics automatically and provide more representative data compared to traditional methods of using pre-course surveys. This research also contributes to reducing the risk of bias associated with current MOOC surveys. It is expected that the findings of this research will promote the generalisation of the methods used in this thesis to identify other learner demographics. In the next chapter, the final conclusions are drawn for this thesis.

Chapter 9

Conclusion

Various key contributions stem from this thesis, especially relating to the area of Learner Profiling (LP). Novel approaches based on cutting-edge ML and NLP approaches were explored and explained in this thesis in Chapters 5, 6, and 7. The purpose was to address the current challenge facing MOOCs regarding the task of obtaining extensive and representative data about learner demographics when learners usually leave demographic surveys incomplete.

Importantly, the new research and implementation area that is proposed in this thesis, LP, is a natural extension of the research area of AP, which seeks to discern writers' attributes in online platforms [15]. LP, on the other hand, seeks to discover learner demographics in a data-driven manner, from educational platforms that have a large number of learners such as MOOCs. The findings of this research have shown that NLP and ML models can achieve encouraging results and potentially contribute to resolving such a challenge.

In this thesis, a large dataset is considered that is rich in terms of size and metadata. Arguably, this makes it unique compared to other public MOOC data and especially valuable for the purpose of this study and its aim. The starting application of this research was to solve a known problem facing MOOCs, namely, the problem of the low response rate that occurs due to the extraction of learners' demographics based on traditional methods, which are non-mandatory, self-reported, and survey-based methods. The research undertaken in this thesis is innovative in nature as there is no study that has examined the automatic extraction of learners' demographics in online educational platforms. This research is the first step toward establishing this new research direction and represents a step to help promote MOOC studies using more representative data.

The outputs of this research have answered the original, high-level research questions, as well as the sub-questions resulting from them. A full explanation of the research answers and its finding has been discussed in Section 8.3.1, Chapter 8. The research started by seeking out how to use DL algorithms explicitly for LP, which was motivated by the popularity of DL techniques at the time; however, DL was appropriately compared with traditional ML approaches, which performed favourably when processing data at the course-level (due to dataset size limitations for the task). The research presented in Chapters 5 and 6, which used DL models, was planned before the pandemic; therefore, an adjustment was made to the last part of this research to undertake some research at the course level for identifying the learners' level of education earlier. It is worthwhile to emphasise that this study is the first to investigate new methodologies, namely, the NLP balancing approach, ensemble learning approaches comparison, RecNN models, and MOOC learners' behaviours in the user profiling area.

It is noteworthy that the focus of this thesis has been the classification of learners based on gender, employment status, and academic level. One possibility that was considered was classifying the gender of learners based on syntactic information extracted from text. Several tree-structured LSTM models were examined and a unique bi-directional composition function was developed for current architectures, assessing 18 different combinations of word-level and sentence-level encoding methods. The findings demonstrated that the proposed bi-directional model outperformed all other models; in particular, the hybrid model based on FFNN and SPINN achieved the greatest accuracy (82.60% classification accuracy). Alongside this, the semantic representation of texts was used to identify learners' employment status, where sequential and the popular parallel ensemble deep learning architectures were compared (CNN and RNN), achieving an average accuracy of 96.3% for the best method, including the NLP balancing method (the Paraphrasing). Finally, rather than relying only on textual metadata, the classification of a learner's academic level was solved using small-sized training data (specifically, training data based on learners' behaviours on a course). The author conjectures that in order to identify a learner's characteristic with fewer textual features, other MOOC platforms' related features can also used, such as those obtained from learner behaviours on the platform. For this particular classification challenge, timestamps, quizzes, and learners' discussions were examined as behavioural data. Even with a simple classifier and regardless of the data size, the innovative solution for this problem achieved high accuracy (89% on average).

The findings showed that using NLP and ML models to extract the demographics provided promising results for the LP area. The role of the available features in the data (predictors) utilised in this research has appeared clearly in the results of the models, as they have shown significant improvements in the models' performance (see Chapters 5, 6, and 7). This highlights the importance of using a large dataset, as well as the importance of investigating MOOC platforms' related metadata. This study includes different approaches of evaluation, where an approach was selected based on the nature of the problem that was addressed. The strengths and weaknesses of the methods were also discussed in each chapter. In addition, an overall discussion of observations about the research methodology is provided in Chapter 8.

To conclude, the researcher believe that the outputs of this thesis are highly significant. They are expected to provide online education researchers with a method for obtaining representative data about learner demographics. In turn, this will help to improve personalisation systems on MOOCs, especially since the future role of these platforms has changed after the COVID-19 pandemic [71], [8].

This thesis adds a novel perspective to the literature on demographic data in an online educational setting. It also offers an advanced computer science methodology to identify learner demographics. This thesis is novel in that no prior studies have investigated the effect of automated methods of extracting learners' demographic information in online educational platforms based on the power of AI techniques. In this research, various techniques are applied to rich data gathered from MOOCs in various disciplines. Furthermore, these AI techniques can produce more demographic data, which delivers more representative data for different MOOC studies, thereby helping to establish a comprehensive understanding of target learners. This could guide further studies in MOOCs and other online educational contexts. It could also lead to the replacement of login surveys with other forms that are less costly and time-consuming – and possibly more reliable, valid, and convincing – using state-of-the-art technology.

Through this research, it appears that such classification models for LP, since they achieve such high levels of accuracy, have made pre-course surveys – along with their high cognitive burden for users and inadequacies in building user profiles – outdated and potentially obsolete. As a result, more valid responses may arise to previous arguments about MOOCs as resources to promote equality in education.

Bibliography

[1] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29, 2008.

[2] A. V. Aho and J. D. Ullman. Optimization of lr (k) parsers. *Journal of Computer and System Sciences*, 6(6):573–602, 1972.

[3] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. IEEE, 2017.

[4] G. Alexandron, L. Y. Yoo, J. A. Ruipérez-Valiente, S. Lee, and D. E. Pritchard. Are mooc learning analytics results trustworthy? with fake learners, they might not be! *International Journal of Artificial Intelligence in Education*, 29(4):484–506, 2019.

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

[6] T. Aljohani and A. Cristea. Training temporal and nlp features via extremely randomised trees for educational level classification. In *International Conference on Intelligent Tutoring Systems*, pages 136–147. Springer, 2021.

[7] T. Aljohani and A. I. Cristea. Predicting learners' demographics characteristics: Deep learning ensemble architecture for learners' characteristics prediction in moocs. In *Proceedings of the 2019 4th International Conference on Information and Education Innovations*, pages 23–27. ACM, 2019.

[8] T. Aljohani and A. I. Cristea. Learners demographics classification on moocs during the covid-19: Author profiling via deep learning based on semantic and syntactic representations. *Frontiers in Research Metrics and Analytics*, 6, 2021.

[9] T. Aljohani, F. D. Pereira, A. I. Cristea, and E. Oliveira. Prediction of users' professional profile in moocs only by utilising learners' written texts. In *International Conference on Intelligent Tutoring Systems*, pages 163–173. Springer, 2020.

[10] T. Aljohani, J. Yu, and A. I. Cristea. Author profiling: Prediction of learners' gender on a mooc platform based on learners' comments. *International Journal of Computer and Information Engineering*, 14(1):29–36, 2020.

[11] O. Almatrafi and A. Johri. Systematic review of discussion forums in massive open online courses (MOOCs). *IEEE Transactions on Learning Technologies*, 12(3):413–428, 2018.

[12] M. E. Alonso-Mencía, C. Alario-Hoyos, I. Estévez-Ayres, and C. D. Kloos. Analysing self-regulated learning strategies of MOOC learners through self-reported data. *Australasian Journal of Educational Technology*, pages 56–70, 2021.

[13] E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[14] R. Alroobaea. An Empirical combination of machine learning models to enhance author profiling performance. *International Journal of Advanced Trends in Computer Science and Engineering*, 9:2130–2137, 2020.

[15] I. Ameer and G. Sidorov. Author profiling using texts in social networks. In *Handbook of Research on Natural Language Processing and Smart Service Systems*, pages 245–265. IGI Global, 2021.

[16] M. Antkiewicz, M. Kuta, and J. Kitowski. Author profiling with classification restricted boltzmann machines. In *ICAISC*, pages 3–13, 2017.

[17] Y. A. Arcia, D. Castro-Castro, R. O. Bueno, and R. Muñoz. Author profiling, instance-based similarity classification. In *Conference and Labs of the Evaluation Forum*, 2017.

[18] E. M. Ardehaly and A. Culotta. Co-training for demographic classification using deep learning from label proportions. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2017-Novem:1017–1024, 2017.

[19] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker. Lexical predictors of personality type. In *Proceedings of The Joint Annual Meeting of The Interface and The Classification Society of North America*, 2005.

[20] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *Third Text*, 23:321–346, 2003.

[21] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, Feb. 2009.

[22] Ç. Arıker. Massive open online course (mooc) platforms as rising social entrepreneurs: Creating social value through reskilling and upskilling the unemployed for after covid-19 conditions. In *Creating Social Value Through Social Entrepreneurship*, pages 284–306. IGI Global, 2021.

[23] S. Ashraf, H. R. Iqbal, R. Muhammad, and A. Nawab. Cross-Genre Author Profile Prediction Using Stylometry-Based Approach Notebook for PAN at Conference and Labs of the Evaluation Forum 2016. 2016.

[24] K. Babić, S. Martinčić-Ipšić, and A. Meštrović. Survey of neural text representation models. *Information*, 11(11):511, 2020.

[25] A. Basile, G. Dwyer, M. Medvedeva, J. Rawee, H. Haagsma, and M. Nissim. N-gram: New groningen author-profiling model. *CoRR*, abs/1707.03764, 2017.

[26] M. H. Baturay. An overview of the world of moocs. *Procedia-Social and Behavioral Sciences*, 174:427–433, 2015.

[27] R. Bayeck. Exploratory study of mooc learners' demographics and motivation: The case of students involved in groups. *Open Praxis*, 8(3):223–233, 2016.

[28] R. Bayot and T. Goncalves. Multilingual author profiling using word embedding averages and SVMs. *SKIMA 2016 - 2016 10th International Conference on Software, Knowledge, Information Management and Applications*, pages 382–386, 2017.

[29] J. Bevendorff, B. Chulvi, G. L. D. L. Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al. Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 419–431. Springer, 2021.

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

[31] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15, 2013.

[32] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[33] J. Bollen, H. Mao, and A. Pepe. Modeling Public Mood and Emotion : Twitter Sentiment and Socio-Economic Phenomena. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453, 2011.

[34] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[35] S.-P. Bravo-Marmolejo, J. Moreno, J. C. Gomez, C. Pérez-Martínez, M.-A. Ibarra-Manzano, and D.-L. Almanza-Ojeda. Identification of age and gender in pinterest by combining textual and deep visual features. In *International Conference on Information and Software Technologies*, pages 321–332. Springer, 2019.

[36] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[37] C. Brooks, C. Thompson, and S. Teasley. Who you are or what you do: Comparing the predictive power of demographics vs. activity patterns in massive open online courses (moocs). In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15, page 245–248, New York, NY, USA, 2015. Association for Computing Machinery.

[38] J. Buda and F. Bolonyai. An ensemble model using n-grams and statistical features to identify fake news spreaders on twitter. In *Conference and Labs of the Evaluation Forum*, 2020.

[39] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[41] G. Chen, D. Ye, Z. Xing, J. Chen, and E. Cambria. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2377–2383. IEEE, 2017.

[42] H. J.-H. Cheng. *Empirical Study on the Effect of Zero-Padding in Text Classification with CNN*. University of California, Los Angeles, 2020.

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

[44] G. Christensen, A. Steinmetz, B. Alcorn, A. Bennett, D. Woods, and E. Emanuel. The mooc phenomenon: Who takes massive open online courses and why? *Available at SSRN 2350964*, 2013.

[45] C. Cieri, D. Miller, and K. Walker. The Fisher corpus: a Resource for the Next Generations of Speech-to-Text. *Proc. LREC*, 4:69–71, 2004.

[46] M. Cisel, R. Bachelet, and E. Bruillard. Peer assessment in the first french mooc: Analyzing assessors' behavior. In *7th International Conference on Educational Data Mining*, 2014.

[47] F. Claude, D. Galaktionov, R. Konow, S. Ladra, and O. Pedreira. Competitive author profiling using compression-based strategies. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 25:5–20, 2017.

[48] J. H. Clear. *The British National Corpus*, page 163–187. MIT Press, Cambridge, MA, USA, 1993.

[49] M. Cliche. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. *arXiv preprint arXiv:1704.06125*, 2017.

[50] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*, 2:1, 2016.

[51] A. I. Cristea, A. Alamri, M. Kayama, C. Stewart, M. Alshehri, and L. Shi. Earliest predictor of dropout in MOOCs: A longitudinal study of FutureLearn courses. In *Information Systems Development*, 2018.

[52] A. Culotta, N. K. Ravi, and J. Cutler. Predicting the demographics of twitter users from website traffic data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 72–78. AAAI Press, 2015.

[53] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[54] J. Deng, L. Cheng, and Z. Wang. Attention-based bilstm fused cnn with gating mechanism model for chinese long text classification. *Computer Speech & Language*, 68:101182, 2021.

[55] L. Deng and Y. Liu. A joint introduction to natural language processing and to deep learning. In *Deep learning in natural language processing*, pages 1–22. Springer, 2018.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of ACL*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[57] R. F. S. Dias and I. Paraboni. Combined cnn+rnn bot and gender profiling. In *Conference and Labs of the Evaluation Forum*, 2019.

[58] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, Oct. 1998.

[59] Y.-S. Dong and K.-S. Han. A comparison of several ensemble methods for text categorization. In *IEEE International Conference onServices Computing, 2004.(SCC 2004). Proceedings. 2004*, pages 419–422. IEEE, 2004.

[60] D. Dowty. Compositionality as an empirical problem. *Direct compositionality*, 14:23–101, 2007.

[61] K. Duan, S. S. Keerthi, W. Chu, S. K. Shevade, and A. N. Poo. Multi-category classification by soft-max combination of binary classifiers. In *International Workshop on Multiple Classifier Systems*, pages 125–134. Springer, 2003.

[62] R. Duran, L. Haaranen, and A. Hellas. Gender differences in introductory programming: Comparing moocs and local courses. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, pages 692–698, 2020.

[63] Y. Epelboin. Mooc: a revolution in teaching? a european view. In *EUNIS 2013 Congress Proceedings*, volume 1, 2013.

[64] D. Estival, T. Gaustad, B. Hutchinson, S. Pham, and W. Radford. Author profiling for english and arabic emails. In *Association for Computational Linguistics*, 2007.

[65] D. Estival, T. Gaustad, B. Hutchinson, S. B. Pham, and W. Radford. Author profiling for english emails. In *In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, 2007.

[66] D. Estival, T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson. TAT: An Author Profiling Tool with Application to Arabic Emails. *Proceedings of the Australasian Language Technology Workshop 2007*, pages 21–30, 2007.

[67] J. R. Evans and A. Mathur. The value of online surveys. *Internet research*, 2005.

[68] D. I. H. Farías, R. M. Ortega-Mendoza, and M. M. y Gómez. Exploring the use of psycholinguistic information in author profiling. In *MCPR*, 2019.

[69] G. Farnadi, J. Tang, M. De Cock, and M.-F. Moens. User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 171–179, New York, NY, USA, 2018.

[70] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

[71] H. Fox. Edtechx summit: Beyond covid-19. *Future Learn*, 2020.

[72] M. Franco-Salvador, N. Plotnikova, N. Pawar, and Y. Benajiba. Subword-based deep averaging networks for author profiling in social media: Notebook for PAN at Conference and Labs of the Evaluation Forum 2017. *CEUR Workshop Proceedings*, 1866, 2017.

[73] M. Franco-salvador, P. Rosso, and F. Rangel. Distributed Representations of Words and Documents for Discriminating Similar Languages. (2010):11–16, 2013.

[74] M. Galesic and M. Bosnjak. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public opinion quarterly*, 73(2):349–360, 2009.

[75] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, 2013.

[76] J. Gardner and C. Brooks. Student success prediction in moocs. *User Modeling and User-Adapted Interaction*, 28(2):127–203, 2018.

[77] B. Garrick, D. Pendergast, and D. Geelan. Introduction to the philosophical arguments underpinning personalised education. In *Theorising personalised education*, pages 1–16. Springer, 2017.

[78] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[79] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[80] F. T. Gianvittorio. Boosted by the pandemic, moocs are here to stay. how can we improve them? *eLearningInside News*, 2021.

[81] M. Gjurković and J. Šnajder. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, ACL*, June 2018.

[82] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press.

[83] S. Granger. The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 37(3):538–546, 2003.

[84] P. J. Guo and K. Reinecke. Demographic differences in how students navigate through moocs. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 21–30, 2014.

[85] J. D. Hansen and J. Reich. Socioeconomic status and mooc enrollment: enriching demographic information with external datasets. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pages 59–63, 2015.

[86] L. Havrlant and V. Kreinovich. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, 46(1):27–36, 2017.

[87] J. Henderson. Discriminative training of a neural network statistical parser. In *ACL'04: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 95–102, 2004.

[88] T. Hennis, S. Topolovec, O. Poquet, and P. de Vries. Who is the learner: Profiling the engineering mooc student. In *SEFI 44th Annual Conference, Tampere, Finland. URL: http://www. sefi. be/conference-2016/papers/Open_and_Online_ Engineer-*

ing_Education__Engineering_Education_ Research/hennis-who-is-the-learner-108_a. pdf, 2016.

[89] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[90] F. Hsieh, R. Dias, and I. Paraboni. Author profiling from Facebook corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), 2018*.

[91] B. Huang and K. M. Carley. Discover your social identity from what you tweet: a content based approach. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 23–37. Springer, 2020.

[92] X. Huang and M. Paul. Neural user factor adaptation for text classification: Learning to generalize across author demographics. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 136–146, 2019.

[93] C. Impey and M. Formanek. Moocs and 100 days of covid: Enrollment surges in massive open online astronomy classes during the coronavirus pandemic. *Social Sciences & Humanities Open*, page 100177, 2021.

[94] A. S. Imran, K. Muhammad, N. Fayyaz, M. Sajjad, et al. A systematic mapping review on mooc recommender systems. *IEEE Access*, 2021.

[95] R. Imran and M. Iqbal. Maponsms18: Multilingual author profiling using combination of features. In *FIRE*, 2018.

[96] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. D. III. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. *Journal of Chemical Information and Modeling*, 53(9):1689–1699, 2013.

[97] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer Texts in Statistics Book, 2013.

[98] Z. Jiang, S. Yu, Q. Qu, M. Yang, J. Luo, and J. Liu. Multi-task learning for author profiling with hierarchical features. 2018.

[99] F. Johansson. Supervised classification of twitter accounts based on textual content of tweets. In *Conference and Labs of the Evaluation Forum*, 2019.

[100] D. Johnson, F. Nafukho, M. Valentin, J. Lecounte, and C. Valentin. The origins of moocs: The beginning of the revolution of all at once-ness. In *Proceedings of 15th International Conference on Human Resource Development research and practice across Europe. Edinburgh, UK: Edinburgh Napier University Business School*, 2014.

[101] Y. Joo, I. Hwang, L. Cappellato, N. Ferro, D. Losada, and H. Müller. Author profiling on social media: An ensemble learning model using various features. *Notebook for PAN at Conference and Labs of the Evaluation Forum*, 2019.

[102] T. Joosten and R. Cusatis. Online learning readiness. *American Journal of Distance Education*, 34(3):180–193, 2020.

[103] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[104] P. Juola and E. Stamatatos. Overview of the Author Identification Task at PAN 2013. *Conference and Labs of the Evaluation Forum 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, 2013.

[105] L. Kaati, E. Lundeqvist, A. Shrestha, and M. Svensson. Author profiling in the wild. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pages 155–158, 2017.

[106] B. Kang. How the covid-19 pandemic is reshaping the education service. *The Future of Service Post-COVID-19 Pandemic, Volume 1*, page 15, 2021.

[107] J. Kapočiūtė-Dzikienė, A. Utka, and L. Šarkutė. Authorship attribution and author profiling of Lithuanian literary texts. In *The 5th Workshop on Balto-Slavic Natural Language Processing*, Sept. 2015.

[108] K. Kavuri and M. Kavitha. A stylistic features based approach for author profiling. In *Recent Trends in Communication and Intelligent Systems*, pages 185–193. Springer, 2020.

[109] S. Kellogg, S. Booth, and K. Oliver. A social network perspective on peer supported learning in moocs for educators. *International Review of Research in Open and Distributed Learning*, 15(5):263–289, 2014.

[110] M. Khamsan and R. Maskat. Handling Highly Imbalanced Output Class Label. *Malaysian Journal of Computing*, 4:(2):304, 2019.

[111] T. Kim, J. Choi, D. Edmiston, S. Bae, and S.-g. Lee. Dynamic compositionality in recursive neural networks with structure-aware tag representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6594–6601, 2019.

[112] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 423–430, 2003.

[113] T. K. Koch, P. Romero, and C. Stachl. Age and gender in language, emoji, and emoticon usage in instant messages. *Computers in Human Behavior*, 126:106990, 2022.

[114] M. Kocher and J. Savoy. Distance measures in author profiling. *Information Processing and Management*, 53(5):1103–1119, 2017.

[115] D. Kodiyan, F. Hardegger, S. Neuhaus, and M. Cieliebak. Author Profiling with bidirectional rnns using Attention with grus: Notebook for PAN at Conference and Labs of the Evaluation Forum 2017. *CEUR Workshop Proceedings*, 1866, 2017.

[116] V. Kovanović, S. Joksimović, D. Gašević, J. Owers, A.-M. Scott, and A. Woodgate. Profiling mooc course returners: How does student behavior change between two course enrollments?. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, L@S '16, page 269–272, 2016.

[117] O. Kramer. K-nearest neighbors. In *Dimensionality reduction with unsupervised nearest neighbors*, pages 13–23. Springer, 2013.

[118] S. Kumar, F. Morstatter, and H. Liu. *Twitter data analytics*. Springer, 2014.

[119] J. S. Laguilles, E. A. Williams, and D. B. Saunders. Can lottery incentives boost web survey response rates? findings from four experiments. *Research in Higher Education*, 52(5):537–553, 2011.

[120] S. R. Lambert. Do moocs contribute to student equity and social inclusion? a systematic review 2014–18. *Computers & Education*, 145:103693, 2020.

[121] L. A. Latif, T. T. Subramaniam, and Z. A. Khatab. Learner profiling towards improving learner success. In *Innovating Education in Technology-Supported Environments*, pages 187–198. Springer, 2020.

[122] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[123] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[124] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[125] F. Liu, J. Perez, and S. Nowson. A language-independent and compositional model for personality trait recognition from short texts. *arXiv preprint arXiv:1610.04345*, 2016.

[126] H. Liu, A. Gegov, and M. Cocea. Ensemble learning approaches. In *Rule Based Systems for Big Data*, pages 63–73. Springer, 2016.

[127] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[128] J. Ma, W. Gao, and K.-F. Wong. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, July 2018.

[129] S. Mac Kim, Q. Xu, L. Qu, S. Wan, and C. Paris. Demographic inference on twitter using recursive neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 471–477, 2017.

[130] N. Mahmoudi, P. Docherty, and P. Moscato. Deep neural networks understand investors better. *Decision Support Systems*, 112:23–34, 2018.

[131] A. Makazhanov and D. Rafiei. Predicting political preference of twitter users. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, page 298–305, 2013.

[132] A. Malko, C. Paris, A. Duenser, M. Kangas, D. Molla, R. Sparks, and S. Wan. Demonstrating the reliability of self-annotated emotion data. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 45–54, Online, June 2021. Association for Computational Linguistics.

[133] P. Malliga. A survey on mooc providers for higher education. *International Journal of Management & Information Technology*, 7(1):962–967, 2013.

[134] R. Mar. I, Me, Mine: The Role of Personal Phrases in Author Profilin. 10456:110–122, 2017.

[135] I. Markov, H. Gómez-Adorno, J. P. Posadas-Durán, G. Sidorov, and A. Gelbukh. Author profiling with doc2vec neural network-based document embeddings. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10062 LNAI:117–131, 2017.

[136] I. Markov, H. Gómez-Adorno, and G. Sidorov. Language- and subtask-dependent feature selection and classifier parameter tuning for author Profiling: Notebook for PAN at Conference and Labs of the Evaluation Forum 2017. *CEUR Workshop Proceedings*, 1866, 2017.

[137] G. Matthews, I. J. Deary, and M. C. Whiteman. *Personality traits*. Cambridge University Press, 2003.

[138] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[139] W. Merrill, L. Khazan, N. Amsel, Y. Hao, S. Mendelsohn, and R. Frank. Finding syntactic representations in neural stacks. *arXiv preprint arXiv:1906.01594*, 2019.

[140] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. volume abs/1301.3781, 2013.

[141] Y. Miura, T. Taniguchi, M. Taniguchi, S. Misawa, and T. Ohkuma. Using social networks to improve language variety identification with neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 263–270, Nov. 2017.

[142] Y. Miura, T. Taniguchi, M. Taniguchi, and T. Ohkuma. Author profiling with word+character neural attention network. In *Conference and Labs of the Evaluation Forum*, 2017.

[143] S. Mojarad, A. Essa, S. Mojarad, and R. S. Baker. Data-driven learner profiling based on clustering student behaviors: learning consistency, pace and effort. In *International conference on intelligent tutoring systems*, pages 130–139. Springer, 2018.

[144] N. P. Morris, B. Swinnerton, and S. Hotchkiss. Can demographic information predict mooc learner outcomes? In *Experience Track: Proceedings of the European MOOC Stakeholder*. Leeds, 2015.

[145] A. Mukherjee and B. Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA, Oct. 2010.

[146] S. Nath. Style change detection using siamese neural networks. In *Conference and Labs of the Evaluation Forum*, 2021.

[147] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, 50(6):1–36, 2017.

[148] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. "How old do you think I am ?”: A study of language and age in Twitter. *Proceedings of the seventh international AAAI conference on weblogs and social media, 8-11 July 2013, Cambridge, Massachusetts, USA*, (January 2013):439–448, 2013.

[149] D. Nguyen, N. A. Smith, and C. P. Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, page 115–123, USA, 2011.

[150] S. M. North, R. Richardson, and M. M. North. To adapt moocs, or not? that is no longer the question. *Universal Journal of Educational Research*, 2(1):69–72, 2014.

[151] M. P. Oakes. Author profiling and related applications. In *The Oxford Handbook of Computational Linguistics 2nd edition*. 2021.

[152] R. M. Ortega-Mendoza, A. P. López-Monroy, A. Franco-Arcega, and M. Montes-y Gómez. Emphasizing personal information for Author Profiling: New approaches for term selection and weighting. *Knowledge-Based Systems*, 145:1–14, 2018.

[153] J. Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 369–378, 2010.

[154] F. M. R. Pardo and P. Rosso. On the identification of emotions and authors' gender in facebook comments on the basis of their writing style. In *ESSEM@AI*IA*, 2013.

[155] B. B. Partee, A. G. ter Meulen, and R. Wall. *Mathematical methods in linguistics*, volume 30. Springer Science & Business Media, 2012.

[156] R. M. Paton, A. E. Fluck, and J. D. Scanlan. Engagement and retention in vet moocs and online courses: A systematic review of literature from 2013 to 2017. *Computers & Education*, 125:191–201, 2018.

[157] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. G. etc. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, Nov. 2011.

[158] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[159] D. D. Pham, G. B. Tran, and S. B. Pham. Author profiling for vietnamese blogs. In *2009 International Conference on Asian Language Processing*, pages 190–194, 2009.

[160] G. Piao and J. G. Breslin. Analyzing mooc entries of professionals on linkedin for user modeling and personalized mooc recommendations. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 291–292, 2016.

[161] J. Pizarro. Using n-grams to detect bots on twitter. In *Conference and Labs of the Evaluation Forum*, 2019.

[162] B. Plank. Predicting authorship and author traits from keystroke dynamics. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 98–104, June 2018.

[163] M. Polignano, M. Degemmis, and G. Semeraro. Contextualized bert sentence embeddings for author profiling: The cost of performances. In *ICCSA*, 2020.

[164] S. R. Porter and M. E. Whitcomb. Non-response in student surveys: The role of demographics, engagement and personality. *Research in higher education*, 46(2):127–152, 2005.

[165] V. Radivchev, A. Nikolov, and A. Lambova. Celebrity profiling using tf-idf, logistic regression, and svm. In *Conference and Labs of the Evaluation Forum*, 2019.

[166] I. Rafique, A. Hamid, S. Naseer, M. Asad, M. Awais, and T. Yasir. Age and gender prediction using deep convolutional neural networks. In *2019 International conference on innovative computing (ICIC)*, pages 1–6. IEEE, 2019.

[167] T. Raghunadha Reddy, B. Vishnu Vardhan, M. GopiChand, and K. Karunakar. Gender prediction in author profiling using relieff feature selection algorithm. *Advances in Intelligent Systems and Computing*, 695:169–176, 2018.

[168] T. Raghunadha Reddy, B. Vishnu Vardhan, and P. Vijayapal Reddy. A survey on Authorship Profiling techniques. *International Journal of Applied Engineering Research*, 11(5):3092–3102, 2016.

[169] F. Rangel, A. Giachanou, B. Ghanem, and P. Rosso. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *Conference and Labs of the Evaluation Forum*, 2020.

[170] F. Rangel and P. Rosso. Overview of the 7th author profiling task at pan 2019: Bots and gender profiling in twitter. In *Proceedings of the CEUR Workshop, Lugano, Switzerland*, pages 1–36, 2019.

[171] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, and B. Stein. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the Conference and Labs of the Evaluation Forum*, 2018.

[172] F. Rangel, P. Rosso, M. Potthast, and B. Stein. Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. *CEUR Workshop Proceedings*, 1866, 2017.

[173] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein. Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. *CEUR Workshop Proceedings*, 1609:750–784, 2016.

[174] S. Rathod. Exploring author profiling for fake news detection. *TechRxiv*. 2021.

[175] J. O. Rawlings, S. G. Pantula, and D. A. Dickey. *Applied regression analysis: a research tool*. Springer Science & Business Media, 2001.

[176] G. S. Reddy, T. M. Mohan, and T. R. Reddy. Author profiling approach for location prediction. In *First International Conference on Artificial Intelligence and Cognitive Computing*, pages 389–395. Springer, 2019.

[177] P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. *Encyclopedia of database systems*, 5:532–538, 2009.

[178] J. Reich, D. Tingley, J. Leder-Luis, M. E. Roberts, and B. Stewart. Computer-assisted reading and discovery for student generated text in massive open online courses. *Journal of Learning Analytics*, 2(1):156–184, 2015.

[179] J. Reutemann. 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*, 2016.

[180] I. Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[181] D. Robson. Online learning: how to acquire new skills during lockdown. *The Guardian*, 2020.

[182] Y. Roh, G. Heo, and S. E. Whang. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[183] C. Romero and S. Ventura. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355, 2020.

[184] S. Rosenthal and K. McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 763–772, Portland, Oregon, USA, June 2011.

[185] M. Sap, G. Park, J. Eichstaedt, M. Kern, D. Stillwell, M. Kosinski, L. Ungar, and H. A. Schwartz. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Oct. 2014.

[186] N. Schaetti. UniNE at Conference and Labs of the Evaluation Forum 2017: TF-IDF and Deep-Learning for author profiling: Notebook for PAN at Conference and Labs of the Evaluation Forum 2017. *CEUR Workshop Proceedings*, 1866, 2017.

[187] N. Schaetti and J. Savoy. Comparison of neural models for gender profiling. In *Proceedings of the 14th international conference on statistical analysis of textual data*, 2018.

[188] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[189] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[190] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[191] B. Seijo-Pardo, A. Alonso-Betanzos, K. P. Bennett, V. Bolon-Canedo, J. Josse, M. Saeed, and I. Guyon. Biases in feature selection with missing data. *Neurocomputing*, 342:97–112, 2019.

[192] E. Sezerer, O. Polatbilek, and S. Tekir. Gender prediction from tweets: Improving neural representations with hand-crafted features. *arXiv preprint arXiv:1908.09919*, 2019.

[193] D. Shah. Highlights from coursera partners conference 2020. *The Report by Class Central*, 2020.

[194] K. Shah, H. Patel, D. Sanghvi, and M. Shah. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5(1):1–16, 2020.

[195] Z. Shao and K. Chen. Understanding individuals' engagement and continuance intention of moocs: the effect of interactivity and the role of gender. *Internet Research*, 2020.

[196] L. Shen. What coursera's post-pandemic future looks like. *Fortune*, 2021.

[197] A. Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[198] L. Shi and A. I. Cristea. Demographic indicators influencing learning activities in moocs: learning analytics of futurelearn courses. *Association for Information Systems*, 2018.

[199] P. Shrestha, N. Rey-Villamizar, F. Sadeque, T. Pedersen, S. Bethard, and T. Solorio. Age and gender prediction on health forum data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3394–3401, 2016.

[200] V. Simaki, P. Simakis, C. Paradis, and A. Kerren. Identifying the authors' national variety of English in social media texts. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 671–678, 2017.

[201] A. Singh and A. Sharma. Acceptance of moocs as an alternative for internship for management students during covid-19 pandemic: an indian perspective. *International Journal of Educational Management*, 2021.

[202] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211, 2012.

[203] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[204] J. Soler-Company and L. Wanner. On the relevance of syntactic and discourse features for author profiling and identification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 681–687, 2017.

[205] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[206] E. Stamatatos, W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola, M. A. Sanchez-Perez, and A. Barrón-Cedeño. Overview of the author identification task at PAN 2014. *CEUR Workshop Proceedings*, 1180:877–897, 2014.

[207] E. Stammatatos, W. Daelemans, B. Verhoeven, P. Juola, A. López-López, M. Potthast, and B. Stein. Overview of the 3rd Author Profiling Task at PAN 2015. *Conference and Labs of the Evaluation Forum 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings*, 1391(31):898–927, 2015.

[208] C. Suman, A. Naman, S. Saha, and P. Bhattacharyya. A multimodal author profiling system for tweets. *IEEE Transactions on Computational Social Systems*, 8(6):1407–1416, 2021.

[209] S. Sun, C. Luo, and J. Chen. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25, 2017.

[210] Z. Sun, A. Harit, J. Yu, A. I. Cristea, and L. Shi. A brief survey of deep learning approaches for learning analytics on moocs. *Springer*, 2021.

[211] K. Surendran, O. Harilal, P. Hrudya, P. Poornachandran, and N. Suchetha. Stylometry detection using deep learning. In *Computational Intelligence in Data Mining*, pages 749–757. Springer, 2017.

[212] C. Swathi, K. Karunakar, G. Archana, and T. Raghunadha Reddy. A New Term Weight Measure for Gender Prediction in Author Profiling. In V. Bhateja, C. A. Coello Coello, S. C. Satapathy, and P. K. Pattnaik, editors, *Intelligent Engineering Informatics*, pages 11–18, Singapore, 2018. Springer Singapore.

[213] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[214] T. Takahashi, T. Tahara, K. Nagatani, Y. Miura, T. Taniguchi, and T. Ohkuma. Text and image synergy with feature cross technique for gender identification: Notebook for pan at conference and labs of the evaluation forum 2018. In *Conference and Labs of the Evaluation Forum*, 2018.

[215] Z. Teng and Y. Zhang. Head-lexicalized bidirectional tree lstms. *Transactions of the Association for Computational Linguistics*, 5:163–177, 2017.

[216] J. Tetreault, D. Blanchard, A. Cahill, and M. Chodorow. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of COLING 2012*.

[217] J. Thanaki. *Python natural language processing*. Packt Publishing Ltd, 2017.

[218] C. Thompson. How khan academy is changing the rules of education. *Wired magazine*, 126:1–5, 2011.

[219] P. Tofighi, C. Köse, and L. Rouka. Author's native language identification from web-based texts. *International Journal of Computer and Communication Engineering*, 1(1):47, 2012.

[220] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[221] P. Upendar, T. M. Mohan, S. L. Naik, and T. R. Reddy. A novel approach for predicting nativity language of the authors by analyzing their written texts. In *Innovations in Computer Science and Engineering*, pages 17–22. Springer, 2019.

[222] K. van de Oudeweetering and O. Agirdag. Demographic data of mooc learners: Can alternative survey deliveries improve current understandings? *Computers & Education*, 122:169–178, 2018.

[223] R. van der Goot, N. Ljubešić, I. Matroos, M. Nissim, and B. Plank. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, 2018.

[224] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[225] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[226] L. Voita. Sequence to sequence (seq2seq) and attention. *Github*, 2021.

[227] S. Volkova and Y. Bachrach. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, behavior and social networking*, 18 12:726–36, 2015.

[228] Y. Wang, S. Li, J. Yang, X. Sun, and H. Wang. Tag-enhanced tree-structured neural networks for implicit discourse relation classification. *arXiv preprint arXiv:1803.01165*, 2018.

[229] Y. Wang, Y. Xiao, C. Ma, and Z. Xiao. Improving users' demographic prediction via the videos they talk about. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1359–1368, Nov. 2016.

[230] WHO. Coronavirus, world health orgnization, 2020.

[231] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[232] X. Xia. Gender differences in using language. *Theory and practice in language studies*, 3(8):1485, 2013.

[233] X. Yan and L. Yan. Gender classification of weblog authors. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pages 228–230. Palo Alto, CA, 2006.

[234] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019.

[235] W. Yin, K. Kann, M. Yu, and H. Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.

[236] F. X. Yu, K. Choromanski, S. Kumar, T. Jebara, and S.-F. Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014.

[237] L. Yuan and S. Powell. Moocs and open education: Implications for higher education. CETIS, 2013.

[238] I. Zafras, A. Kostas, and A. Sofos. Moocs & participation inequalities in distance education: a systematic literature review 2009-2019. *European Journal of Open Education and E-learning Studies*, 5(1), 2020.

[239] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[240] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. *arXiv preprint arXiv:1509.01626*, 2015.

[241] X. Zhao and C. Li. *Deep Learning in Social Computing*, pages 255–288. Springer Singapore, Singapore, 2018.

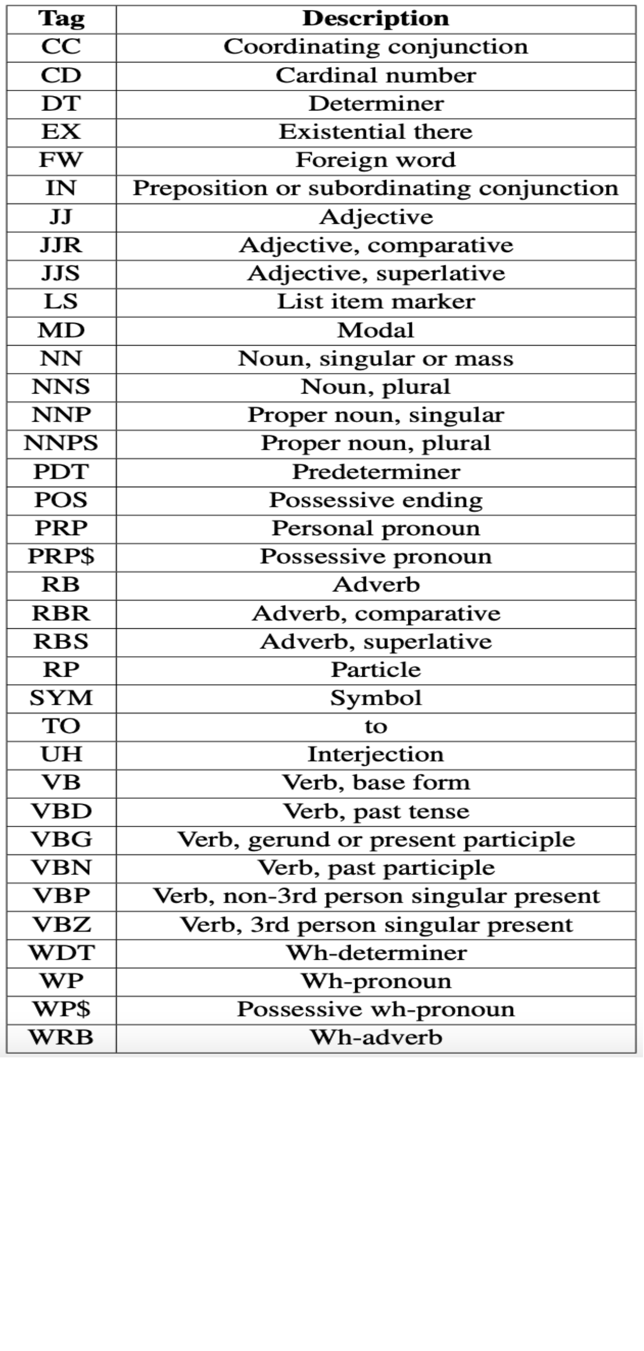
[242] C. Zhenghao, B. Alcorn, G. Christensen, N. Eriksson, D. Koller, and E. J. Emanuel. Who's benefiting from moocs, and why. *Harvard Business Review*, 2015.

[243] X. Zhu, P. Sobihani, and H. Guo. Long short-term memory over recursive structures. In *International Conference on Machine Learning*, pages 1604–1612. PMLR, 2015.

[244] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, 2015.

Appendix A

POS Tags: Description

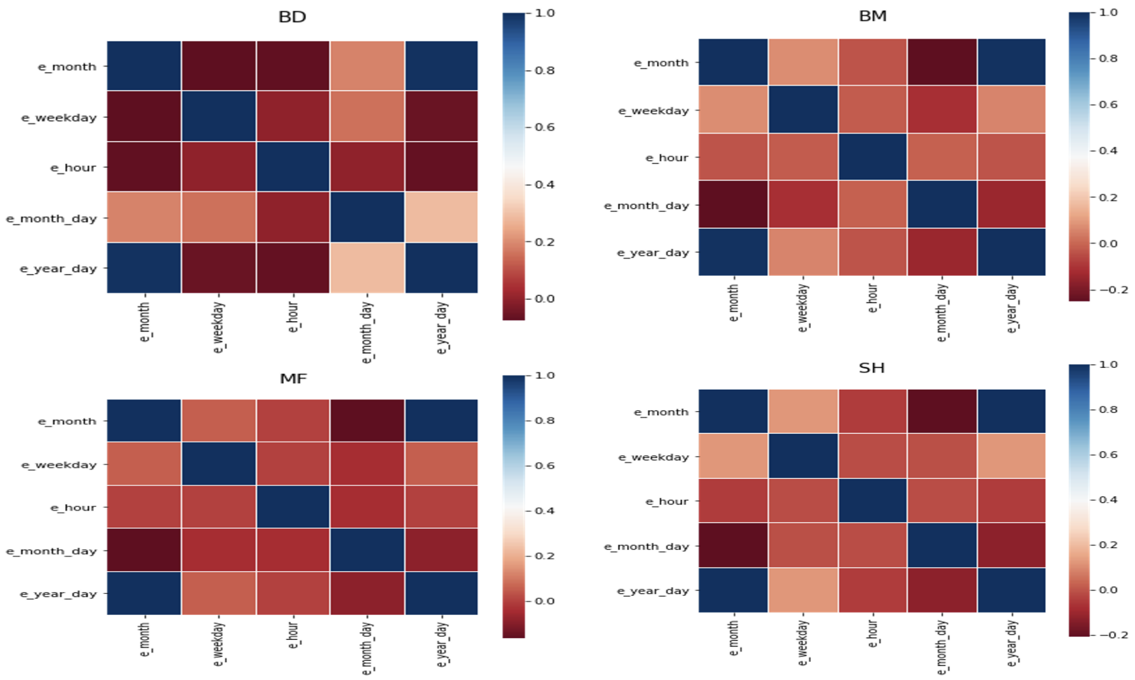


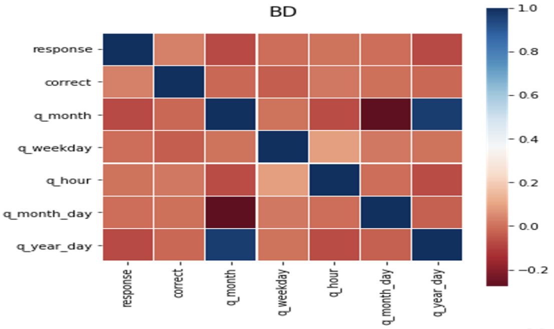
Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP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

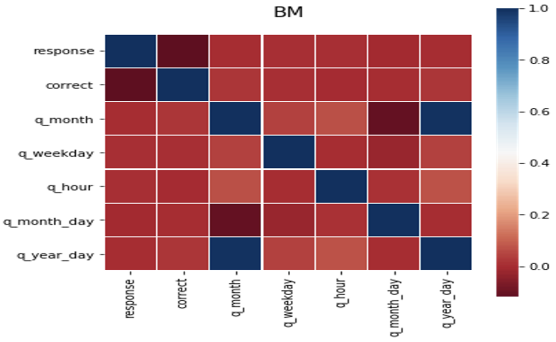
Appendix B

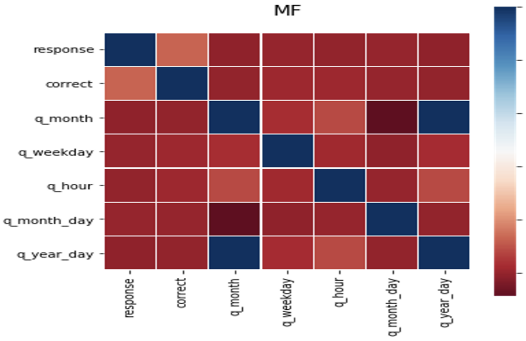
Heat Maps

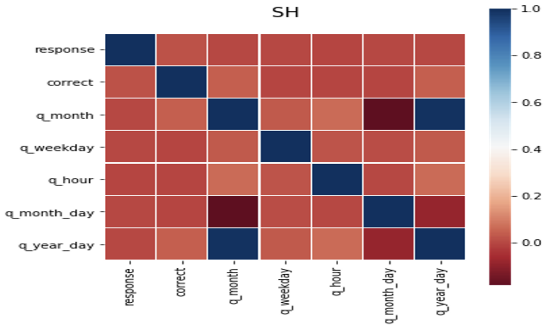
Pearson's correlation coefficient was used in Chaptr 7 to obtain a statistical measure of the relationship between the utilised features. This further assisted in eliminating correlated features. To present the results visually, heat maps were used.

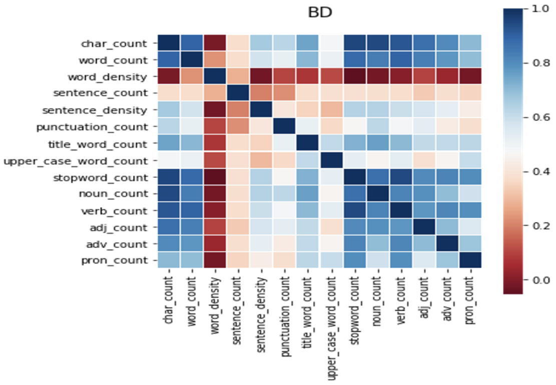


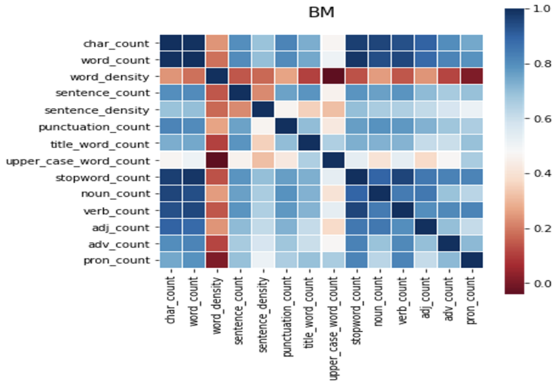

BD
response
correct
q_month
q_weekday
q_hour
q_month_day
q_year_day
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

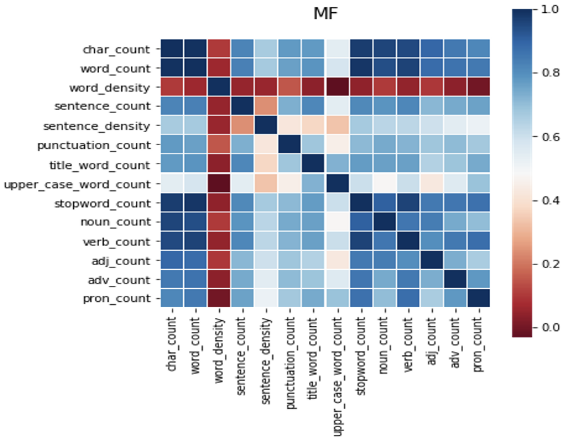

BM
response
correct
q_month
q_weekday
q_hour
q_month_day
q_year_day
1.0
0.8
0.6
0.4
0.2
0.0

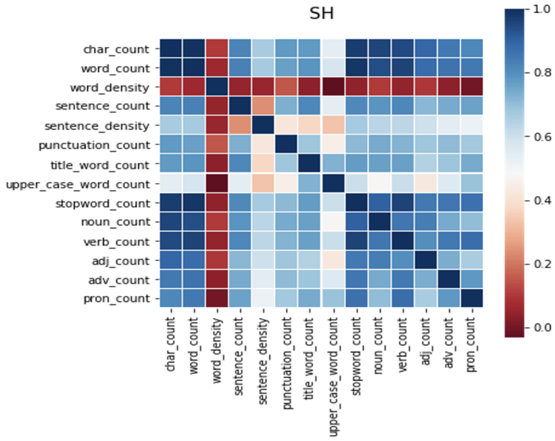

MF
response
correct
q_month
q_weekday
q_hour
q_month_day
q_year_day
1.0
0.8
0.6
0.4
0.2
0.0


SH
response
correct
q_month
q_weekday
q_hour
q_month_day
q_year_day
1.0
0.8
0.6
0.4
0.2
0.0


BD
char_count
word_count
word_density
sentence_count
sentence_density
punctuation_count
title_word_count
upper_case_word_count
stopword_count
noun_count
verb_count
adj_count
adv_count
pron_count
1.0
0.8
0.6
0.4
0.2
0.0


BM
char_count
word_count
word_density
sentence_count
sentence_density
punctuation_count
title_word_count
upper_case_word_count
stopword_count
noun_count
verb_count
adj_count
adv_count
pron_count
1.0
0.8
0.6
0.4
0.2
0.0


MF
char_count
word_count
word_density
sentence_count
sentence_density
punctuation_count
title_word_count
upper_case_word_count
stopword_count
noun_count
verb_count
adj_count
adv_count
pron_count
1.0
0.8
0.6
0.4
0.2
0.0


SH
char_count
word_count
word_density
sentence_count
sentence_density
punctuation_count
title_word_count
upper_case_word_count
stopword_count
noun_count
verb_count
adj_count
adv_count
pron_count
1.0
0.8
0.6
0.4
0.2
0.0