UNIVERSITY OF KWAZULU-NATAL

Developing a predictive model using Twitter dataset for recruiting job-fit candidates in higher education institutions

by

Junior Vela Vela

210535115

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Management, Information Technology and Governance

College of Law and Management Studies

Supervisor: Prof. Prabhakar Rontala Subramaniam

2022

DECLARATION

I, Junior Vela Vela, declare that

i. The research reported in this dissertation/thesis, except where otherwise indicated, is my original research.

ii. This dissertation/thesis has not been submitted for any degree or examination at any other university.

iii. This dissertation/thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

iv. This dissertation/thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   a) Their words have been re-written but the general information attributed to them has been referenced;
   b) Where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

v. Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and fully referenced such publications.

vi. This dissertation/thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation/thesis and in the References sections.

Signed:  Date: 18/08/2023

DEDICATION

I dedicate this doctoral degree to my father, Boniface Vela Buabua. Thank you for teaching me the value of education and the ways of Jesus.

ACKNOWLEDGEMENT

First and foremost, I would like to express my gratitude to the Almighty God, Jesus Christ, for his protection and guidance. This journey would be impossible without your help, oh Lord. Secondly, I want to express my heartfelt gratitude to my parents, Vela Buabua and Mbewa Laza, for all the sacrifices and support. To my siblings, Lady Vela, Julia Vela, Dorcas Vela, Katsia Vela, Kerren Vela, Johana Vela, and Herman Bamata, thank you for your support and prayers.

I would like to express my heartfelt gratitude to my supervisor, Prof. Prabhakar, for his guidance and support. Thank you very much, Sir.

A thank you to all my friends and colleagues, Mursi, Phomolo, and Ncamiso for the support and prayers.

LIST OF ABBREVIATIONS

AI- Artificial Intelligence

ANN- Artificial Neural Network

API – Application Programming Interface

CBOW- Continuous Bag-of-Words

CIPD- Chartered Institute of Personnel and Development

CUP – Capture Understand Present

DP- Deep Learning

FAIR-Facebook AI Research

Glove - Global Vectors

HR- Human Resources

HRIS- Human Resource Information Systems

HRPA- Human Resources Predictive Analytics

ICUP – Identify Capture Understand Present

IM- Impression management

LIWC- Linguistic Inquiry and Word Count

LSTM- Long Short-term Memory Network

ML- Machine Learning

NLP- Natural Language Processing

SC- Soft Computing

TF-IDF- Term Frequency-Inverse Document Frequency

UTAUT- Unified Theory of Acceptance and Use of Technology

VADER- Valence Aware Dictionary and sEntiment

Word2Vec – Word to Vector

ABSTRACT

Organisations in a variety of industries are being confronted with challenging issues and trends like population changes, globalisation, and high-performance expectations. Thus, in such a competitive market, organizations have begun to pay particular attention to the recruitment and selection process, as people are their most precious assets. Employees are the most important part of any organisation as they offer values and perspectives. Employees are generally products of universities and colleges. The growth of any university depends on its ability to recruit and select qualified employees in terms of skills, knowledge, behaviour, and attitudes at all levels. However, the key aspects involved in the staff selection process, have not been thoroughly investigated. The selection process that includes interview sessions has not attracted much research. It is argued that a job interview may fail to provide a true picture of the suitability of a job candidate. As some candidates use deceptive ingratiation by claiming to correspond to the interviewers and/or organization's values, beliefs, opinions, or attitudes to appear more appealing or pleasant, thus misleading the interviewers into selecting them for the jobs. Interview faking appears to be hard to detect, and methods for reducing it are hardly available. Nevertheless, it is important to assess the attitudinal suitability of potential academics. This can be done through various techniques such as machine learning and deep learning using social media platforms such as Twitter. Social media is an important aspect of people's lives nowadays. As people increase their digital presence on social networking sites, the use of social media as a recruiting channel is slowly gaining momentum. This study aimed at determining how the suitability of academics can be classified using Twitter dataset. To this end, the design and development of deep learning job-fit predictive artefacts using Twitter dataset followed rigorous steps of the design science methodology. The results of this study reveal that academic suitability can be predicted using deep learning methods. This study recommends that Universities, Higher education departments consider using artefacts based on social media datasets as supplement tools to enhance the recruitment and selection process.

Keywords: Artificial Intelligence, Deep Learning, Machine Learning, Human Resources, Recruitment & selection, social media, Dataset, LSTM, ANN, BiLSTM, Twitter.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION AND BACKGROUND

1.1 Introduction

Big data analytics (BDA) is now regarded as a critical investment for businesses to remain competitive (Verma, Singh, & Bhattacharyya, 2020). In the domain of Big Data, artificial intelligence (AI), machine learning, and deep learning constitute a new paradigm for learning. Machine/Deep Learning has become a significant resource and investment for firms looking to improve their business processes in today's business world. Data scientists are developing evidence-based algorithms to help people make better decisions, especially when human judgment is not the most critical aspect (Cappelli, Tambe, & Yakubovich, 2020). These evidence-based algorithms are usually presented to business people through data visualisation tools. Data visualisation is an essential component of the scientific process. Effective visualisations enable a scientist to comprehend their own data as well as communicate their findings to others (Waskom, 2021). Within organisations, the Human Resource (HR) department has developed over time. Human resources are regarded as highly advantageous for a company (Sharma, Dashora, & Saxena, 2021).

The importance given to human resources is getting momentum in every organisation (Ekwoaba, Ikeije, and Ufoma, 2015). Organizations have begun to pay close attention to the recruitment and selection process in such a competitive environment since people are their most valuable asset (Rozario, Venkatraman, & Abbas, 2019). Employees are the most important part of any organisation as they offer values and perspectives to the organisation. The success of an organisation depends on the people who work there. Employing the right people for a job can bring success to an organisation, but also, hiring the wrong people can be costly (Djabatey, 2012). Employees of any organisation are generally the products of universities and colleges. Hence, the appropriate people for the job need to be produced by the right people who are equipped and have the capacity to do so. Higher education institutions are very important in the development of every country. Therefore, universities should ensure that the right people are employed (Otoo, Assuming, and Agyei, 2018). The growth of any university depends on its ability to recruit and select qualified employees in terms of skills, knowledge, behavior and attitudes at all levels (Maynooth, 2016). However, the key aspects involved in the staff selection process, have not been thoroughly investigated. Previous research on the recruiting and selection process has mostly focused on employees' performance and the parameters for attracting the best talents, resulting in employees' retention and

organisational effectiveness (Rozario *et al*., 2019). The selection process, which includes interview sessions has not attracted much research. It is argued that what you see in the interview may not be what you get on the job (Barrack, Shaffer, Jonathan, & DeGrassi, 2009). People behave or display their traits in social encounters in an effort to shape the impressions others create of them (Leary & Kowalski, 1990; Leary, 2019). Throughout a job interview, impression management (IM) is prevalent (Peck & Levashina, 2017). Impression Management is described as a goal-directed activity that is either conscious or unconscious, deceitful or authentic (Gardner and Martinko, 1988; Bozeman and Kacmar, 1997; Bolino, Long, and Turnley, 2016). Faking appears to be difficult to detect, and there are few strategies for reducing interview faking (Melchers, Roulin, & Buehl, 2020). Thus, some organisations decided to improve their recruitment and selection process by adopting and integrating Artificial Intelligence in their recruitment and selection process (Albert, 2019).

Organizations that are willing to invest in analytics at every stage of the hiring process will have access to a long-term database that will help them employ individuals with the right talents to achieve their goals (Chen, Cheng, Collins, Charbria, & Cheong, 2018). However, it is not clear whether recruiters are prepared to make the transition from traditional recruitment methods to using artificial intelligence when recruiting (Fernandez, 2019). In 2018, Deloitte's worldwide human capital survey indicated that only 38% of organisations use AI-enabled services at any stage of the recruitment or selection processes (Deloitte Global Human Capital Trends, 2018; Kerey, 2021). Another study indicated that the utilisation proportion is even lower, finding that only 22% of North American enterprises had adopted AI applications in their HR operations (Tambe, Cappelli, & Yakubovich, 2019).

Many firms start sourcing and attracting talent by casting a wide net and keeping themselves linked in the candidates' pool. Social media platforms have recently enabled the access to previously unheard-of volumes of data about employment skills and experiences usable in today's business environment (Fernandez, 2019). Social media has been increasingly used in the recruitment process (Koch, Gerber, and Klerk, 2018). Thirty seven percent of companies in the United State of America were using social media to screen potential employees in 2012, and the percentage of corporations that use social media to screen job candidates increased to 60% in 2016 (Careerbuilder, 2012; Careerbuilder, 2016). Organisations should not limit their use of social media to background checks, but it can also be used to analyse and predict the suitability of the behaviours, personality, and attitudes of job fit candidates. Identifying and

attracting the right job fit candidates should be a key element of the talent management strategy of any organisation.

Universities are facing a challenge in identifying and attracting new academics as the current workforce is aging (Larkin & Neumann, 2012; Udjo, & Erasmus, 2014; Newssa, 2016; Kissoonduth, 2017; Maslen, 2019). Hiring academics with the right interpersonal skill is also crucial to the universities as the attitude of academics affect the classroom (Adesina, Raimi, Bolaji and Adesina, 2016). Academics and practitioners have generally asserted that when a person and the person's occupation are compatible, the work the person does is more likely to be joyful and helpful to both the individual and the society (Barrick & Mount, 1991; Holland, 1966; Kern, McCarthy, Chakrabarty, & Rizoiu, 2019). Various studies have used social media dataset for sentiment analysis and prediction of personality traits, and behaviour. This study designed and developed a deep learning job-fit predictive model using design science methodology. The model predicts the suitability of academics based on their Twitter digital footprint. The design science methodology aims to expand the expectations of human and organizational abilities by producing fresh and cutting-edge artefacts (Hevner, March, and Park, 2004).

1.2 Background

The effectiveness and efficiency of a country's higher education system is one of the key factors in its development (Otoo, Assuming, Agyei, 2018). As higher education nurtures and produces the skills and expertise needed for the development of any nation, there is a need to select and recruit appropriate academics to produce people with appropriate skills (Berry, Petrin, Gravelle, and Famer, 2011). The process of selecting the most qualified individuals to fill open positions in an organization is known as recruitment. These candidates must necessarily meet organizational hiring criteria, such as experience, knowledge, skill, qualification(s), and attitude requirements for the position (Otoo *et al*., 2018, 203). Oke, Okunola, Oni, and Adetoro (2010, 124) assert that the attitudes of the employees or members of an organization have a significant role in determining the success or failure of any organization. Therefore, having employees with the right skills, experience, knowledge, and attitude is arguably the most important asset of any organisation (Sinha and Thaly, 2013). Identifying, attracting, and retaining the best talent is considered one of the important elements of organizational efficacy (Kehinde, 2012; Thunnissen, 2016). As a result, the human resource function is getting

particular attention from most organisations. In this regard, Universities cannot be an exception.

Academics are both teachers and researchers, and their long-term contributions to knowledge production, innovation, and skill development, both individually and nationally, are vital. Academic recruitment and retention is a global issue that affects both poor and developed countries (Kissoonduth, 2017). Since 2004, Graeme Hugo has been reporting on the serious consequences of the aging academic workforce. He has warned that the shrinking academic workforce will have a significant impact on institutions, claiming that, "Between a fifth and a third of their employees will leave in the next decade" (Hugo 2005: 20). Fullan and Scott (2009: 6) cite similar findings, claiming, the inevitable retirement of the baby boomer generation will have a significant impact on the staff and administration of our universities.

According to Larkin and Neuman (2012), Australia's universities were facing an unparalleled human resource crisis due to its aging academic workforce. In support of the former, Tham, & Holland (2018) assert that the academic working force in Australia and New Zealand are aging. In South Africa, universities are still facing challenges. The minister of Higher Education, Blade Nzimande delivered a speech saying the following:

> *"The challenge is multi-faceted, having to do with the slow pace of transformation, regeneration, and change, the aging workforce, developments in higher education worldwide that demand ever greater levels of expertise from staff, the relatively under-qualified academic staff workforce, and low numbers of postgraduate students representing an inadequate pipeline for the recruitment of future academics" (National Institute for humanities and social sciences conference, 2016).*

In the same perspective, Aubbaye (2017) claims that with a dwindling professoriate that is not being replaced at the rates necessary to sustain the growth of the higher education sector, South African institutions in particular face significant and unprecedented problems. Furthermore, Kim (2023) asserts that staffing shortages in higher education would certainly worsen in the next years. Future academic understaffing will mostly result from the postsecondary workforce's ageing at a rapid rate.

Otoo *et al*., (2018) said that there could be loopholes in the recruitment and selection process in the educational system due to the fact that not all schools have appropriate academics in Africa. Moreover, Herschberg, Benschop, and van den Brink (2018) assert that researchers on academic staff evaluation have discussed multiple pertinent processes that make both recruiting

and selecting ambiguous endeavours. The practice of scouting, in which candidates are purposefully encouraged to apply using formal or informal networks, which takes place in closed but also in some open-recruiting situations, is an illustration of such opaqueness in recruitment (Van den Brink, 2010, p. 115). Furthermore, Otoo *et al*., (2018) claimed that there is a need to evaluate the recruitment and selection process. An effort is made by Human resource experts to address the issue of the attraction, recruitment, and selection of talent.

Presently, the selection process involves screening applications, checking of curriculum vitae, checking references, and in other cases background checks of potential talent (Van Esch, Black, & Ferolie, 2019; Lebedeva, Golubeva, Chaykina, Egorov, & Romanovskaya, 2022). Furthermore, the image the potential candidates portray during the interview, via appearance, impression management may play a noteworthy part in the selection process (Barrick, Shaffer, and Degrassi, 2009). However, the face-to-face interview may not be sufficient to identify the right candidates with the required attitudes. The capability of candidates is measured through just thirty to sixty minutes of the interview that may result in recruiting inappropriate candidates. Oke, Okunola, Oni, and Adetoro (2010) claim that university administrators are often surprised by the academics' carefree attitude in carrying out their duties. This implies that there may be loopholes in interview processes, leading to recruiting the wrong individuals for the work. Melchers *et al*. (2020) claim that some candidates use deceptive ingratiation by claiming to correspond to the interviewers and/or organization's values, opinions, beliefs, or attitudes to appear more appealing or pleasant. Thus, it is important to assess the behaviour, attitudes of potential academics through various techniques such as artificial intelligence, machine learning and deep learning. In artificial intelligence (AI) and machine learning, deep learning models represent a new learning paradigm. Recent breakthroughs in natural language processing and image analysis have sparked widespread interest in this topic, as applications in a variety of other domains using big data appear to be feasible (Emmert-Streib, Yang, Feng, Tripathi, & Dehmer, 2020). Artificial intelligence (AI) has been discussed in modern civilization in almost every facet of human endeavour. The use of general-purpose artificial intelligence is still a long way off, at least for the time being. The best chances are in building evidence-based models to enhance decision-making, particularly in cases where human judgment is no longer the most important factor (Cappelli *et al*., 2020).

As people increase their digital presence in social media platforms, the use of social media as a recruiting channel is slowly gaining momentum. Recruiters are already utilising social media to post job openings (Ouirdi, Ouirdi, Segers, and Pais, 2016). Social media is an important

aspect of people living nowadays. According to Swanepoel *et al*. (2014) and Hosain, Manzurul Arefin, & Hossin, (2020), e-recruitment is fast gaining popularity due to its effectiveness and its capacity to reach out to younger audiences. Melanthiou, Pavlou, & Constantinou (2015) note that using social media to recruit has additional advantages, such as enhancing a company's image. This means that, within a proper governance framework, recruitment strategies must permit such adaptability. Therefore, for human resources to ignore social media as a component of e-recruitment would be equivalent to disregarding email two decades ago (Landers and Schmidt, 2016). Ouirdi *et al.* (2016) assessed the adoption of social media in employee recruitment in central and Eastern Europe using the Unified Theory of Acceptance and Use of Technology (UTAUT) framework. Results showed that the core hypotheses of UTAUT were supported. Melanthiou, Palou, and Constantinou (2015) assert that a well-designed system and strategic utilization of social media in the recruitment of potential personnel may significantly assist the recruitment and selection process. Bigsby, Ohlmann, and Zhao (2019) used binomial regression to predict the relationship between recruits' Twitter activities with coaches' scholarship offer decisions. Furthermore, the study demonstrated the expanding significance of social media as a tool for recruiting. According to a study by Kern, McCarthy, Chakrabarty, and Rizoiu (2019), different professions have unique psychological profiles that can be accurately predicted by using unobtrusively obtained linguistic data from social media. Kern *et al*. (2019) autonomously classify user personalities and graphically correlate the personality traits of various professions based on 128,279 Twitter users reflecting 3,513 lines of work. The findings show how social media can be used to match people with their dream jobs. According to Kosinski, Stillwell, and Graepel (2013), easily accessible digital recordings of behaviour, such as Facebook Likes, can be used to forecast a variety of highly sensitive personality traits. There is, thus, an opportunity to explore social media data in evaluating job-fit candidates for positions by recruiters.

1.3 Research problem

Academic positions require a passion for teaching, researching, and supervising. As a result, universities strive to recruit academics with a passion for teaching and researching. Selecting the appropriate candidate is vital to any university, as hiring inappropriate employees can be costly to an organization (Djabatey, 2012). It is critical to acknowledge that human judgment, as it is utilised in recruitment today, is defective. Human resource teams today have access to a series of tests and selection processes that cannot be deemed "fact-based" indicators of

effectiveness (Fernandez, 2019). In selection interviews, impression management (IM), particularly deceptive IM (faking), is a source of concern (Roth, Klehe, & Willhardt, 2021).

The evolution in technology has prompted changes in the lifestyle of human beings, which has led to increasing their presence on social media. Furthermore, the rapidly changing requirements for recruitment based on employers' expectations have prompted the recruiters to add new ways in the recruitment and selection process. Therefore, the relevance of social media in the recruitment process cannot be ignored anymore. Job seekers are already using social media to research firms for which they might want to work for. Recruiters, on the other hand, are using social media to check the background of job candidates as an additional tool of the selection process (Melanthiou, Pavlou, and Constantinou, 2015). However, the use of artificial intelligence methods in recruitment and selection such as predictive models is still in its infancy phase. Furthermore, Mishra et al., (2016) argue that without predictive analytics in human resource management, businesses will not last in the long run.

Barrick, Shaffer, and Degrassi (2009: 1394) assert that "what you see during an interview may not be what you get on the job". He further said that during interviews, candidates may portray an image that may not reflect the candidate's true selves. The traditional job interview may not be sufficient in identifying the right candidates with preferred interpersonal skills, behaviours, or attitudes. Deficiency in recruitment whereby identifying suitable interpersonal skills, behaviours, and attitudes are not given much attention can be overcome if the suitability of candidates can be predicted. This poses the main question: *How can the Twitter dataset be used to predict the suitability of academics for recruitment?*

1.4 Research questions

The main research question was broken down into four research question. The following questions focus on the development of the artefact that will guide the building of the predictive model.

1. How can the Twitter dataset be pre-processed for data labelling?
2. How can the resulting pre-processed data be labelled for classifying job-fit candidates?
3. How can the labelled data be used for the predictive model?
4. How can the performance of the job-fit predictive model be evaluated?

The researcher's aim in this study was to build a predictive model for classifying job-fit candidates using the Twitter dataset.

1.5 Research objectives

This study's primary research goal was to ascertain how the suitability of academics can be classified using Twitter data set. To address this, the following specific research objectives were developed:

1. To identify and implement pre-processing techniques to clean the dataset according to job-fit attitudes criteria.
2. To establish how the pre-processed data can be labelled.
3. To establish how the labelled data be used for predictive model
4. To evaluate the performance of the built predictive models.

1.6 Research rationale

The emergence of social media has shaped how different industries are functioning. Many companies are incorporating the usage of social media in their business strategies. The evidence of businesses using social media in human resources is beginning to emerge. Companies are mostly using social media for marketing and screening potential candidates for their job openings. The high level of competitiveness among potential candidates in terms of their qualifications and experiences has made it difficult for human resources managers to select the right candidate who are the most aligned with the company values. Potential candidates are also using social media to present themselves. Through this study, the researcher designed a predictive model that predicts job-fit candidates based on their Twitter activities. The importance of the model developed in this study is that it can be used as an additional tool in the recruitment process.

Human resource predictive analytics has the potential to revolutionise human resource management (Mishra, Lama, and Pal, 2016). The use of predictive analytics in human resources brings much more clarity and data-driven decision-making. Industries will not survive in the long run without predictive analytics in human resource management (Mishra *et al*., 2016). However, Human Resource Predictive Analytics is still in its infancy stage. Thus, this study contributes to the Human Resource Predictive Analytics processes in adding another aspect to it. Especially in the impression management of potential job candidates.

1.7 Scope of the research

This research focuses on developing artefacts for training data, testing data, and predicting suitable job-fit academic candidates and thereby developing a recruitment model using Twitter

dataset. The study focuses on using Twitter dataset described in section 3.6 and 4.2. Data collection focused on randomly selecting academics and non-academics from South Africa and the rest of the world on Twitter. Furthermore, in this study, a process model that describes steps to predict suitable academics from Twitter dataset was designed. The researcher used Analytical tools and techniques including Machine Learning described in sections 3.7; 4.3.1; 5.2.1, and 6.3 to address the objectives of this study. Moreover, in line with the research problem, to address the research objectives, the researcher designed and implemented two types of classification in this study: binary and multi class classifications.

1.8 The significance of the study

During interview sessions, individuals can deceptively portray themselves as fitting the job's values, and interpersonal skills requirements of the organisation. Thus, it is important to find ways to tackle it. As a lot of people use social media to display their true selves, it is then an important platform to explore as one of the ways to supplement the recruitment interview sessions. In this study novel ways to supplement the recruitment process (interview process) were proposed. The models provide the possibility to predict the suitability of academics in relation to the attitudes and behaviour aspects. The outcomes of this study are particularly important for the human resources department of higher education and the department of higher education in general, as the researcher proposes solutions that can mitigate the error in the interview process. In the long run, this study can benefit the human resource department of any field as the study will set the ground for other specific human resource fields to deepen the research or replicate this study.

1.9 Dissertation structure

The entire dissertation is made of eight chapters as follows:
Chapter one: Introduction and background

In this chapter, the issue of higher education institutions' challenge of finding the right job applicant with the desired interpersonal abilities, attitudes, or behaviours via the conventional job interview process is introduced and presented to contextualise the study. Moreover, a background to the study, rationale and scope of the study are also provided.

Chapter two: Literature review

In the context of the available knowledge, this chapter situates the current research problem. In this chapter, the researcher situates the use of social media data for predictive purposes in the

body of knowledge. Furthermore, the recruitment and selection process, the use of analytics techniques in human resource management and social media are reviewed.

Chapter three: Research methodology

The study's design and methodology are described in this chapter. Furthermore, this chapter covers the research process, strategy, sample method, data analysis procedures, and ethical issues. The strategy and methodology used to address the study's research problem are also described in this chapter and the process model of this study is discussed.

Chapter four: Academic data pre-processing

In chapter four the researcher addresses the first research objective. All the techniques used to pre-process the data and present the inputs and outputs during this process are described in this chapter. Also, the rules used to pre-process the data are discussed and justified.

Chapter five: Academic data labelling

This chapter addresses the second research objective. The researcher discusses the techniques used for data labelling and presents the labelled dataset. The rules used to label the data are also discussed in this chapter.

Chapter six: Academic suitability classification using deep learning

In this chapter, the third research objective is addressed. Also, the implemented artefacts are discussed, and presented accordingly, and all the techniques used in the experiments are discussed in detail. The inputs and outputs about this step are presented in this chapter.

Chapter seven: Performance evaluation

In this chapter, the researcher addresses the fourth research objective. The evaluation methods used in this study to gauge how well the designed artefacts performed are discussed by the researcher. Also, different performance evaluation metrics used in this study are discussed in this chapter.

Chapter eight: Conclusion and Recommendations

The research is wrapped up and the findings are summarized in chapter eight. In this chapter, an updated research process model is also provided. The chapter concludes by offering potential directions for future studies.

1.10 Summary of the chapter

In this chapter the researcher presents the study's context, the problem that led the researcher to conduct this study, the research objectives and questions, the study's scope, and the study's justification. A brief summary of the chapters of this research study is also provided. The importance of human resource analytics and the usage social media analytics are described in this chapter. The reviewed literature is discussed in the next chapter.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

A literature review should not simply consist of a list of recently published articles. Instead, diverse debates on the topic, arguments, and counter arguments should be thoroughly compared in order to discover research gaps and fill those gaps with new research (Machi and McEvoy, 2016:5). Furthermore, a literature review, according to Hart (2018:31), tries to put a "research endeavour into context by linking concepts and theories to problems and questions" in order to justify its value for constructive debates on the topic. This chapter discusses the interaction between social media and recruitment. With to the enormous growth in the capacities and capabilities of social media and the Internet, the number of sourcing opportunities and activities has increased significantly in recent times. The usage of social media platforms for recruitment is gaining more steam as the number of users on those platforms rises. Thus, this chapter provides a thorough discussion on the social media, recruitment and selection process, academic characteristics, classification techniques, and highlights the gap in the literature.

2.2 Recruitment and human resource

The HR department's primary function is to recruit, and the recruitment process is the first step in achieving competitive quality and a competitive edge in recruitment for the organization (Hamza *et al*., 2021). The recruitment process is a systematic procedure that starts with finding candidates and concludes with setting up and facilitating interviews. It requires a significant amount of time and resources (Muscalu, 2015). The process of discovering, screening, shortlisting, and employing possible human resources for the goal of filling roles within companies is known as recruitment (Hamza, Othman, Gardi, Sorguli, Aziz, Ahmed, & Anwar, 2021). It is the most important aspect of human resource management. It is also regarded as the process of choosing the approrpiate individual for the right job at the right time. Recruitment occurs both internally, i.e. within the organization, and externally, i.e. with outside sources (Abdullah & Othman, 2019). Internal and external recruitments have distinct characteristics such as skill levels, the size of the organisation, the recruiting policy, the organisation's value, and the image of the job (DeVaro, 2020). Internal recruitment appears to be a realistic choice and a crucial source because it gives employees room for advancement and personal development (Villeda, McCamey, Essien, & Amadi, 2019). By giving current employees the option to advance their careers and making use of their existing personnel base, the enterprise

will also be in a better position to develop (Alwi, ul Hasan, & Zaman, 2022). On the other hand, External recruitment also provides benefits. It infuses the organization with new ideas. Outsiders are less sensitive to unproductive pressures from peers and subordinates since they are not entangled in organizational politics (DeVaro, 2020). Each has its own set of limitations that only apply to specific companies and are more effective in specific situations, but organizations can use them all to find the right answer. To establish which channel or recruiting metrics is best for the organization, recruiters should gather real-time recruitment metrics, since doing so will help them gain information that will make the solution more viable (Alwi et al., 2022). The recruitment of the best candidates is crucial for any organisations regardless of the approach taken.

The process of recruiting the best candidate for the job is known as selection. It is the process of interviewing candidates and analysing their abilities for a given job, after which candidates are chosen for the appropriate roles (Hamza *et al*, 2021). The organisation will be able to attain its desired goals and objectives if the proper applicants are chosen for the correct positions (Dany & Torchy, 2017). When hiring personnel, it is critical to make sure they have the credentials, skills, and abilities needed to do the job obligations in a well-organized manner (Hamza *et al*., 2021).

According to Abdullah and Abdul Rahman (2015), in a typical recruitment process, an individual goes through four steps, namely, (1) "position clarification to be filled", (2) "update of the job description and job specification", (3) "identification of possible sources of qualified candidates", and (4) "selection of the most appropriate way to communicate". In the first step, the job vacancy must be defined in order to decide how to fill it or what kind of employee is required to fill it. Once the vacant post has been determined, job descriptions and specifications must be updated, which is the second step (Hameed & Anwar, 2018). Candidates can learn more about the job opening in the job description, while the job specification lists the qualifications needed to do the job (Sharma, 2019). The third step is to specify possible sources of qualified candidates, which entails creating a shortlist, filtering candidates, and determining the position's requirements, to determine the most qualified candidate to fill the position (Anwar & Ghafoor, 2017). This is important because when a company needs an employee, the organisation must find the right candidate for the right job (Hamza *et al*., 2021). The last step is to communicate with the candidate(s). The applicant or applicants must feel confident and comfortable when applying for the open position, making this the most important step (Top & Ali, 2021).

Many processes are involved in the selection process, including preliminary application receipt, screening, interviewing, testing, medical testing, references, and the final employment decision (Hamza *et al*, 2021). Every organization prepares its selection strategy based on its requirements. According to Anwar (2016), certain firms place a high value on various examinations, while others place a higher value on interviews and reference checks. According to Abdulla *et al*. (2017), the hiring process includes procedures like resume and application reception, CV screening, writing tests, interviews, background and reference check, medical and physical examination, placement, and final selection. As it is evident, the majority of the steps do not emphasize on the attitude and characteristics of the candidate. How to determine whether an applicant fits the values and attitudes of the position remains a big gap.

2.3 Impression management and interview selection session: A review

People are concerned with how others perceive them, and this is especially important in organizations (Leary 1995, Leary & Kowalski 1990). When applying for jobs, a strong self-presentation is thought to be essential, and once hired, employees frequently worry about presenting the best possible image to their superiors, co-workers and subordinates. Indeed, how people are perceived by others has a significant impact on their level of likeability as well as on whether they are perceived as competent and dedicated or incompetent and not dedicated to their jobs. Consequently, this would have an impact on the rewards they receive, and how quickly they advance within the organization (Bolino, Kacmar, Turnley, & Gilstrap, 2008).

As it affects people and organizations in a significant way, impression management is a crucial organizational phenomenon (Bolino et al. 2008). Perceptions of employees have a significant impact on hiring choices, performance reviews, promotions, and other personnel activities (Ferris, Hochwarter, Buckley, Harrell-Cook, & Frink, 1999; Bolino et al. 2008; Langer, König, & Scheuss, 2019). As a result, impression management is crucial in determining the type and use of human resources within a company, which may ultimately have an impact on the company's ability to survive and prosper (Becker & Gerhart 1996, Bowen & Ostroff 2004, Huselid 1995).

Although the initial theoretical and empirical research on impression management was done by sociologists and social psychologists (e.g., Goffman, 2002, Jones 1964, Schlenker 1980), organizational researchers started to concentrate on this subject in the 1980s. Over the past few decades, research into applicant impression management (IM) techniques used in interviews

has exploded. One of the most widely used selection tools in corporations is the employment interview; almost every business uses interviews to determine which applicants to hire (Amaral, Powell, & Ho, 2019). Despite the fact that interviews are employed to assess jobseekers' capabilities and abilities, they can be influenced by unintended factors, such as IM tactics (Huffcutt, Van Iddekinge, & Roth, 2011; Amaral et al., 2019).

Researchers have identified two types of IM strategies in interviews, despite the fact that candidates may use a variety of tactics: honest and deceptive impression management strategies (Levashina & Campion, 2006, 2007; Peck & Levashina, 2017; Bourdage et al., 2018). Candidates who use genuine, good past experiences and values that align with the company to demonstrate their competence and fit are engaging in honest IM tactics; whereas dishonest IM strategies are employed when candidates attempt to create a favourable first impression by exaggerating or faking credentials, experiences, or fit that they do not have.

2.3.1 Categories of impression management

According to Barrick et al. (2009), The verbal (verbalizations or words) aspect of impression management is a part of the larger category of self-presentation behaviours that candidates may exhibit during the interview.

Additional categories of "self-presentation" during job interviews were listed by Barrick et al. (2009) and included verbal behaviours like word frequency, pitch, and fluency as well as nonverbal behaviours like nodding, cheering, and hunching forward. Verbal impression management techniques have typically been divided into two categories in the context of interviews, namely: self-focused or self-promotion and other-focused. Four different subtypes of self-promotion include “exemplification”, “entitlements”, “enhancements”, and “self-promoting statements” (Stevens & Kristof, 1995; Ellis, West, Ryan, & DeShon, 2002; Barrick et al. 2009). Other-focused impression management, also known as ingratiation, is the verbal act of trying to appear likeable to the interviewer or organization. It has the following three subtypes: “other-enhancement”, “opinion conformity”, and “fit with the organization” (Stevens & Kristof, 1995; Ellis et al. 2002; Barrick et al. 2009). Candidates may deliberately attempt and can be trained to exert control over their verbal and nonverbal behaviour in social situations like the interview (Barrick et al. 2009). Candidates who can successfully use verbal and nonverbal cues, can cause the interviewer to give them an emotionally motivated rating.

2.3.2 The effect of deceptive impression management on interview performance

Some recommendations to avoid interviews due to the possibility of faking were made (Koenig, Parrish, Terregino, Williams, Dunleavy, & Volsch, 2013). However, the available research findings regarding whether and how deceptive impression management affects actual interview result such as interview performance assessments or interview outcome, which includes getting an offer of employment or being invited to the next stage of selection, shows mixed results. Interview performance appraisal have frequently been employed as criteria to assess the effectiveness of IM strategies (Melchers, Roulin, & Buehl, 2020). Researchers that studied the relationships between the outcomes of actual interviews and self-reported lying found correlations between the two, varying from small and negative to moderate and positive (Levashina & Campion, 2007; Roulin et al., 2014; Buehl & Melchers, 2017; Bourdage et al., 2018; Amaral et al., 2019). Peck and Levashina (2017) conducted a meta-analysis of the correlation between IM and interview performance and found that, when the interviewer graded the performance, there were correlations between self-focused impression management and interview performance. Additionally, four studies looked at correlation between interview performance in mock interviews and self-reported dishonest IM. Two of the studies discovered minor to moderate positive correlations, (Ingold, Kleinmann, König, & Melchers, 2015; Buehl & Melchers, 2017) while the other two discovered insignificant correlations (Swider et al., 2011; Bourdage et al., 2018).

Honest IM strategies generally result in higher interview ratings, while dishonest IM strategies typically result in lower interview ratings (Kristof-Brown et al., 2002; Bourdage et al., 2018; Amaral et al., 2019; Melchers et al., 2020). According to Melchers et al., (2020), there are few methods for reducing interview faking or deceptive IM, however, it seems nearly impossible to identify interview faking. Moreover, given how difficult it is for interviewers to spot deceptive IM techniques, they are probably less linked to interview performance appraisal (Roulin, Bangerter, & Levashina, 2015). As it can be seen, spotting deceptive IM during interview sessions is still a challenge for recruiters. Thus, this study proposes artefacts using social media data as supplement tools to aid in tackling deceptive impression management during the selection process in the context of academics' selection.

2.4 Recruitment of academics

In terms of achieving its strategic goals, any educational institution, whether public or private, needs an effective, efficient, and qualified faculty (Valenzuela, 2019). The faculty, as one of

the most significant contributors to educational performance, is expected to fulfil its primary responsibilities of teaching, conducting research, and development (Otoo *et al*., 2018). As educators are the heart and soul of the educational system, the best candidates must be recruited, selected, and allowed to shape and nurture the youth of a country (Valenzuela, 2019). Thus, it is critical to make sure the correct personnel are hired (Otoo *et al*., 2018; Valenzuela, 2019). The most frequently stated attributes companies' value most in job seekers are qualifications, work experience, and communication or interpersonal skills (Otoo *et al*., 2018). Work experience and qualifications are indicators of competence with regard to a candidate's technical abilities, whereas communication skills seem to be a broad concept that encompasses a wide range of abilities. Leadership skills, team skills, the capacity to engage with or influence others, problem-solving abilities, organizational skills, crisis managerial skills, and presentation skills are all part of communication skills (Brink & Costigan, 2015; Otoo *et al*., 2018:201; Coffelt & Smith, 2020:3). Acculturation, emotional intelligence, and language proficiency are examples of other communication skills (Otoo *et al*., 2018; Coffelt & Smith, 2020). To effectively manage a diverse workforce, a corporation must choose the best-qualified individual for the job while also keeping in mind the need to establish a staff that is representative of the larger corporate world. This could be accomplished by employing more effective and inclusive recruitment and selection procedures. The main objective of recruitment efforts is to draw in enough qualified applicants to submit applications for open positions in an organization (Saks, 2017:48). By contrast, the primary goal of selection activities is to discover the best candidates and persuade them to take a position within the organization (Otoo *et al*., 2018).

To fulfil its mandates of teaching, engaging in research and providing other consular services, public sector organizations such as public universities must maintain successful recruitment and selection methods (Otoo *et al*., 2018:201). This is because if the selection process turns out badly, the organisation will incur both financial and time costs.

2.4.1 Characteristics of academics

The core elements of an academic position are teaching and conducting research, with some organizational and administrative duties thrown in for good measure (Turk & Ledić, 2016). Academics have a high level of autonomy and decision-making power over themes and methods of teaching and research, as well as the assessment of students' performance and the provision of recreational activities. Academic independence, variety, the potential for

innovation, and personal development are intrinsically motivating elements for people who choose such careers (Erez and Shneorson, 1980). Although academics in the practical sciences and the humanities have different goals and motivational features, there is a significant commonality between them (Erez and Shneorson, 1980).

Academic conduct is evaluated and rated in a variety of ways, from a variety of angles (Nowakowski, 2013; Nowakowski, 2018). When evaluating the conduct of an academic, several individual factors can be considered (Nowakowski, 2018). In addition to examining the attitudes themselves, it is appropriate to consider the intents, motives, and consequences of actions. In a study by Nowakowski (2018), integrity is regarded as the principal attitude. This, together with the moral conduct of a teacher, the capacity to apply innovative approaches when dealing with issues, authority, and communication competencies are presented as the main characteristics to be considered when evaluating an academic.

Banach (1995) as cited by Nowakowski, (2013) categorised some traits of academics namely, personal traits, intellectual traits, teaching traits, educational traits, and external traits. The term "traits" refers to the features and characteristics that make a person distinctive (Allport, 1927: Allport, 1931: Asendorpf, 2009). Thus, traits of academics can be defined as the characteristics that make an academic distinctive.

Personal traits entail being ‘accessible’, ‘outgoing’ and ‘direct’, ‘respecting the dignity of students’, being ‘tolerant’, ‘tactful’, ‘warm-hearted’, ‘friendly’ and ‘agreeable’. Also ‘having a sense of humour’, ‘having serenity’, being ‘socially committed’, ‘trustworthy’, ‘patient’, ‘emotionally balanced’ and ‘optimistic’ (Banach, 1995 as cited by Nowakowski, 2013:56).

Intellectual traits entail being ‘wise’, ‘intelligent’, ‘creative’, ‘passively adaptive’ and ‘perceptive’ (Banach, 1995 as cited by Nowakowski, 2013:56).

Teaching traits entail, ‘having the capacity to apply various teaching methods and teaching aids’, being ‘hardworking’, ‘familiar with students, their needs and abilities, ‘fond of them’, ‘impartial and fair in assessing students’ as well as ‘showing reproductive attitude’. They also entail ‘being ambitious’, ‘striving to enrich the educational base of the university’, ‘recognising and highlighting the achievements and behaviour of students’ as well as ‘creating a sense of security and conditions for assisting students to succeed’ Being communicative and expressive’ (Banach, 1995 as cited by Nowakowski, 2013:56).

Educational traits entail being, ‘open students’ problems’, ‘eagerly involved in work with students’, emotionally connected with them’, ‘forgiving’, ‘caring’ and ‘having the capacity to avoid ridiculing students’, (Banach, 1995 as cited by Nowakowski, 2013:56).

External traits f entail having ‘a neat appearance’ and ‘good overall presentation’, taking care of the physical and mental health’, being ‘well mannered’, and ‘pleasant’ (Banach, 1995 as cited by Nowakowski, 2013:56).

According to Toledo-Pereyra (2012), a researcher possesses the following attributes: curiosity, motivation, inquisitiveness, commitment, sacrifice, excelling, knowledge, recognition, the capacity to apply academic approach and integration. The mustard-research enumerate 10 qualities required to be a good researcher, namely, ‘having an analytical mind’, ‘being a people person’, ‘the ability to stay calm’,’ intelligence’, ‘curiosity’, ‘being a quick thinker’, ‘commitment’, ‘excellent written and verbal communication skills’, ‘sympathetic’, and ‘being systematic’. Furthermore, Banach (1995) as cited by Nowakowski (2013:56) provides some quality of a good researcher, namely: being ‘friendly to respondents’, ‘free from prejudice’, ‘truthful’, ‘observant’, ‘careful in listening’, ‘economical’, as well as ‘accuracy’ and being ‘free from hasty statements’.

Although other factors of the supervisory relationship were regarded as essential, it was the supervisors' interpersonal traits that affected the sense of supervision quality (Andriopoulou & Prowse, 2020). Moreover, although the supervisor-supervisee relationship is not the primary source of learning, it is unquestionably the vehicle through which successful supervision and learning can take place (Roach, Christensen, and Rieger, 2019). Students view supervision to be of good quality when supervisors are available, approachable, attentive, and supportive (Halbert 2015). According to a study by Roach et al., (2019), students ranked academic honesty, constructive criticism, good communication, and bonding as the most favoured supervisory traits.

2.5 Lexicons-based behavioural trait detection: A Review

One of the most important and constant characteristics that can be identified from data on people's behaviours is personality. The capacity to identify personality and behavioural traits by analysing the contents of consumer-generated text has drawn a lot of interest as social media usage has increased (Han, Huang, & Tang, 2020). Previous studies have used deep learning, machine learning techniques, and even broad psychological lexicons to predict personality. A person's personality is made up of a variety of distinctive patterns based on their feelings,

attitudes, and behaviours that aim to identify the enduring characteristics of an individual that can explain and foretell observed behavioural differences. Personality has been shown to influence a variety of human behaviours, including professional aptitude, financial decisions, interpersonal skills, and even health. Personality qualities, have an impact on job fit, product recommendation efficacy, and investment policy (Judge, Higgins, Thoresen, & Barrick, 2006; Zabkar, Arslanagic-Kalajdzic, Diamantopoulos, & Florack, 2017; Mayfield, Perdue, & Wooten, 2008).

To assess individuals' emotions, attitudes, and actions, several methods have been developed. One of the methods is through the analysis of user-generated content (UGC) (Han *et al*., 2020). UGC has been demonstrated to reflect many parts of a user's attitude, behaviour, and emotions. In particular, social media provides a variety of digital records, including text, pictures, and videos, that can be used to discover more about users' personalities (Tadesse, Lin, Xu, & Yang, 2018). Text is one of the most essential data sources for detecting personality traits among other sorts of valuable content, as a huge variety of textual elements have been discovered to strongly suggest an individual's personality traits (Ren, Shen, Diao, & Xu, 2021).

Many previous studies have used Linguistic Inquiry and Word Count (LIWC), a well-known social psychological lexicon, to categorize texts in a document into various linguistic categories and then examine the connection between word categories and personality traits (Yarkoni, 2010; Golbeck, Robles, Edmondson, & Turner, 2011). Other than techniques that use psychological lexicons, systems relying on the Bag-of-Words (BOW) concept are a typical class of personality recognition techniques. Such techniques gather keywords from subscribers' blogs or comments, feed the text into classification models using machine learning or deep learning, train the models, and then forecast the user's personality (Ren *et al*., 2021). Han et al. (2020) outline a method for automatically developing a personality lexicon for social media ecosystems. The term frequency-inverse document frequency (TF-IDF) method is employed to retrieve keywords from microblogs. Word embedding and prior knowledge, also known as a lexicon, are combined to more accurately represent the semantic meaning of words in the domain of personality analysis. According to Park et al. (2015), who created a personality prediction model that relied on the user's language, language-based evaluation may be a useful tool for assessing personality.

Panicheva, Ledovaya, & Bogolyubova, (2016) used lexical, morphology, and semantic to correlate the dark triad personality traits in Russian Facebook textual data. Using data from

Twitter, Gaddis and Foster (2015) looked at the correlation between dark side personality traits and key leadership behaviours. To detect suitable academics, and also using academic lexicons from diverse literature to predict job fit academics. More details about the techniques and methods used are provided in chapters 4, 5, 6, and 7.

2.6 Social media and big data

The amount of raw data produced by governments, companies, research, and internet platforms has exploded in the last ten to fifteen years. Digitisation is the main trigger of this exponential growth of data called big data (Qi, 2020). As a result, new techniques for capturing, processing, and analysing these complex and large data sets have been created (Reddy, Reddy, Lakshmanna, Kaluri, Rajput, Srivastava, & Baker, 2020). Social Media are online websites that are spread over long distances between computers. Social media is used by millions of individuals worldwide to upload images, and videos, update their status, and leave regular comments (Arora, Bansal, Kandpal, Aswani, & Dwivedi, 2019). Social media such as Instagram, Twitter, WhatsApp, and Facebook are among platforms that generate structure and unstructured data (Aragao, & El-Diraby, 2021). The following five Vs of big data define social networks: value, volume, variety, velocity and veracity (Abkenar, Kashani, Mahdipour, & Jameii, 2020). The use of big data analytic methods and frameworks in social media analysis is therefore common (Sivarajah, Irani, Gupta, & Mahroof, 2020).

2.6.1 Studying human behaviour through social media big data

Social data analysis has received a lot of attention in an effort to recognize and comprehend user communication patterns (Abkenar et al., 2020). Recently, academics and policymakers have focused on social network research (Tufekci, 2014). This has led to the publication of numerous research papers using social network datasets. Large datasets, commonly referred to as "big data," have been used to study how people think and act, as well as human behaviour (Paul, Ahmad, Rathore, & Jabbar, 2016). Social media sites have generated a great number of behavioural datasets. Politicians, corporations, scholars, journalists, and governments are using social media datasets to extensively analyse, study human behaviour (Boyd and Crawford, 2012). Historically, Information about the preferences and behaviour of users was gathered using questionnaires. Although social science still uses this approach frequently, the growth and acceptance of social networks have made it possible for us to gather information on users' activities in previously unheard-of ways, such as reliably from subscribers' social platform accounts (Abkenar et al., 2020). Now, social scientists use the Application Programming

Interface (API) to submit related requests to online social networks to extract a significant volume of data from users (Perriam, Birkbak, & Freeman, 2020:277). To be able to handle these emerging data formats in methodologically advanced ways, digital researchers create new methods inspired by internet technologies. Digital approaches are distinguished by the lack of distinction between qualitative and quantitative methods. For example, a hyperlink allows a text to be interpreted qualitatively while still being placed in a system of link networks that can be analysed using quantitative methods (Perriam *et al*., 2020). Big data social media has a similar effect on the study of human behaviour as did the invention of the microscope in the study of biology (Tufekci, 2014).

2.6.2 Characteristics of social media big data

The 5 Vs of big data namely velocity, variety, volume veracity, and value are considered as the challenges of analysing social media datasets. One of the challenges is to be able to access, assemble, analyse the big data to get meaningful insights (Santhosh Kumar & D 'Mello, 2020; Hariri, Fredericks, & Bowers, 2019). Tackling inconsistencies in the datasets, data reductions are also part of the challenges (Uthayasankar, Mustafa, Zahir, Vishanth, 2017). Every second, a significant amount of data can be generated, which is referred to as volume (Gandomi and Haider, 2015). The term "velocity" refers to the rapid creation of data, sometimes known as "streaming data" (Kitchin, 2014; Younas, 2019). Structured, unstructured, and semi-structured data, such as photos, movies, and texts, are represented by variety (Sagiroglu and Sinanc, 2013; Pei, Li, & Tong, 2018). The term "veracity" refers to the honesty of the data being analyzed as well as the correctness of any information (Bello-Orgaz, Jung, & Camacho, 2016). The term "value" refers to the useful information collected for commercial purposes as well as the actual data values (Peng, Wang, Zhou, Wan, Wang, Yu, & Niu, 2017). The most widespread use of big data, or big social data, is in the field of social media, where all five of these characteristics are present (Duan, Edwards, & Dwivedi, 2019).

Trend detection, opinion mining, social media analytics, and sentiment analysis are some of the most common big data applications for social media (Ghani, Hamid, Hashem, & Ahmed, 2019). Liu, Luo, Gong, Xuan, Kou, & Xu, (2016) used big social media data for event detection. Their approach can be used to generate brief text automatically, detect events, and summarize data for big data analysis. Barnaghi, Ghaffari, & Breslin, (2016) used the Twitter dataset for opinion mining and text categorisation. Adarsh & Ravikumar (2018) used social media data to perform sentiment analysis in the aviation sector.

Machine learning techniques are used in the majority of extant approaches to social media big data analysis (Cambria, Rajagopal, Olsher, & Das, 2013; Cambria, Wang, & White, 2014). Classification (Aggarwal and Zhai, 2012, Reuter and Cimiano, 2012), clustering (Liu, Morstatter, Tang, & Zafarani, 2016), and deep learning are some of the most prominent techniques (Jansson & Liu, 2017; Imran, Castillo, Diaz, & Vieweg, 2018).

2.6.3 Human resource and social media

A common source of HR big data for candidate screening is data from social media, particularly LinkedIn (Davison, Maraist, Hamilton, & Bing, 2012; Hamilton & Sodeman, 2020). However, a collection of social media comments from actual users of the firm's product or service might be utilized to answer more general queries (Hamilton & Sodeman, 2020). According to Marr (2017), some companies use social media to figure out which items or features are gaining traction from consumers, and also to learn whether and why product marketing and advertising are effective. As most social media data is geotagged, reviews about goods could be contrasted to Radio Frequency Identification (RFID) tracking data about the specific plants or divisions that make those products. By examining the plants/divisions that best satisfy customer needs and recognizing better plants as knowledge sources to train other locations, big data could be leveraged to improve human resource training capabilities (Hamilton & Sodeman, 2020).

Considering social media data analytics leverages a publicly available data source to specifically address strategic human capital challenges, a good working relationship between HR and marketing stakeholders is a must. Big data from social media platforms like LinkedIn, Twitter, blogs, and Facebook plays a significant part in HR procedures like staffing. It provides a big volume of easily accessible data that helps businesses make informed decisions (Hamilton & Sodeman, 2020; Verma, Singh, & Bhattacharyya, 2020). Many firms are turning to social media platforms (Facebook, Twitter, and LinkedIn) in order to acquire unique information about the existing and future employees. Artificial intelligence and machine learning's growing capabilities may have an influence on the simplicity and validity of leveraging social media data from applicants and workers to make hiring decisions (Harrison & Hartwell, 2022).

2.6.4 Twitter as a social media

After achieving stronger than anticipated growth in 2020, Twitter's global user base will only increase by 6.7 million, reaching 345.3 million in 2022 (InsiderIntelligence, 2022). On Twitter, a microblog item is referred to as a tweet and has a character restriction of 280. With a character limit of 280, Twitter is a popular medium for sharing news, current events, beliefs, and user

behaviours, making it valuable for social monitoring. Making sense of public perspectives on current issues on social media and personality studies can be done using this rich user-generated data (Kursuncu, Gaur, Lokala, Thirunarayan, Sheth, & Arpinar, 2019; Ayo, Folorunso, Ibharalu, & Osinuga, 2020). Twitter provides two different types of data. The first is live streaming current data, while the second is historical data. These data can also be acquired by signing up for a Twitter developer account, which was done for this study, and completing the authentication steps, or by purchasing them from businesses that have partnered with Twitter (Hino & Fahey, 2019). Twitter users can interact with one another through tweeting, retweeting, and making comments. Additionally, many users add hashtags to their tweets to enable other users to easily find their contents (Lowe & Laffey, 2011; Dutta, Chetan, Joshi, & Chakraborty, 2018).

2.7 Evolution and types of human resource analytics

Human resource analytics have been around for a long time. The first book on "How to Measure Human Resources Management" was authored by the pioneer of human resource analytics Jac Fitz-Enz in 1984 (Fitz-Enz, 1984). with the growing strategic importance of HR analytics for businesses and the acceptance of digital technologies, the definition and process of HR analytics have evolved significantly over time. Looking at the overall development of business analytics, there are three fundamental stages, each with various levels of difficulty, usefulness, and intelligence. There are three main analytics approaches, namely, descriptive, predictive and prescriptive (Lepenioti, Bousdekis, Apostolou, & Mentzas, 2020; Akerkar, 2013; Sivarajah et al., 2017; Krumeich, Werth, & Loos, 2016).

The descriptive approach of HR analytics is concerned with the generation of ratios, metrics, graphs, and findings on human resources using internal and external organizational data, as well as workplace and administrative HR information, with focus on the past (Margherita, 2021). This kind of analytics, sometimes referred to as business reporting, interprets previous data and extrapolates it to provide light on significant company changes and historical events. The major outcome is that it clarifies the raw data for the various parts of the business (Jabir, Falih, & Rahmani, 2019).

HR Predictive analytics is about making decisions based on data, and it includes statistical approaches, data mining, and complex algorithms that can evaluate process/workflow data and produce forecasts and scenarios. Predictive analytics is a tool that HR professionals use to

implement future business planning in order to identify issues before they arise, to find new services and explore more options to cut down on waiting time, to boost productivity, and to reduce risks (Jabir et al., 2019).

Predictive analytics have since given way to HR prescriptive analytics, which are based on the availability of huge and diverse HR data. Predictive analytics give HR options for decisions that will enhance performance and completely transform the way it makes decisions for human resource management (Mishra et al., 2016; Fitz-Enz & Mattox II, 2014). When compared to Human Resource Management (HRM) predictive analytic, HRM prescriptive analytic is a more sophisticated analytical method (Meijerink, Boons, Keegan, & Marler, 2021). The development of HR analytics should be viewed as a maturity evolution instead of a chronological one, even though technology has advanced to cater for increasingly cutting-edge types of analytics (Margherita, 2021).

The need for, and opportunity for developing specialized human resources analytics proficiencies has increased as more organizations use big data, digital technologies, and data science techniques in human resources situations (Edwards, 2018; Chitra & Srivaramangai, 2018; Angrave et al., 2016; Kryscynski, Reeves, Stice-Lusvardi, Ulrich, & Russell, 2017). Particularly, data analytics can be applied throughout the hiring process as well as the entire cycle of workforce management and planning, comprising attraction, procurement, development, and retention (Isson & Harriott, 2016).

2.7.1 Human resource predictive analytics

The growing importance of business analytics as an organizational strategic competency has prompted the creation of data-driven human resource management and cutting-edge analytics solutions that can link employee performance to company value drivers (van der Togt, Huselid, & Ulrich, 2001; Van der Togt & Rasmussen, 2017; Becker, Huselid, & Ulrich, 2001; LinkedIn, 2018). The workplace is undergoing a transformation. To locate, acquire, and reward people, HR professionals are increasingly relying on talent intelligence and automation. There are various advantages to this transformation, including increased productivity, enhanced corporate performance, and Gross Domestic Product (GDP) growth. However, establishing a data-driven role that spans all aspects of HR can be difficult (Chen, Cheng, Collins, Chharbria, & Cheong, 2018).

Human resource management is experiencing significant transformation as a result of digitization. In today's competitive world, a talented employee is unquestionably one of the

company's most significant assets. In the realm of HRM, predictive analytics help organizations to achieve their goals (Likhitkar and Verma, 2020).

Human resource analytic is getting momentum in every aspect of human resource function namely recruiting, succession planning, engagement, retention, training, and development, benefits, and compensation (Molefe, 2014). HR professionals are using analytics as part of their overall HR role, with an increasing number focusing solely on HR analytics. These individuals work on teams dubbed "talent analytics" and "people analytics" (Chen et al., 2018). Employees are viewed as an investment by every firm, as their performance has an impact on the organizational effectiveness (Schraeder, & Jordan, 2011; Likhitkar and Verma, 2020).

Predictive analytics has the potential to transform the success cycle of a business. It has the potential to transform an organization's human resource practices, particularly in the fields of recruitment and selection, training and development, and talent retention (Tamizharasi & Rani, 2014). Human resource analytics involves the use of if scenarios that predict the consequences of changing conditions and / or policies (Mishra, Lama, & Pal, 2016). Since its humble beginnings as a simple administrative task, human resources analytics has developed to offer sophisticated analytical and predictive capabilities (Edwards & Edwards, 2019), which can increase employee retention and engagement while also generating incentives for entire organizations (Deloitte, 2017). Turnover and cost per hire are the main focus of traditional human resource analytics (Angrave et al., 2016). Advancements in technology, accessibility of human resource big data through cloud storage, talent pursuit, and protection have steered human resource analytics to another level of analytics called Human Resource predictive Analytics (Ladimeji, 2013).

Human resource predictive analytics makes use of predictive modelling techniques and generates insights that cannot be attained through traditional human resource analytics (Bassi & McMurrer, 2015). Several organisations have proactively opted predictive modelling for their different business functions (Fitz-Enz & John Mattox, 2014). Employee profiling, employee attrition and loyalty analysis, appropriate recruitment profile selection, employee fraud risk management, worker sentiment analysis and predicting of human resource capacity, and recruitment needs are the areas identified from which human resource predictive analytics can create values (Malisetty, Archana, and Kumari, 2017). Mishra et al., (2016) proposed a decision-making model for HR in organisations using predictive analytics techniques. HRM algorithms help people make better decisions by providing predictions about how a current

decision will affect future outcomes. These so-called predictive algorithms use regression-based forecasting techniques to assist managers in forecasting which workers are likely to leave the company, and thus helping to plan on how to make retention decisions or predicting a job candidate's future performance, and thus assisting hiring managers with selection decisions (Cheng & Hackett, 2021; Leicht-Deobald, Busch, Schank, Weibel, Schafheitle, Wildhaber, & Kasper, 2019).

Human Resource Information Systems (HRIS) provide a wealth of data about employees, but there are still a few best practices for leveraging this data for better HR decision-making. According to the Chartered Institute of Personnel and Development (CIPD) report, there is still a large gap in HR professionals' ability to make data-driven and evidence-based decisions (Ekawati, 2019). Human resource managers still prefer to make decisions based on their personal experiences (Ekawati, 2019). According to the 2016-2017 CIPD HR outlook survey, only 47% of HR leaders employ data analytics in the areas of attraction, recruitment, and selection (HR Outlook, 2017).

2.7.2 Artificial intelligence in human resources

Another area of innovation with the potential to be disruptive is artificial intelligence (AI). Unlike the first application of AI in human resource management (Lawler & Elliot, 1996), which evaluated the effects of expert systems on job evaluation by considering both efficiency and psychological results, AI's potential can now be investigated across a variety of contexts. Turnover prediction, staff rostering, applicant search, HR sentiment analysis, and résumé data collection with extraction of information and worker self-service are just a few examples (Strohmeier & Piazza, 2015). Advanced analytics can be assisted by AI applications to uncover rich, practical valuable information and forecasts about human resources (Margherita, 2021).

According to Van Esch et al. (2019), Global businesses are starting to integrate AI capabilities into their hiring and selection procedures. Advanced technologies are starting to be introduced to replace mundane duties like interview scheduling, email communication, and resume screening. AI has, however, occasionally fallen short in the hiring and selection processes. Vallance (2022), claims that AI tools in recruitment and selection failed to reduce bias. Furthermore, Dastin (2018), claimed that Amazon has abandoned a secret AI recruiting tool that was biased towards women. The AI tool implemented was discriminating women in the recruitment and selection process.

A framework was developed by Jia, Guo, Li, Li, and Chen (2018) to show how AI applications can be merged and used to improve human resource planning, hiring, screening, training and development, reward systems, wage assessment, and employee relationship management. AI-enhanced processes are used in talent acquisition, selection, training, performance evaluation, advancement, retention, and employee compensation (Cappelli & Tavis, 2018). In a study conducted by Johansson and Herranen (2019), findings revealed that AI in recruiting is still a relatively young field, with few organizations utilizing it in all aspects of their hiring process. An example of the application of analytics in human resources is the usage of online labour platforms. The so-called gig economy is made up of online labour platforms, which are for-profit labour market intermediaries that leverage technology to build online, multi-sided marketplaces. Despite the lack of agreement on a definition, the term "gig economy" is commonly used to define an economic system that comprises online labour platform firms that link up requesters, that is organizations or consumers with on-demand gig workers in industries such as transportation (for example Bolt and Uber) cleaning (for example Helpling); and food delivery (such as Uber eats and Deliveroo) (Duggan, Sherman, Carbery, & McDonnell, 2020; Meijerink & Keegan, 2019).

2.8 Data preparation for labelling

The practice of discovering and eliminating anomalies in a dataset that may have a detrimental impact on prediction is known as data cleaning (Kalra & Aggarwal, 2017). Data pre-processing can be done in a variety of ways. Data normalization, stemming, tokenization, lemmatization, and stop word removal are a few examples (Haryanto & Mawardi, 2018).

2.8.1 Text cleaning and pre-processing

Pre-processing is the first task in text classification, and using the right pre-processing methods can increase classification effectiveness (Symeonidis, Effrosynidis, & Arampatzis, 2018). Pre-processing tweets entails preparing them for subsequent tasks such as event detection, bogus information detection, and emotion valance estimation. People typically follow their own sequence of informal language rules when using social media. Due to this, each Twitter user has a distinctive style of writing, which includes abbreviations, non-standard punctuation, and misspelled texts (Naseem, Razzak, & Eklund, 2021). Tweets frequently contain informal language, acronyms, URLs, hashtags, and user mentions. The noise introduced by these language flaws can affect automatic categorization performance (Orellana, Arias, Orellana, Saquicela, Baculima, & Piedra, 2018). According to research conducted by Fayyad, Piatetsky-

Shapiro, & Uthurusamy, (2003), noise in Twitter datasets can reach up to 40%, which can have a substantial impact on classification performance. Using appropriate text pre-processing algorithms in a way that pre-processing does not deteriorate, but instead improves classification performance is one of the common challenges in handling noise and the absence of structure in tweets (Naseem, Razzak, & Eklund, 2021).

Saif, Fernández, He, and Alani (2014) investigated whether eliminating stop words enhances or harms the accuracy of sentiment classification systems on Twitter. Six alternative stop word detection systems were tested using data from Twitter from six distinct datasets. Research findings showed that using pre-compiled lists of stop words has a negative influence on sentiment categorization algorithms. Symeonidis et al. (2018) used four major machine learning methods, including Linear SVM, Nave Bayes, Logistic Regression, and Convolutional Neural Networks, to evaluate 16 widely employed pre-processing strategies on two Twitter datasets for sentiment analysis. Furthermore, they discovered that some methods, such as lemmatization, deleting digits, and substituting contractions, enhance accuracy while others, such as punctuation removal, do not. Naseem et al., (2021) investigate the impact of twelve alternative pre-processing approaches on three pre-classified Twitter datasets on hate speech classification tasks. The study also provided a systematic method to text pre-processing that employs a variety of pre-processing approaches to preserve features without losing information.

On the other hand, Kuyumcu, Aksakalli, and Delil (2019) propose that deep neural network classifiers combined with word embedding are part of the solutions for eliminating pre-processing requirements. The study results reveal that Fast Text is substantially more accurate than the other methods and is more robust in terms of pre-processing. Elnagar, Al-Debsi, and Einea (2020) developed models that are totally based on deep learning and do not necessitate any pre-processing. Their experimental findings demonstrated that all artefacts performed suitably on the SANAD corpus, with convolutional-GRU achieving the lowest accuracy of 91.18 percent and attention-GRU achieving the highest accuracy of 96.94 percent.

2.8.2 The common pre-processing tasks of Twitter data

Researchers used various pre-processing components to prepare the Twitter data for classification, sentiment analysis, and labelling (Singh & Kumari, 2016; Gull, Shoaib, Rasheed, Abid, & Zahoor, 2016; Jianqiang, & Xiaolin, 2017; Rane & Kumar, 2018; Palomino

& Aider, 2022). In this section, the discussion is based on some of the pre-processing steps done with Twitter data by some researchers.

Lowercasing: On Twitter, capitalization is challenging because users prefer not to use it as they believe capital letters add a casual tone to the discussion and make the conversations sound more like speech (Tait, 2016). As a result, lowercasing everything appears to be the most practical solution (Palomino & Aider, 2022).

Removal of URLs and Twitter features: This refers to completely removing URLs, website, image, or anything else related to a URL. Furthermore, it entails removing the hashtag character but not the keywords or words contained within the hashtags, for example, removing the # character from #glad but not the text glad because it can have a sentimental polarity. Additionally, the acronym RT that appear at the start of tweets denoting retweeting or reposting of another person's tweets is removed.

Removal of punctuations: It is typical to remove punctuations before sentiment analysis. However, some punctuation characters are connected to emoticons that convey sentiments, so removing them might make the sentiment analysis less accurate. As such, references to emoticons used to express feelings can be found in punctuation sequences. They include the following: :), :D, ;), or <3 (Kim, 2018; Palomino & Aider, 2022). Thus, the researcher has to decide on whether to remove them or not, especially when performing sentiment analysis.

Negation handling: Negations are words that express the opposing meaning of other words or phrases, such as no, not yet, and never. It is typical procedure to substitute a negation followed by a text for sentiment analysis with the word's antonym. For instance, the word bad, which is an antonym of good, is used in place of the phrase not good. However, stop-word lists frequently include negative words like “no”, “not”, and “never” as well as negative contractions like “mustn't”, “couldn't”, and “doesn't”. Therefore, the researcher can swap out all contractions and negative words with not, and after stop words are eliminated as part of misspelling correction, this issue can be fixed. (Symeonidis, Effrosynidis, & Arampatzis, 2018; Palomino & Aider, 2022).

Stop-word removal: Stop-words are extremely prevalent words that aren't very useful for addressing an information need. Examples include conjunctions, pronouns, and definite and indefinite articles (Manning, 2008; Zhai & Massung, 2016; Palomino & Aider, 2022). Usually, these words are eliminated from the analysis to minimize processing. Every term in the corpus is processed according to frequency, and the most frequent terms are then manually chosen for

their semantic content when creating a stop-word list. Following that, the terms on this list are removed from further consideration (Makrehchi & Kamel, 2017). NLTK, a popular Python library, has a stop words list that can be used to remove stop words (Palomino & Aider, 2022).

Emoticons and emojis translation: An emoticon is a pictorial depiction of a visual cues, like a smiling face or smirk, made up of keyboard characters. Emojis are tiny digital images or icons that are used to convey ideas or feelings (Taggart, 2015; Fernández-Gavilanes, Juncal-Martínez, García-Méndez, Costa-Montenegro, & González-Castaño, 2018).

Acronym and slang expansion: Microtext is slang and acronyms that have emerged as a result of computer-mediated communication. Microtext includes phrases such as "c u later2day" (see you later today), which are frequently used in tweets and short message service (SMS) texts but are not found in standard English (Satapathy, Guerreiro, Chaturvedi, & Cambria, 2017; Palomino & Aider, 2022). Additionally, Palomino, Grad, & Bedwell, (2021) have discussed how crucial it is to reveal hidden messages with sentimental repercussions by expanding acronyms. The *SMS and Slang Translator dictionary* can be used to expand acronyms and slang (Verma, 2022).

Spelling correction: Misspellings may lead to language features being overlooked. Using tools that detect and correct misspelled words improves classification performance (Mullen & Malouf, 2006; López-Hernández, Almela, & Valencia-García, 2019). Despite the fact that no spellchecker is perfect, some of them have shown to be fairly accurate (Symeonidis, et al., 2018).

Removal of unnecessary spaces: Since spaces are regarded as word boundaries. Therefore, it is imperative to recognize the spaces between characters. Unfortunately, splitting a space-filled sequence of characters can also split what should be considered a mono "token". This frequently occurs with foreign language expressions such as deja vu as well as names like New Orleans, The Netherlands, and Côte d'Ivoire (Palomino & Aider, 2022).

Lemmatization: In linguistics, lemmatisation is the method of bringing together a word's inflected forms so that they can be examined as a specific item identified by the word's lemma, or dictionary form. For instance, words such as marking, marks, or marked, are shortened to mark. In other words, this is the procedure for returning words' inflection to their original form. The process of lemmatization is dependent on vocabulary and word form. The goal of this procedure is to eliminate inflectional endings and restore words to their dictionary-basis forms (Qorib, Oladunni, Denis, Ososanya, & Cotae, 2023).

Stemming: Stemming is the method of determining the immutable stem of a specific word (stemma), which does not always correspond to its morphological root. In the process of stemming, a word's final portion is simply removed to reveal the word's stem (Qorib et al., 2023). For instance, the word greater may be over- or under-stemmed, and It could be altered to gre or great or just left as greater. The purpose of stemming and lemmatization is to return inflected versions of words to their standard form. Lemmatization algorithms are based on the usage of dictionaries and morphological analysis, whereas stemming algorithms function without comprehending the context or the differences between words. Lemmatization is a more precise and resource-intensive procedure than stemming algorithms, yet stemming techniques have their own benefits, such as speed and simplicity. Furthermore, the lack of accuracy in identifying stemmas may not always be of vital importance (Izatovich, 2021). Both Lemmatization and stemming are part of data preparation (data pre-processing).

Tokenization: is a pre-processing procedure that divides a text stream into texts, short sentences, symbols, or other constructive elements known as tokens (Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown, 2019; Gupta & Malhotra, 2015). Tokens are text segments that have been identified as meaningful segments for the intent of text analysis (Mullen, Benoit, Keyes, Selivanov, & Arnold, 2018). Individual words, as well as bigger or smaller chunks such as word sequences, word subsequences, paragraphs, phrases, or lines, are considered tokens. Tokenization decisions have a considerable impact on subsequent analysis (Denny, & Spirling, 2018). Language plays a big role in tokenization (Mullen *et al*., 2018). Text classification and text mining both necessitate the use of a parser to handle the tokenization of the texts (Kowsari *et al*., 2019).

Park, Lee, Jang, and Jung (2020) discovered that a hybrid approach of morphological segmentation followed by Byte Pair Encoding (BPE) works best in machine translation of Korean to English or English to Korean, and also in natural language understanding tasks. Hiraoka, Shindo, and Matsumoto (2019) developed an artfact that combines a language model for unsupervised tokenization with a text classifier and then trains both models at the same time. They sampled sentence segmentation stochastically during training to render the model robust and reliable against uncommon tokens, which enhanced text classification performance.

2.9 Data labelling process for academic suitability

To tackle classification problems, supervised machine learning classifiers need a significant amount of labelled training data. Deep learning classifiers for classification problems have

increased recently, which has led to the emergence of many ad-hoc labelling tools (Sager, Janiesch, & Zschech, 2021). Therefore, producing meaningful labels is an essential prerequisite for every Machine Learning/Deep Learning-based application because it affects how well the model performs. As a result, the labelling process represents a crucial method of interaction between people and computers in which the learning base of the ML/DL system is projected with tacit human expert knowledge (Rapson, Seet, Naeem, Lee, Al-Sarayreh, & Klette, 2018; Sager et al., 2021). The labelling procedure currently constitutes a significant time and financial investment in the creation of ML/DL-based systems (Rapson *et al*., 2018).

2.9.1 Sentiment analysis

The technique of classifying people's opinions represented in a text as neutral, negative or positive is known as sentiment analysis (SA) (Zad, Heidari, Jones, & Uzuner, 2021). People are producing massive amounts of thoughts and reviews about goods, services, and everyday activities due to the quick expansion of internet-based applications such as websites, social networks, and blogs. Businesses, governments, and researchers can extract and evaluate the public mood and perspectives using sentiment analysis in order to acquire business insights and improve decision-making (Birjali, Kasri, & Beni-Hssane, 2021). Sentiment analysis uses three main classification levels, namely: sentiment analysis at the document, sentence, and word/phrase or aspect levels (Almatarneh & Gamallo, 2018).

At the document-level sentiment analysis, the entire corpus is evaluated to ascertain its polarity. Document level has the benefit of giving us an overall polarity of the particular entity, but the drawback is that every statement that expresses an opinion is subjective. Furthermore, opinions about various entities cannot be retrieved separately (Shirsat, Jagdale, & Deshmukh, 2017). According to Khan, Durrani, Ali, Inayat, Khalid, & Khan, (2016), after performing sentence level analysis, it is more precise to examine each sentence separately in order to get results that are more accurate. The subjective sentences are retrieved for sentiment analysis, and only the objective sentences are discarded. The entire document is viewed as a single fundamental piece of data. It is presumed that the document only expresses one viewpoint regarding a specific subject. This approach is useless if the document includes viewpoints on a wide range of subjects, such as Twitter, Instagram or Facebook datasets.

In sentence-level sentiment analysis, each sentence is processed and evaluated to ascertain polarity. It expresses a neutral, negative, or positive opinion about the sentence (Shirsat et al., 2017). The document has to be divided into sentences for sentence level sentiment analysis.

Subjectivity classification is the process of dividing a document into sentences. Subjectivity classification tasks and sentiment classification are both a part of sentiment analysis at the sentence level (Vanaja & Belwal, 2018).

Feature-based sentiment analysis is another name for phrase level sentiment analysis. In this step of analysis, a product's feature terms are the primary focus. It is primarily based on evaluations, feedback, criticisms, and complaints. Its applications range from movie reviews to hotel reviews and online store reviews (Shirsat et al., 2017; Vanaja & Belwal, 2018).

2.9.2 Lexicon detection

For lexicon-based techniques, dictionaries are created manually or automatically, utilizing seed-words as the starting point. This method starts by selecting a small number of words to use as a dictionary. Then, one can use WordNet, a thesaurus, or an online dictionary to enlarge that dictionary (Gupta & Agrawal, 2020). After the dictionary has been constituted, it can be used to detect the presence of each word at a sentence level. More details about this technique are provided in chapter five.

As this study opted for a supervised learning approach, data labelling was therefore, an important step. To the researcher's best knowledge, at the time when this study was being conducted no study had labelled Twitter data as suitable academic and non-suitable academic. To label the data of this study, two main processes were involved, namely: lexicon detection and sentiment analysis. As shown in chapter five, Lexicon detection was done using an academic dictionary. The second process was sentiment analysis using TextBlob. To label the data, a set of rules were set based on the outcomes of the lexicon detection experiment and sentiment analysis experiment outcome. As this study opted for binary classification and multiclass classification, the dataset was labelled in two ways, namely, with academic suitability binary classification labelling (two labels) and academic suitability multiclass (four labels), the process of labelling is discussed in detail in chapter five.

2.10 Text Classification

Classification or text classification, which is the allocation of free-text documents to specified classes, is another fundamental NLP application (Otter *et al*., 2020). Over the last few decades, text classification problems have been extensively researched and handled in a variety of real-world applications. Many scholars are currently involved in building applications that use text classification algorithms, especially in light of recent achievements in Natural Language

Processing (NLP) and text mining. In text classification often supervised, unsupervised, and semi-supervised techniques are used (Jiang, Song, Wang, Zhang, & Sun, 2017; Torres & Vaca, 2019). It is thought that trait detection is a classification issue. Trait detection can be used to classify individuals as suitable or non- suitable academics. For training, text classification uses labelled data. Feature extraction, dimension reductions, classifier selection, and evaluation are the four processes that most text classification systems go through (Otter *et al*., 2020; Minaee, Kalchbrenner, Cambria, Nikzad, Chenaghlu, & Gao, 2021).

Feature Extraction: Texts and documents are generally in unstructured formats. When employing mathematical modelling as part of a classifier, these unstructured textual data sequences must be translated into an organized feature space. To begin, data must be cleaned to remove any extraneous characters or words. Formal feature extraction methods can be used after the data has been cleaned. Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), Global Vectors for Word Representation (GloVe), Word2Vec, and FastText are some of the most commonly used feature extraction techniques (Goldberg & Levy, 2014; Pennington, Socher, & Manning, 2014; Otter *et al*., 2020). Dzisevič & Šešok, (2019) investigated the results of using three different text feature extraction approaches while using a neural network to classify short sentences and phrases into classes in order to ascertain which approach is efficient in capturing text features and enables the classifier to achieve the highest level of accuracy. The findings demonstrate that when working with larger datasets, the TF-IDF feature extraction method outclasses other methods.

Dimensionality Reduction: When text or corpus datasets contain a lot of unique terms, data pre-processing operations may be time-consuming and memory-intensive. Using low-cost algorithms is a typical approach to this problem. However, with some datasets, these low-cost algorithms do not perform well as expected. to make their applications less complex in terms of time and memory, and prevent performance degradation, many scholars prefer to use dimensionality reduction. Pre-processing with dimensionality reduction could be more efficient than building low-cost classifiers (Zebari, Abdulazeez, Zeebaree, Zebari, & Saeed, 2020; Otter *et al*., 2020).

Classification Techniques: The selection of the optimal classifier is the most crucial phase in the text classification pipeline. Devoid of a complete conceptual grasp of each classifier, selecting the most effective algorithm for a text classification application can be challenging (Otter *et al*., 2020). For text classification, Liu & Guo (2019) used an attention-based

bidirectional long short-term memory model with a convolution layer, and their model outperformed the then existing state-of-the-art text classification methods in terms of classification accuracy. Support Vector Machines (SVM) was used by Luo (2021) to categorize English texts and documents. The Rocchio classifier performs better than other classifiers when the extracted feature size is small. However, experimental results of the classification conducted on 1033 text documents showed that SVM performed better than the other artefacts.

Evaluation: The pipeline for text classification ends with the evaluation stage. Understanding a model's functioning is necessary for both using and creating text classification approaches. There are many ways to evaluate supervised artefacts. The simplest way of evaluation is accuracy calculation, but it does not function well with unbalanced datasets (Otter *et al*., 2020). Some other metrics such as F1, and precision can be utilised to evaluate imbalanced datasets (Sambasivam, & Opiyo, 2021). Sambasivam, & Opiyo, (2021) used F1, recall, precision and accuracy to evaluate cassava disease detection classification. A detailed discussion of evaluation is provided in section 2.15.

2.11 Machine Learning based text classification

A subset of artificial intelligence (AI) known as machine learning permits computers to learn and improve without being specifically programmed (Samuel, 1959; Jordan & Mitchell, 2015). It concentrates on building algorithms that can analyse and forecast data (Yamak, 2018). Machine learning is focused on the conception of computer programs capable of accessing data and learning on their own. The learning process begins with data observations in order to identify patterns and make better decisions. Numbers, photographs, or text, pictures of people or even bakery items, repair records, time-series data from sensors, or sales reports are the starting point for machine learning. The data is acquired and prepared for use as training data, or the information that will be utilized to train the machine-learning model (Brown, 2021). By constructing a model from sample inputs, such algorithms bypass strictly static program instructions by making data-driven predictions or judgments (Yamak, 2018).

The primary goal is for computers to learn instantaneously without human involvement or guidance and then adjust their actions accordingly. The focus of the field of machine learning is on two interconnected questions: How does one develop computer systems that evolve over time? What are the fundamental statistics, information, and conceptual laws that regulate all learning systems, which include computers, humans, and institutions? (Jordan, & Mitchell, 2015). Machine learning research is important for addressing these fundamental questions in

science and engineering as well as for the incredibly useful computer software it has produced and has been used in a variety of applications. A machine learning system's function might be descriptive or predictive. Descriptive means the system utilizes data to explain what happened, while predictive means the system utilizes data to forecast what will happen; or prescriptive, implying that the system will make recommendations based on the data (Malone, Rus, & Laubacher, 2020).

Machine learning has developed over the past 20 years from a research interest to a practical technology with numerous commercial applications. Machine learning has proven to be a useful technique for creating useful software for robot control, speech recognition, computer vision, natural language processing, and other artificial intelligence applications (Athey, 2018). Computer science and a variety of industries that deal with data-intensive problems, such as consumer services, problem diagnosis in complex systems, and logistics chain control, have benefited greatly from machine learning. In order to evaluate high-throughput experimental data in novel ways, machine-learning approaches have been developed. This has had a similarly broad spectrum of effects across empirical sciences, from biology to astrophysics and sociology.

Chatbots and predictive text, language translation tools, Netflix recommendations, and how your social media feeds are presented are all powered by machine learning. It makes self-driving cars and robots that analyse photos of patients' conditions possible (Brown, 2021). With machine learning's increasing ubiquity, everyone in business may come across it and will require some working knowledge of the subject. According to a Deloitte poll from 2020, 67 percent of businesses are utilizing machine learning, and 97 percent are expecting to employ it in the coming year (Brown, 2021).

The three disciplines of machine learning are supervised, unsupervised, and reinforcement learning. Machine learning techniques such as supervised learning infer a function from labelled training data (Raschka & Mirjalili, 2017). Supervised machine learning models are trained using labelled data sets, allowing the models to learn and improve over time (Rustam, Reshi, Mehmood, Ullah, On, Aslam, & Choi, 2020). Unsupervised machine learning is a type of machine learning in which a program searches for commonalities in unlabelled data. Unsupervised machine learning can detect trends and patterns that humans are not looking for (Carrillo-Larco, & Castillo-Cara, 2020). By developing a reward system, reinforcement machine learning educates computers to take the best action through trial and error. By letting

the machine know when it has made the right decisions and providing feedback, reinforcement learning can be used to teach systems to play video game or driverless cars to drive. This enables the machine to gradually learn the proper actions (Brown, 2021).

Machine learning text classification involves utilizing machine learning technique to classify input from a corpus into a set of predetermined classes (Luo, 2021). Text classification includes both short and long text. Supervised machine learning techniques include automatic text classification (Kadhim, 2018). To classify English text and documents, Luo (2021) employed a supervised Support Vector Machines (SVM) classifier. He deduced from the experimental research that when using more than 4000 variables, the classification rate surpasses 90%. Waheeb, Ahmed Khan, Chen, and Shang (2020) employed machine learning approaches to classify sentiment analysis for discharge summaries. Their new approach offers a flexible and efficient way to assess the quality of the treatment based on the positive, negative, and neutral terms used at each discharge summary sentence level.

2.12 Deep learning-based text classification

Deep learning is a branch of artificial intelligence that concentrates on creating massive neural network models that can accurately draw conclusions from data (Kelleher, 2019). Deep learning is particularly suited to contexts where the data is complex and where there are large datasets available. Artificial intelligence and machine learning research led to the development of deep learning. Deep Learning enhances classical Machine Learning by adding more "depth" (complexity) to the model and modifying the data with various functions that allow data representation in a hierarchical manner, through multiple levels of abstraction (Schmidhuber, 2015; LeCun and Bengio, 1995). Deep Learning can help address more complex issues notably well and quickly due to the use of more complex models that allow for massively parallel processing (Panand & Yang, 2010). If there are sufficiently substantial datasets characterizing a problem, these complicated models used in Deep Learning can improve classification accuracy or decrease errors in regression problems (Kamilaris, & Prenafeta-Boldú, 2018). Depending on the network architecture employed, Deep Learning consists of numerous distinct components such as convolutions, pooling layers, fully connected layers, gates, memory cells, activation functions and encode/decode schemes. Deep Learning models are highly hierarchical structures. Their vast learning capacity as well as their potential for being adaptive to a broad range of very complicated tasks in terms of data analysis enable them to perform particularly well in classification and prediction (Pan and Yang, 2010; Minae et al., 2021).

Deep learning is now used by the majority of Internet businesses and high-end consumer devices. Facebook utilizes deep learning to analyze language in online interactions, among other things. Google, Baidu, and Microsoft utilise deep learning for image search and machine translation. Moreover, deep learning is now the standard technology for speech recognition, as well as face detection on digital cameras, on all modern smartphones.

Besides, deep learning is used to interpret medical pictures including X-rays, Computerized Tomography (CT) and Magnetic Resonance Imagery (MRI) scans as well as diagnosing health issues in the healthcare industry. Deep learning is also utilized in self-driving cars for geolocation and mapping, motion planning and steering, and perceptions of the surroundings, as well as monitoring driver's condition (Kelleher, 2019).

Although Deep Leaning is best recognized for dealing with raster-based data such as videos and photos, it may also be utilised with audio, voice, and natural language (Sehgal, Gupta, Paneri, Singh, Sharma, & Shroff, 2017; Kamilaris, & Prenafeta-Boldú, 2018). Text classification can be done using deep learning algorithms and techniques (Lavanya, & Sasikala, 2021). Nguyen, Nguyen, Tojo, Satoh, and Shimazu (2018) used recurrent neural network-based models to recognize parts in Japanese legal texts, that were necessary and had some kind of effectuation. Ji, Tao, Fei, and Ren (2020) investigated information extraction in the legal field using deep learning techniques, with the goal of extracting evidence information from court record documents. The classification task was initially combined with the extraction task to form a multi-task learning issue. To jointly address the two tasks, the study offers a brand-new end-to-end model.

2.13 Application of machine learning and deep learning in social media

Deep Learning is based on the philosophy of connectionism. In this section, the social media problem domains where machine and deep learning has been employed as a key methodology and techniques to solve the problems are discussed.

The contemporary living society is made up of many different entities, humans being one of them. Human conduct can be intuitively divided into two categories: individual behaviour and group behaviour. Humans respond differently in various social contexts as users of society (Dogan, Martinez-Millana, Rojas, Sepúlveda, Munoz-Gama, Traver, & Fernandez-Llatas, 2019; Vespignani, 2009; Lazer, Pentland, Adamic, Aral, Barabasi, Brewer, & Van Alstyne, 2009). Certain atmospheric changes, environmental occurrences, or societal influences affect

social behaviour (Lazer *et al*., 2009). It is necessary to become familiar with individual social behaviour to gain an understanding of a society's well-being, and social developments.

Furthermore, the impact of social influences on user behaviour should be investigated. As previously defined, social media is a prime form of connecting people in society and relies heavily on user-generated content. People share their ideas on a wide range of topics through social media, which has become an important aspect of modern culture (Fan, Du, Dahou, Ewees, Yousri, Elaziz, Elsheik, Abualigah & Al-qaness, 2021). As a result, Deep Learning provides enthralling tools for analysing users' behaviour and learning correlations between their previous and current attributes based on social media, hence its use to predict human behaviour in social networks in a variety of research.

Tommasel, Rodriguez, and Godyoy (2018) used hybrid Support Vector Machines and Recurrent Neural Network model to detect aggressive content in multiple social media. The model analysed a broad range of word, character, sentiment, word embedding, and irony features. To train the classifier, the K-nearest neighbour algorithm, the Naive Bayesian algorithm, and the Support Vector Machine algorithm were employed. Another work by Ranganathan, Hedge, Irudayaraj, and Tzacheva (2018) used decision tree classification to detect emotions on Twitter. They created a corpus of tweets and related fields in which each tweet is identified by emotion using vocabulary and emoticons.

Fan et al. (2021) built an artefact to identify and classify toxicity in social networks using the Bidirectional Encoder Representations from Transformers (BERT). Malhotra and Jindal (2020) proposed a real-time deep learning-based system to analyse online social media to detect onset depression, self-harming, and suicidal behaviour. Xue, Wu, Hong, Guo, Gao, Wu, Zhong, & Sun (2018) proposed a deep learning approach for personality recognition from social media posts.

2.14 Text representation techniques for classification purpose

Learning word representation is a critical step in many natural language processing tasks. Distributed representations, also referred to as word embedding, have made enormous strides in learning a transformation of each word from raw text data to a dense, lower-dimensional vector space. The majority of methods currently in use utilise contextual data from corpora (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014) as well as supplementary data such as sub word information (Cao and Lu, 2017), implicit syntactic dependencies (Shen, Zhou, Long, Jiang, Pan, & Zhang, 2018a; Shen, Zhou, Long,

Jiang, & Zhang, 2018b), and semantic relations (Bollegala, Alsuhaibani, Maehara, & Kawarabayashi, 2016; Liu, Huang, Zhang, Gao, Xuan, & Lu, 2018). The aforementioned context-aware word embedding perform well in conventional evaluations including word analogy and similarity because semantic information is crucial to these tasks. However, in practical applications such as information retrieval and text classification, word contexts alone are unlikely to guarantee success in the dearth of task-specific attributes. The sections that follow provide detailed discussions of the different text representation techniques used for text classification.

Word Embedding

A feature learning technique called word embedding maps each word or phrase in a vocabulary to an N-dimensional vector of real integers. Even though the model has syntactic word representations, this does not necessarily mean that it accurately captures the semantic meaning of the words. Bag-of-word models, on the other hand, disregard the word's semantics. For instance, the words “car”, “vehicle”, and “automobile” are frequently used interchangeably. In the bag-of-words model, however, these words' corresponding vectors are orthogonal. This issue poses a significant challenge to comprehending sentences within the model. The bag-of-words also has the flaw of disregarding the phrase's word order. Since the n-gram cannot solve this issue, each word in the sentence must be compared to another to find similarities. To tackle this challenge, many academics focused on word embedding. With regard to their Skip-gram and Continuous Bag-of-Words (CBOW) models, Mikolov, Chen, Corrado, and Dean (2013) present a straightforward one-layer architecture based on the internal product of two-word vectors.

The conversion of unigrams into understandable input for machine learning classifiers has been accomplished using a variety of word embedding techniques. This sub-section focuses on reviewing a few of the most common techniques used for deep learning.

Word2Vec

Mikolov, Sutskever, Chen, Corrado, & Dean, (2013), and Mikolov et al., (2013) proposed “word to vector” representation as a more efficient architecture for word embedding. To construct a high-dimensional vector for each word, the Word2Vec method employs deep neural networks with two hidden layers, the Skip-gram model, and CBOW. The Skip-gram model examines a document of words represented by *w* as well as their context represented by *c*. The aim is to enhance the probability.

Figure 2.1 depicts a straightforward CBOW model that attempts to locate the text based on prior text, whereas the Skip-gram model looks for texts that might appear close to each other. The input layer and output layer weights represent $v \times N$ as a matrix of w (Rong, 2014).

This approach is quite useful for discovering text corpus correlations including word likeness. For instance, in the vector space that it allots to word to vector, this embedding would regard two words like "big" and "bigger" as being close to one another.

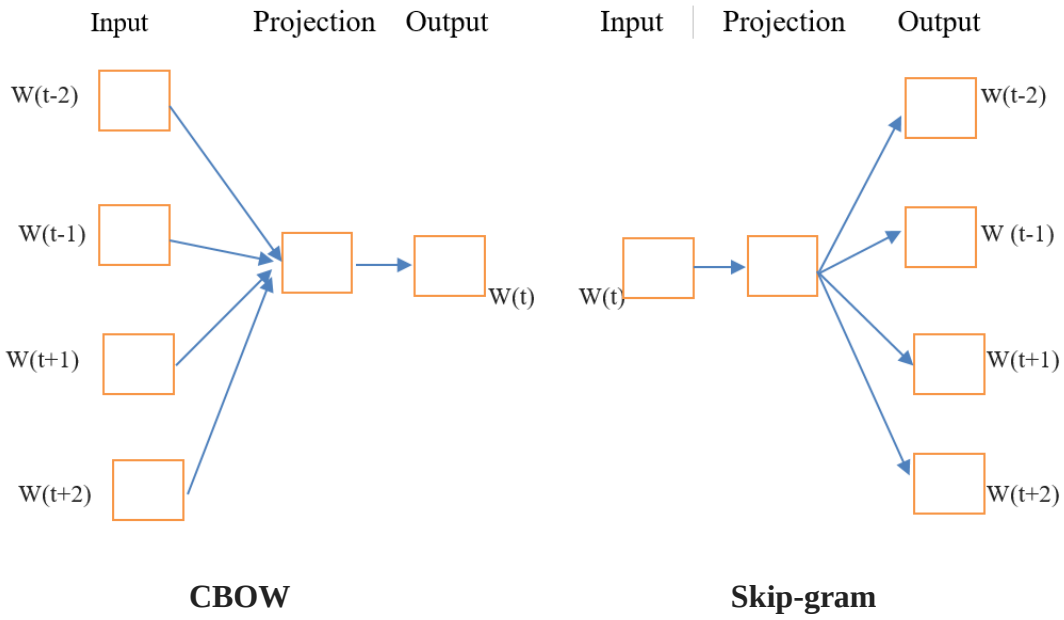

Figure 2.1: The continuous bag-of-words (CBOW) architecture.

Source: Mikolov et al., (2013)

The continuous bag-of-words model

Numerous words are used to represent a particular target of texts in the continuous bag-of-words model. For instance, the context words "airplane" and "army" for the target word "air-force". This involves making several attempts to duplicate the input to hidden layer connections in order to match the quantity of context words (Mikolov *et al*., 2013). As a result, the bag-of-words model is commonly utilised to represent an unordered group of texts as a vector. The first step is to establish a lexicon, which includes all of the corpus's unique words. The shallow neural network's output will be w_i and the task will be *"forecasting the word given its context"*. The amount of words used is determined by the window size choice (the most frequent size is 4–5) (Kowsari *et al*., 2019).

The continuous skip-gram model

The continuous skip-gram model is very similar to the CBOW architecture model (Mikolov *et al.*, 2013); Though, rather than predicting the current word contextually, the goal of this model is to optimize the classification of a word based on another word within the identical phrase. "For machine learning classifiers, the continuous bag-of-words model and continuous skip-gram model are employed to maintain the syntactic and semantic information of phrases" (Kowsari *et al.*, 2019:8).

Global vectors for word representation (GloVe)

Global Vectors (GloVe) is a robust word embedding technique being utilized for text classification (Pennington, Socher, & Manning, 2014). The method is comparable to the Word2Vec approach, in which each word is represented by a high-dimensional vector and trained using the adjacent words throughout a large corpus (Pennington et al., 2014; Kowsari *et al.*, 2019). "The pre-trained word embedding that is employed in many works is built on the 400,000 vocabularies that were trained using Wikipedia 2014 and Gigaword 5 as corpora, along with 50 dimensions for word representation" (Kowsari *et al.*, 2019:9). GloVe also offers pre-trained word vectorizations with 100, 200, and 300 dimensions that have been trained on even larger documents, such as Twitter data. The equation 2.1 denotes the GloVe:

$$f\left(w_i - w_j, \widetilde{w}_k\right) = \frac{P_{ik}}{P_{jk}} \qquad \text{Equation (2.1)}$$

where $w_i$denotes the word vector of word *i* and P_{ik} denotes the likelihood of word *k* occurring in the context of word *i*.

FastText

Bojanowski, Grave, Joulin, & Mikolov, (2017) assert that many alternative word embedding models neglect the morphology of words by giving a different vector to each word. By developing a new word embedding method dubbed FastText, the Facebook AI Research group introduced a revolutionary technique to overcome this issue (Kowsari *et al.*, 2019). Every word, *w*, is denoted as an n-gram bag of characters. For instance, given the word "institute" and n = 3, FastText will generate the representation below, which is made up of character tri-grams (Kowsari *et al.*, 2019).

"in, ins, nst, sti, tit, itu, tut, ute, te"

Term Frequency-Inverse Document Frequency

Inverse document frequency (IDF) was proposed by Jones (2004) as a tool to be used in conjunction with term frequency to reduce the impact of impliedly common words in a document. Words having a high or low-frequency term in the corpus are given higher weight by the IDF. Term Frequency-Inverse Document Frequency is the name given to the combination of TF and IDF (TF-IDF).

The equation formula of the weight of a term in a document by TF-IDF is presented in the following equation 2.2:

$$W(d, t) = TF(t, d) * log\ (\frac{N}{df(t)}) \quad \text{Equation (2.2)}$$

Where *N* is the number of documents in the corpus, and *TF* is the term frequency, *df* is the document frequency. *Df (t)* represents the document frequency of the term *t*. The first term in the equation enhances word embedding precision, while the second term enhances recall. TF-IDF attempts to address the issue of common terms in the document, however, it still has significant descriptive limits. Specifically, TF-IDF is unable to account for the likeness between the words in the document because each word is presented separately as an index. However, as more complicated models have been developed recently, other methods such as word embedding, which can combine notions such as word similarity and part of speech tagging, have been introduced.

2.15 Performance evaluation of text classification models

Data-driven AI systems often function in a complex ecosystem made up of a wide range of actors and components, between which a large number of signals and messages are transmitted. Significant signals encompass the predicted target value for a specific data case as guesstimated by the following:

(i) A model, for example, a class label in classification or a real number in regression,

(ii) The variance or ambiguity in these estimates, for example, confidence intervals, calibrated class probabilities, and

(iii) Performance measurements of a machine learning model on a test data set, for example classification accuracy or F1 (Flach, 2019). Our primary focus in this section will be on the latter category of signals.

While designing and using machine learning tools to address issues in text classification or other areas that will enable new developments in these fields, it is critical to understand that without appropriate tools for assessing the new approaches, there is no way to know whether or not they are performant. An essential component of machine learning is performance evaluation. But it's a difficult task. As a result, careful execution is required for machine learning applications to be trustworthy (Japkowicz & Shah, 2015). Over the past twenty years, performance assessment metrics for machine-learned classifiers have improved.

At least in supervised machine learning, it could seem that performance metrics and its associated signals are well established. Each of the evaluation metrics have a precise technical interpretation that can be connected to specific use cases. Furthermore, many of the evaluation metrics have clear correlations between them (Flach, 2019). A performance evaluation metric in machine learning is crucial for assessing how well a machine learning model performs on a new dataset (Dia, Ahvar, & Lee, 2022). A performance evaluation metric determines whether or not the machine learning model will successfully solve the problem for which it was trained. The performance of machine learning models for both classification and regression can be measured using a variety of performance evaluation criteria.

When assessing the effectiveness of a regression-based machine learning model, the R2 score is a crucial indicator. The coefficient of determination, or R squared, is another name for it. It operates by counting how much variance in the predictions the dataset can account for. To put it plainly, it is the discrepancy between the model's predictions and the dataset's sample data (Kharwal, 2021; Chicco, Warrens, & Jurman, 2021).

A confusion matrix is a tool for assessing how well a classification model is working. The goal is to determine how frequently cases of class 1 are categorized as class 2. For instance, one can use the confusion matrix to determine how frequently the classification model confused the dog and cat photos (Kharwal, 2021; Zeng, 2020).

One of the performance assessment metrics for a classification-based machine learning model is a classification report. It shows the recall, precision, F1, and support of the model. It gives a clearer picture of the overall effectiveness of the trained model. The ratio of genuine positives

to the total of both true and false positives is known as precision. Recall is determined by the proportion of true positives to the total of true positives and false negatives.

Moreover, F1 represents the weighted harmonic mean of recall and precision, whereas support denotes the quantity of instances of the class that actually occur in the dataset.

F1-score just describes the performance evaluation procedure and does not vary between models (Kharwal, 2021; Dansana, Kumar, Bhattacharjee, Hemanth, Gupta, Khanna, & Castillo, 2020).

The fraction of the variability of a machine learning model's prediction is measured using the explained variance. It is, in a nutshell, the discrepancy between the anticipated value and the expected value. Understanding how much information we can lose by recombining the dataset requires a strong understanding of the idea of explained variance (Camacho, Smilde, Saccenti, & Westerhuis, 2020; Kharwal, 2021). Performance evaluation is discussed in detail in chapter seven.

2.16 Positioning the research

Informal language, limited context, and noisy scant material characterize social media text (Virmani, Juneja, & Pillai, 2018). The process of extracting this information entails finding important, accurate, and usable facts from a haphazard and loud textual material. Without effective methodology, approaches, techniques, individuals, organisations, and governments are struggling to gain accurate information from social media data. Thus, this study has produced a novel methodology, approach and tools as supplement tools for assisting with extracting valuable information out of Twitter data in the context of academics' recruitment and selection. This study followed a supervised approach. Supervised learning, often known as supervised machine learning, is a subclass of artificial intelligence and machine learning. It is characterized by the use of labelled datasets to train algorithms that accurately identify data or predict outcomes (Kadhim, 2019; Jiang, Gradus, & Rosellini, 2020). The Twitter dataset had to be pre-processed, labelled based on some rules.

Literature informed the setting up of the rules used in this study. After data extraction and collection, the next step was to identify appropriate tools, techniques and requirements for helping to address the research objectives of this study. Literature from the departments of Human resources, Education, Psychology and Information technology, specifically with regard to artificial intelligence, machine learning, deep learning, and natural language pre-processing

was consulted to establish the set of rules applied to this study. The textual data was then pre-processed following a set of rules described in detail in chapter four. For this study, pre-processing rules such as converting text to lower case, the removal of URLs, @, #, special characters, numbers, and punctuation, tokenization and lemmatization proved to be suitable for this study. The data labelling took place after data pre-processing to get the data ready for classification (supervised learning).

Several studies have been conducted to measure and identify the behaviour of social media users. Also, several studies have been carried out on personality detection using well-known lexicons such as WordNet, SentiWordNet, SenticNet, and others (Das & Das, 2017). However, only few researchers linked personality with a career using social media data (Kern et al., 2019), and to this researcher's best knowledge, no study has ever been conducted before to use Twitter data to design and implement classification artefacts for predicting academics' suitability. This study assembled lexicons pertaining to academics' characteristics, specifically, interpersonal skills, attitudes, and behaviours. To the researcher's best knowledge, there is no study with dictionaries pertaining to the following academics: lecturers, supervisors, and researchers. Thus, in this study, the researcher assembled keywords pertaining to academics to be used as dictionaries to detect academic traces in social media. A set of words that best describe academics to find traces of academics in the Twitter dataset were used. Some terms with academic connotations that are most luckily to be found in online posts were identified. The lexicon dictionary assembled for this study is described in detail in chapter five. It was important to develop the academic lexicon dictionary for this study for data labelling purposes. After developing the dictionary to be used in this study, it was important to label the data since this study opted for a supervised learning approach. To label the data, a combination of sentiment analysis and/or lexicon detection was used. As this study opted for two types of classification, namely binary and multi-class classification, labelling was done in two ways: binary labelling; with two labels or classes and multi class labelling with four labels.

Researchers have used social data to design and implement artefacts that detect personality and profile different types of jobs. However, due to the paucity of studies on the use of Twitter data to detect academic suitability, in this study the researcher opted for deep learning methods to classify suitable academics. Moreover, this study provides a process model for detecting traces of academics on social media (Twitter).

2.17 Summary of the chapter

A growing corpus of studies has studied the association between personality and career success. Furthermore, researchers have studied the misrepresentation and dishonesty practised by job applicants during the interview process and during algorithm-based data screening. However, as mentioned before, the links between social media data and academic suitability have hardly been explored. Thus, to position the current study the researcher reviewed literature on the recruitment process, academics, social media, machine learning and deep learning techniques. Various perspectives of researchers and practitioners on the relevant topics were outlined before highlighting the gap. The methods and design used in this research are described in the next chapter.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Introduction

Research methodology is a methodical approach to solve a problem. It is a science of studying how to conduct research (Kumar, 2019). Description, explanation, and prediction of the phenomena are procedures involved in the research methodology (Rajasekar, Philominathan & Chinnathambi, 2013). Many academics make a distinction between research methods and methodology, with methods denoting the manner in which the research was carried out and methodology denoting the broad science or philosophy that serves as the study's foundation.

According to Creswell and Creswell (2017), research methods are research procedures and strategies, whereas methodology is the guiding structure that combines methods with the study's findings. This chapter presents the research methodology and methods employed to address the study's goals and objectives. The methods used to acquire and analyse data are explained in this chapter. The methods used guided this study in designing a process model. This chapter also discusses the designed process model which is based on the design science methodology and the CUPP framework. The following research questions were considered when deciding on which approach and procedures to use:

(i) How can the Twitter dataset be pre-processed for data labelling?
(ii) How can the resulting pre-processed data be labelled for classifying job-fit candidates?
(iii) How can the labelled data be used for the predictive model?

3.2 Common research methodologies

Information technology and information systems (IT/IS) have evolved, as have research methodologies to assist the field. This has bolstered the battle of research methodologies among researchers, which has been fuelled primarily by the foundations of social sciences and computer technology, which combine to form the business and technological application of information systems and information technology (Warfield, 2010). Research methodology refers to a systematic approach to problem solving. It is a science that studies how research should be conducted. Furthermore, it refers to the procedures by which researchers go about their work of describing, understanding, and predicting occurrences (Kassu, 2019; Goundar, 2012). According to Nabukenya (2012) and Mishra & Alok, (2017), there are a variety of

research designs, namely: case studies, action research, grounded theory research, surveys design science research, experimental design, and attitude research among others.

By combining the processes of acting and conducting research simultaneously and connecting them through critical reflection, action research aims to bring about transformative change. Action research is defined as research that arises from an investigator's interaction with participants of an institution on a topic that is of genuine importance to them and in which the members of the organization intend to take action as a result of their intervention. Action Research is largely used to better understand the innovation process in welfare systems (Argyris, Putnam, & Smith 1985; Mac Naughton, 2020).

Survey research (SR) is a term used to describe surveys that are carried out in order to increase scientific understanding. SR is particularly well-suited to addressing queries about what, how much, and how many, as well as inquiries about how and why to a greater extent than is commonly understood. In summary, survey methodology is based on measuring variables by asking people questions and then examining correlations between variables (Ghazi, Petersen, Reddy, & Nekkanti, 2018).

Case study research (CSR) is described as an empirical investigation into a current phenomenon in its real-life setting, especially when the distinctions between phenomenon and context are blurred. It is qualitative and observational in nature, with a predetermined scope of the study (Ebneyamini & Sadeghi 2018). Case study research, on the other hand, has some drawbacks: First, designing and scoping a CSR project to ensure that the research question(s) can be effectively and satisfactorily answered is challenging. Second, because businesses and other organizations are not always willing to participate in CSR, the availability of relevant case study sites may be limited. CSR reporting can be challenging since the rigor of the method used to get at the results, as well as the validity of the facts and conclusions reached, must be established. As a result, CSR is frequently regarded as lacking rigour (Nabukenya, 2012; Vega, 2018).

Grounded Theory (GT) is an inductive methodology for theory development that permits a researcher to create a theoretical explanation of a topic's general properties while also basing the account in empirical observations or facts. Rather than gathering data to test a theory or hypothesis, iGrounded Theory is conducted to generate a theory from data.

Design Science Research (DSR) is a problem-solving process aimed at developing new concepts, processes, technical skills, and solutions. Design Science Research produces a

systematic technique for designing and developing artefacts that address issues, making it particularly helpful in the actual world (Vom Brocke, Winter, Hevner, and Maedche, 2020). The nature of the research problem which is to build a job-fit predictive model using deep learning algorithms has prompted the researcher to opt for a design science approach. Deep learning algorithms have the capacity to handle large number of features, especially when working with unstructured data. Thus, this study adopted the design science approach as design science research produces artefacts in terms of processes. The former is further discussed in the section that follows:

3.3 The design science research approach

Design science research attempts to produce prescriptive knowledge regarding the design of information systems (IS) artefacts such as applications, procedures, models, or ideas (Hevner, March, Park, & Ram, 2004). It also guarantees that the predictive models are developed in a methodical way. Thus, in this study, the design science approach allowed the designing and implementation of the suitable academic models to follow a rigorous process of development.

Given the exploratory character of this research, and as design science research seeks to produce knowledge about how objects can and should be designed and implemented, typically by human initiative, to attain a desired set of objectives, design science approach is employed (vom Brocke et al., 2020). This research followed a design science methodology. Peffers, Tuunanen, Gengler, Marcus, Rothenberger, and Chatterjee (2007) proposed a design science process model with explicit steps that match the setting of this study. DSR projects' final goals or research products are referred to as artefacts (Hevner & Chatterjee, 2010). Models, methods, instantiations, and constructs are all examples of artefacts. The design science research approach is made of six phases, namely, problem identification and motivation, objectives for a solution, design and development, demonstration, evaluation, and communication.

3.3.1. Problem identification and motivation

Identification of the problem and motivation is the first phase of DSR process (figure 3.1). In this phase, the research problem is contextualised and actions to be taken in order to achieve the objectives are understood (Peffers *et al*., 2007). This study identified that Higher education institutions are facing challenges in recruiting and selecting academics who have the right attitude and passion for the job. The interview sessions seem not to provide the interviewers enough rooms to fully comprehend the available candidates. The identified problem motivated this study to provide a solution by designing and developing an additional tool to assist and

improve the recruitment and selection process of higher institutions, specifically in the selection process. Chapter one (section 1.3) discusses and emphasizes the problem.

3.3.2 Objectives for a solution

According to Peffers *et al.* (2007), the objectives should be rationally deduced from the problem formulation. Understanding the current status of the problem and current solutions, as well as their usefulness, if any, are essential resources in this phase. The objective for a solution is the second phase of the design science research approach (Figure 3.1). As most people spend a lot of time and display their true self on social media, and with the relatively small dataset, the Twitter data of this study can be used to tackle the issue of tracing suitable academics on social media (Twitter).

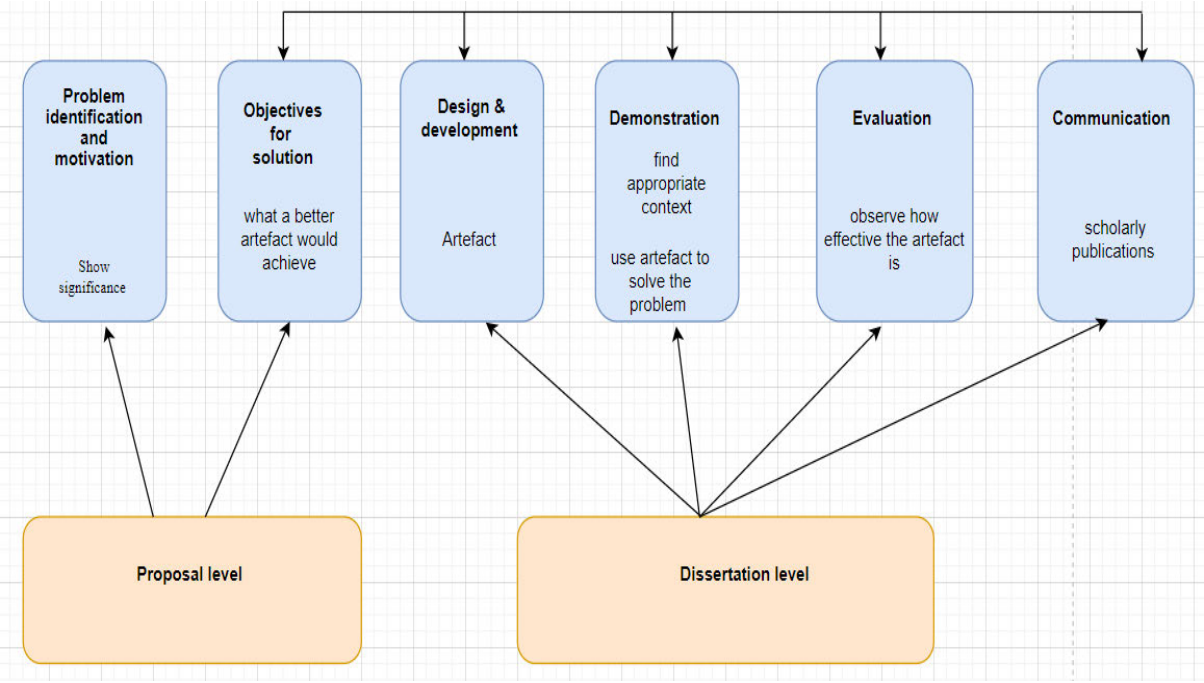

Figure 3.1: Research methodology pertaining to this research

3.3.3 Design and development

The design and development phase as depicted in Figure 3.1 starts with cleaning Twitter data, and data transformation as part of data pre-processing. Data pre-processing is a significant step that improves the quality of data and makes it easier to extract valuable insights from it (Huang, Li, & Xie, 2015; Zelaya, 2019). The primary objective of data pre-processing is to get the data

ready for the mining process. Data from the real world is commonly ambiguous, inaccurate, and incorrect (due to errors or outliers). The Twitter dataset in this study was cleaned based on a set of rules pertaining to the academic suitability described in section 4.3.1.1.

The second task in the design and development phase of this study is data labelling. Data labelling and annotation include the extraction of patterns and knowledge and then assigning labels to the dataset. The practice of labelling data so that machine-learning or deep learning classifiers can interpret and memorize the input data is known as data annotation (Ryazanov, Nylund, Basu, Hassellöv, & Schliep, 2021; Baur, Heimerl, Lingenfelser, Wagner, Valstar, Schuller, & André, 2020). Since this study opted for supervised learning, it was paramount to label the pre-processed data in line with the academic suitability requirement of this study. Thus, after being pre-processed, the Twitter dataset was labelled according to the academic suitability criteria for binary and multi-class classification. Sections 5.3.3.1 and 5.3.3.2 provide more details about the labelling experiments and rules applied.

The third task in the design and development was the word embedding experiment. Word embedding is a sort of word representation that allows for the depiction of words with comparable meanings (Aklouche, Bounhas, & Slimani, 2018; Mikolov *et al*., 2013). GloVe, FastText, and Keras tokenizer were used in this study as word embedding methods to vectorise the cleaned and labelled data for training and testing purposes as described in section 6.5.

The academic suitability model design and implementation was the final stretch in the design and development stage of this study. During the design and implementation of academic suitability artefacts, model specifications and the selection of appropriate algorithms were made. The best experiments were selected for this study after a variety of experiments were performed at this step to meet the academic suitability requirement. More details about this step are provided in section 6.11.

3.3.4 Demonstration

One of the elements necessary for the demonstration is a thorough comprehension of how to use the artefact to solve the problem (Peffers, Tuunanen, & Niehaves, 2018). The demonstration phase as depicted in figure 3.1 entails the demonstration of the experiments. The data cleaning and pre-processing experiments are presented in chapter four. The dataset was cleaned in accordance with the objectives of this study. After cleaning the raw data, the labelling experiment presented in chapter five took place. The labelling of the Twitter dataset was done in accordance with the criteria for academic suitability. The data was labelled for

classification purpose. The classification experiments of this study are presented in chapter six. Four artefacts, namely, BiLSTM.G.2, BiLSTM.F.2, ANN.4, and LSTM.4, implemented in this study are presented in chapter six section 6.11. The implemented artefacts were evaluated in chapter seven sections 7.4; 7.5; 7.6 and 7.7.

3.3.5 Evaluation

The fifth phase of the DSRM process is evaluation (Figure 3.1), which involves comparing a solution's aims to real results acquired during the demonstration of the use of the artefact. The former necessitates a thorough understanding of important measurements and analytic methods (Venable, Pries-Heje, & Baskerville, 2012). After evaluation, the researcher can decide whether to go back and enhance the artefact's design and development or go on to communication and leave further changes to future studies (Deng & Ji, 2018). The performance of the academic suitability artefact was evaluated in this phase. The aim was to find out how well the academic suitability artefacts are performing in addressing the research objectives.

3.3.6 Communication

Presentation and communication as depicted in Figure 3.1 is the last phase of the design science methodology. The results of DSRM research should be properly communicated to both scientific and non-scientific audiences (Peffers et al., 2018). As the primary means of communication to an academic audience for this study, the researcher implements the design science research's concluding phase through a written dissertation. Moreover, the journal articles from the dissertation are the outcomes to be communicated to the reviewers and public audience.

Given the processes involved in designing and implementing the artefacts for this study, it was important to incorporate a social media framework to guide the study. The social media framework fits into the dissertation level of the DSR model which covers the design, demonstration, evaluation and communication phases. The next section discusses the available social media frameworks in literature and the selected framework for this study.

3.4 Social media methodological frameworks

Social media analytics entails using frameworks and techniques to "gather, analyse, evaluate, and visualise social media data" (Zeng, Chen, Lusch, and Li, 2010; Holsapple, Hsiao, & Pakath, 2018). The efficiency of social media analytics is predicated on the techniques and methods

utilised in this process. Thus, researchers have proposed multiple frameworks to get it done by removing various roadblocks along the way (Holsapple et al., 2018; Singh & Verma, 2020). Various Social media frameworks are discussed in the sections that follow.

3.4.1 Process flow diagram of system using CUP SMA

Singh and Verma, (2020), proposed a climbable, ultrafast, and fault-tolerant framework for processing real-time data to extract hidden insights while avoiding the overhead of big data technology. This framework is flexible enough to handle batch processing as well as real-time data streams in distributed and parallel contexts. This framework has five blocks namely, data/ Streaming API, Database, pre-processing block, processing block, and the visualisation unit. The first stage Data/ Streaming API focuses on capturing the data, the database component focuses on storing the collected data and analytics outcome and the pre-processing block focuses on applying the pre-processing of the collected data for transformation purposes. Moreover, the processing block focuses on applying various analytics techniques to extract meaningful insights and the visualisation unit focuses on plotting and giving visual sense to the analysed data.

The process flow diagram of system using CUP SMA (Figure 3.2) could not be used for this research as it focuses on live stream data. Furthermore, not only does the process lack iteration within the steps, but it also concludes with data analysis visualisation.

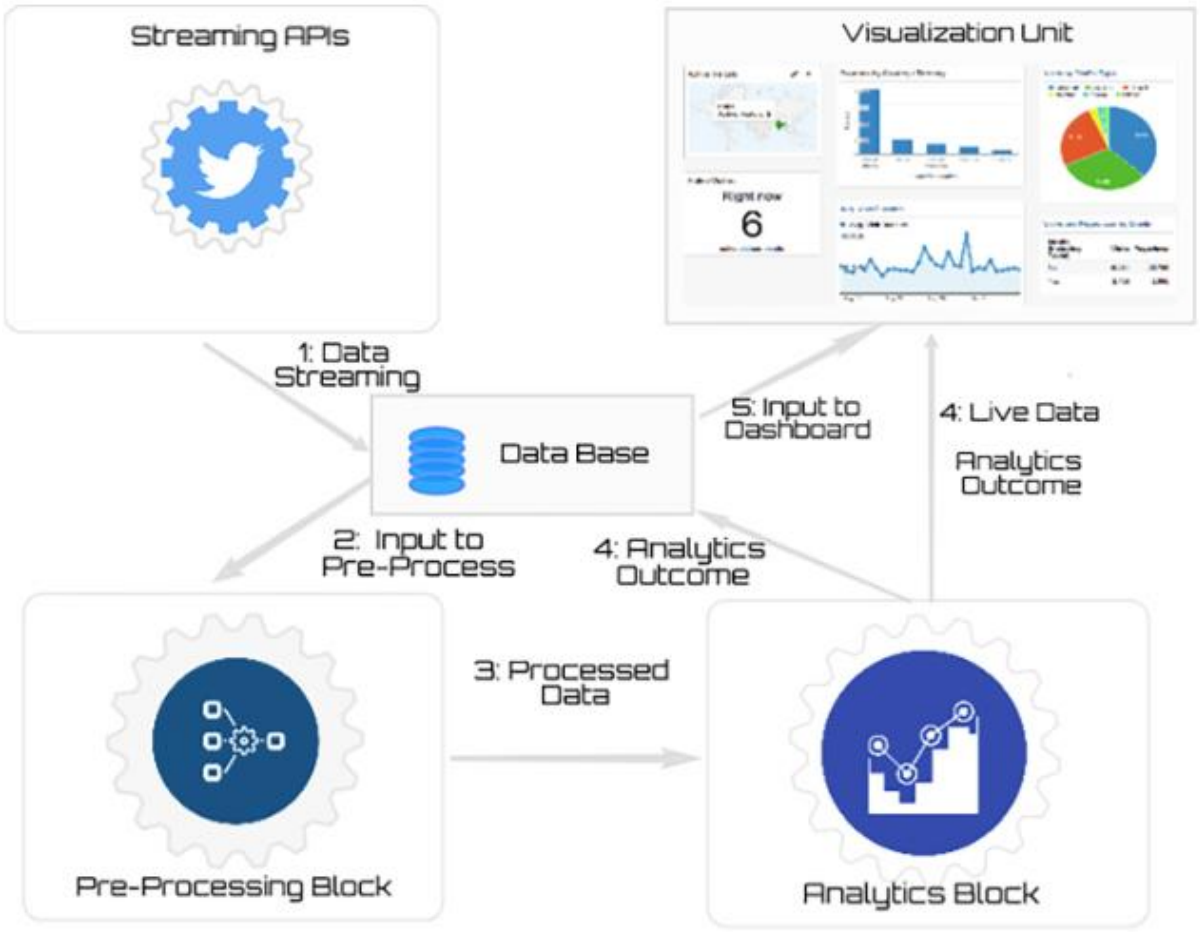

Figure 3.2: Process flow diagram of system using CUP SMA

Source: Singh and Verma, 2022

3.4.2 CUP framework

Fan and Gordon (2014) designed a social media analytics framework called the CUP framework (Figure 3.3), which has 3 phases, namely: "capture," "understand," and "present". Capture is the process for gathering social media information from various social networking sites, and channels. It can be done in-house or by a third-party service provider. A corporation utilizing social media analytics can use the capture stage to find interactions on social media sites that are pertinent to its interests and activities. Using news feeds and APIs, huge amounts of pertinent data are scraped or pulled from a plethora of social media platforms.

Following 'Capture,' the next step is to figure out (understand) how to choose relevant data for analysis while reducing noise from the data, applying various processing methods, and getting insight. When a firm obtains opinions on its products and procedures, it must figure out what those opinions imply (Fan & Gordon, 2014). The ‘understand’ stage, lies at the heart of social

media analytics (Fan and Gordon, 2014). Its outputs have the potential to have a direct impact on present data and indicators, as well as the progress of future corporate decisions and activities. Some analyses can be pre-processed offline depending on the methodology used and the information required (Fan & Gordon, 2014).

The present phase is focused on presenting the 'understanding' stage's findings in a meaningful manner. The results of numerous analytics are collated, analysed, and displayed to users in a simple and straightforward manner. Visualization techniques can be used to present information (Fan & Gordon, 2014).

This study made used of a modified CUP framework presented in section 3.3.4. The CUP framework as it stands, does not emphasises on the 'identify' and 'classification' steps. In this study, the researcher emphasizes the importance of including the 'identify' and 'classification' steps as these two steps are important in social media analytics.

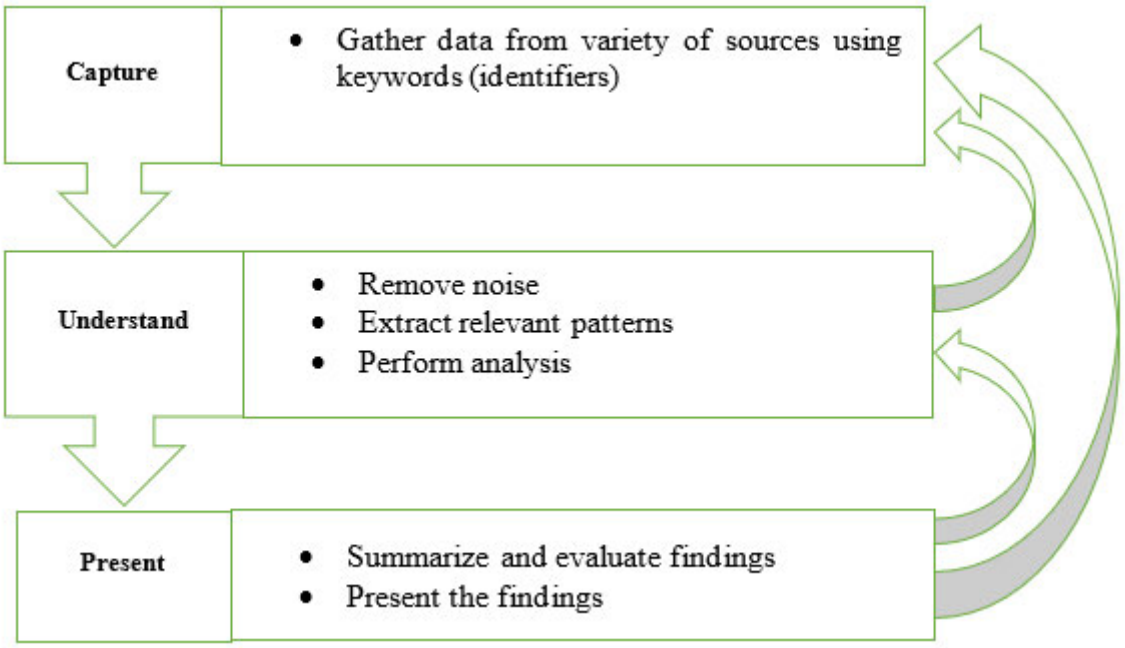

Figure 3.3: CUP Framework

(Fan & Gordon, 2014)

3.4.3 ICUP framework

Fan and Gordon's CUP framework was improved by Oh, Sasser, and Almahmoud (2015) by including an identity phase to identify tweets prior to the capture stage (figure 3.4). As a result, the modified CUP framework's four steps are as follows: Identify, gather, understand, and present. The initial step is to identify tweets by looking for keywords that are related to the

research purpose. Some keywords are specific, while others are gathered indirectly from other elements. The second phase is "capture," which entails the following two tasks: Twitter dataset download and pre-processing. The third phase is the “understanding” phase, which entails extracting and exploring the relevant data as well as data analysis. “Present” is the framework's last phase. In the SMA framework, it entails the process of summarizing and reporting findings. The lack of iteration within the steps is the main flaw of the ICUP framework. Given the dataset's nature, social media analysis is an ongoing and iterative process. Techniques used don't always equate to cleaner data or more accurate classifiers. Therefore, iteration is necessary to guarantee better classification outcomes. Thus, this framework could not be used for this research as it is.

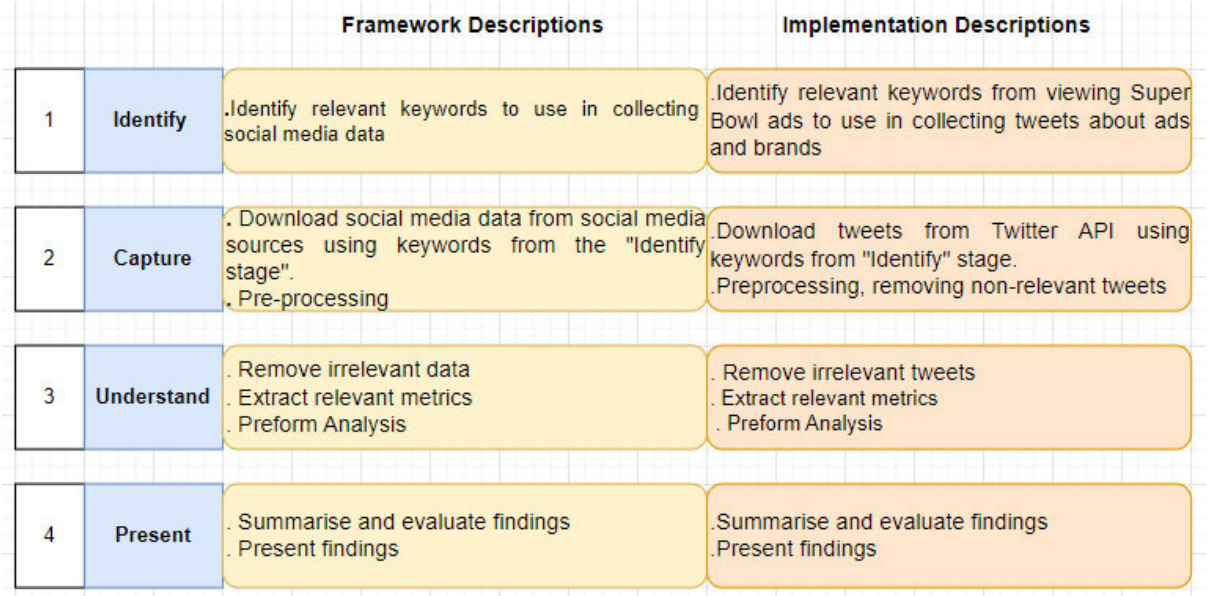

Figure 3.4: ICUP Framework

Source: Oh, Sasser, & Almahmoud, 2015

3.4.4 Modified CUP

This study has opted for the modified CUP Social Media Analytics (SMA) methodological framework (figure 3.5). The CUP model was originally designed by Fan and Gordon (2014) then modified by OH, Saaser, and Almahmoud (2015) to be suitable for their research's setting. For this study, one more stage was added to the modified CUP SMA to be suitable for their research's setting. The choice of this SMA model was based on this model’s capacity to fit the context of this study which is to design a predictive artefact using the Twitter dataset. The Social Media Analytics model helped this study to build predictive artefacts that provide insight for better decision-making in the recruitment process. Other models such as the model designed

by Stieglitz, Dang-Xuan, Bruns, Neuberger (2014) do not provide steps that fit the context of this study. The steps involved in the modified CUP are discussed in the sections that follow:

(i) Identify academics and non-academics

In this stage, the identification of all identifiers (Twitter handles, keys words) is made. Since the aim is to analyse and predict, the researcher should at least have an idea of keys words from the literature that can identify job-fit candidates for the proposed job opening. For this study, identify stage was used to identify academics and non-academics Twitter handles users.

(ii) Capture academics and non-academics

After identifying keywords or identifiers, there is a need to collect relevant data from social media. The data collected from social media are archived to be exported to an analysis application. The data collected must go through various pre-processing steps which include data modelling, part of speech tagging, stemming, feature extraction, and other syntactic and semantic operations that support analysis (Jurafsky & Martin, 2014). The main focus in this stage is to collect data that address the main purpose which is to build a predictive model for job-fit candidates.

In this stage, some challenges related to pre-processing should be addressed. The researcher should pay attention to the context and structure of data from social media. Mosley (2012) argues that social media data is informal and the main challenge is to connect the right set of data to be able to comprehend the context of the conversation. Best, Greenhalgh, Lewis, Saul, Carroll, & Bitz, (2012) supported that claim by saying that most messages are brief and that the context of each message should be understood. Data from Twitter was obtained for this study using the Application Programming Interface (API). R version 3.6.3 was used to extract data from Twitter through their API. Furthermore, the library rtweet was used for data extraction.

(iii) Understand the suitability of academics

After collecting the appropriate data, in the next stage, the "understand" stage, the extraction of relevant measures and the analysis of data are the main activities (Chong, Zhang, Mak, & Pang, 2015; Gordon & Fan, 2014). Techniques such as statistics, natural language processing, text and data mining, network analysis, and machine learning, are used in this stage to understand the data. This stage provides facts based on analysis. At this stage, effective metrics and trends can be generated. Data for this study were pre-processed, analysed, and labelled at this stage.

(iv) Suitability classification

One needs to have a good understanding of the current trend to be able to implement the classification phase. After, the understanding stage, preparing data for classification modelling has to be carried out using classification modelling techniques. This stage is added to the SMA framework designed by Gordon. In this study, the classified data was then used to build the job fit predictive model.

(v) Present the suitability of academics

The outcomes from the stage before are compiled and presented in this stage in a relatable format. Visualization techniques are used in this stage to present useful information. The common interface used in this stage is the dashboard (Gordon & Fan, 2014).

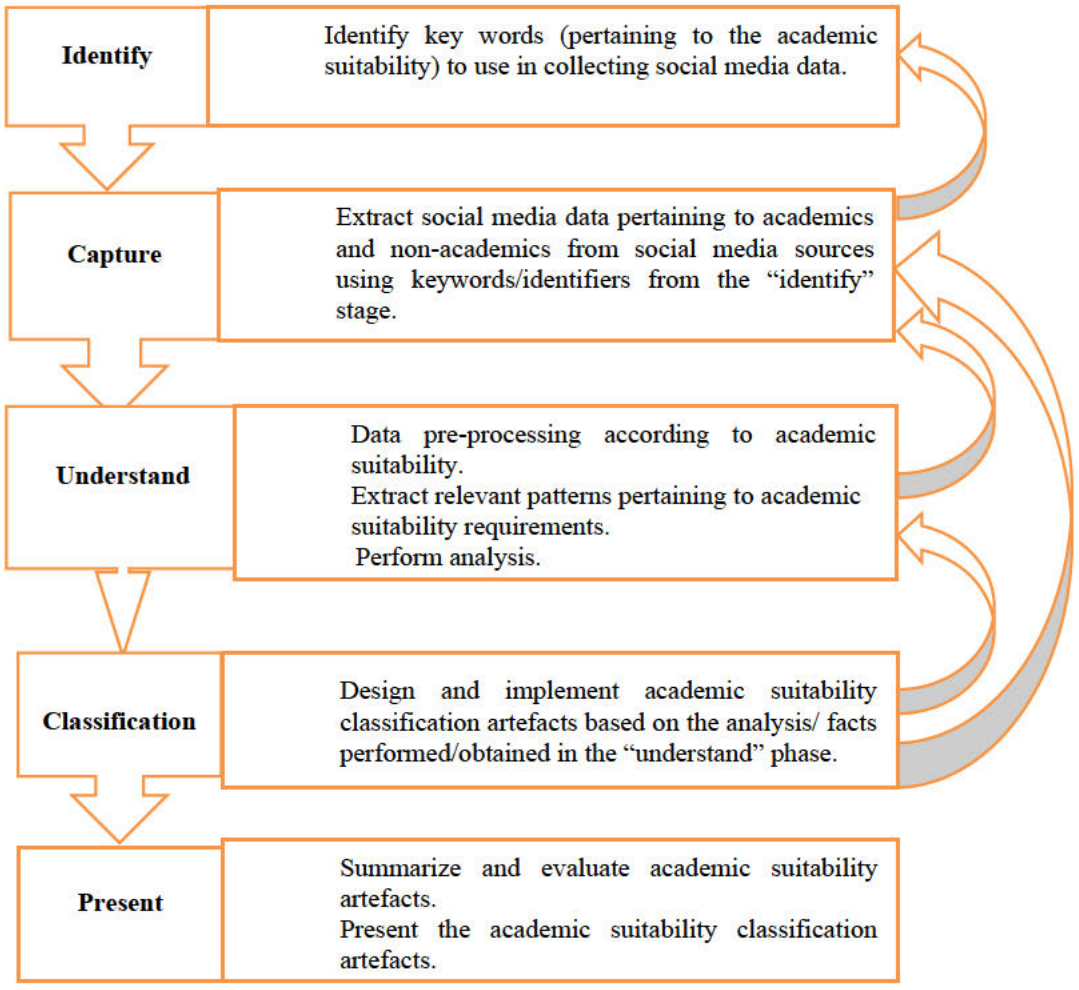

Figure 3.5: The modified CUP Social Media Analytics

Source: Researcher

3.5 The process model

The aim of this study is to identify the suitability of academics from Twitter using Machine Learning techniques. To this end, the modified CUP framework was chosen as a guiding framework to address the research objectives of this study (section 1.5). A process model is a simplified representation of the process involved in the artefact development (Kooij, Zacher, Wang, & Heckhausen, 2020; Brown, Von Daniels, Bocken, & Balkenende, 2021). Thus, a process model illustrated in Figure 3.6 was designed for this study to implement the modified CUP framework. The modified CUP research model (Figure 3.5) is implemented through five primary steps in the process model (Figure 3.6), namely, data extraction, rule-based data pre-processing, academic data labelling, and academic suitability model building and evaluation.

The first step of the process model is called data extraction which is within the "identify" and "capture" phases of the modified CUP research model (Figure 3.5). In this step, the task was to identify and extract available data for this study. Data was captured from Twitter and represented as raw data. In this step, raw data is the output of the "identifying" and "capturing" phases. Raw data serves as the input variable for the next step, that is, the rule-based data pre-processing step.

Data quality is a metric that assesses how well a dataset is fit for its intended use (Merino, Caballero, Rivas, & MSerrano, 2016; Heinrich, Hristova, Klier, Schiller, & Szubartowicz, 2018; Cichy, & Rass, 2019). The rule-based data pre-processing (step 2) is the first block within the "understand" phase of the CUP framework. A set of rules were used to clean, tokenize and lemmatize the data as described in section 4.3.1.1. Finding anomalies, spotting them, and fixing them are all tasks in the rule-based data pre-processing step. After the dataset has been cleaned, all tweets contain only plain words. The cleaning task removes noise, resulting in clean, and high-quality data for the future steps. After data cleaning, Tokenization is the next task.

Tokenization is the process of employing a tokenizer to break down tweets into texts, ciphers, or other expressive elements called tokens (Orbay & Akarun, 2020). The cleaned data served as the input variable and the tokens as the output variables during Tokenization. Words in each tweet were broken down to tokens. The token occurrences in a document are used directly to vectorise the corpus for classification purposes. Moreover, the tokens served as the input variables during lemmatization. During the lemmatization task, the aim is to distil a word/token down to its most basic form. The pre-processed data is the output of the rule-based data pre-processing step (step 2).

Academic data labelling (step 3) is the second block within the “understand” phase of the modified CUP framework. In this step, the input variable is the pre-processed data, and as described in sections 5.4.1 and 5.4.2, the labelling is done in two ways, namely: binary labelling and multiclass labelling. To achieve labelling, two main tasks were completed, namely, sentiment analysis and lexicon detection. The classes/labels serve as a description of the data's object class, and deep learning classifiers use them to recognize that specific class of tweets when given fresh data. The output variable of this step is the labelled data. The third step's goal is to prepare the data for the classification phase.

The academic suitability model building, and evaluation are step 4 and 5, respectively. These steps are the two steps within the “classification” phase of the modified CUP framework. In steps 4 and 5, the aim is to build a predictive model by training the labelled data, then testing and evaluating the trained classification artefacts. To do so, the first task was to vectorise the labelled data. This was done through a selection of three embedding methods as described in section 6.5. Furthermore, in step 4, different word embedding techniques and model specifications were used to find traces of academics in the tweets. More details about model classification are provided in chapter six. The last step is model evaluation whereby the performance of trained artefacts is, validated, evaluated and compared with existing artefacts. More details about model evaluation are provided in chapter seven. The output of the classification is presented in the form of a chapter and journal articles which are aligned with the last component of the modified CUP, that is, the “present” phase.

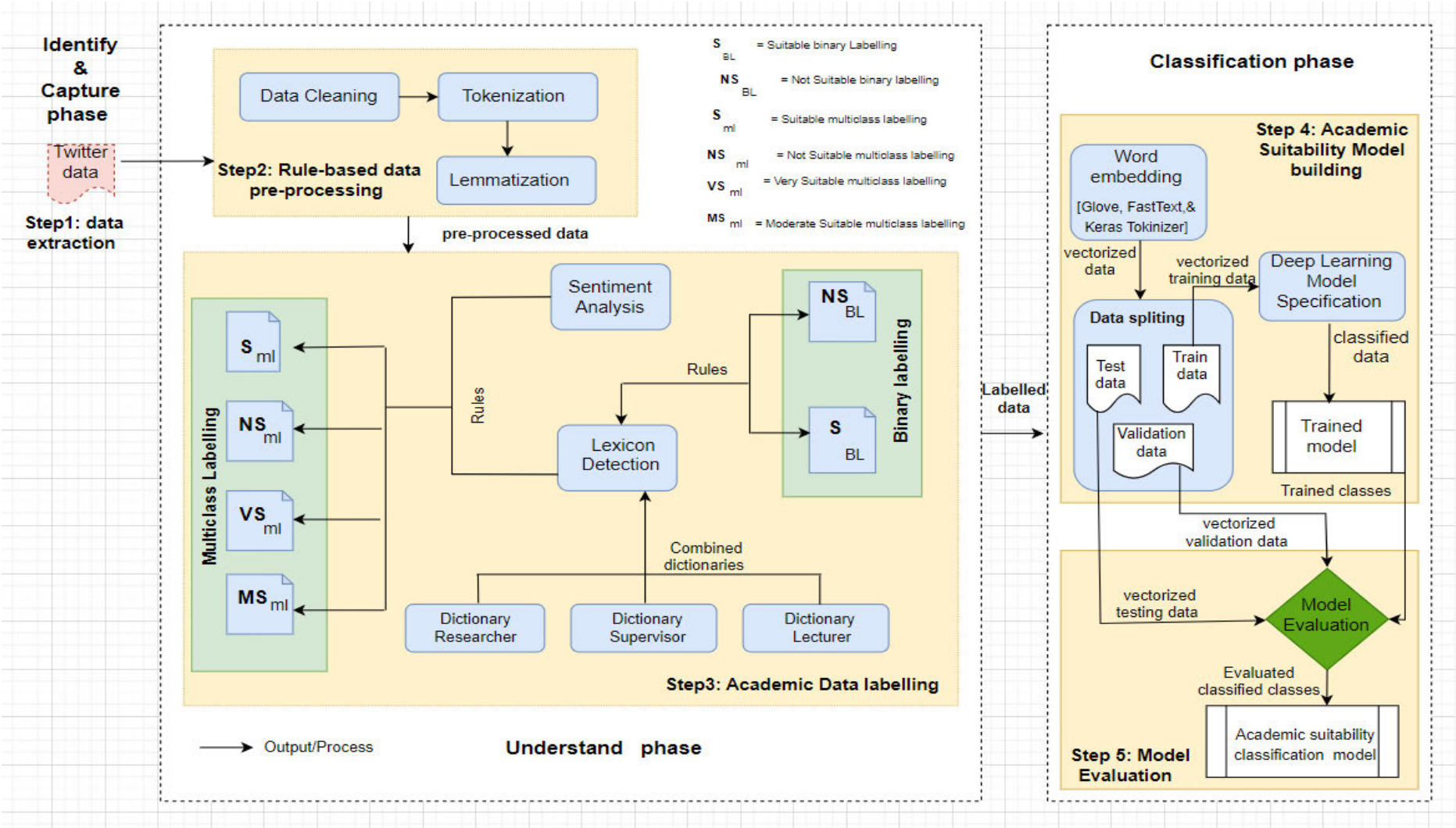

Figure 3.6: The process model for predicting job-fit candidates in academia

Source: Researcher.

3.6 Sampling

Sampling is the process of selecting units such as organisations and people from a population the researcher is interested in for inclusion in a study. The results drawn from the sample may be generalised (Taherdoost, 2016; Alvi, 2016). Yin (1994) defines a sample as a small group of the entire population chosen to partake in a study. In this study, the dataset extracted has 324.000 rows. Data was collected from 650 users, both academic and non-academic users. The researcher identified academics through their description. Any description such as Ph.D., master, researcher, or academic was considered as an indication of an individual being an academic.

3.7 Data collection

Data collection is the method of gathering information from all relevant sources in order to address the problem at hand, test the hypothesis, and evaluate the results (Clark, & Vealé, 2018; Neelankavil, 2015). There are two types of data, namely: secondary and primary (Hasan, 2021; Heap and Waters, 2019; Collis and Hussey, 2014). The research problem under investigation prompted this study to opt for Twitter data. The researcher applied for a developer account through Twitter API. The access allows the researcher to access tweets, retweets, likes, favourites, and tweet counts. The data extracted is generally in textual format. The Python and R applications were used to directly access Twitter API to extract tweets and retweets through queries.

3.8 Techniques and tools used for data analysis

Data analysis is the procedure of converting data into information that can be utilised to clarify and make decisions about a particular scenario (Anderson, 2015; Babak, Babak, Myslovych, Zaporozhets, & Zvaritch, 2020). Data was analysed using Python. Python was chosen for its widespread use in the data science community. Moreover, the choice of python was prompted by its performance in machine learning and deep learning. Various libraries were utilised for data analysis such as pandas, NumPy, NLTK, Scikit-learn, matplotlib, seaborn, and Textblob.

Numerical Python (NumPy) is a library that contains multidimensional array objects as well as a collection of functions for manipulating those arrays (Morra, 2018). Pandas is a data manipulation and analysis software package for the python programming language (Chen, 2017). Scikit-learn is a python-based machine-learning library. Furthermore, Scikit-learn is a community-maintained package that includes a range of well-documented and well-

implemented machine learning methods (Hao, & Ho, 2019). The Natural Language Toolkit, or NLTK for short, is a set of modules and programs for metaphorical and statistical natural language processing in English that are Python-based (Elbagir & Yang, 2019). Numpy, Scikit-learn, Nltk, Seaborn libraries were used during the pre-processing phase of the process model of this study.

For the labelling phase of the process model, TextBlob, Numpy, Matplotlib, and regular expressions (regex) were used to label the data. Matplotlib is a graphing library for Python and its NumPy numerical math extension. It offers an object-oriented API for embedding graphs into applications utilizing Tkinter and other general-purpose GUI toolkits (Lemenkova, 2020). Textblob is a text-processing package for Python 2 and 3. It offers a basic API for doing standard natural language processing (NLP) activities like part-of-speech tagging, classification, sentiment analysis, noun phrase extraction, and translation, among others (Hermansyah, & Sarno, 2020). Regular expressions (regex) are a technique of describing strings that follow a pattern (Michael, Donohue, Davis, Lee, & Servant, 2019).

For the classification phase of the process model of this study, ANN, LSTM, BiLSTM, sigmoid, Fastext, Glove, Keras tokenizers were used as tools and techniques for classification. More details on these tokenizers are provided in chapter six and seven.

Furthermore, a Dell computer with Windows 10 Enterprise as an operating system, Intel ® Core ™ i5-8500 CPU @3.00HHz, 3000Mhz, 6 Core(s), 6 Logical processor(s), Physical Memory (RAM) of 16 GB was used for the analysis. In addition, Google Collaboration (Colab) was also used as an online Jupiter notebook for data analysis. Python 3.6 is the version used for the data analysis of this study. Chapters 4, 5, 6, and 7 provide more details on the analysis and how the tools were used.

3.8 Reliability and validity

Validity is the capacity of an artefact to measure what it is prearranged to measure. Criterion validity is the degree to which a measuring tool accurately predicts behaviour or ability in a particular area. The measuring instrument is called “criteria”. It is of two types, namely: predictive validity, and concurrent validity. The degree to which results from a test or scale can be used to predict results from other criterion measures is known as predictive validity (Hoffmann, Wiben, Kruse, Jacobsen, Lembeck, & Holm, 2020). Concurrent validity is a form of evidence that can be used to support the application of a test for forecasting other outcomes. It is an element used in sociology, psychology, and other behavioural or psychometric sciences

(Ng, Collins, Hickling, & Bell, 2019). Reliability is the ability of an instrument to create reproducible results. Whenever reused, similar scores should be obtained (Deepan, Shamaz, Chandrashekar, and Joe 2015). This study used a cross-validation technique to measure validity and reliability.

Cross-validation is a resampling technique for evaluating machine-learning artefacts on a small sample of data (Alakus, & Turkoglu 2020; Ko, Chung, Kang, Kim, Shin, Kang, & Lee, 2020). According to Berrar (2019), one of the most extensively used data resampling approaches for estimating the genuine prediction error of models and tuning model parameters is cross-validation. The latter is a method of resampling data that is used to assess the generalizability of prediction models and prevent overfitting. Techniques such as regularization and early stopping were used as cross validation techniques for reliability and validity techniques of the deep learning artefacts in this study. Early stopping is a type of regularization employed in deep learning to minimize overfitting when using an optimization technique such as gradient descent to train a learner (Song, Kim, Park, & Lee, 2019). Metrics such as precision, F1 score and validation loss were also employed to measure the validity and reliability of the artefacts.

3.9 Ethical considerations

A researcher should abide by a set of moral principles during the research process to prevent any misconduct (Resnik, 2015). In order to safeguard the participants' anonymity and privacy, in this study the researcher used Twitter datasets that have been anonymized by removing usernames and Twitter handles during the data pre-processing phase. In order to extract data from Twitter, the first step was to apply for a Twitter API account for developers. The developer account allows researchers/developers to get consumer or search keys for extracting data/public tweets from Twitter (Appendix 2). The ethical clearance approval from the research ethics committee at the University of KwaZulu-Natal was then applied for. Data extraction was carried out only after obtaining the ethical clearance (Appendix B).

Anonymity and confidentiality were applied to protect the research subjects' identities and well-being. Anonymity refers to the safeguarding of study participants' identity so that not even the researcher cannot link them to the data gathered (King, Horrocks, & Brooks, 2018).

When neither the researcher nor the readers can associate a particular response with a particular respondent, anonymity in the research study is insured (Babbie, 2015; Babbie, 2020). The

anonymity of the participants in this study was guaranteed by removing their usernames and geolocation information. A research study maintains confidentiality if the researcher recognizes a specific respondent's responses but commits not to reveal them publicly (Babbie, 2015:35). In this study, the researcher did not reveal any of the respondents' personal information. All of the participants were kept anonymous. To ensure anonymity, user names, Twitter handles were omitted during the data cleaning phase.

3.10 Summary of the chapter

The necessity to obtain information from social media networks as a supplement to conventional media has grown since the advent of social media usage. Social media networks provide access to a vast public database of textual data that can be used to gather critical information. In chapter three, the researcher has explained how social media data can be utilised methodologically for scientific research. Furthermore, the research methodology, design, approach, process model, social media frameworks, population, data analysis techniques and the ethical considerations were discussed in detail in this chapter. The modified social media iterative framework as well as the process model integrated in the social media analysis framework were also discussed. The process model is made up of steps/processes that are intended for finding academic traces in Twitter dataset. The next chapter presents the data pre-processing procedure that was employed in this study.

CHAPTER FOUR: ACADEMIC DATA PRE-PROCESSING

4.1 Introduction

Data mining is a technique for extracting useful patterns and trends from data. Missing values, noisy data, missing information, data inconsistency, and outlier data are common in raw data. Data pre-processing is a crucial step in increasing computational effectiveness. Preparing and translating data into an appropriate format is one of the most frequent tasks involved in data mining. The percentage of decreased noise in the datasets determines the efficacy of pre-processing (Mehanna & Mahmuddin, 2021). Based on the discussion in chapter two sections 2.8; 2.8.1 and 2.8.2, text pre-processing in this study was modelled as a rule-based information extraction procedure that gets tweets ready for labelling. The rules selected for this study were dimed appropriate for the purpose of this study, which was focussed on determining the suitability of academics. Thus, in this chapter the steps involved in the data preparation for labelling are discussed and presented. The researcher discusses in detail the data cleaning, tokenization and lemmatization involved in the rule-based pre-processing step of the process model discussed in section 3.5. In order to achieve the goal of having reliable and accurate data ready for labelling, the information extraction step comprises rules that are explained in this study. This chapter addresses the first research objective, which is as follows:

To identify and implement pre-processing techniques in order to clean the dataset according to job-fit attitudes criteria.

4.2 Data extraction and Twitter dataset

The tremendous rise of information technology in recent decades has made information transmission more crucial (Troussas, Virvou, & Mesaretzidis, 2015). Internet users now have the ability to express and share their opinions on a variety of topics and events thanks to the growth of social media. Social media websites like Twitter, Instagram, Reddit, LinkedIn, and Facebook have become more important in recent years, and they tend to be a cornerstone in the diffusion of data and information because users can access a great amount of information when using them (Krouska, Troussas, & Virvou, 2016; Troussas, Virvou, & Espinosa, 2015). Twitter, for example, provides a platform for detecting discussions on a variety of issues faster than other traditional information sources (Krouska, *et al*., 2016). Twitter is a well-known microblogging and social media platform service that allows users to share, broadcast, and interpret 140-character posts known as tweets (Singh & Kumari, 2016). Users use Twitter to

express their opinions (Wagh & Punde, 2018). The need to discover digital footprints of academics from textual data and predict their academic suitability, has been the most important aspect that triggered the choice of Twitter. Thus, data extraction started after obtaining all required permissions.

The activity or operation of extracting data from data sources for subsequent processing or storage is known as data extraction (Raza & Gulwani, 2020). Data for this study was extracted from Twitter. Twitter dataset can be obtained through API or from commercial enterprises that sell Twitter data such as Audience, Risetag, or Tweetdeck (Desai, 2018). R libraries such as rtweet and tidyverse were used to scrape data through Twitter API. Timelines of identified Twitter handles were extracted. Several researchers have used Twitter data, Liu and Homan (2019) created a Twitter Job/Employment Corpus, which is a compilation of tweets annotated using a supervised learning framework with humans in the loop. Banda, Tekumalla, Wang, Yu, Liu, Ding, & Chowell, (2021) created a curated dataset of over 1.12 billion tweets related to COVID-19 chatter. Initially, this study used 324215 tweets from 650 users. After, different pre-processing activities, the dataset remained with 308986 tweets. The study made use of Twitter API through developer account to extract and collect data.

The extracted Twitter dataset consists of 3 columns. Table 4.1 and Table 4.2 present some of the columns, and/or attributes within the extracted dataset.

Table 4.1: The extracted Twitter sample dataset

	user_id	text	description
0	992523275291918336	The real name of my degree: https://t.co/gEYbY...	PhD with @LizHadly and @PetrovADmitri. genomic...
1	992523275291918336	How community-led conservation can save wildli...	PhD with @LizHadly and @PetrovADmitri. genomic...
2	992523275291918336	Please nominate exception grad students for a ...	PhD with @LizHadly and @PetrovADmitri. genomic...
3	992523275291918336	Can you help our park partner @HaleakalaNPS ?\...	PhD with @LizHadly and @PetrovADmitri. genomic...
4	992523275291918336	NOT HAPPY ABOUT THIS\nJackass killed a bunch o...	PhD with @LizHadly and @PetrovADmitri. genomic...
5	992523275291918336	Does anybody have examples of successful @NSF_...	PhD with @LizHadly and @PetrovADmitri. genomic...
6	992523275291918336	@LangurLover https://t.co/ZLiOyOGNV7	PhD with @LizHadly and @PetrovADmitri. genomic...
7	992523275291918336	Also important to note - we recently learned t...	PhD with @LizHadly and @PetrovADmitri. genomic...
8	992523275291918336	Yeah cuz um, geometry I learned in 7th grade a...	PhD with @LizHadly and @PetrovADmitri. genomic...
9	992523275291918336	@ThatLionLady im gonna do it this week i swear...	PhD with @LizHadly and @PetrovADmitri. genomic...

Table 4.2: A sample of attributes of the dataset

User_id	**"992523275291918336"** The Identity given to a specific user
Text	**"the real name of my degree"** Refers to the tweet(text) made by the user
Description	**"Phd with @LizHadly and @PetrovADmitri.Genomic"** Refers to how a user describe him/herself

4.3 Data preparation

Large volumes of data are becoming easier for businesses to store and collect. These datasets can help with better decision-making, more in-depth analytics, and increasingly, they can supply training data for machine learning (Anderson, & Roth, 2018). Finding and correcting inaccurate data is a problem that data analytics faces on a regular basis, and failing to do so can lead to erroneous analytics and unreliable conclusions. A crucial method in current textual data analytics tasks is data cleaning. Data cleaning entails modifying text before analysis by determining which units to use, deleting content that is useless for particular tasks, combining semantically related phrases to reduce data sparsity and improve predictive power, and enhancing the quantity of semantic information recorded (Hickman, Thapa, Tay, Cao, & Srinivasan, 2020).

Data cleaning, on the other hand, might eliminate useful information, for example, deleting stop words that are crucial to a research subject, introduce errors into the analysis and radically affect subsequent results (Boyd, 2016). In practice, achieving this trade off simpler data with minimal information loss is a challenging task, and scholars have spent a lot of time figuring out the best method to go about it (Denny & Spirling, 2018).

4.3.1 Data cleaning techniques and tools

Data cleaning is a critical phase in Natural Language Processing (NLP). Twitter data is messy by nature, with typos, idiomatic expressions, slang, and undesired information like URLs and idioms (Pereira, 2017; Saini, Punhani, Bathla, & Shukla, 2019). Data cleaning's major goal is to find and fix mistakes and anomalies in order to enhance data quality for analysis and decision-making (Ridzuan & Zainon, 2019). Techniques such as regular expressions are used as tool of data cleaning.

Regular expressions (regex) are a technique of describing strings that follow a pattern (Davis, 2019). Regexes are often used to handle problems like input validation and find/replace in most

mainstream programming languages (Nield, 2017; Davis, 2019). Regular expressions are a type of keyword search that allows you to identify text by utilizing a pattern rather than an exact string. Regexes are frequently used to parse text in general-purpose languages, validate content entered into web forms, and search text files for a certain pattern (Chapman & Stolee, 2016). Some of the data cleaning rules in this study were implemented using regex. The Regex was used in this study for its performance and widespread use in the data science community. Regex helped in removing and replacing characters (words, numbers, and emoticons) in the Twitter dataset.

Natural Language Toolkit (NLTK): is a collection of Python-based computer modules, libraries, and other content modification techniques for the English language**.** Bird and Loper of the University of Pennsylvania developed NLTK. The NLTK toolbox can help in a variety of NLP tasks, including tokenization, lemmatization, stemming, parse tree representations, labelling, parsing, chunking, and Named Entity Recognition "NER". "Corpus readers", "tokenizers", "stemmers", "taggers", "chunkers", "parsers", "wordnet", and NER all have test codes in NLTK (Yogish, Manjunath, & Hegadi, 2018). NLTK was used in this study for stop words removal, tokenization, and lemmatization. NLTK was used for its performance and its widespread use in the data science community. Furthermore, Yogish, Manjunath, & Hegadi, (2018) claim that the Natural Language Toolkit is the most effective way to get started with textual data cleaning as part of Natural Language pre-processing. The NLTK is preloaded with a set of stop words that proved to be effective for this study.

4.3.2 Rule-based data cleaning

Data cleaning is a required step in the data preparation process for text classification, sentiment analysis, word detection, and topic modelling (Krouska, *et al*., 2016). Twitter data is semi-structured/unstructured and contains a lot of extraneous information that is irrelevant to the prediction. Large datasets also require more time for training, and stop words decrease predictive performance. In order to conserve computational resources and improve prediction accuracy, text pre-processing is necessary (Umer, Ashraf, Mehmood, Kumari, Ullah, & Sang Choi, 2021). The general data cleaning cycle rules consist of the following phases: Punctuation removal and non-alphabetic character removal, removal of URL, removal of @mention, lowercasing, stop word removal, removing whitespace and replacing elongated words.

Punctuation removal and non-alphabetic character removal:

When determining how to pre-process a dataset, the first decision a researcher must make is deciding on which classes of characters and mark-up to deem meaningful text. The non-letter characters and mark-up may be useful in certain analyses, but they are regarded as non-informative in many systems. As a result, eliminating them is a standard procedure. Punctuations are the most frequently eliminated of these character classes. The first data cleaning action we take is whether or not to maintain or delete punctuation (Umer *et al*., 2021). In this study, all punctuations `''‘! ()- []{};:'"\,<>./?$%^&#*_~'''` were removed as they did not contribute to the study. Regex was used in this study to remove unnecessary punctuation.

The text has been cleared of any numbers or other special characters (Denny & Spirling, 2018). Numbers are deleted from Tweets as they have no significance for text mining and deleting them reduces the intricacy of the training of the model. Statistical power is increased. Reduces the capacity to capture the style of speech and the authenticity of the data (Hickman *et al*., 2020). All numbers in the corpus were removed as the study focused on text only and numbers did not have much meaning for the purpose of this study which is finding traces of academics in text. The following equation: ['text'].("[^a-zA-Z]", " ") depicts the regex used in this study to replace special characters. This regex is matching any tweet (character) in the text column of the dataset that is not a-z or A-Z. This equation matches only alphabetical characters and remove non-alphabetical characters. This rule is used in step 2.4.1 of procedure 4.1.

['text'].("[^a-zA-Z]", " ") Equation (4.1)

Removal of URLs:

The removal of URLs is done to get rid of any URLs in the tweet. It involves the removal of URLs that begin with HTTP, pic: \\, and https (Pradha et al., 2019). All URLs were removed using regex as URLs did not yield any useful information for this study. The equation ['text'].('(https?://[^>]+)') depicts the regex used in this study to extract any url. This regex is matching and extracting any urls found in any tweet in the text column. This rule was is in step 2.2.1 of procedure 4.1.

['text'].('(https?://[^>]+)') Equation (4.2)

The equation 4.3 depicts the regex syntax used in this study. The regex equation is matching and removing any RT found in a tweet in the text column of the dataset. This rule is used in step 2.3.1 of procedure 4.1.

['text '].("[RT]", " ") Equation (4.3)

Removal of the symbol "@"

Removing the symbol "@" makes it easier to get rid of user mentions because they do not supply any useful information about the text (Pradha et al., 2019). All @ symbols were removed using Regex from the corpus as they did not supply any useful information. The equation 4.4 depicts the regex used in this study to match and remove any tag from the text column within the dataset. This rule is used in step 2.5.1 of procedure 4.1

['text']. ("@[\w]*") Equation (4.4)

Lowercasing

The process of changing all letters in a dataset to lowercase letters is known as lowercase conversion. During text mining, researchers recommend always using lowercase conversion (Kobayashi *et al*., 2018b; Banks *et al.*, 2018). The reasoning behind this is that whether or not a word's first letter is uppercase, for example, when it begins a phrase, it rarely affects its meaning (Umer *et al*., 2021). As computers can tell the difference between capital and lowercase letters, the same text written in capital or lowercase could be quantified differently if not unified. Furthermore, lowercasing can increase statistical power (Hickman, Thapa, Tay, Cao, & Srinivasan, 2020). Furthermore, Hickman *et al*., (2020) recommend to apply lowercasing unless there are relevant terms in the corpus that differ semantically. In this study, texts were lowercased as the researcher wanted all terms in the corpus to be quantified same way. Regex was used to convert all text into lower case. The equation 4.5 depicts the regex used in converting in tweet found in the text column of the dataset. This rule is used in step 2.1.2 of the procedure 4.1.

['text']. apply(lambda x: x.lower()) Equation (4.5)

Stop words removal

Stop word removal are commonly used in NLP to exclude terms that are so widely but contain little meaningful information (Banks, Woznyj, Wesslen, & Ross, 2022). This decreases the size of the corpus without sacrificing crucial information (Krouska, *et al*., 2016). Stop words were removed from the corpus as their removal did not affect the meaning of the text in the corpus. The NLTK was used to remove all stop words. The NLTK was used for its widespread use in NLP and performance.

Removing whitespace

Whitespace has no importance in the text, thus it is deleted for computational reasons (Pradha *et al*., 2019). After removing certain parts of the text in a tweet, generally there are white spaces left. For computational reason all extra white spaces were deleted in this study using strip () method. Strip () is a python method that deletes any leading and trailing characters or spaces at the beginning and end of each line. This rule is used in step 2.7.1 of the procedure 4.1

Replacing elongated words, abbreviations and slang

An elongated word is one that has a character that is frequently, yet wrongly repeated one or more times. For instance, "*wonnnderfullllllll", "yesssssss".* For the following two major reasons, some studies found it necessary to substitute words such these with their source words: first, to be comprehended by the classifier and second, to avoid exclusion due to uncommon occurrence. However, these lengthened words can sometimes convey an emotion, whether positive or negative (Mehanna & Mahmuddin, 2021). In this scenario, substituting the original word for the lengthened term will significantly lessen the sentiment's intensity. To address this issue, Agarwal, Xie, Vovsha, Rambow, & Passonneau, (2011) proposed replacing repetitive characters with only three characters, such as '*fullllllll*' by '*fulll*'. Due to their uncommon occurrence in this study, the Replacement elongated words were ignored.

Twitter's character limits discourage online users from using natural language and instead encourage them to utilise slang, acronyms, and abbreviated sentences. An abbreviation is a word that has been shortened or an acronym of a word. Slang is another informal means of communicating ideas or meaning that is sometimes limited to a certain audience or situation. It is also seen as casual. Therefore, it is essential to handle such informal insertions in the tweets by changing them to their true word meaning, since this improves the performance of automatic classifiers without information loss (Naseem, Razzak, & Eklund, 2021).

Stemming and lemmatization

This includes eliminating suffixes from words in order to return words to their initial form (Hickman *et al*., 2020). Stemming algorithms function by deleting the word's suffix according to grammatical norms. Stemming algorithms decrease vocabulary size by removing morphological prefixes and suffixes from words using heuristic methods, leaving just the word stem (Denny & Spirling, 2018). In a nutshell, lemmatization is the act of determining the root of any word in order to understand a phrase, clause, sentence, or other piece of content. If there are any errors, it reduces the validity (Hickman *et al*., 2020). Moreover, caution is advised because combining different word forms into one root can erase differences in speech style and/or miscategorise words if done incorrectly. Lemmatization and Stemming are two very distinct procedures that produce very different results (Hickman *et al*., 2020).

Tokenization

Tokenization means dividing the text into specific tokens that are the atomic pieces of information in the language representation of choice (Ofer, Brandes, & Linial, 2021). This technique divides the corpus into words/terms and creates a bag-of-words word vector. Tokenization has a considerable impact on the future analysis's performance, thus it must be accurate and efficient (Denny & Spirling, 2018). The corpus was divided into tokens using NLTK.

Rules that were used in this study

A clean tweet can only be written in lowercase letters, and should not have URL (http) links, special characters ^a-zA-Z", punctuations, hashtags, user handles (@user), stop words, and extra white spaces. The rules are presented using procedure 4.1.

After extracting and collecting the raw data from Twitter, the next step was to clean the data. Procedure 4.1 depicts the data cleaning rules of this study. The raw data from Twitter is referred in this study as Raw Tweet R_t . The R_t is the input variable of this data cleaning procedure 4.1. The output variable for procedure 4.1 is a clean tweet. In this study a clean tweet is referred to as C_t.

Procedure 4.1: Data cleaning

Input: R_t

Output: C_t

---

1. Initialize an empty string C_t to store the result of output and store the result in C_t
2. **For** each R_t

2.1 If a R_t is in uppercase,

2.1.2. convert the R_t to lowercase using regular expression (4.5).

2.2 *Else* if a R_t has URLs

2.2.1 Replace all URLs with the word 'URL' using regular expression (4.2) and store the result in C_t.

2.3 ***Else*** if a R_t has RT

2.3.1 Remove all using regular expression methods (4.3) and store the result in C_t.

2.4 *Else* if a R_t has any special characters ((: \ | [ ] ; {} - + ( ) < > ? ! % *,)

2.4.1 Remove using regular expression methods (4.1) from the tweet Store result in C_t.

2.5 Else if a R_t has any @

2.5.1. Remove the word '@,' *using* regular expression *(4.4)* and store the result in C_t

2.6 Else if a R_t has stop words

2.6.1 Remove any stop words using NLTK

2.7 Else if

2.7.1. Remove any extra white spaces using strip (), and store the result in cleaned text

End if

End For

The results of procedure 4.1 are displayed in Table 4.3. The process was demonstrated using a sample tweet. For this procedure, a sample of the raw text is: The real name of my degree: https://t.co/gEYBY

Table 4.3: A sample of output data

Input	Action	Output
The real name of my degree: https://t.co/gEYBY	Tweet Lowercasing	the real name of my degree: https://t.co/geyby
the real name of my degree: https://t.co/geyby	Removal of URL	the real name of my degree:
the real name of my degree:	Removal of alphanumerical characters	the real name of my degree
the real name of my degree	Stop words removal	real name degree
real name degree	Tokenization	'real', 'name', 'degree'
'real', 'name', 'degree'	lemmatization	'real', 'name', 'degree'

4.4 Tokenization

Special characters like apostrophes and commas are frequently removed during the tokenization process. This separates the text data into parts also known as or tokens. The tokens are usually made up of a single word or an N-gram, which is a token made up of N consecutive words (Purnamasari & Suwardi, 2018). The corpus is split into words/terms, and a bag-of-words word vector. It significantly affects how well future analyses turn out, thus it must be accurate and efficient (Denny & Spirling, 2018). The tweets were split into tokens using NLTK in python. Tokenization was important in this study to get the data ready for vectorisation. Since computer systems/classifiers work well with numbers, the embedding methods used in this study involved assigning each token a numerical value. Procedure 4.2 depicts the tokenization procedure employed in this study. The input was C_t and the output is the Tokens tweet referred as T_t in this study**.**

Procedure 4.2: Tokenization

Input: C_t
Output: T_t

---

1. Initialize an empty string of T_t to store the result of output/tokens.

2. For each C_t

   2.1 Split the C_t into tokens using NLTK library

   2.2 Remove white spaces in between tokens

   2.3 Return and store split C_t (tokens) in T_t

   End For

Clean texts were divided into specific tokens which are the atomic pieces of information. Table 4.4 depicts the tokens produced in this study.

Table 4.4: The tokens produced for the sample input

Input	**Output**
real name degree	'real', 'name', 'degree'

4.5 Lemmatization

After tokenizing the clean tweets, the next step was to lemmatize the tokens. Lemmatization was performed in this study to determine the root of any word in order to understand a phrase, clause, sentence, or other pieces of content. Lemmatization was done through NLTK in python. Lemmatization was necessary to reduce the margin of computational error in the analysis of the text. Procedure 4.3 depicts the lemmatization procedure of this study.

After tokenizing the clean tweets, the next step was to lemmatize the Tokens tweet T_t. The input was T_t and the output is the pre-processed data referred to as P_t in this study**.**

Procedure 4.3: Lemmatization

Input: T_t
Output: P_t

1. Initialize an empty string of P_t to store the result of output
2. For each T_t

   2.1 Perform lemmatization on Tokens using NLTK library

   2.2 Remove any white spaces in between lemmas

   2.3 Return and Store lemma in P_t

End For

Table 4.5 depicts a sample output of the lemmatization procedure of this study.

Table 4.5: Lemmatization input and output

Input	Output
['was', 'close', 'finishing', 'full', 'tube', 'chapstick', 'without', 'losing']	['was', 'close', 'finish', 'full', 'tube', 'chapstick', 'without', 'lose']

4. 6 Summary of the chapter

Data pre-processing involves applying cleaning techniques on raw data to get it ready for analysis (Kadhim, 2018). No matter how effective the trained model is, the root of the issue may still be the data itself because machine learning models are only as good as their data (Tae, Roh, Oh, Kim, & Whang, 2019). If the data is noisy or contains redundant and unnecessary information, knowledge discovery may produce inaccurate findings. As a result, not all pre-processing methods are appropriate for all text classification issues because some may affect the classification outcomes. The method used in this study is displayed in Figure 4.1.

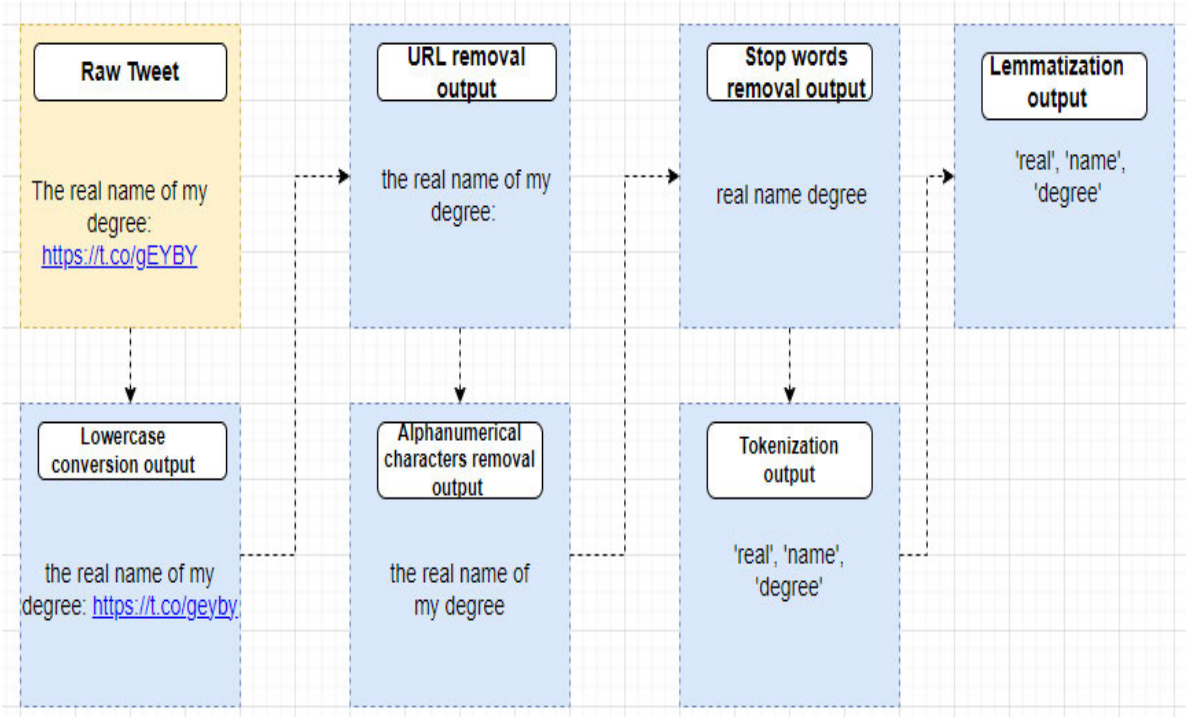

Figure 4.1: The rule-based techniques' output of this study

The activities involved in the first phase of the process model, namely, data cleaning, tokenization and lemmatization are presented and discussed in detail in this chapter. In the quest to get the dataset ready for labelling, a rule-based approach was selected and implemented. In this chapter the researcher discussed various data pre-processing techniques, emphasizing the rationale behind each technique selection. The textual data was converted to lowercase words, any special characters were removed, and then the data was tokenized and lemmatized. The output of this chapter serves as the starting point for chapter 5, in which the researcher explains the procedures carried out with regard to labelling the pre-processed data.

CHAPTER FIVE: ACADEMIC DATA LABELLING

5.1 Introduction

This chapter addresses research objective 2: that is, "to establish how the pre-processed data can be labelled". Labelling data is a vital yet time-consuming process in the development of supervised machine learning artefacts (Fredriksson, Mattos, Bosch, & Olsson, 2020). Due to the nature of data, every labelling activity has a range of difficulties (Desmond, Duesterwald, Brimijoin, Brachman, & Pan, 2021). Despite recent labelling paradigm modifications and major algorithmic breakthroughs, human-in-the-loop data labelling remains the most reliable technique of getting labelled data (Desmond, Muller, Ashktorab, Dugan, Duesterwald, Brimijoin & Pan, 2021). Many real-world AI systems, such as conversational agents, need that labellers deal with enormous label sets comprising tens, hundreds, or even thousands of labels. With more labels, the labeller's decision-making becomes more challenging, and the operations of labelling become more tedious and time-consuming (Desmond *et al*., 2021). The use of human and Artificial Intelligence collaboration is a prevalent paradigm for assisting and optimizing human decision-making processes (Cai, Winter, Steiner, Wilcox, & Terry, 2019; Wang, Weisz, Muller, Ram, Geyer, Dugan, & Gray, 2019; Wolf & Blomberg, 2019). Crowdsourcing, semi-supervised learning, and active learning are examples of well-known labelling techniques (Fredriksson *et al.,* 2020).

The aim of this study was to find the suitability of academics from textual data of the users. In this chapter, the researcher discusses in detail the data labelling stage of the process model presented in chapter three. The collected data needed to be labelled after being pre-processed. Thus, to label the collected dataset from Twitter, this study used different techniques, namely: binary labelling and multi class labelling. To establish the relationship between academics and sentiment in the virtual environment, lexicons relevant to academics were chosen for the dictionary based on the human resources, education, and psychology literature discussed in Chapter 2. The techniques used are further discussed in sections 5.6.1 and 5.6.2.

5.2 Sentiment analysis

Sentiment analysis is a method of determining a customer's likes, dislikes, opinions, or feedback on a piece of material, which can be classified as neutral, negative or positive (Budiharto & Meiliana, 2018). Hasan, Moin, Karim, and Shamshirband (2018) describe sentiment analysis, also referred to as opinion mining, as a study area that looks at people's attitudes, perceptions, assessments, and emotional responses toward things like goods, systems, institutions, people, issues, events, and themes, as well as their attributes. The term sentiment analysis was first introduced by Nasukawa & Yi, (2003). At several degrees of granularity, sentiment analysis has been considered as a Natural Language processing function (Hasan *et al.,* 2018).

Social networking sites like Twitter, Instagram, and Facebook have attracted a lot of users recently (Hasan *et al*., 2018). Twitter has seen a prevalent user adoption and fast development in communication volume amongst numerous social media sites (Zimbra, Abbasi, Zeng, & Chen, 2018). The majority of people use social media to communicate their feelings, ideas, or views about things, locations, or people (Hasan *et al*., 2018). Twitter users can use status update texts, or tweets, to inform their followers of what they are saying, doing, or what is happening in the world (Wang, Niu, & Yu, 2019). According to Zimbra *et al.,* (2018), these interactions provide excellent chances to access and comprehend users' viewpoints on topics of interest, and they contain data that can be used to explain and predict business and social phenomena including product sales, stock returns, and election results. The sentiments expressed by users in their text communication is central to these analyses.

Wang *et al*., (2019) claim that due to the sheer vast amount of data available on Twitter, extracting people' emotion polarities conveyed in Twitter tweets has become a major research issue. It has long been the most popular social media platform for extracting sentiment-rich data (Kumar & Jaiswal, 2020). It is the qualitative and quantitative foundation for assessing sentiments because of its interconnectivity with a diverse user-base and full involvement from the users. Data gathering, feature selection, sentiment classification, and sentiment polarity detection are all part of the general sentiment analysis. Sentiment analysis has three different levels, with each level unit having its own polarity: phrase/sentence level, document level, and aspect/feature level (Mehanna & Mahmuddin, 2021). According to Alsaeedi & Khan, (2019), at document level, a document can only be categorized as "positive," "negative," or "neutral", at Phrase level, each sentence is categorised as "neutral," "positive," or "negative", whereas at

aspect level, sentences/documents can be categorised as "positive," "negative," or "non-partisan" based on some aspects of the sentences/archives. The latter is commonly referred to as "perspective-level assessment grouping", since each tweet has a sentiment polarity and was written in a specific context. For this study, the researcher opted for sentence level sentiment analysis.

After pre-processing the Twitter dataset as discussed in detail in chapter four, it was important to perform sentiment analysis for the multiclass labelling purpose. The sentiment analysis was regarded helpful in contextualising the lexicon detected in a tweet. Section 5.2.2 offers more information on this.

5.2.1 Sentiment analysis techniques and tools

The rising dimensionality, complexity, and fuzziness of user-generated Twitter data necessitate the development of new and enhanced sentiment analysis methods (Kumar & Jaiswal, 2020). Soft Computing (SC) is one such area of study that takes advantage of a combination of new computational methods that emulate consciousness and cognition in many key ways: they can learn quickly, generalize into realms where direct experience is lacking, and perform modelling from inputs to outputs using parallel processing architectures that mimic biological processes (Kumar & Jaiswal, 2020; Aggarwal, 2018). Soft computing approaches provide a non-trivial solution to situations that are inherently imprecise and unpredictable in nature (Kumar & Jaiswal, 2020). Soft computing's guiding idea is to exploit the use of tolerance for inaccuracy, vagueness, and partial truth in order to gain manageability, resilience, cheap solution cost, and a better relationship with reality (Balas & Fodor, 2013; Kumar & Jaiswal, 2020).

Machine learning, lexicon-based techniques, and hybrid approaches are commonly employed in polarity determination (Ravi and Ravi, 2015, Medhat, Hassan, & Korashy, 2014; Hasan *et al*., 2018). Dictionary and corpus-based approaches are the two basic types of lexicon-based polarity determination methods. In the former, a dictionary of terms is utilized to categorise a text based on its polarity score (Borg & Boldt, 2020). The probability of sentiment words combined with a positive or negative set of terms is used in a corpus-based technique by searching through vast amounts of text, such as Google search results (Ravi & Ravi, 2015). This might be explained by starting with a set of words with opinions and then searching a huge corpus for related opinion words, culminating in a context-specific word set (Borg & Boldt, 2020). Bahrainian and Dengel (2013) proposed a hybrid method for polarity recognition of subjective texts in the consumer-products area that combines sentiment lexicons with a

machine learning classifier. Many methods, such as Linguistic Inquiry and Word Count (LIWC), now allow complex features to be extracted from texts. However, the majority of these tools necessitate some programming expertise. The Valence Aware Dictionary and sEntiment Reasoner (VADER) was utilised in a study by Elbagir and Yang (2019) to assess the sentiment of tweets and categorise them using multiclass sentiment analysis. VADER sentiment is a sentiment analysis framework that uses a rule-based and lexicon-based approach, as well as support for intensity estimate. Furthermore, when compared to seven sentiment analysis lexicons, VADER sentiment performed better or as well (Hutto & Gilbert, 2014).

TextBlob is a purely Python-based natural language processing package. TextBlob is utilized to determine the data polarity and subjectivity scores (Ahuja & Dubey, 2017). It provides a straightforward API for carrying out typical natural language processing (NLP) operations, including noun phrase extraction, part-of-speech tagging, sentiment analysis, classification, translation, and more (Loria, 2018; TextBlob, 2020; Manguri, Ramadhan & Amin, 2020). Gujjar & Kumar, (2021) used TextBlob to understand the emotion of an email thread. Ahmed, Rabin, and Chowdhury (2020) utilised TextBlob to extract sentiments from tweets about reopening. TextBlob is suitable for textual data such as Tweets as it can be used to measure the polarity and subjectivity at sentence level, therefore, this study used TextBlob to measure the polarity of each tweet at sentence level in the corpus.

5.2.2 Experiment set up: sentiment analysis

Finding each tweet's sentiment was the initial stage in the sentiment analysis process. Python's TextBlob was used for sentiment analysis. The pre-processed text presented in chapter 4 served as the input variable (P_t). TextBlob produces a sentence's polarity and subjectivity. Polarity is expressed as -1, 0 or 1, where -1 denotes a negative emotion, 0 denotes a neutral emotion and 1 denotes a positive/good emotion. In this study, sentiment polarity was either -1, 0 or 1. For the sentiment analysis experiment conducted, (P_t) refers to the pre-processed input variable text and (s_P) refers to the output variable sentiment polarity. The procedure for sentiment analysis is shown in procedure 5.1.

Procedure 5.1: sentiment analysis

Input: P_t

Output: s_P

1. Initialize temporary column s_P to store the s_P

2. For each P_t

    2.1 TextBlob computes s_P of each tweet at phrase level.

    2.2 if s_P> 0 then:

    2.3 return 1

        2.3.1 label it as positive

    2.4 Else if s_P= 0 then

    2.5 return 0

        2.5.1 label it as neutral

    2.6 else

    2.7 return -1

        2.7.1 label it as negative

    End if

End For

In this experiment, the polarity count as described in Table 5.1 and Figure 5.1 reveals that 116173 tweets had a positive (1) polarity, 138661 tweets had a neutral polarity (0) and 54152 tweets had negative polarity (-1).

Table 5.1: Polarity Count

Input	**Output**		
Pre-processed Tweet	**Polarity count of Positive (1)**	**Polarity count of Neutral (0)**	**Polarity count of Negative (-1)**
	116173	138661	54152

```
df['Sentiment'].value_counts()

 0    138661
 1    116173
-1     54152
Name: Sentiment, dtype: int64
```

Figure 5.1: Code procedure of value counts

Table 5.2 depicts a sample of data with negative polarity, neutral polarity and positive polarity. The tweets were the inputs and polarities, either -1. 0 or 1) are the output.

Table 5.2: A sample of texts with negative, neutral and positive polarity

Tweet	Polarity
[‘against’, ‘outsource’, ‘service’]	-1
[‘machine’, ‘ngeke’, ‘nglunge’,’mina’]	0
[‘best’, explanation’, ‘chinese’, ‘language’]	1

5.3 Academic lexicon development

Textual data is produced every day in every industry, including medicine, finance, journalism, politics, education, and corporate, to name a few. This data can be easily inspected, let alone processed, by a single person (Park, Kim, Lee, Choo, Diakopoulos, & Elmqvist, 2017). As a result, automated text analysis methods like probabilistic topic modeling, document summarization, and sentiment analysis are growing in popularity (Pang & Lee, 2008; Park *et al*., 2017). Most of these techniques center on textual concepts, which are described as a collection of semantically related words that describe a particular thing, phenomenon, or subject. Developing a lexicon for a certain concept typically takes a lot of time and effort; therefore, only a few concepts created by humans are available, each with a limited number of terms (Park *et al*., 2017).

Several attempts have been made before to develop lexicons that express emotions, intentions, opinions, and attitudes (Dang, Zhang, & Chen, 2009; Pang & Lee, 2008; Gitari, Zuping, Damien & Long, 2015). Corpus-based approaches and Dictionary are the two basic methodologies for producing opinion lexicons (Gitari *et al*., 2015; Huang, Xie, Rao, Feng, & Wang, 2020). In corpus-based techniques, words with opinion with preferred syntactic or co-occurrence patterns are found using a domain corpus. The semantic orientation of terms and

expressions to be incorporated in an opinion lexicon is determined using syntactic, structural, and sentence level elements and natural language processing rule-based procedures (Gitari *et al*., 2015). In academic study, text analysis using dictionary categories has a rich history (Fast, Chen, and Bernstein, 2016).

A bootstrapping process employing a limited collection of seed words with opinion and an online dictionary such as WordNet and SentiWordNet are employed in several of the suggested methods to construct a dictionary (Esuli, & Sebastiani, 2006). There are numerous dictionaries of opinion vocabulary composed primarily from adjectives, but also nouns, adverbs, and verbs (Taboada, Anthony, & Voll, 2006; Gitari *et al*., 2015). Fast *et al*., (2016) proposed Empath, a technique that uses cutting-edge word embedding to proficiently build a high - level semantic lexicon for a concept. Esuli, & Sebastiani, (2006) used a semi-supervised technique with WordNet word associations such as synonym, hyponymy, and antonym to autonomously build a lexical resource. However, according to Park et al., (2017), using a pre-built lexicon for document analysis without considering the document context and keyword usage trends could easily result in a content interpretation error.

As a result, Park et al., (2017), developed CONCEPTVECTOR1, a visual analytics system that merges a user-driven lexicon-building process with tailored document analysis in a very efficient and adaptable manner. Scholars have put in a lot of effort to manually construct lexicons with consistent semantics. The Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, & Booth, 2001) is an illustration of a manually constructed lexicon that defines a variety of concepts**.** Similarly, the General Inquirer research project develops lexica in a variety of ideas (Park *et al*., 2017). Although these manually constructed databases provide high-quality data for a variety of natural language comprehension and text analysis activities, the fundamental issue is the enormous human work required to generate and evaluate them.

In this study the researcher used an approach that consists of three phases, namely: developing a suitable academic dictionary, conducting sentiment analysis on the pre-processed data, and detecting traces of academic interests/words in the dataset followed by labelling them based on the sentiment score and/or the lexicon dictionary score.

5.3.1 Phase one: developing suitable academic dictionary

In the first phase, the task was to generate lexicon dictionaries suitable for academics. To this end, a profile of three categories of academics was constituted, namely: lecturers/teachers, researchers, and supervisors. This phase was carried out in three steps. The first step was to

identify the interpersonal skills of the three categories of the academics, and it involved identifying and analysing interpersonal skills requirements for academics from different educational institutions in South Africa, Europe and America. Universities' vacancy pages and some web portals were screened to gather the needed requirements. The second step was to identify the personal trait requirements of academics from psychology and education research (Carifio, & Hess, 1987; Barnes, Williams, & Archer, 2010; Azure, 2016; Chiresche, 2017; Davis, 2020). This stage involved identifying key published studies to profile a suitable academic. The third step was to search lexicon databases such as LIWC, Harvard Inquirer, and Stanford to identify any lexicon aligned with the suitability of the academic profile of this study.

Dictionary S which represents the supervisor category, Dictionary T which represents the Teacher or lecturer category, and Dictionary R which represents the researcher category are the dictionaries identified and used for this study. These dictionaries intend to describe their interpersonal skills, their traits and their academic jargons.

DictionaryS =['compassion','sympathy','insight','understanding','caring',
'care','commiseration','help','intuition','clearly','talk','happy','thi
nk','know','consider','cause','should','would','guess','philosopher','p
hilosophic','philosophical','evidence','address', 'amendment','approval
','argument', 'biography', 'coin', 'collaboration', 'criticism', 'aspir
ation','decision','fulfilment','innovation','inspirational', 'meritorio
us', 'proceed', 'reward', 'purposeful','examiner', 'reviewer','frank','
openhearted','outspoken','unreserved','straightforwardness', 'straightf
orward', 'accepting', 'supervise', 'supervision', 'evaluation', 'candid
ate', 'chancellor', 'intellect','intellectual','intelligence','junior',
'knowledge','doctor','adviser','advisor','advocate','education','educat
ional','yield','lead','persevere','prolong','tend','misrepresent','perp
lexity','get','phd','happen','swift','direct','willingness','eager','li
vely','coaching','optimistic','details','formal','expertise','trustwort
hy','accomplishments','facts','effective','logic','clear','analytical',
'accuracy','logical','predictable','cautious','challenging','quality','
stability','tactful','growth',]

DictionaryR =['collect','information','data','identify','critically','exa
mine','determine','ascertain','factual','outcomes','outcome','
revolution','partially','totally','action','prevent','perfectionist','p
erfection','analysing','analysis','scrutinizing','scrutinise','integrat
es','integrate','components','component','evaluating','evaluate','vario
us','options','option','science','project','research','instruction','sk
ills','master','lessons','theory','across','waters','harvest','museum',
'forest','hidden','ancient','experiment','colleague','laboratory','lab'
,'literalness','literary','literature','experimental','master','theory'
,'usual','furthermore','moreover','thus','ineffective','ineffectiveness
','ineffectual','ineffectualness','inefficiency','insufficiency','mista
ken','enhance','Sustain','recovery','pinnacle','output','obtain',]

DictionaryT =['accentuate','accompany','achievement','acquire','train','support','active','action','adapt','adopt','eagerness','enthusiastic','entrust','gratitude','opinionated','positivity','resolute','resolved','upbeat','vigilant','vigilance','wilful','teach','essay','exam','hear','feel','view','see','touch','listen','area','bend','‘explanation’, 'arrive','go','scope','upcoming','test','study','learnt','learn','facilitate','participation','cover','briefly','later','move','textbook','chapter','integrate','advance','collaborate','expand','manage','participate','continuity','desire','hardworking','student','course','credit','dean','degree','discourse','examine','faculty','form','grade','graduate','graduation','grammar','attendance','assignment','teach','bachelor','beginner','being','benefactor','classmate','learner','letter','major','mathematics','matriculate','instruct','instruction','instructor','encourage','kindness','character','constant','steadfast','steadiness','underway','continue','inevitable','keep','permanent','absentee','back','breakdown','default','drop','erroneous','fail','give','illiterate','illogical','inaccuracy','mind','feelings','expressive','recognition','appraiser','’evolution’,'friendly','social','counsellor','happiness','enthusiastic','courteous','cooperation','appease','casual','patient','sincere','security','open','warm','calm','trust','loyalty','consistent','informal','empathetic','appease',]

5.3.2 Phase two: academic lexicon detection

In phase two, the three dictionaries were combined into one dictionary for academics. The combined dictionary is referred in this study as Academic Dictionary (A_D). In this study the researcher opted to use one combined dictionary as using three different dictionaries was yielding different labels thus resulting in an extremely imbalanced dataset for classification. Thus, to minimise the extreme imbalance labelled dataset, combining the three dictionaries into one proved to be more effective. Procedure 5.2 depicts the academic lexicon detection of this study. The pre-processed text P_t was used as the input variable and Lexicon detection L_D is the output variable for this experiment.

Procedure 5.2: Lexicon detection

Input: P_t

Output: L_D

---

1. Define the A_D
2. Initialize a temporary column L_D to store labels
3. For each row in the dataset

    3.1 If lexicon from A_D is detected using regular expression

    3.2 label it True

    3.3 Else Label it as False

    End if

End For

As depicted in Figure 5.2, the lexicons which were not found in the tweets (labelled as 0 or False) account for 208632 tweets and the lexicons which were found in the tweets (labelled as 1 or true) account for 100354 tweets.

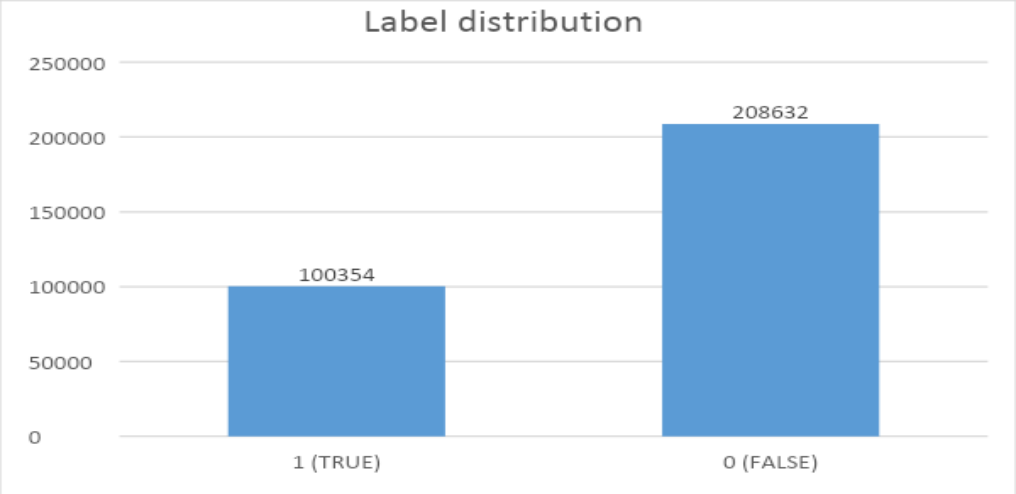

Figure 5.2: Academic Lexicon Label distribution

5.4 Phase three: labelling suitable academics

There are mainly three type of machine learning classifications, namely, binary, multiclass and multilabel classification (Alswaina & Elleithy, 2018; He, Yang, Gao, Liu, & Yin, 2019). This study opted for binary and multiclass classification as the multilabel classification was not part of the scope of this study as this study intends to sequentially classify academic suitability. Thus, the data of this study is labelled as binary labelling (section 5.4.1) and labelled as multiclass labelling (section 5.4.2).

5.4.1 Academic binary labelling

The first task consisted of finding traces of the combined dictionary A_D and labelling the text either 1 (Suitable) or 0 (Not Suitable). The aim was to have two labels for the binary classification presented and discussed in chapter 6. Table 5.3 depicts this binary labelling input and output. As depicted in Figure 5.2 and Table 5.4, label 1 (True) represents 100354 tweets and label 0 (False) represents 208632 tweets. In this study, the label 1 (True) stands for "suitable" and the label 0 (False) stands for "not suitable" for the binary classification experiment 6.1 and 6.2 in chapter six.

Table 5.3: Sample Academic binary labelling

Input (Tweet)	Output (Label)
['dagga', 'aka', 'marijuanas', 'khathu', 'weed']	0
['lil', 'wayne', 'evolution', 'tha', 'goat']	1

Table 5.4: Value count of Academic binary labelling

Label	Value count
0 (Not Suitable)	208632
1 (Suitable)	100354

5.4.2 Academic multiclass labelling

In this step, sentiment polarity s_P and Lexicon detection scores L_D were used to label the data. The use of sentiment analysis in labelling data in this approach is justified by the fact that

sentiment analysis provides context to the lexicon detected in the text. For instance, if user1 tweets "research is a nightmare" and user2 tweets "research is my love language". The word research which is found in the two examples does not have the same connotation, therefore, the polarity score from sentiment analysis can help in this case to label this tweet differently. To do so, a set of conditions shown in Table 5.6 were established.

The multiclass academic labelling is depicted in procedure 5.3. The input variables were the pre-processed text P_t , the sentiment polarity s_P the lexicon detection L_D. The output variable is Academic Label A_L

Procedure 5.3: Multi class Academic labelling

Input: $P_t , s_P, and\ L_D$

Output: Academic Label A_L

---

1. Initialize temporary column Label to store Academic Label A_L
2. **For** each P_t

2.1 If sentiment score s_P = 0 and Lexicon detection score L_D = 1 then label it as S

2.2 Else if sentiment score s_P =0 and Lexicon detection score L_D = 0 then label it as NS

2.3 Else if sentiment score s_P =1 and Lexicon detection score L_D = 1 then label it as VS

2.4 Else if sentiment score s_P =1 and Lexicon detection score L_D = 0 then label it as NS

2.5 Else if sentiment score s_P =-1 and Lexicon detection score L_D =1 then label it as MS

2.6 Else if sentiment score s_P =-1 and Lexicon detection score L_D = 0 then label it as NS

End if

End For

Table 5.5 provides a sample of inputs, namely: pre-processed data, sentiment polarity, Lexicon detection score, and the output academic label as well as the description of the output variables.

As depicted in Figure 5.3 and Table 5.6, the value count in the multiclass labelling, the label 0 (Not Suitable) has most of the counts followed by the label 2 (Very Suitable) then the label 1 (Suitable) and the label 3 (Moderately Suitable).

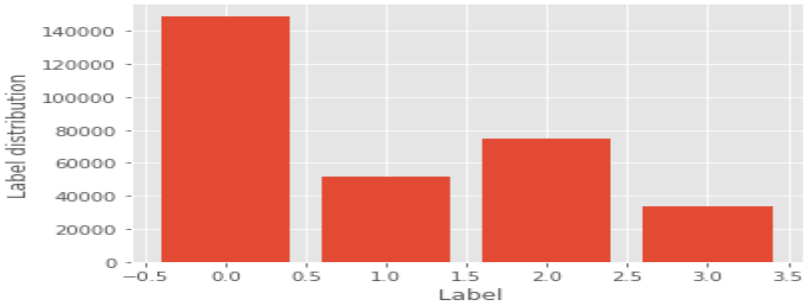

Figure 5.3: Four labels' distribution

Table 5.5: A sample of labelling conditions for academic multiclass labelling

Input variables			**Output variable**	
Pre-processed text P_t	**Sentiment polarity** s_P	**Lexicon detection score** L_D	**Label** A_L	**Description of the output variables**
['take', 'back', 'day', 'revolution', 'gave', 'banger']	0 [neutral]	True [detected]	1	1= Suitable [**S**]
['money', 'smell', 'bush']	0 [neutral]	False [not detected]	0	0 = Not Suitable [**NS**]
['orlando', 'pirates', 'one', 'teams', 'psl', 'simply', 'get', 'certificate', 'participation', 'win']	1 [positive]	True [detected]	2	2 = Very Suitable [**VS**]
['street', 'safe']	1 [positive]	False [not detected]	0	0 = Not Suitable [**NS**]
['broke', 'girl', 'think', 'lady', 'spend', 'money', 'boyfriend', 'stupid']	-1 [negative]	True [detected]	3	3 = Moderately Suitable [**MS**]
['assholes', 'drivers', 'guy', 'green', 'car']	-1 [negative]	False [not detected]	0	0 = Not Suitable [**NS**]

Table 5.6: Value count of academic multiclass labelling

Label	**Value counts**
0 (Not Suitable)	149120
1 (Suitable)	51464
2 (Very Suitable)	74440
3 (Moderately Suitable)	33962

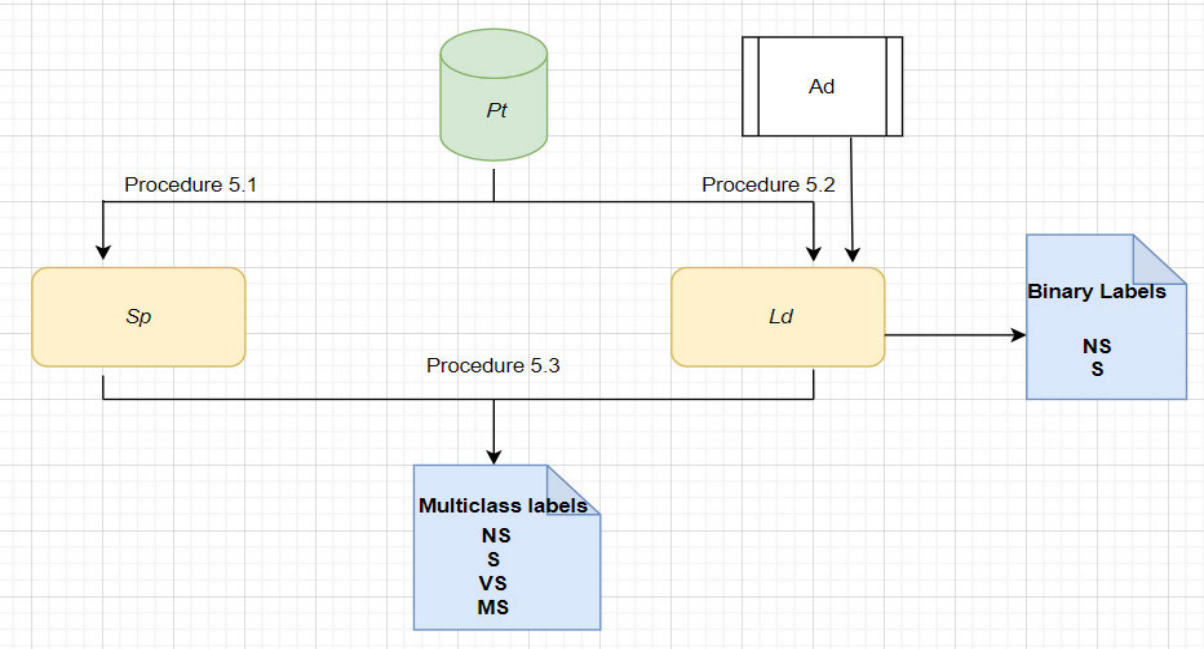

Figure 5.4: Data labelling approach

Figure 5.4 summarizes how the process of labelling data was conducted. The pre-processed text is the input variable for sentiment analysis and lexicon detection meanwhile academic dictionary is also an input variable for lexicon detection process. The input variables are processed through some rules provided as procedure 5.1, 5.2. and 5.3. These procedures are used to fit the suitability of academic requirements. The outputs are binary labels and multi class labels.

5.5 Summary of the chapter

People's thoughts, attitudes, and emotions toward services, products, people, and topics are studied using sentiment analysis. This chapter discussed the labelling processes undertaken in this study. The lexicon construction, lexicon detection, sentiment analysis, and the labelling approaches were discussed in this chapter. For its performance and ability to handle sentence level analysis of twitter data, the Textblob sentiment analysis tool was preferred compared to VADER, and LIWC for sentiment analysis. Thus, this chapter presented the sentiment analysis techniques, tools and experiments used and performed in this study. Furthermore, a dictionary (lexicon) constructed based on the literature review as part of the labelling process was also presented.

The output of this chapter in the process model is the academic labelling step that served as input in chapter six. Furthermore, in chapter six academic suitability classification using the deep learning model is discussed.

CHAPTER SIX: ACADEMIC SUITABILITY CLASSIFICATION USING DEEP LEARNING

6.1 Introduction

This chapter addresses the research objective 3, namely: To establish how the labelled data can be used for a predictive model.

A classification model helps to obtain meaning from training data. Training and testing are the two stages of classification. In the training phase, classifiers are fed training data to acquire understanding about data patterns. The testing phase will then employ the trained classifier to forecast the test data. This study used supervised deep learning classifiers to classy suitability. The classifiers were selected for this study due to their suitability for classifying texts. To build the academic suitability models, this study used the pre-processed and labelled data described in chapters four and five. In this chapter, the classification stage of the process model discussed in section 3.5 is implemented.

Furthermore, this chapter presents the effective classification models developed after conducting extensive experimentation with the Twitter dataset. The approaches, tools and techniques used to classify people as suitable or not suitable based on their tweets are covered in this chapter.

6.2 Classification

As mentioned before, there are basically three types of classification, namely: binary classification, multiclass classification and multi label classification. Binary classification is the technique of predicting variables, in which the output is limited to two. The objective of *binary classification* is to divide a set of elements into two categories using a classification rule (Vyas & Uma 2018; Jiang, 2019). The main aim of this study which was to forecast academic suitability prompted the researcher to opt for binary classification to classify job applicants as suitable or not suitable academics based on the Twitter footprint.

Multiclass classification is the task of categorising instances into one of three or more classes. The former presupposes that there are more than two classes to choose from, but the text must be allocated to just one (Haralabopoulos, Anagnostopoulos, & McAuley, 2020; Imane & Mohamed, 2017). Despite the relatively small size of the dataset utilised in this study, the researcher also wanted to classify different levels of academic suitability, hence the choice of

multiclass classification in this study. The dataset used for this experiment is from Twitter as described in Chapters 3, 4, and 5. The dataset was pre-processed and then labelled as described in Chapters 4 and 5. The dataset was labelled in two different ways as described in sections 5.6.1; and 5.6.2.

Multi-label classification is utilised when there are two or more classes, and the classified data could belong to all of the classes simultaneously or none of them. Data sets with more than one target variable can be classified using multi-label classification (Al-Salemi, Ayob, & Noah, 2018). The multi label classification was not selected in this study as it is not suitable for the research problem of this study.

6.3 Tools

Libraries and software packages that would aid in addressing research objectives were chosen for this study. Exploring potential tools to see how they might aid or hinder the experiments of this study was an important task. Python 3.7 as well as Python 3.6.9 in Google Colab were used in this study. The libraries utilized for data analysis included NumPy, Panda, and Keras. Moreover, Keras as a python platform was chosen to build the suitability of academic classification models due to its extensive use in deep learning and performant data analytics ecosystems. Keras was chosen because it provides user-friendly and performant deep learning libraries that allowed this study to achieve satisfactory results. Keras is a Python interface for artificial neural networks that is open-source software. It serves as a front-end for the TensorFlow library, a library of highly powerful and abstract building elements for developing deep neural networks (Ketkar, 2017). Keras is a valuable tool for fast prototyping ideas because it supports both CPU and GPU processing (Ketkar, 2017; Manaswi, 2018). In addition, Manaswi (2018) describe Keras as a high-level Python library for deep learning that runs on top of TensorFlow and is compact and easy to learn. It allows developers to concentrate on the core principles of deep learning, such as layer creation for neural networks, while the nitty-gritty details of tensors, their forms, and their mathematical subtleties are taken care of. The sequential API and the functional API are the two most common types of frameworks. The sequential API is built on the concept of a layer sequence; it is the most typical application of Keras and the most straightforward portion of the framework. A linear stack of layers can be thought of as the sequential model.

6.4 The Data split

Critical steps in the machine learning process include training, testing, and validation. Training is the process of picking a batch of input data and feeding it to a classifier to learn and detect patterns (Pirina & Cöltekin, 2018; Lee, Gui, Manquen, & Hamilton, 2019). The model optimizes its weights throughout training to improve its output. The training data for a model must include an example of an entity to classify as well as the accurate categorization of the entity (Lee *et al*., 2019). The initial stage in developing a classification model using basic deep learning techniques is to appropriately prepare the training data, followed by setting the parameters of the deep learning classifiers (Pirina & Cöltekin, 2018; Khalaf, Hussain, Alafandi, Al-Jumeily, Alloghani, Alsaadi, & Abd, 2019; Lee *et al*., 2019). Each trained model is validated using the validation set after being built using the training set and various model parameter settings. Predictions on the validation set allow the developer to evaluate model accuracy because the validation set consists of samples with known provenance, but these parameters are unknown to the artefact (Xu & Goodacre, 2018). Based on the Pareto principle which states that 80 percent of the output of a given scenario or system is determined by 20 percent of the input, in this study, the dataset was randomly divided into 80-20 basis. The Pareto principle was adopted in this study due to its widespread use in the machine and deep learning community. It was assumed that 20% of testing set (input) can determine the 80% of the outputs of the trained model. The dataset was split into the following categories: "X_train" representing 80% of the dataset, "X_test" representing 10 % of the testing set and "Val_set" representing 10 % the testing set. The validation set were used to fine-tune the classification models' parameters to reach the best result using the Twitter dataset of this study. Furthermore, for this study, the researcher also divided the dataset into training and testing categories on a 70-30 basis. However, the difference in results was not significant.

6.5 Word embedding/feature vector

Data is represented by a set of characteristics that can be binary, categorical, or continuous. Characteristics are interchangeable with input variables or attributes. It's critical to find a suitable data representation that can be used to accurately measure features. The difficulty of identifying the most compact and informative feature set is solved via feature extraction (Wang, Su, & Yu, 2020). Building feature vectors remains the most frequent and convenient technique of data representation for classification and regression issues (Sammons, Christodoulopoulos, Kordjamshidi, Khashabi, Srikumar, & Roth, 2016). Transforming textual

data into meaningful vectors is a mean to connect with machines so that they may do Natural Language Processing activities and solve mathematical problems (Singh & Shashi, 2019). Researchers in the field have developed a variety of vectorization models, ranging from the most basic to the most complex, to aid in the solution of NLP problems. For this study, word embedding techniques such as Keras Tokenizer for multiclass classification, FastText, and Glove for binary classification were adopted.

Word embedding converts words into a vector space (Köhn, 2015; Nelson, 2021). In semantic vector space models of language, each word is represented by a real-valued vector (Pennington et al., 2014; Egger, 2022). These vectors can be utilized as features in a range of applications, including information extraction (Manning *et al*., 2008; Egger, 2022), text classification (Sebastiani, 2002; Stein, Jaques, & Valiati, 2019), question answering (Tellex, Katz, Lin, Fernandes, & Marton, 2003; Esposito, Damiano, Minutolo, De Pietro, & Fujita, 2020), named entity recognition (Turian et al., 2010; Wu, Xu, Jiang, Zhang, & Xu, 2015) and machine translation (Botha and Blunsom, 2014) and processing (Belinkov, Lei, Barzilay, & Globerson, 2014; Socher, Perelygin, Wu, Chuang, Manning, Ng, & Potts, 2013).

In neural network artefacts for NLP tasks such as sequence tagging (Ma and Hovy, 2016; Lample et al., 2016) and text classification (Kim, 2014), pre-trained word embedding have shown to be quite effective (Qi, Sachan, Felix, Padmanabhan, & Neubig, 2018). Several pre-trained word embedding models have been made public in recent years (Pennington et al., 2014; Chiu, Crichton, Korhonen, & Pyysalo, 2016; Zhang, Chen, Yang, Lin, & Lu, 2019). "Word2Vec" (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov et al., 2013), and "Global Vectors" (GloVe), and FastText (Athiwaratkun, Wilson, & Anandkumar, 2018) are among the popular deep learning methods for word embedding (Pennington, Socher, & Manning, 2014).

The Facebook AI Research (FAIR) lab developed the machine learning library FastText for efficient word representation learning and text classification. FastText's algorithm is based on two papers: "Enriching word vectors with sub word information" (Bojanowski et al., 2017) and "a bag of tricks for efficacious text classification" (Joulin, Grave, Bojanowski, & Mikolov, 2016). FastText is a word2vec model extended version that symbolizes each word as an n-gram of characters rather than learning word vectors explicitly (Mestry, Singh, Chauhan, Bisht, & Tiwari, 2019). FastText implies that a word is made up of n-grams of characters, with n ranging from one to the length of the word (Kuyumcu, Aksakalli, & Delil, 2019). The advantage of

FastText is that it can identify vector representations for texts that are not explicitly available in the dictionary because FastText preserves the word vectors as n-grams of characters. Furthermore, FastText provides the benefit of being able to comprehend suffixes, prefixes, and shorter words more quickly and effectively (Malik, Aggrawal, & Vishwakarma, 2021), FastText is described in the equation 6.1:

$$-\frac{1}{\mathrm{N}} \sum_{n=1}^{N} y_n \log\left(f(B\ Ax_n\ )\right) \qquad \text{Equation (6.1)}$$

Where x_n is the n-th document's normalized bag of features, y_n is the label of n-th document, N is the number of documents, *f* is the SoftMax function and A and B are the weight matrices. The probability distribution across predefined classes are computed with the SoftMax function *f* as illustrated in the equation 6.1. FastText as a word embedding method was used in this study as one of the word representation approaches for binary classification of this study.

The Glove was developed by Pennington, Socher, & Manning, (2014), and it mixes information about global statistics with information about local context. The co-occurrence probability matrix is presented as the main feature. Glove is described in the equation 6.2:

$$F\left(w_i, w_{j,} \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}} \qquad \text{Equation (6.2)}$$

Where P_{ik} is the number of times the words of order *i* and *k* occur together in single window. The GloVe model's main principle is to concentrate on the co-occurence probabilities of words within a corpus of texts in order to embed those words in useful vectors. The purpose of the GloVe model is to construct a function *F* that can forecast the co-occurrence of probabilities ratios given two word vectors w_i and $w_{j,}$ and a context word vector $\tilde{w}_k$ as inputs, provided that the ratio *Pik/Pjk* is dependent on three words, *i*, *j*, and *k* as depicted in equation 6.2 (Pennington, Socher, & Manning, 2014)

The Glove as a word embedding method was used in this study to vectorise the pre-processed and labelled data. The GloVe was used for the binary classification. There are three steps in the Glove word embedding process that produce word vectors for each document's individual word (Mohammed, Jacksi, & Zeebaree, 2020). First, a vocabulary function stores and computes each recurrent word (token) without repeating it in the initial step. Then the word-

to-word co-occurrence approach constructs a matrix to assess the frequency of each pair of words occurring, ascribe weight to it, and store it in a matrix, with the matrix dimension equating the number of unique words appearing in the dataset. In the third step, the glove uses the matrix of word co-occurrences to generate a word vector for each word that contains the frequency weights.

Keras' Tokenizer class is used to vectorize a text corpus. Each text input is transformed into an integer sequence or a vector with a coefficient for each token in the form of binary values for this purpose. Keras Tokenizer has four methods, namely, fit_on_texts, texts_to_sequences, texts_to_matrix and sequences_to_matrix, (Tensorflow.org; Kathuria, Gautam, Singh, Khatri, & Yadav, 2019; Vinayakumar, Alazab, Jolfaei, Soman, & Poornachandran, 2019).

The Keras tokenizer class includes the fit on texts method, which is used to update the internal lexicon for the texts list. The texts_to_sequences method convert text corpus tokens into an integer sequence. The texts_to_matrix method of the tokenizer class is useful for transforming a document into a NumPy matrix. The sequences are converted into NumPy matrix form using the Keras tokenizer class function sequences to matrix (Kathuria *et al.*, 2019; Vinayakumar et al., 2019). This study used the Keras tokenizer class as a word embedding method for multiclass classification. The choice of Keras tokenizer was prompted by the results it yielded. Glove, and FastText were used for binary classification.

6.6 Deep Learning functions

In this section, a presentation of some of the deep learning functions, namely, Sigmoid, ReLU, Leaky ReLU, and Softmax activation functions, Loss function cross-entropy is provided. In addition, the section specifies which parameters were used in this study and provides the reason for selecting such parameters. In the training of neural networks, the activation function is crucial as it gives the model the essential nonlinearity to learn complicated representations. Also, the activation function determines whether to activate a neuron by computing the weighted sum and then adding bias to it.

A sigmoid function consists of the following equation 6.3:

$$S(x) = \frac{1}{(1 + e^{-x})}$$ Equation (6.3)

Where *e* is the Euler's number, *x* is a real valued number and S(x) equals the Sigmoid function. All real numbers fall within the domain of sigmoid functions, and the return value is typically monotonically increasing but may also be decreasing. Sigmoid functions typically display a return value (y axis) between 0 and 1. A second typical range is from -1 to 1. Sigmoid activation is implemented in Keras as Keras. activations. sigmoid(x). For the binary classification in this study, the sigmoid activation function was employed. As shown in the equation 6.3, Sigmoid was used to predict the probability of the two classes of the binary classification, namely: NS (0) and S (1).

The ReLU employs the equation 6.4:

$$f(x) = \max(0,x) \qquad \text{Equation (6.4)}$$

Where f is the function, x is any input value, and max is the maximum. When the function is given a negative value as input, it returns 0 as an output, but when it is given any positive value x, it returns that value as an output. The output therefore has a range from 0 to infinity. ReLU is a genuine nonlinear function that also happens to be an excellent activation function. It not only enhances performance but also helps to reduce the number of computations during the training phase. This is due to the output having a 0 value when x is negative, which deactivates the neuron (Moolayil, Moolayil, & John, 2019). The ReLU activation function was utilised for both the binary classification models and multi class classification models. ReLU enabled the designed models of this study to take into account non-linearity and interactions. In addition, it assisted the models in accounting for interaction effects.

The Softmax activation function is used to calculate the relative probabilities. This signifies that the ultimate probability value is calculated using the values of Z21, Z22, and Z23. The SoftMax function returns the probability of each class, much like the sigmoid activation function (Cao, Su, Yu, Chang, Li, & Ma, 2018). The SoftMax activation function has the following equation 6.5:

$$Softmax\ (z)_i = \frac{exp(z_i)}{\Sigma_{j=1}^{k} \quad exp(z_j)} \qquad \text{Equation (6.5)}$$

Where Z represents the input vector, $exp(z_i)$ represents the standard exponential function for the input vector. The exponential acts as the non-linear function. Also, $exp(z_j)$ represents the standard exponential function for output vector, whereas k represents the number of classes in the multi-class classifier. The Softmax unit *j* produces the probability that the input fits into class *j* (Kagalkar & Raghuram, 2020). The SoftMax activation function was utilised for the multi class models of this study. SoftMax was used in the output layer to calculate the relative probability of each class [NS (0), 1 (S), 2(VS), 3(MS)] of the Twitter dataset used in the multiclass classification.

Loss functions are used to calculate the amount that a model should try to minimise during training. Cross-entropy, which measures the gap between the model's real and projected suitability category *y*, was introduced as an evaluation metric. Generating strong predictions on the training data equates having a minimal loss (Ruby & Yendapalli, 2020). The cross-entropy loss function is illustrated in the following equation 6.6:

$$Loss = -\sum_i y_i \log \hat{y}_i \qquad \text{Equation (6.6)}$$

Cross-entropy was used for the binary classification to minimise the loss of training features of the dataset during the training phase. During the training phase, the classifiers usually lose training features.

Where y_i is the true label taking a value 0 or 1 and $\hat{y}_i$ is the SoftMax probability for the i^{th} class. Loss function cross-entropy was used in this study during the training to minimise the loss of training features.

Leaky ReLU is an improved version of ReLU. A new activation function was proposed to bypass some of ReLU's shortcomings: Leaky ReLU produces a slightly slanted line instead of a horizontal line when the negative value is negative, which aids in successfully updating the weights via backpropagation (Moolayil *et al*., 2019). The following equation 6.7 is used to calculate Leaky ReLU:

$$f(x) = \max(0.01*x, x) \qquad \text{Equation (6.7)}$$

Where *f* is the function, *x* is any input value, and *max* stands for maximum. For a negative input(x), we specify the ReLU activation function as a very small linear component of x instead

of defining it as 0. If the input is positive, this function returns *x*; however, if the input is negative, it returns a very little value equal to 0.01 times *x*.

6.7 Optimization techniques

In this section, the researcher presents and discusses the Adam optimizer and the Root Mean squared propagation.

Adam optimizer is a deep neural network training algorithm that uses an adaptive learning rate optimization mechanism. Adam optimizer is a deep learning training algorithm that replaces stochastic gradient descent. Adam optimizer mixes the finest features of the AdamGrad and RMSProp methods to create an optimization technique for noisy issues with sparse gradients (Zhang, 2018; Ali, Sarowar, Rahman, Chaki, Dey, & Tavares, 2019). To identify distinct learning rates for each parameter during the training phase of binary classification models, Adam optimizer uses the adaptive learning rate approach. The Adam optimizer was used for the binary classification models of this study, namely, BiLSTM.G.2 and BiLSTM.F.2.

Root Mean Squared Propagation (RMSProp), is a gradient descent extension and the AdamGrad version of gradient descent that utilizes a dying average of partial gradients to adapt the step size for each parameter. The RMSProp optimizer limits oscillations in the vertical position. As a result, it can boost the learning rate, allowing the algorithm to take greater horizontal steps and converge faster (Xu, Zhang, Zhang, & Mandic, 2021; Yu, Zhang, Chen, & Qin, 2021; Zou et al., 2019). Based on its performance with the Twitter dataset used in this study, the RMSprop was used for the multiclass classification to boost the learning rate during the training phase.

6.8 Deep Learning (DL) layers

In hierarchical network topology, a layer is described as a combination of neurons or a conceptually isolated group. As DL grew in popularity, numerous network architectural experiments were done to increase performance for a range of use case scenarios. Regular supervised techniques such as classification, regression, computer vision, natural language processing, speech recognition, and a combination of multiple domains are some of the use case scenarios. Keras gives us different sorts of layers, including Dense, Dropout layers, as well as various ways to connect them, to make the model creation process easier. The performance of a model also depends on complexity of the number of layers used. Thus, for

optimal performance with the dataset used in this study, three layers were used. The discussion bellow focuses on the Keras layers used in this study.

6.8.1 Dense layer

A dense layer is a standard Deep Neural Network layer that joins every neuron in one layer to every neuron in the one before it (Javid, Das, Skoglund, & Chatterjee, 2021; Onan, 2021). All the artefacts implemented in this study used dense layers to connect neurons in one layer to every neuron in the previous one. For the ANN.4 implemented model, the dense layer was utilised to connect the flatten layer and the output layer. For the BiLSTM.G.2 and BiLSTM.F.2 implemented model, the dense layer was used to connect the BiLSTM layer and dropout layer.

6.8.2 Embedding layer learning

Embedding layer learning is accomplished through word embedding learning. The Embedding Layer (EML) in the Keras framework enables the mapping of texts into vectors. When given an integer as input, the EML searches its internal dictionary to produce the corresponding dense vector. Its starting weight is chosen at random. For the binary classification artefacts of this study, the FastText and Glove were used to vector the tweets/ tokens first, then the vectors were used during the embedding layer. The embedding layers were used in the binary classification to connect the input layer and the bidirectional LSTM layer. Regarding the study's multiclass classification artefacts, the Keras tokenizer were used to vectorise the tweets/tokens, then the vectors were process at the embedding layer of the models. The embedding layer was also employed to connect the input layer with the LSTM layer for the multi class implemented artefacts.

6.8.3 Dropout layer

By introducing regularization and generalization capabilities into the model, the dropout layer in Deep Learning aids in reducing overfitting. The dropout layer minimizes computation in the training process by dropping out a few neurons or setting them to 0. Since it can capture more randomness, the dropout can decrease over-fitting. The basic concept is to add noise into the network layer's output value, shattering the inconspicuous inadvertent mode (Zhang, Pan, Sun, & Tang, 2018; Wu, Li, Wang, Meng, Qin, Chen, & Liu, 2021). Dropout layers were used in this study, specifically for the two binary classification artefacts to reduce the overfitting of the models.

6.9 Cross-validation

In this section, the researcher discusses the early stopping technique used in this study as a cross-validation strategy. The early stopping technique is a type of cross-validation strategy. The concept is that whenever the model's performance on the validation set deteriorates, we stop training it. Callback is the function used in Keras to end the training. The callbacks for each iteration contain records of the validation loss value, training loss value, validation accuracy rate, and training accuracy rate. Callbacks have two significant parameters: "monitor" and "patience", using "monitor" to track modifications to the output value and patience being used to supervise the improvement of validation loss. Once the validation loss is not improving, patience stops the training (Ashqar, & Abu-Naser, 2019). This study used early stopping for both the binary classification models and multiclass classification artefacts. The early stopping as a cross-validation is a strategy used in this study to ensure reliability and validity of the artefacts developed.

6.10 Performance metrics

In order to appreciate the classification of suitable academics, and considering the highly unbalanced distribution of labels, the following performance metrics are used: accuracy, precision, recall, and F1. The accuracy of a model is a measure that describes how well it performs throughout all classes (Richens, Lee, & Johri, 2020; Bikmukhametov, & Jäschke, 2020), and it is expressed in the following equation 6.8:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Equation (6.8)}$$

The ratio of accurately categorized positive samples (True Positive) to the total number of categorized positive samples, categorised either correctly or incorrectly, is known as Precision (Juba, & Le, 2019; Berger & Guda, 2020), and it is expressed in the following equation 6.9:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Equation (6.9)}$$

Furthermore, recall is measured as the proportion of positive samples that were accurately classified as positive to the total number of positive samples (Juba, & Le, 2019; Berger & Guda, 2020). Recall is expressed using the following equation 6.10:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Equation (6.10)}$$

The harmonic mean of precision and recall (F1) provides a better indication of detection performance for unbalanced datasets. F1 gets its best value at 1 and worst at 0 (Al-Qatf, Lasheng, Al-Habib, & Al-Sabahi, 2018), and it is expressed in the following equation 6.11:

$$\text{F1} = 2 \text{ x} \frac{Precision \; x \; Recall}{Precision+Recall} \qquad \text{Equation (6.11)}$$

“True positives” (TP) is the model's number of correct positive predictions of a class, whereas. “True Negatives” (TN) is the number of correct negative predictions provided by the model for a class. “False positives” (FP) is the number of false positive predictions provided by the model for a class and “False Negatives” (FN) is the model's inaccurate predictions of negative outcomes. F1 conveys the balance between “precision” and “recall”.

Other metrics used in this study are validation loss and loss. Validation loss is the value of the cost function for the cross-validation data, and the value of the cost function for the training data is loss (Zhang, & Sabuncu, 2018; Li, Dong, Wang, & Xu, 2020).

6.11 Results and discussion of the implemented artefacts in this study

In this section, the discussion is about the learning classifiers implemented in this study, namely, Artificial Neural Network, Long Short-Term Memory (LSTM), Bi Long Short-Term Memory (BiLSTM). As indicated in Table 6.1, the binary classification used the pre-processed data as input variable and the label BL as target variable, whereas multi class classification, the input variable was pre-processed data and the target variable was the label ML.

Table 6.1: The input and target variables

Variable	**Description**
Pre-processed data:	Cleaned and labelled tweets.

[Input Variable]	
Label BL: (NS and S) **[**Target Variable]	The target variable is divided into two categorisations. The label BL accounts for the binary labelling described in chapter 5. NS= Not Suitable. S= Suitable.
Label ML: (NS, S, VS, MS) **[**Target Variable]	The label ML accounts for the multiclass labelling described in chapter 5. NS= Not Suitable. S= Suitable. VS= Very Suitable MS= Moderate Suitable.

6.11.1 Binary classification

A sequence processing model called a bidirectional LSTM, also referred to as a BiLSTM, consists of two LSTMs, one of which moves the input forward and the other backward (Zhang, Wang, & Zhang, 2018). Bidirectional LSTMs can be used to increase model performance on sequence classification issues. BiLSTM significantly improves the quantity of data available to the network, giving the algorithm better context (Ansari, Du, & Naghdy, 2020).

The binary classification approach was adopted in experiment 6.1 and 6.2. As also shown in section 5.4.1, in these experiments the dataset consisted of two labels. The dataset was split into the training set, which amounted to 247188 tweets, representing 80% of dataset and the testing set, which amounted to 61798 rows, representing 20% of the dataset. FastText and Glove were used to vectorize the tokens (textual data). The experiment 6.1 with BiLSTM was done using the pre-trained embedding algorithm FastText to vectorize the tokens and this model is call **BiLSTM.F.2.** The experiment 6.2 with BiLSTM was done using Glove to vectorize the tokens and this model is called **BiLSTM.G.2**

Experiment 6. 1: BiLSTM.F.2 model

The next step was classification after the Twitter data had been pre-processed and labelled. In this phase, data was vectorised using FastText, and the dataset was split into training, testing and validation data. The procedure 6.1 depict the BiLSTM.F.2 classification model of this study. C_L for pre-processed data was used as the input variable and the performance metrics *A* for accuracy, *P* for precision, *R* for Recall, *L* for Loss and *F1* for F1 score are the output variable for this experiment. Procedure 6.1 has five steps, namely, load dataset, load the pre-trained word vector, training set specification, testing set specification and the output specification. For this binary classification BiLSTM.F.2, the parameters presented in Table 6.2 were used.

Table 6.2: Parameters specification for BiLSTM.F.2

Parameters	Value
Embed_dimension	300
Weight_decay	1e-4
Batch_size	256
Num_epochs	30
Dropout	0.4
LSTM	32
Activation function	Relu
Optimizer	Adam
Activation function	Sigmoid
Patience	3

Procedure 6.1: Binary classification model BiLSTM.F.2.

Input: C_L
Output: ***A, P, R, L, F1***

---

1. **Load dataset** (pre-processed data)
   1.1 ***If*** loading dataset is true, then
   1.2 read and display the dataset as panda table,
   1.3 pre-process data,
   1.4 proceed to lexicon detection,

1.5 label the dataset.

End if

2. **Load the pre-trained word vector (FastText)**
   2.1 train the word vector using FastText, then save the results.
   2.2 If the word is included in the model, return the relevant word vector.
   2.3 Else, return 0.
3. **Training set specification**
   3.1 Build an embedding layer with an output shape of (sample 31, 300).
   3.2 Build a bidirectional LSTM with output shape of (sample, 64).
   3.3 Set a dense layer with an output shape of (sample, 32) with ReLU as an activation function (6.4).
   3.4 Build a dropout layer with an output shape of (sample, 32).
   3.5 Build a Dense_1 layer with Sigmoid as an activation function (6.3).
   3.6 Compile the model with Adam as optimizer, metrics equals to 'accuracy'.
   3.7 Save the model
4. **Testing set specification**
   4.1 Use the val set to validate the saved trained model.
   4.2 Use the test set to test the saved trained model.
5. Return *A* (6.8), *P* (6.9), *R* (6.10), *L*, and *F1* (6.11)

The BiLSTM.F.2 produced the following results:

The accuracy of the BiLSTM.F.2 was 83%, the label N/ (0) had a precision score of 83%, a recall score of 79% and a F1 score of 85%. The Label S/ (1) had a precision score of 90%, a recall score of 75% and a F1 score of 83% (Table 6.3).

Table 6.3: Classification report of model BiLSTM.F.2

		Label			
Model name	**Metric**	**NS/(0)**	**S/(1)**	**Accuracy (%)**	**Weighted F1-score (%)**
BiLSTM.F.2	Precision	0.79	0.90	83	83
	Recall	0.92	0.75		
	F1	**0.85**	**0.82**		

As Figure 6.1 reveals, the label S has a better precision score than the label NS. However, the label NS has better recall and F1 scores than the label S.

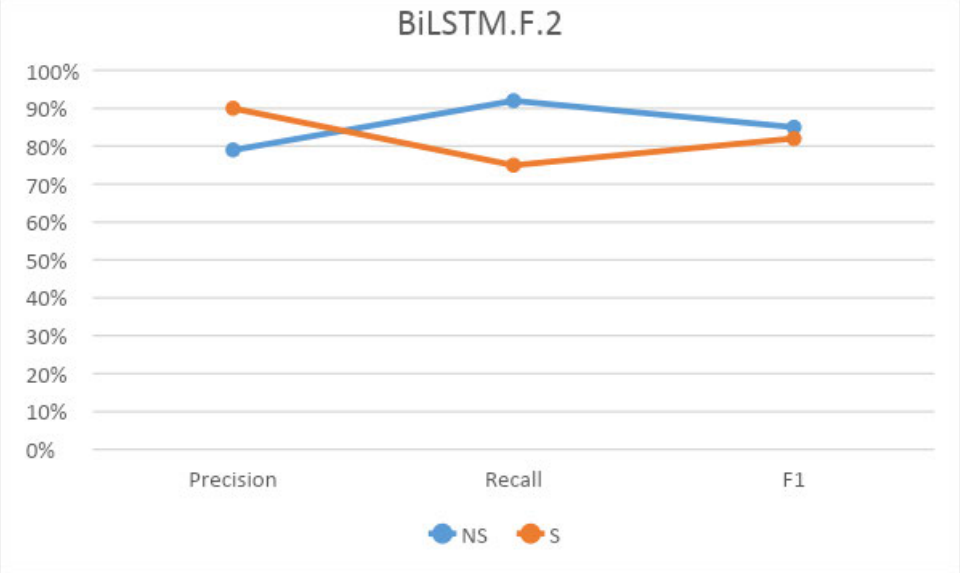

Figure 6.1: BiLSTM.F.2 labels' performance comparison

Experiment 6. 2: BiLSTM.G.2

Following the labelling and pre-processing of the Twitter data, the next step was classification phase. In this phase, data was vectorised using FastText, and the dataset was split into the following categories: training, testing and validation data. The procedure 6.2 depict the BiLSTM.G.2 classification model of this study.

C_L for pre-processed data was used as the input variable and the performance metrics *A* for accuracy, *P* for precision, *R* for Recall, *L* for Loss and *F1* for F1 were the output variables for

this experiment. Procedure 6.2 has five main steps, namely: load dataset, load the pre-trained word vector, training set specification, testing set specification and output specification. For this binary classification BiLSTM.G.2, the following parameters were used (Table 6.4):

Table 6.4: Parameters specification for BiLSTM.G.2

Parameters	Value
Embed_dimension	300
Weight_decay	1e-4
Batch_size	256
Num_epochs	30
Dropout	0.4
LSTM	32
Activation function	ReLU
Optimizer	Adam
Activation function	Sigmoid
Patience	3

Procedure 6.2: binary classification model BiLSTM.G.2.

Input: C_L
Output: ***A , P , R , L , F1***

---

1. Load dataset (pre-processed data)
   - ***1.1*** ***If*** loading dataset is true, then
   - ***1.2*** read and display the dataset as panda table,
   - ***1.3*** pre-process data,
   - ***1.4*** proceed to lexicon detection, and
   - ***1.5*** label the dataset.

   End if

2. **Load the pre-trained word vector (GloVe)**
   - 2.1 train the word vector with GloVe, and save the result,
   - 2.2 If the word is in the model, return the corresponding word vector,

2.3 Else, return 0.

3 **Training set specifications**

3.1 Build an embedding layer with an output structure of (sample 31, 300).

3.2 Build a bidirectional LSTM with output structure of (sample, 64).

3.3 Set a dense layer with an output structure of (sample, 32) with ReLU as an activation function (6.4).

3.4 Build a dropout layer with an output shape of (sample, 32).

3.5 Build a Dense_1 layer with Sigmoid as an activation function (6.3).

3.6 Compile the model with Adam as optimizer, metrics equals to 'accuracy'.

3.7 Save the model.

4 **Testing set specification**

4.1 Use the Validation set to validate the saved trained model.

4.2 Use the test set to test the saved trained model.

5. Return *A* (6.8), *P* (6.9), *R* (6.10), *L*, and *F1* (6.11).

The BiLSTM.G.2 produced the following results:

The accuracy of the BiLSTM.G.2 was 86.83%, the label N/ (0) had a precision score of 85%, a recall score of 97% and a F1 score of 91%. The Label S/ (1) had a precision score of 92%, a recall score of 65% and a F1 score of 76% (Table 6.5). In general, the model BiLSTM.G.2 showed a better precision score for the label S compared to the label NS. However, the label NS had better recall and F1 scores than the label S.

Table 6.5: Classification report of model BiLSTM.G.2

		Label			
Model name	**Metric**	**NS / (0)**	**S / (1)**	**Accuracy (%)**	**Weighted F1-score (%)**
BiLSTM.G.2	Precision	0.85	0.92	86.8	86
	Recall	0.97	0.65		
	F1	**0.91**	**0.76**		

For both binary classification models, namely, BiLSTM.F.2, and BiLSTM.G.2, the label with lesser size (S) has a better precision. However, the label S has lesser Recall and F1 score. This is due to the fact that the label NS has a bigger size than the label S.

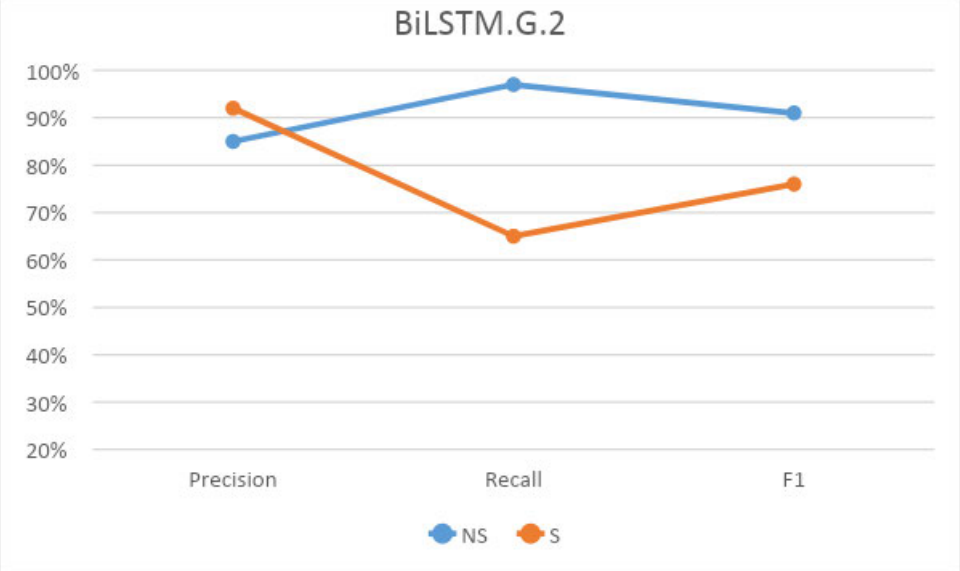

Figure 6.2: BiLSTM.G.2 Labels performance comparison

The figure 6.3 depicts the BiLSTM.G.2 and BiLSTM.F.2 architecture as both BiLSTM.G.2 and BiLSTM.F.2 used identical parameters. The architecture depicts the layers and the input and output shapes of the BiLSTM.G.2 and BiLSTM.F.2 models.

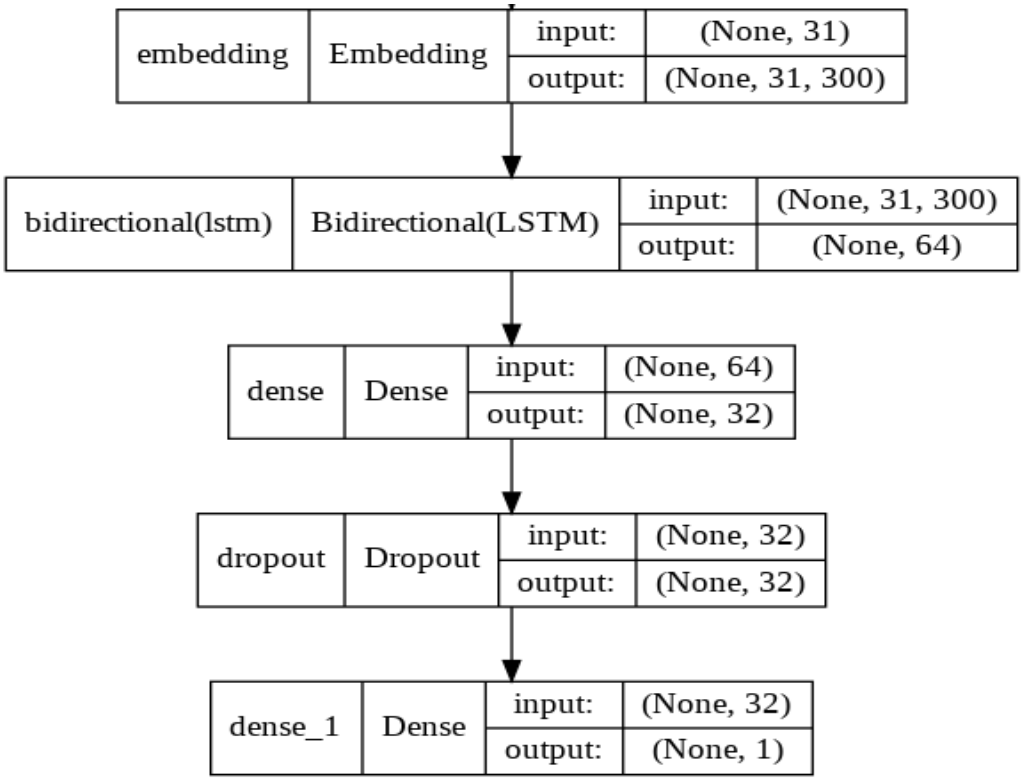

Figure 6.3: The BiLSTM.G.2 and BiLSTM.F.2 architecture

6.11.2 Multi class classification

The multiclass classification findings of this study are presented and discussed in this section. To build the multiclass classification artefacts, the dataset described in section 5.4.2 (Figure 5.3) was initially used. However, after multiple experiments, it was noticed that the cause of poor results was the extremely imbalanced classes. Thus, the researcher ended up using under-sampling techniques to fix the imbalance of classes. The size of the classes used in the multiclass classification is depicted in Table 6.6. For the multi class classification, two classification models were used for this study, namely **LSTM.4** and **ANN.4.**

Table 6.6: Value counts of target variable for multi class classification

Label	**Value counts**
0 (Not Suitable)	44680
1 (Suitable)	31448
2 (Very Suitable)	46695
3 (Moderately Suitable)	21953

6.11.2.1 LSTM.4

For the purpose of this study, the classification model that used LSTM as a learning classifier is named LSTM.4.

Long Short-Term Memory networks, or "LSTMs," are a type of RNN that can understand long-term dependencies. LSTMs are specifically developed to prevent the issue of long-term dependency (Park, Choi, Choi, Ryu, & Kim, 2020). All recurrent neural networks are made up of a series of repeated neural network modules. LSTMs have a chain-like structure as well, although the recurring module is structured differently (Kim, & Cho, 2019; Zhao, Mao, & Chen, 2019). Only with a few tiny linear interactions, it flows straight down the entire chain. It's incredibly easy for data to simply travel along with it unaltered. The LSTM can delete or add information to the cell state, which is carefully controlled by structures called gates (Colah, 2015; Yuan, Li, Shardt, Wang, & Yang, 2020).

The LSTM classifier was used in the LSTM.4 model as the learning classifier. Due to the LSTM capability to learn long-term dependence of the input data, this study opted for it to enhance the performance of the academic suitability classification.

Experiment 6.3: LSTM.4

This experiment involved carrying out the classification procedure for the previously processed and labelled Twitter data. In this phase, data was vectorised using the Keras Tokenizer, and the dataset was split into categories, namely: training, testing and validation data. The classification model for experiment 6.3 was named **LSTM.4**. The C_L for pre-processed data was used as the input variable and the performance metrics *A* for accuracy, *P* for precision, *R* for Recall, *L* for Loss and *F1* for F1 score are the output variable for this experiment. The procedure has five main steps, namely, load dataset, load Keras Tokenizer, training set specification, testing set specification and the output specification.

For multiclass classification LSTM.4, the following parameters were used (Table 6.7): Procedure 6.3 depict the LSTM.4 classification model of this study.

Table 6.7: Parameters specification for LSTM.4

Parameters	Value
Max_Lenght	500
Vocab_size	128
Batch_size	256
Epochs	30
Loss	Categorical_crossentropy
Flatten	128
Activation function	ReLU
Optimizer	RMSprop
Activation function	SoftMax
Patience	5
Mode	Max
Monitor	Validation loss

Procedure 6.3: The multi class classification model LSTM.4

Input: C_L

Output: ***A, P, R, L, and F1***

---

1. Load dataset

1.1 ***If*** loading dataset is true, then

1.2 read and display the dataset as panda table,

1.3 pre-process data,

1.4 proceed to lexicon detection and

1.5 label the dataset.

End if

2. **Load Keras Tokenizer**

2.1 Use Keras tokenizer to vectorize the tokens.

2.2 Divide the dataset into training (X_train; y_train; Val_train) and testing (X_test; y_test; val_test) sets.

2.3 Set maximum sentence length = 500, to uniform text length.

3 **Training set specifications**

3.1 Set model parameters.

3.2 Set an embedding layer with an output shape of (500, 128).

3.3 Build an LSTM layer with an output structure of (sample, 16).

3.4 Build a dense layer with an output structure of (sample, 16) with activation function ReLU (6.5).

3.5 Build another dense layer with an output shape of (sample, 8) with SoftMax activation function (6.6).

3.6 Compile the model with RMSProp as an optimizer, metrics equal to 'acc' and loss equals to categorical_crossentropy.

3.7 Save the model.

4. Testing set specifications

4.1 Use the test set to test the saved trained model.

4.2 Use the validation split to validate the saved trained model'

5. Return *A* (6.9), *P* (6.10), *R* (6.1), *L*, and *F1* (6.11).

A multi-class classification approach was chosen in experiment 6.3. In this experiment, the dataset consists of four labels mentioned in section 5.6.2. The dataset was split into the following sub-sets: training, which had 115820 tweets, representing 80% of the total dataset and testing, which accounted for 28956 tweets, representing 20% of the total dataset. Keras tokenizer was used to vectorize the tokens (textual data). The LSTM.4 produces the following results:

The accuracy of the **LSTM.4 was** 95.98%, and the weighted F1 was 96%. The Label NS in the **LSTM.4** had a precision score of 100%, a recall score of 99% and a F1 score of 99%, whereas the label S had a precision score of 97%, a recall score of 95% and F1 score of 96%. Also, he

label VS had a precision score of 96%, a recall score of 95%, and a F1 score of 96%. Lastly, the label MS had a precision score of 87%, a recall score of 94%, and a F1 score of 90% (Table 6.8).

Table 6.8: Classification report of model LSTM.4

		Label					
Model name	**Metric**	**NS**	**S**	**VS**	**MS**	**Accuracy (%)**	**Weighted average F1-score (%)**
LSTM.4.	Precision	1.00	0.97	0.96	0.87	95.98	96
	Recall	0.99	0.95	0.95	0.94		
	F1	**0.99**	**0.96**	**0.96**	**0.90**		

Findings showed that the label NS had better precision, recall, and F1 scores than the labels S, VS, and MS (Table 6.4). The Labels S and VS have better precision, recall and F1 scores than the label MS. Figure 6.4 depicts the performance comparison of precision, recall and F1 scores of the four labels. The label MS has a lesser performance compared the other three, namely, NS, S and VS.

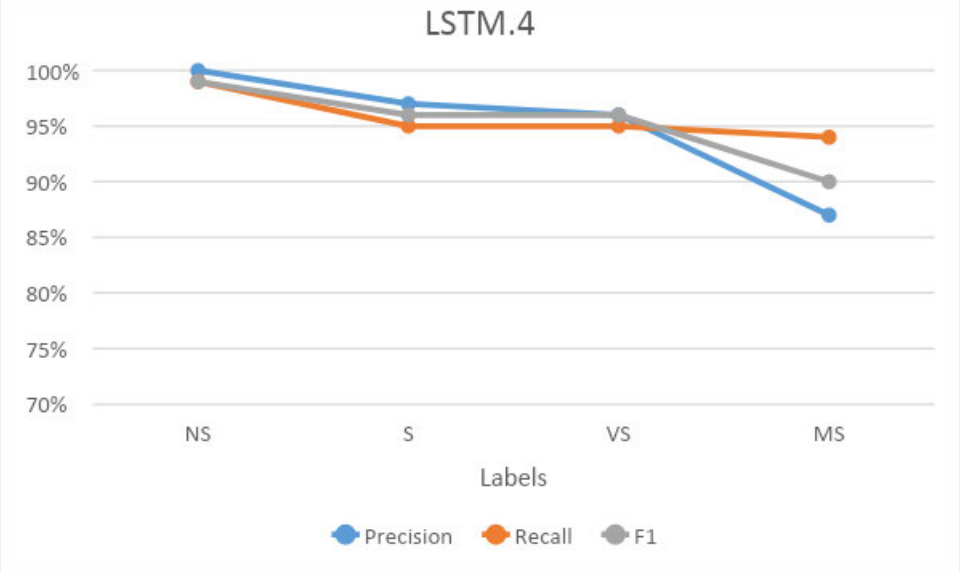

Figure 6.4: LSTM.4 Labels performance comparison

Table 6.9 provides the LSTM.4 model's summary, showing its five layers, namely: the input layer, the embedding layer, the LSTM layer and two dense layers.

Table 6.9: LSTM.4 model summary

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 500)]             0

 embedding (Embedding)       (None, 500, 128)          9723136

 lstm (LSTM)                 (None, 16)                9280

 dense (Dense)               (None, 16)                272

 dense_1 (Dense)             (None, 4)                 68

=================================================================
```

6.11.2.2 The Artificial Neural Network4 (ANN.4)

The classification model that utilized ANN as a learning classifier is referred to as ANN.4 for the purpose of this study.

Since the 1970s, researchers have studied and developed artificial neural networks, which were influenced by the biological neural network of a human (Guo, Nguyen, Vu, & Bui, 2021). ANN is split into the following three segments: the input layer, the hidden layer(s), and the output layer (Yegnanarayana, 2009). Neurons are the information processing units of each layer ANN layer (Guo *et al.,* 2021). The artificial neurons in one or more hidden layers receive inputs from the ANN, which are evaluated and processed to determine the output to the next layer (Tu, 1996; Shenfield, Day & Ayesh, 2018). The collection of weights and biases for the hidden layer and output layer neurons can be adaptively changed using a "learning rule" (typically gradient descent based back-propagation) in ANNs. Due to their self-adaptive nature, ANNs may capture highly complicated and non-linear interactions between dependent and independent variables (Shenfield, Day & Ayesh, 2018). There are four labels on this model, namely: NS, S, VS, and MS. The ANN sends inputs to artificial neurons in one or more hidden layers, which are analysed and processed to generate the output to the next layer. ANNs were able to capture complex and non-linear interactions between dependent and independent factors. In this experiment 6.4, a variety of settings are used.

Experiment 6.4: ANN.4

The next phase was classification after the pre-processing and labelling of the Twitter data. In the classification phase, data was vectorised using Keras Tokenizer, and the dataset was split into training and testing data. The classification model for experiment 6.4 is named **ANN.4**. C_L for pre-processed data was used as the input variable and the performance metrics *A* for accuracy, *P* for precision, *R* for recall, *L* for loss and *F1* for F1 score are the output variable for this experiment. The procedure has five main steps, namely: load dataset, load Keras tokenizer, training set specification, testing set specification and the output specification.

For this multiclass classification ANN.4, the following parameters in Table 6.10 were used:

Table 6.10: Parameters Specification for ANN.4

Parameters	Value
Max_Lenght	500
Vocab_size	128
Batch_size	256
Epochs	30
Loss	Categorical_crossentropy
Flatten	128
Activation function	ReLU
Optimizer	RMSprop
Activation function	SoftMax
Patience	5
Mode	Max
Monitor	Validation loss

Procedure 6.4 depict the ANN.4 classification artefact of this study.

Procedure 6.4 for the multiclass classification model ANN.4

Input: C_L
Output: ***A, P, R, L, F1***

---

1. Load dataset (pre-processed data).
   1.1 ***If*** loading dataset is true, then
   1.2 read and display the dataset as panda table,
   1.3 pre-process data,
   1.4 proceed to lexicon detection, and
   1.5 label the dataset

   End if

2. **Load Keras Tokenizer**
   2.1 Use Keras tokenizer to vectorize the tokens.
   2.2 Divide the dataset into training (X_train; y_train) and testing (X_test; y_test) sets.
   2.3 Set maximum sentence length = 500, uniform text length.
3. **Training set specification**
   3.1 Set model parameters.
   3.2 Set an embedding layer with an output shape of (500, 128)
   3.3 Build a flatten layer, with an output structure of (sample, 64000)
   3.4 Build a dense layer with an output structure of (sample, 32) with activation function Relu (6.4).
   3.5 Build a classification (dense) layer, in which the academic suitability classification is obtained by SoftMax activation function (6.5) with the weight continuously updated by the loss function (6.6), and the final output shape is (sample, 4).
   3.6 Compile the model with Adam as optimizer, metrics equals to 'acc' and a loss equals to categorical_crossentropy.

3.7 Save the model.

4. **Testing set specification**

4.1 Use the validation split to validate the saved trained model.

4.2 Use the test set to test the saved trained model.

5 Return *A* (6.8), *P* (6.9), *R* (6.10), *L*, and *F1* (6.11)

The multi-class classification approach was chosen in experiment on four. In this experiment, the dataset consists of four labels also mentioned in section 5.4.2. The dataset was split into the training set, which had 247188 tweets, representing 80% of the total dataset and the testing set, which had 61798 tweets, represents 20% of the total dataset. Keras tokenizer was used to vectorize the tokens (textual data). The ANN.4 produced the following results:

The accuracy of the ANN.4 was 93%, and the weighted F1 score was 93%. Precision, recall, and F1 scores for the label NS were 97%, 99%, and 98% respectively. Label S in the ANN.4 had a precision score of 93%, a recall score of 91%, and an F1 score of 92%. Label VS had a precision score of 92%, a recall score of 93%, and an F1 score of 92%, whereas Label MS had a precision score of 86%, a recall score of 85%, and an F1 score of 86% (Table 6.11).

Table 6.11: Classification report for ANN.4

		Label					
Model name	**Metric**	**NS**	**S**	**VS**	**MS**	**Accuracy (%)**	**Weighted F1-score (%)**
ANN.4	Precision	0.97	0.93	0.92	0.86	93	93
	Recall	0.99	0.91	0.93	0.85		
	F1	**0.98**	**0.92**	**0.92**	**0.86**		

Figure 6.5 depicts the graphical performance comparison of precision, recall and F1 of the four labels, namely, NS, S, VS and MS. The label NS has the better score compared to the rest. The label MS has the lesser performance compared to the label NS, S and VS.

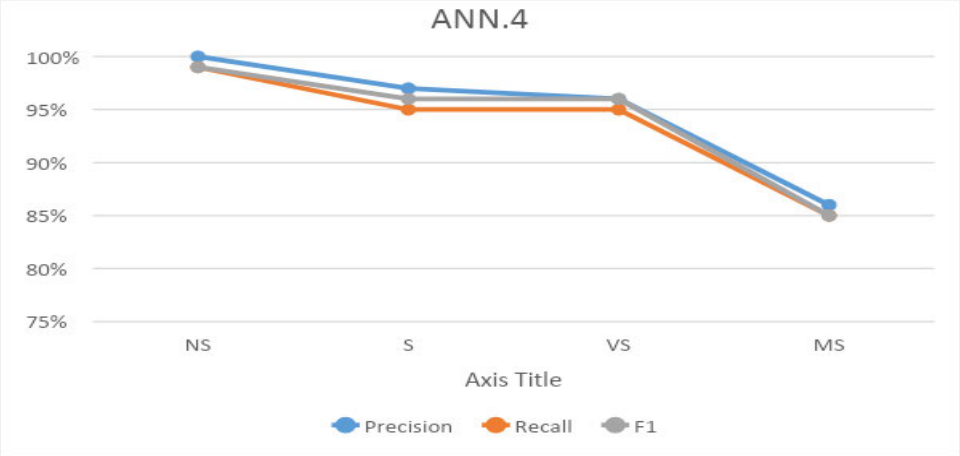

Figure 6.5: ANN.4 labels' performance comparison

Figure 6.6 shows the ANN.4 architecture. The architecture depicts the layers and the input and output shapes of the ANN.4 classification model. The input layer, the embedding layer, the flatten layer, and the two dense layers make up the five layers.

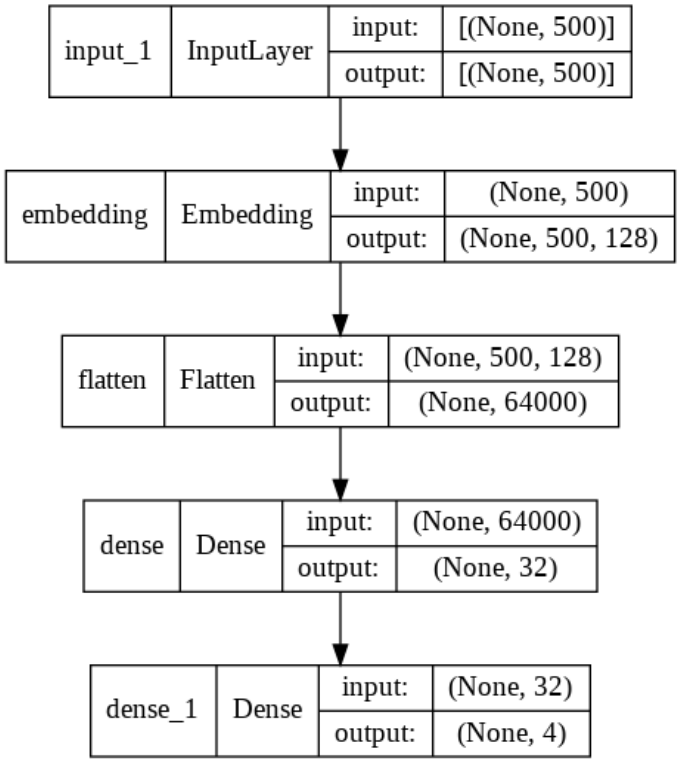

Figure 6.6: ANN.4 classification model architecture

6.12 Summary of the chapter

In prior studies, textual data patterns were found to be reliable indicators of attitude, behaviour, and personality on social media. With the increased use of social media, the data provided by users can be used to understand their opinions and interests, allowing for the recommendation of services and facilities or the prediction of their suitability. Thus, for this study, the researcher designed and implemented models that trace academic interest in the tweets of users and predict their academic suitability. First, the textual data had to be transformed into a learnt representation of text, where words with similar meanings were represented similarly. To this end, three word embedding techniques were used for this study. The FastText and GloVe pre-trained models were used as word embedding techniques for binary classification and Keras tokenizer for the multiclass classification. To determine the best model for classifying academic suitability, various deep learning classifiers were trained. Using the stratified sampling method, data were partitioned into the training set which accounts for 80% of the total dataset and the test set which accounts for 20% of the dataset to maintain the balance of classes across sets prior to modelling. This chapter presented the different results of the experiments of this study. For the multi class classification, two models were designed and presented in this chapter. The chapter established how the labelled could be used for classification by designing and implementing four models namely, BiLSTMF.2, BiLSTMG.2. LSTM.4 and ANN.4 Procedures for the different models were also provided in this chapter. In the next chapter, the models are discussed with regard to their performance evaluation based on the presented experimental results.

CHAPTER SEVEN: PERFORMANCE EVALUATION

7.1 Introduction

Evaluation is crucial in Design Science Research for establishing an artefact's believability and any required changes (Peffers, Tuunanen, & Niehaves, 2018). The results of the extensive experiments presented in chapter six are discussed in this chapter. The experiments conducted on different Deep learning (DL) classifiers such as LSTM, Bi-LSTM, and ANN are evaluated in this chapter using performance metrics described in section 7.2. As described in chapter 6, different types of classifications, embedding techniques (section 6.5), labelling (sections 5.4.1 and 5.4.2), and parameters were used in this study. In addition, a comparison of the artefacts developed in this study with the existing artefacts from the literature is made. This chapter addresses research objective 4, which states as follows:

"To evaluate the performance of the built predictive models".

7.2 Performance evaluation metrics

Performance metrics are important during model training and validation (Erickson & Kitamura, 2021). The following are the most widely used classification performance metrics: accuracy, F1 score, precision, recall, training and validation loss, and validation accuracy.

The percentage of correct predictions the model makes out of all possible predictions is known as accuracy in classification problems. Accuracy is a good metric to consider when the target variable classes in the data are approximately balanced (Behera, Kumaravelan, & Kumar, 2019). Precision is the fraction of truly positive classes from all classes the model predicted positive, whereas recall is the fraction of positive classes predicted as positive. Also, F1 score is the harmonic mean of positive predictive value and sensitivity (recall) (Erickson & Kitamura, 2021). Training loss indicators what degree of fit the model has with the training data, while validation loss measures how efficiently it fits novel data. Validation loss is a metric for evaluating a deep learning model's performance on the validation data (Abbas & Tap, 2019; Seetha & Raja, 2018).

7.3 A brief description of input and target variables

The pre-processed text which was transformed into a feature vector were used as the input variable and the labels as described in section 5.4.1 and 5.4.2 were used as the target variable. For the binary classification, the target variables were 1 (suitable) and 0 (not suitable) (refer

section 5.4.1 & figure 5.2). The multiclass classification consists of four labels namely 0 (Not Suitable), 1(Suitable), 2 (Very Suitable), and 3 (Moderately Suitable) (refer section 5.4.2 & figure 5.2). The Figures 7.1 and 7.2 depict the value counts of the datasets used in these experiments. For the binary classification, label 0 (not suitable) has 208632 rows in the dataset, and label 1 (suitable) has 100354 (Figure 7.1). The label 2 (very Suitable) has 46695 rows in the dataset, label 0 (Not Suitable) has 44680 rows, label 1 (suitable), and label 3 (moderately suitable) has 21953 rows in the dataset (7.2). The four labels (0, 1, 2, 3) represents the target variables in the dataset for the multiclass classification. The labelling process resulted in an imbalanced dataset shown in figures 7.1 and 7.2.

```
df['Label'].value_counts()

0    208632
1    100354
Name: Label, dtype: int64
```

Figure 7.1: Value counts of dataset with 2 labels

```
df['Label'].value_counts()

2    46695
0    44680
1    31448
3    21953
Name: Label, dtype: int64
```

Figure 7.2: Value counts of dataset with 4 labels

7.4 Comparison of binary classification artefacts

For the binary classification, two models were implemented namely, The BiLSTM.F.2 and BiLSTM.G.2. The two models are compared and evaluation in this section. In addition, the overall performance is assessed using a weighted average of *P*, *R*, and *F1*. As revealed by Table 7.1 and Figure 7.3, the BiLSTM.G.2 which used the Glove pre-trained model to vectorise the pre-processed text performed slightly better than the BiLSTM.F.2 which used the FastText pre-trained model to vectorise the text inputs. In general, despite the fact that BiLSTM.G.2 which used Glove has a slightly better precision, F1 and accuracy score, it does not indicate its

superiority or inferiority. It has merely demonstrated that the BiLSTM.G.2 is better suited to the dataset employed in this study. In a study conducted by Dharma, Gaol, Warnars, & Soewito, (2022), it was concluded that the difference in accuracy between models using Glove and FastText is not crucially significant. In fact, the accuracy score of models using Glove and FastText depends on the applied datasets. The findings of this experiment support the finding of the researchers indicated in this paragraph.

Table 7.1: Model accuracy with respect to word embedding used

Model	**Weighted Average %**			**Accuracy %**
	P	*R*	***F1***	
BiLSTM.F.2	0.84	0.83	**0.83**	**83.23**
BiLSTM.G.2	0.87	0.87	**0.86**	**86.83**

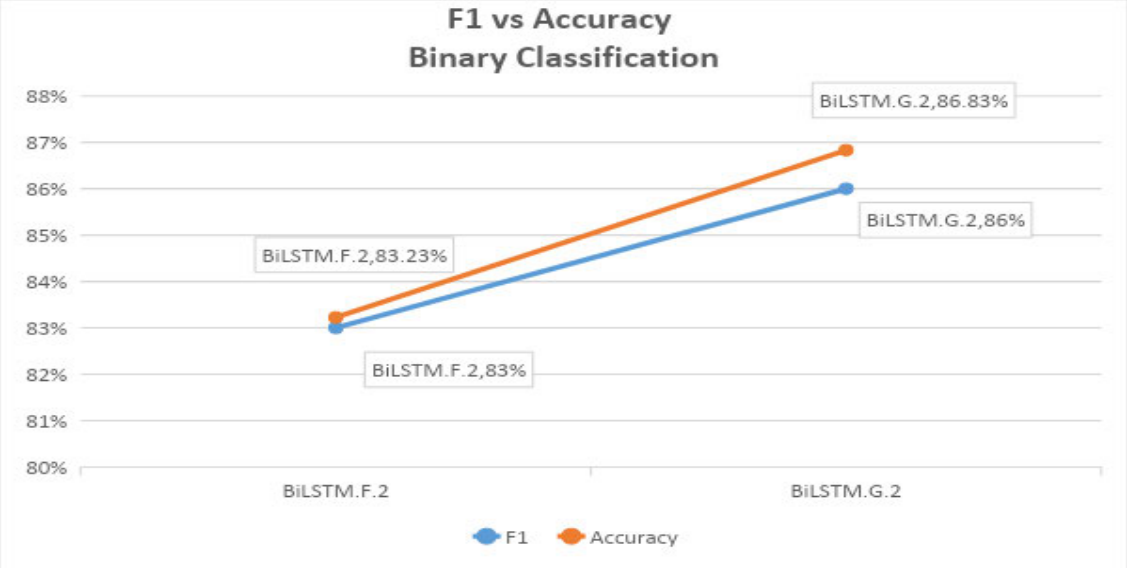

Figure 7. 3: F1 and Accuracy comparison for binary classification

To evaluate how well the model performs on unseen data, a validation set was used. This model used the early stopping as a technique for cross-validation. The validation loss has demonstrated that the BiLSTM.F.2 has a validation loss score of 0.34 while BiLSTM.G.2 has a validation loss score slightly above 0.31 (Figures 7.4, and 7.5). Moreover, Figure 7.3 depicts that BiLSTM.G.2 performs better than BiLSTM.F.2 in terms of accuracy (86,83% vs 83,23%) and F1 score (86% vs 83%).

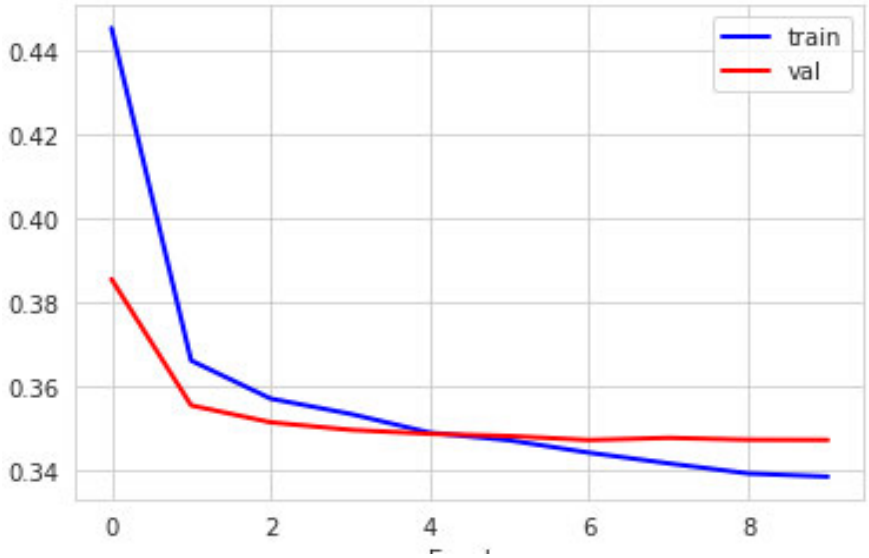

Figure 7.4: Validation loss of BiLSTM.F.2

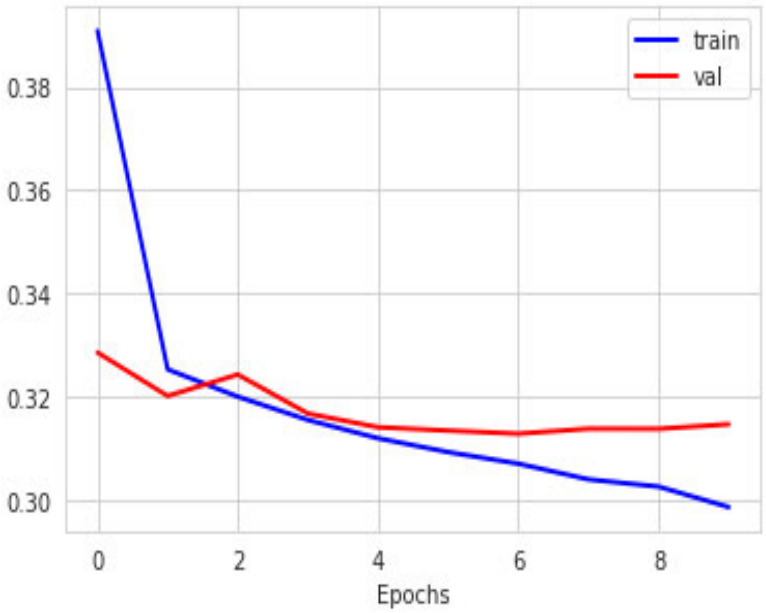

Figure 7.5: Validation loss of BiLSTM.G.2

As depicted in Figures 7.6 and 7.7, the comparison of the training validation shows that the BiLSTM.G.2 slightly performed better than the BiLSTM.F.2. The validation set offered an objective assessment of the model fit on the trained classification model.

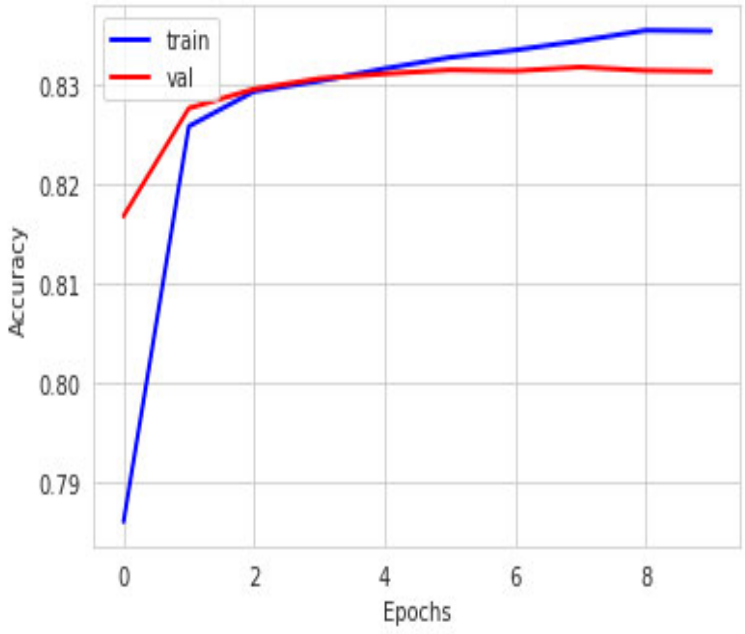

Figure 7.6: Training and validation accuracy of BiLSTM.F.2

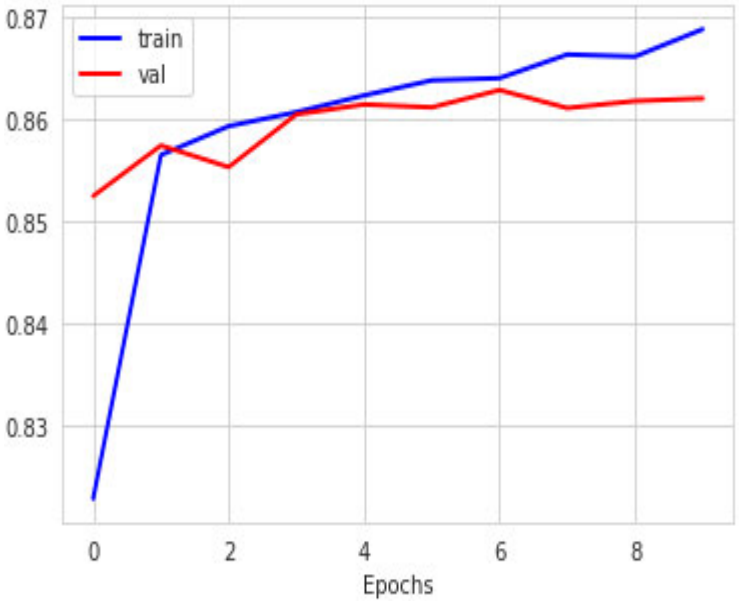

Figure 7.7: Training and validation accuracy of BiLSTM.G.2

7.5 Comparisons of multiclass classification artefacts

A multiclass classification was chosen for this experiment, the dataset employed for this artefact is described in Figure 7.2. The *P*, *R*, and *F1* metrics were used to measure the effects of each class, while accuracy was used to measure the overall classification effect of the model. Two classification models, namely: LSTM.4 and ANN4, described in chapter 6 were implemented here and both models used similar parameters except the learning classifiers LSTM and ANN. Performance-wise, the LSTM.4 model outperforms the ANN.4 model. The weighted average of *F1* was 0.93 for the ANN.4 model and 0.96 for the LSTM.4, and accuracy was 96 (Table 7.2). The LSTM.4 seemed to work better for the dataset used. Hu, Ni, & Wen, (2020) claim that extension of ANN improved the accuracy of prediction of certain models such as copper price prediction. Thus, as also depicted in this experiment, LSTM being an extension of RNN which is an extension of ANN has the capability to produce better accuracy. It can be concluded that for this dataset, LSTM performed better than ANN which supports the study by Hu *et al*., (2020). In his study, Ma, (2020) concluded that LSTM performs better than ANN. The reason for the latter might be because of the LSTM learning classifier's improvement of the vanishing gradient issue.

Table 7.2: The models' accuracy with respect to 4 Labels

Model	**Weighted Average %**			**Accuracy %**
	P	*R*	***F1***	
ANN.4	0.93	0.93	0.93	**93**
LSTM.4	0.96	0.96	0.96	**96**

When measuring the validation loss of the two models, namely: ANN.4 and LSTM.4, LSTM.4 has a minimal loss of 0.06 and a minimal validation loss of 0.12 while ANN.4 has a minimal loss of 0.08 and a minimal validation loss of 0.2 (Figures 7.8 and 7.9)

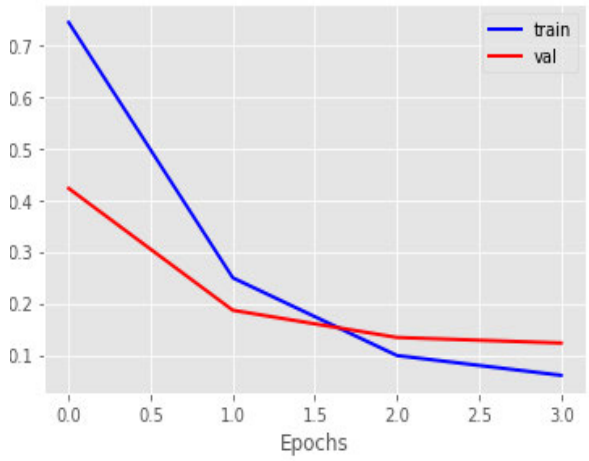

Figure 7.8: Training and validation loss of LSTM.4

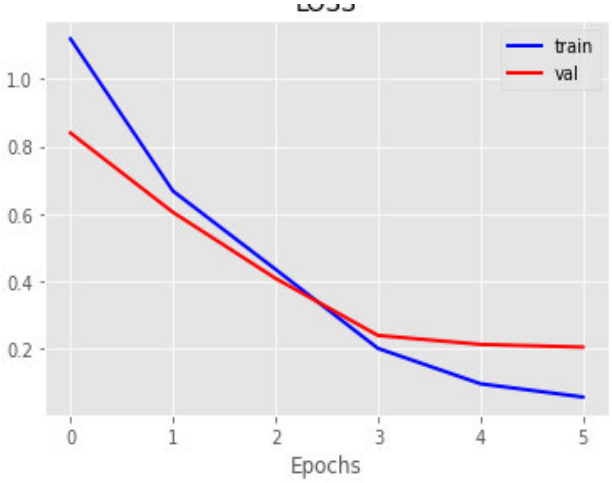

Figure 7.9: Training and validation loss of ANN.4

7.6 Overall comparison of the models of this study

The models depicted in Table 7.3 are binary classification and multiclass classification. The goal was to review the performance metrics of all the artefacts implemented in this study. As described in Figure 7.2, two binary classification artefacts namely BiLSTM.F.2, and BiLSTM.G.2 used an imbalanced dataset of two classes, the ANN.4 and LSTM.4 also used an imbalanced dataset as an input as described in Figure 7.1. The LSTM.4 had better metrics score with 0.96 weighted average of F1 and 96% accuracy, whereas ANN.4 had 0.93 weighted average of F1 and 93 % accuracy. Also, for the BiLSTM.G.2, the weighted average of F1 was 0.86 and accuracy was 87%. The weighted average f1 score of BiLSTM.F.2 was 0.83 and accuracy was 83% (Table 7.3 and Figure 7.4). The size of the input datasets, the embedding

word technique used, the parameters, and the classifiers used to train the models have some impact on the models' performance implemented in this study.

For the LSTM.4, and ANN.4 models did not perform well when a pre-trained word embedding was used to vectorise the tokens. Furthermore, when using Adam optimizer, the two models had slightly higher accuracy but poor validation loss scores as the two models were extremely over fitting. For the BiLSTM.F.2 and BiLSTM.G.2, when the size of the label 0 (Not suitable) was reduced to match the label 1 (Suitable) size (0=100354 & 1= 100354), accuracy was slightly decreased as depicted in Table 7.4.

Table 7.3: Overall Performance Comparison

Model	**Performance metrics Weighted average %**			
	Precision	*Recall*	*F1*	*Accuracy*
BiLSTM.F.2	0.88	0.81	0.83	83
BiLSTM.G.2	0.87	0.87	0.86	86.83
ANN.4	0.93	0.93	0.93	93
LSTM.4	0.96	0.96	0.96	96

Table 7.4: Comparison of BiLSTM.F.2 and BiLSTM.G.2 with reduced dataset size

Model	**Input size (rows)**	**Performance metrics Weighted average %**			
		Precision	*Recall*	*F1 score*	*Accuracy*
BiLSTM.F.2	**100354**	0.84	0.83	0.83	82.5
BiLSTM.G.2	**100354**	0.82	0.82	0.82	82

7.7 Comparison of performances with related research

In the large domain of Machine Learning, the goal is to forecast an outcome based on the data given. When the output reflects distinct classes, the prediction is referred to as a “classification problem”, but when the outcome is a numeric measurement, it is referred to as a "regression problem" (Grandini, Bagli, & Visani, 2020). Deriving a prediction model from a labelled set of raw training data is one of the most prevalent data science problems. Data scientists working on such a problem are surrounded by options and possibilities right from the start (Swearingen, Drevo, Cyphers, Cuesta-Infante, Ross, & Veeramachaneni, 2017). This is attributable to the enormous number of various algorithms available, as well as the complexity of deploying, fine-tuning, and interpreting them (Luque, Carrasco, Martín, & de Las Heras, 2019). Furthermore, Kirsch, Van Amersfoort, & Gal (2019) claim that data efficiency is a major issue in deep learning. Different researchers (Smola & Vishwanathan, 2008; Wardhani, Rochayani, Iriany, Sulistyono, & Lestantyo, 2019; Grandini *et al*., 2020; Agostinho & Mendes-Moreira, 2022) have used different metrics such as accuracy, confusion matrix, precision, recall, *F1* score, sensitivity, specificity, percent correction classification, ROC curve and AUC to evaluate the performance of their predictive models. However, the common metrics which have been widely used are accuracy, precision, recall, F1 score and loss validation. Various experiments were conducted in this study to investigate the effects of dataset size, embedding methods, and classifiers, as well as other parameters, on the performance of the models. This section provides a discussion on the comparison of the artefacts implemented in this study with the existing researches from the literature that used Twitter dataset. Due to the fact that this study used the Twitter dataset, the discussion in this section will first compare the study's artefacts to those from journal papers that also used the dataset, with an emphasis on recruitment. Secondly, the discussion focuses on the study's artefacts to those from scholarly studies that employed the Twitter dataset to examine user behaviour, emotion, attitude, sentiment, and opinion on the social media platform.

Kern et al. (2019) use 128,279 Twitter users from 3,513 different vocations to autonomously identify user personalities and visually correlate the personality traits of various professions. For their study, ten professions, namely: school principal, superintendent, data scientist, software engineer, executive chef, athletics director, teacher, agent, manufacturer and campaigner were selected. Findings demonstrate how social media may be utilised to connect people with their ideal careers. The indicated study was focused on the big five personality

traits, aligning them with occupations. However, the researchers failed to categorize specific words in line with the corresponding occupations/jobs. It would have been better to develop different dictionaries for different occupations. The researchers used different models and the XGBoost with 74% accuracy performed better. However, when compared to all the experiments of this study, the models described in Table 7.3; 7.5, and figure 7.10 outperform their models.

Menon & Rahulnath, (2016) estimated emotional intelligence through Twitter data to evaluate and rank job candidates. They designed a method that automatically assesses a candidate's aptitude and eligibility during the hiring process. To assess the emotional intelligence and aptitude of users, the researcher used two main techniques, namely: meta-attributes extraction and multi label classification. They also used the Big five model to explore the emotional intelligence of users. However, the researchers failed to specify which occupation/job is aligned with the emotional intelligence or with the personality trait identified.

Utilizing algorithm-independent approach for classification, the researchers used Naïve Bayes, support vector machine, and Decision tree. For the classification using dependent approach, they used Random Forest and K-Nearest Neighbour. The findings reveal that the Naïve Bayes classifier scores were as follows: 87.3% accuracy, 80.67% precision, and 87.1% recall. Furthermore, the Support Vector machine classifier scores were as follows: 86.1% accuracy, 69.83% precision and 75.2% recall, whereas the Decision Tree classifier scores were 84.8% accuracy, 82.93 precision and 85.83% recall. Also, the Random Forest classifier scores were 64.18% accuracy, 92.09% precision and 89.13% recall, whereas, the K-Nearest Neighbour scores were 25.30% accuracy, 83.24% precision and 80.22% recall. In terms of accuracy, the Naïve Bayes performed better (87.3%) than the other classifiers. However, the LSTM.4 detailed in table 7.3 outperforms their best model.

Putra, Wasmanson, Harmini, & Utama, (2020) conducted a study for emotion classification using Sundanese Twitter dataset. KNN, RF, NB, and SVM were used in the study. The study was conducted using 2518 Tweets, and the dataset contained four unique emotions/labels, each with a balanced amount of data. The small size of the dataset (2518 Tweets) used in the study by Putra et al., (2020) can be considered as a shortcoming because a bigger dataset could have been more representative of the Twitter users. The SVM model had the highest accuracy of 95%. When compared to the LSTM.4 described in Table 7.3, the LSTM.4 performed slightly better than their SVM model. Jianqiang, Xiaolin, & Xuejun, (2018) conducted a study to

measure sentiment analysis using the Stanford Twitter Sentiment Gold dataset (STSGd) which has 2034 Tweets (1402 negative Tweets and 632 Tweets). Both the Glove-CNN artefact of Jianqiang et al., and the BiLSTM.G.2 artefact of this study used Glove as a word embedding method to vectorise data. However, the small size of the dataset used by Jianqiang et al., can be considered as a shortcoming of their study as the dataset size plays a role in the performance of classifiers (Chu, Hsu, Chou, Bandettini, & Lin, 2012). Furthermore, the experiments of this studydemonstrate that the size of the dataset impact the performance of the artefacts. The Glove-Convolution Neural Networks (Glove-CNN) model of the Jianqiang et al., (2018) has 87% accuracy, and when compared to the BiLSTM.G.2 depicted in Table 7.1, their model and the BiLSTM.G.2 have similar accuracy. Kunte and Panicker (2020) used XGBoost and Ensemble classifiers to predict the personality of Twitter users. There were 9,918 tweets in their dataset. With their ensemble classifier, the artefact was able to reach an accuracy of 82.59 %. When compared with artefact of this study, LSTM.4 performed better. Moreover, LSTM.4 used a bigger dataset compared to the dataset used by Kunte and Panicker (2020). Reshi, Rustam, Aljedaani, Shafi, Alhossan, Alrabiah, and Ashraf (2022) used a worldwide Twitter dataset to measure covid-19 sentiment analysis. There is a total of 208 locations in the datasets collected. Their proposed model LSTM-GRNN achieved an accuracy of 95 %. When compared to the artefact of the current study depicted in Table 7.5, LSTM performed better.

Table 7.5: A comparison of the performance of this study's artefacts with other researchers' artefacts

Author	**Classifier**	**Accuracy (%)**
Putra *et al* (2020)	SVM	95
Jianqiang *et al* (2018)	Glove-CNN	87
Kunte & Panicker (2019)	XGBoost	83
Reshi *et al* (2022)	LSTM-GRNN	95
Kern *et al* (2019)	XGBoost	74
Menon & Rahulnath, (2016)	Naïve Bayes	87.3
Researcher own work	BiLSTM.G.2	87
Researcher own work	LSTM.4	96

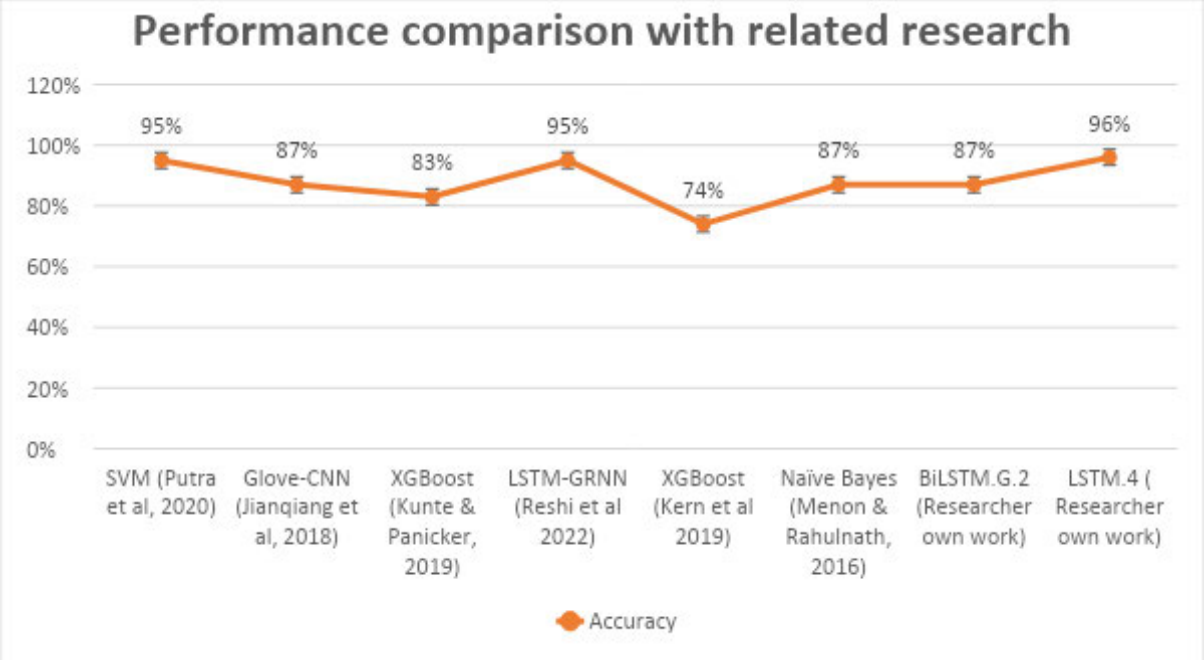

Figure 7.10: A comparison of the performance of this study's artefacts with other researchers' artefacts

7.8 Summary of the chapter

To properly classify texts in a variety of applications, as well as due to the incredibly rapid increase in the number of complex corpus and texts, it is now necessary to have a better comprehension of machine learning methodologies. The ability of these learning classifiers to comprehend intricate models and non-linear interactions within data determines their effectiveness. Nevertheless, it can be challenging for academics to identify appropriate designs, structures, and evaluation methods for text classification. In this chapter the researcher has discussed in detail the performance evaluation of the different experiments/classification models of this study. A variety of evaluation measures, such as accuracy, precision, recall, F1 score, and validation loss were employed to evaluate the classification performance of the artefacts. Moreover, a comparison between the different artefacts implemented were presented in this chapter. Furthermore, in this chapter, the researcher has also evaluated this study's experiments in the light of other researchers' previous studies. This study's experiments showed that the dataset size, parameters, and the embedding strategies all have an impact on how well machine/deep learning models perform. Different parameters produce different results. Thus, the results suggest that researchers should pay attention to the type of embedding techniques and model parameters when using Twitter data. Furthermore, this chapter addressed research objective 4. The next chapter provides the conclusion and recommendations.

CHAPTER EIGHT: CONCLUSION AND RECOMMENDATIONS

8.1 Introduction

This chapter address the main research objective of this study. Furthermore, the chapter provides a summary of the responses to the questions posed in section 1.4. The main research objective of this study was to establish how the suitability of academics can be classified using Twitter data set. The study's main findings and a summary of the dissertation are presented in this chapter. The study's contribution to the corpus of knowledge, a discussion of the study's shortcomings as well as some helpful recommendations for future studies are also provided in this chapter. The process model that guided this study included three main phases, namely: identify and capture, understand, and classification. The first phase included the identification and capturing of data from Twitter, the second phase consisted of data cleaning, tokenization, lemmatization, sentiment analysis, lexicon detection and data annotation. The third phase consisted of the vectorization, data splitting, model specification, training, testing, validation, and evaluation.

Experiments were conducted on two forms of classifications, namely: binary classification and multiclass classification. Also, the following four artefacts: BiLSTM.F.2, BiLSTM.G.2, ANN.4, and LSTM.4, were developed, implemented, and evaluated through various experiments. An evaluation and comparison of the classification models was discussed in detail in chapter seven.

8.2 The updated process model

A process model (Figure3.6) was designed to address the research questions. The initial process model (Figure 3.6) was then updated (Figure 8.1) after the experiments and recommended for academic suitability classification.

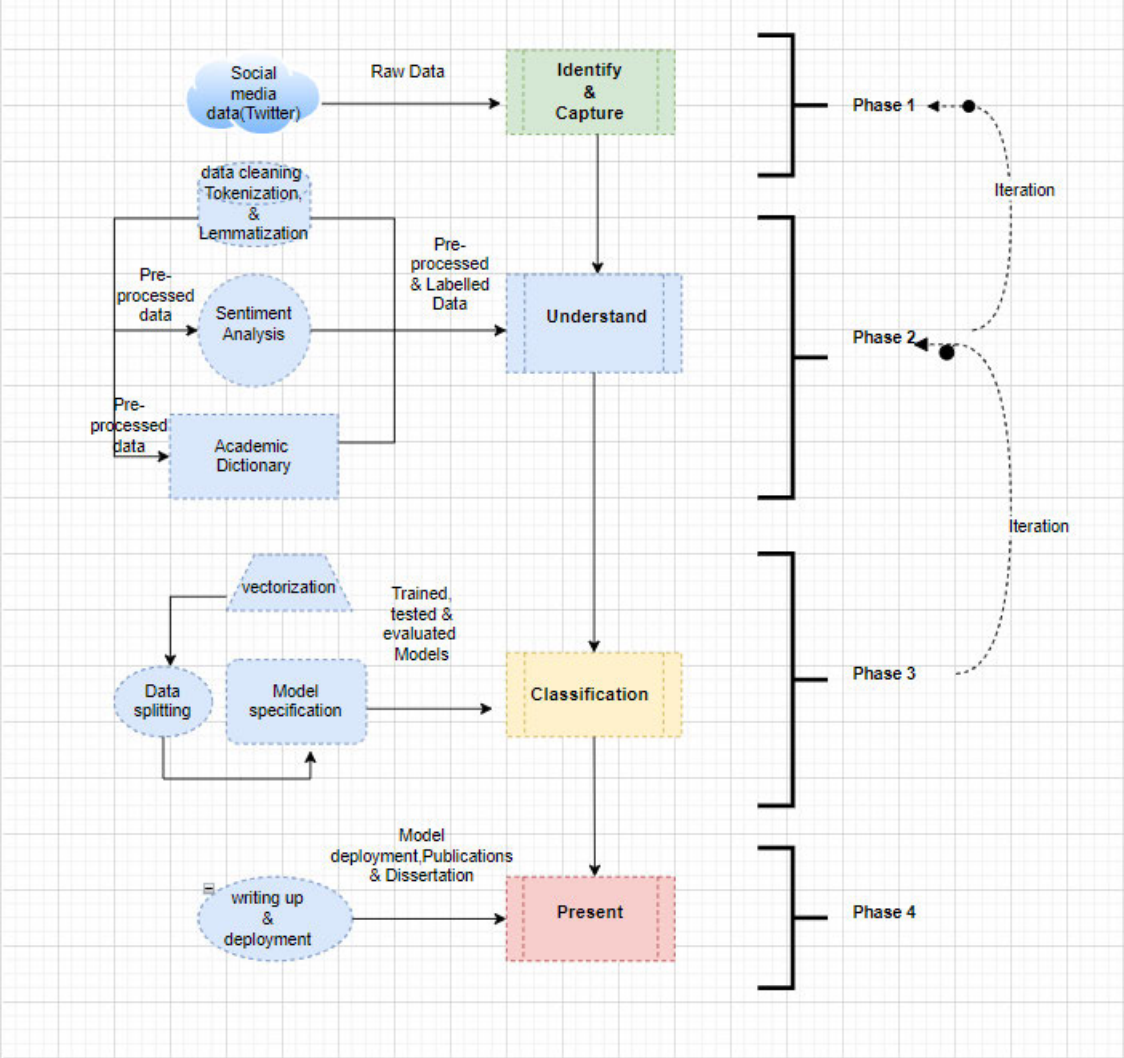

Figure 8.1: The updated process model

The multiple steps required to answer the research questions stated in chapter one, section1.4 are depicted in the updated process model. The process model is divided into four phases, each of which must be completed before moving on to the next phase. Iteration occurs between phases 1, 2, and 3 depending on the outcomes of the classification model and the study's objectives. In the updated process model, phase 2 is simplify for more clarity. The updated process model does not specify the details of the two labelling methods used in this study. The labelling steps are simplified in the updated process model to allow scalability.

8.2.1 Identify and Capture

Although the first phase does not address any research question, it was important to show how data for the study was identified and captured. Through a Twitter developer account, the researcher requested authorization to utilize the Twitter dataset, and it was approved. A

researcher must apply for and register a developer account to access Twitter data. The developer must register the Twitter application by providing a name, description, and domain after creating a developer account. The Twitter API authenticates the users and provides them with the required access key and access token via the app management dashboard. After, getting all the authorization for this study, academics and non-academics Twitter users were identified based on the description on their profiles. Their publicly available timelines data were extracted using R libraries described in chapter three. The following is the script used in R to extract data:

```
## load rtweet package
Install.packages("rtweet")
library('rtweet')
## authentication token
consumer_key = "XXXXXXXXXXXXX",
consumer_secret = "XXXXXXXXXXXX",
access_token = "XXXXXXXXXXXXXX",
access_secret = "XXXXXXXXXXXXX"
setup_twitter_oauth (consumerkey, consumersecret, accessToken,
accessTokensecret)
searchTwitter(" ", n=1000, lang='en')
```

8.2.2 Understand

This phase addresses the first and second research questions/objectives. The understand phase was implemented in chapter four and chapter five. The first research objective was "To identify and implement pre-processing techniques to clean the dataset according to job-fit attitudes criteria". As stated in chapter four and illustrated in phase two of the process model detailed in chapter three (section 3.5), multiple data pre-processing approaches were utilized to clean data to achieve this research objective. The researcher pre-processed the Twitter data in this study using rule-based techniques. Python was used as a platform to implement the rule-based techniques. The pre-processing procedure consisted of identifying and deleting texts/tweets that were not meaningful for classification. According to the findings, appropriate text pre-processing procedures should be chosen based on the research purpose.

Pre-processing might either remove useful information or introduce errors into the analysis as well as affect the classification findings (Boyd, 2016; Alam, & Yao, 2019; Mishra, Roger, Marini, Biancolillo, & Rutledge, 2021). Despite different text pre-processing approaches being available, not all of them are appropriate for academic suitability classification. Thus, appropriate text pre-processing procedures must be selected depending on the research goal. The following pre-processing techniques for academic suitability classification are recommended: Tokenization, conversion of textual data to lower case, removing the alphanumeric characters and removing the @ and # symbols. These techniques are applied because they do not change the meaning or order of the words in the dataset. In this study, the outcome of this step is a pre-processed data presented in chapter four. The pre-processed data serves as the input for the next step; the labelling step.

The second research objective was "to establish how the pre-processed data can be labelled ". According to Castiglioni, Rundo, Codari, Di Leo, Salvatore, Interlenghi, & Sardanelli, (2021), the process of annotating a dataset into specific categories is known as data labelling. The raw data from social media is unorganized and devoid of labels (Uddin, Bapery, & Arif, 2019).

This researcher suggests using two methods of labelling data after pre-processing it, that is binary labelling and multiclass labelling. Binary labelling consists of detecting the lexicon and then labelling as suitable (S) or non-suitable (NS) for the binary classification. The second method is to use sentiment analysis and lexicon detection to label the data based on some rules described in detail in chapter five for multiclass classification. Sentiment analysis was performed first to get the polarity score of each tweet. The next step was to perform lexicon detection using the developed dictionary. The third step was to use a set of rules described in chapter five (procedure 5.3) to label each Tweet based on sentiment and lexicon detection scores. In this method, sentiment is important to contextualise the user's text. The dataset was labelled into Suitable (S), Not Suitable (NS), Very suitable (VS) and Moderate suitable (MS). Various python libraries were used to achieve this as described in detail in chapter five. The outcome of this step is the labelled data presented in chapter five. The labelled data served as the input for the next step; the classification step.

8.2.3 Classification

This phase addresses the third and fourth research objectives. The third research objective was "To establish how the labelled data be used for predictive model". The classification phase was achieved in chapter six and seven. The goal of this study was to find academic footprints on

social media. To achieve this, various methods, tools and techniques such as binary classification and multiclass classification, Glove and FastText embedding methods and Keras platform were used and conducted. Various steps were involved in this phase as depicted in the process model. After labelling the data, different word embedding techniques were used to convert the tweets into numerical values for classification purpose. FastText, Glove and Keras tokenizer as word embedding methods for different classification models were employed in this study.

Model specification was the next step. Different deep learning model specifications such as activation function, dropout, batch_size, optimizer and patience were experimented. The choice of the number of layers, the activation functions, the classifier and, the choice of optimizer are some of the specification steps in the classification phase described in detail in chapter six. Also, LSTM, BiLSTM and ANN were used classifiers, SoftMax was used as an output activation function, and ReLU was used as an activation function in the hidden layers. The model specifications/parameters used in this study are described in detail in chapter six. The experiments that had better results, namely, BiLSTM.G.2, BiLSTM.F.2, ANN.4, LSTM.4, were presented in chapter six in details.

The fourth research objective was "To evaluate the performance of the built predictive models". To address this, accuracy, *F1*- score, precision, and validation loss were the metrics used to evaluate the performance of the experiments of this study. Furthermore, findings of experiments were compared with the findings of the experiments that were conducted by other researchers before.

8.2.4 Present

This is the last phase of the updated process model, here the outcomes of the study are presented as chapters of a dissertation, and also in the form of research papers for publication and for conferences. The deployment of the model is also part of the present phase. The model can be deployed to be used in a specific context for a specific university or department. Chapter four, five, six, and seven present different results of this study. The results of each step, namely, data pre-processed data, labelled data, binary and multiclass classification models and evaluation were presented in the respective chapters.

8.3 Research model

The researcher recommends using the research model shown in Figure 8.2 for detecting indications of suitable academics on social media. Furthermore, the researcher also posits that the relationship between the variables tweets and digital footprint of academics is influenced by sentiment polarity and lexicon detection. To find traces of academics, researchers need to consider two important variables, namely, sentiment polarity and lexicon detection for the multi class classification of academic suitability. Sentiment polarity can provide context to tweets. Thus, it is important to conduct a sentiment analysis.

This model can serve as a guideline for researchers using social media data, especially Twitter dataset. The model provides guidance to what variables to take into consideration when using Twitter data for the classification purpose.

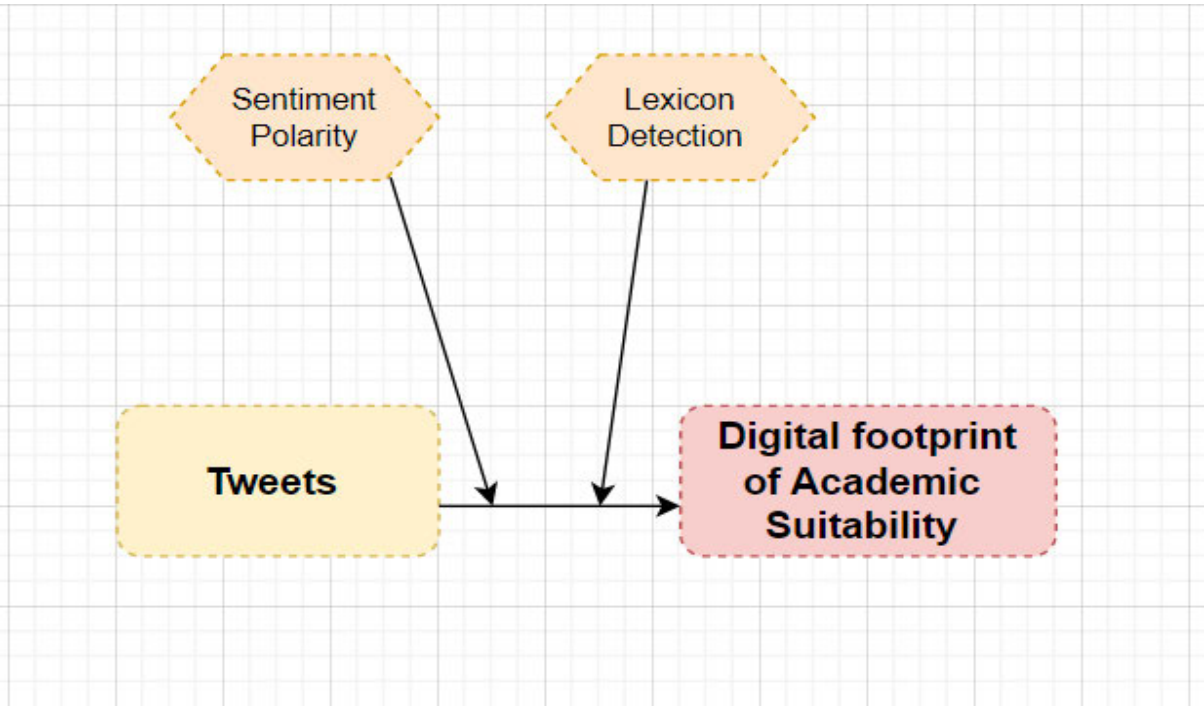

Figure 8.2: Model for predicting academic suitability

8.4 Research limitations

It is worth noting that, due to the pervasiveness of technology, solutions for academics' suitability on social media will continue to evolve. This study's objective was to develop artefacts that could classify academic suitability using Twitter dataset. While the various

experiments conducted in this research were deemed satisfactory, they had a few limitations. The trends of user activity on Twitter may differ from those on other social media platforms, making it difficult to generalise the findings beyond Twitter. However, the dataset used in this study provides a good starting point for solving the generalisation issue, due to structural differences, a model built with this training set is unlikely capable of performing well when using other social media data such as YouTube, Facebook, WhatsApp, or Instagram. Furthermore, while most individuals spend a lot of time and express their actual self on social media, it may be argued that some people display fraudulent behaviour on social media, which can be regarded a drawback in using Twitter dataset.

The techniques used for data pre-processing yielded good results. However, there is room for improvement as the pre-processed techniques used were not optimal for the lack of better techniques. This study used only the text variable in the dataset. It is recommended that the use of more variables available in the dataset such tweet frequency, retweet, and/or follower count to compare the results be explored in future studies. In addition, the scarcity of literature on academic profile, characteristics of academics and the social media analytics methodology is considered as a limitation to this study. The data labelling techniques used in this study yielded good results. However, there is still room for improvement. There is a need for more studies on textual data labelling techniques. The researcher experimented the binary and multiclass classification using the Twitter dataset but was not successful in experimenting the multi label classification with this dataset. This is identified as a limitation of this study. Furthermore, data labelling yielded an imbalance dataset for classification. Due to the imbalanced label of the dataset, obtaining an optimal validation loss score was difficult. Imbalanced dataset is a big limitation in machine learning, and it is highlighted as one of the limitations of this study.

8.5 Contribution to knowledge

Firstly, findings add to the literature by elucidating the relationship between users' digital footprints and the higher education institutions' values/interpersonal skills requirements. Secondly, considering the research objectives that guided this study, findings suggest that a model that tracks academic traits can be included in the higher education institutions' recruitment procedures. This study enhances social media behavioural studies by presenting a process model that may be used in behavioural social media research. The original CUPP framework was expanded into an integrated process model. The process to develop and implement a suitability model was integrated into the CUPP framework developed by Fan and

Gordon (2015). Furthermore, this study has added more models to the fast field of text classification. Artefacts developed can be starting points for more performant models in the human resource department using social media data in the field of recruitment and selection. Moreover, the findings of this study can contribute to the human resources research by adding valuable insights on recruitment and selection using social media data. Human resource can utilise the outcomes of this study to improve the recruitment and selection process and strategies. This study contributes to the scarce literature on academic characteristics, lexicon construction and data labelling.

8.6 Recommendations

The researcher puts forth some recommendations for different entities, namely, researchers, governments, universities, and individuals as follows:

8.6.1 Recommendation for universities, higher education departments, and governments

Various aspects of higher education require changes (Silander and Stigmar 2019). Special attention must be paid to a few essential factors in order to ensure optimal educational quality (Kremer et al. 2013). Governments in both wealthy and developing countries aim to improve education quality. In order to achieve this, educational institutions should have the right personnel. In the light of the latter, in this study the researcher explored new avenues in supplementing and improving the recruitment and selection process of academics. This study demonstrated that there is potential for exploring social media data to supplement the recruitment process. Thus, it is recommended that universities, the higher education department, and governments should provide platforms for more research in this field as it bears possibilities. The implementation of the various research outputs in this field are also recommended. Furthermore, it is recommended that governments should design and implement policies in collaboration with social media platforms owners that allow researchers to effectively collect data on social media for research purposes as these data contains useful information that can improve the effective recruitment of academics.

8.6.2 Recommendation for the human resources departments

Human resource management is being transformed by AI technology, which is delivering new features and redefining the way human resources are managed in organisations. Machine learning, bots, virtual reality, robotics, deep learning, the internet of things, cognitive conversation, natural language processing (NLP) and augmented reality are all examples of AI

technology (AIT) that are transforming the way HR managers work. This study recommends the usage of social media data to be included in human resources' strategies. Human resources can use AI application that uses social media data to improve certain human resources functions such as the recruitment and selection processes. Furthermore, human resources should fully embrace human resource predictive analytics as it has the potential to revolutionise the human resource field.

8.6.3 Future research

As a result of the study's limitations, the following research areas are recommended for future research:

More research may be done to include more training data to discover traces of academics on social media. Furthermore, research can include exploring other variables besides the tweets. In this study, a Twitter dataset was used to demonstrate the utility of the embedding techniques used. However, future research can try out different approaches to evaluate the outcomes. Furthermore, the different classifiers used in this study demonstrated their suitability for the data used. Future studies can use novel classifiers to find traces of academics. Data from different social media platforms can be included as well. While the dataset employed in this study was limited in terms of developing a generalizable classifier, the model's capabilities might be used for additional purposes with a more robust and varied dataset. Traditional questionnaires and unstructured human resource textual data, as well as the digital footprints of various users, could be included in future study.

Additionally, researchers can apply transfer and reinforcement learning using the results of this study in the future. Findings of this study can be a starting point for addressing similar problems in the recruitment and selection processes. Additionally, care should be taken to strike a balance between the exploration of uncharted territory and the utilization of current knowledge regarding the suitability of academics when using the Twitter dataset in future studies. The unsupervised approach should also be considered in finding traces of academic suitability as using approaches of machine learning can yield different results.

8.6.4 Recommendation for job seekers

The fact that social media data, especially Twitter data, can be used to assess the suitability of individuals for specific jobs has been demonstrated. Additionally, a number of organizations are now beginning to incorporate social media data at various levels in their recruitment and

selection strategies. As a result, it will be appropriate for job seekers to use social platforms not only for networking with friends and family but also to market their professional capabilities.

8.7 Summary of the study

Given the fact that social networking sites have become important aspects of people's life, and also since they have grown in popularity, organisations are starting to pay attention to them. The key research question for this study was “How can the Twitter dataset be used to predict the suitability of academics”? This question prompted the researcher to design and implement artefacts for detecting and identifying academic footprints based on tweets for higher institutions. To this end, a model for identifying and detecting markers of academic suitability using tweets was proposed. Binary and multiclass academic suitability prediction models were designed in this study using publicly available Tweets to find digital footprint suitable academics on Twitter. Various deep learning features and approaches were tested in the research to identify a mix that works well for the objectives of this study. The BiLSTM.G.2 (87%) and LSTM.4 (96%) are the two artefacts that achieved high accuracy for the binary classification and multiclass classification. Chapters three and eight described the prediction workflow and the general plan for addressing the research objective. The proposed solution proves to be effective for this study. The experiments determine the efficiency and effectiveness of the methodology employed in this study to classify academic suitability using data from social media as it produced better results than some other studies. Furthermore, this study also provides some recommendations for higher institutions and human resources departments. This study has demonstrated that using social media data (Twitter), recruiters can effectively tackle the issue of negative impression management during the interview sessions. In summary, this study improves to the body of knowledge by incorporating novel approaches to the problem of negative impressions management during interview session in the recruitment and selection process. Given that social media platforms have ingrained themselves into people's lives and have grown in popularity, future research could analyse how other social media data can be used to address the problem of negative impressions management in the recruitment and selection of academics or any other profession. Such studies will help in detecting the applicants’ suitability for job openings.

9. REFERENCES

Abbas, W., & Tap, M. (2019). Adaptively weighted multi-task learning using inverse validation loss. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1408-1412). IEEE.

Abdullah, N. N., & Abdul Rahman, M. (2015). The Degree of Openness in Turkey's Public Expenditure. *International Journal of Administration and Governance*, *12*(1), 8-12.

Abdullah, N. N., & Othman, M. B. (2019). Examining the effects of intellectual capital on the performance of Malaysian food and beverage small and medium-sized enterprises. *Technology (Ijciet)*, *10*(2), 135-143.

Abkenar, S. B., Kashani, M. H., Mahdipour, E., & Jameii, S. M. (2020). Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telematics and Informatics*, 101517.

Adarsh, M. J., & Ravikumar, P. (2018). An effective method of predicting the polarity of airline tweets using sentimental analysis. In *2018 4th International Conference on Electrical Energy Systems (ICEES)* (pp. 676-679). IEEE.

Adesina, O. J., Raimi, S. O., Bolaji, O. A., & Adesina, A. E. (2016). Teachers' attitude, years of teaching experience and self-efficacy as determinants of teachers' productivity in teachers' professional development programme in Ibadan Metropolis, Oyo State, Nigeria. *Journal of Emerging Trends in Educational Research and Policy Studies*, *7*(3), 204-211.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)* (pp. 30-38).

Aggarwal, C. C. (2018). Neural networks and deep learning. *Springer*, *10*, 978-3.

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining text data*, 163-222.

Agostinho, S. P. L., & Mendes-Moreira, J. (2022). Probabilistic Metric to measure the imbalance in multi-class problems. In *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications* (pp. 151-162). PMLR.

Ahmed, M. E., Rabin, M. R. I., & Chowdhury, F. N. (2020). COVID-19: Social media sentiment analysis on reopening. *arXiv preprint arXiv:2006.00804*.

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

Akerkar, R. (2013). Advanced data analytics for business. *Big Data Computing. Boca Raton, FL: Chapman and Hall/CRC*, 377-9.

Aklouche, B., Bounhas, I., & Slimani, Y. (2018). Query Expansion Based on NLP and Word Embeddings. In *TREC*.

Alakus, T. B., & Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. *Chaos, Solitons & Fractals*, *140*, 110120.

Alam, S., & Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, *25*, 319-335.

Albert, E. T. (2019). AI in talent acquisition: a review of AI-applications used in recruitment and selection. *Strategic HR Review*.

Ali, M. N. Y., Sarowar, M. G., Rahman, M. L., Chaki, J., Dey, N., & Tavares, J. M. R. (2019). Adam deep learning with SOM for human sentiment classification. *International Journal of Ambient Computing and Intelligence (IJACI)*, *10*(3), 92-116.

Allport, G. W. (1927). Concepts of trait and personality. *Psychological Bulletin*, *24*(5), 284.

Allport, G. W. (1931). What is a trait of personality?. *The Journal of Abnormal and Social Psychology*, *25*(4), 368.

Almatarneh, S., & Gamallo, P. (2018). Automatic construction of domain-specific sentiment lexicons for polarity classification. In *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection-15th International Conference, PAAMS 2017 15* (pp. 175-182). Springer International Publishing.

Al-Qatf, M., Lasheng, Y., Al-Habib, M., & Al-Sabahi, K. (2018). Deep learning approach combining sparse autoencoder with SVM for network intrusion detection. *Ieee Access*, *6*, 52843-52856.

Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of Twitter data. *International Journal of Advanced Computer Science and Applications*, *10*(2).

Al-Salemi, B., Ayob, M., & Noah, S. A. M. (2018). Feature ranking for enhancing boosting-based multi-label text categorization. *Expert Systems with Applications*, *113*, 531-543.

Alswaina, F., & Elleithy, K. (2018). Android malware permission-based multi-class classification using extremely randomized trees. *IEEE Access*, *6*, 76217-76227.

Alvi, M. (2016). A manual for selecting sampling techniques in research. From https://mpra.ub.uni-muenchen.de/70218/ [12/08/2021]

Alwi, S. K. K., ul Hasan, S. W., & Zaman, S. U. (2022). Internal vs. External Recruitment: The Impact of Operational and Financial Factors. *KASBIT Business Journal*, *15*(2), 115-129.

Amaral, A. A., Powell, D. M., & Ho, J. L. (2019). Why does impression management positively influence interview ratings? The mediating role of competence and warmth. *International Journal of Selection and Assessment*, *27*(4), 315-327.

Anand, M., & Eswari, R. (2019). Classification of abusive comments in social media using deep learning. In *2019 3rd international conference on computing methodologies and communication (ICCMC)* (pp. 974-977). IEEE.

Anderson, H. S., & Roth, P. (2018). Ember: an open dataset for training static pe malware machine learning models. *arXiv preprint arXiv:1804.04637*.

Anderson, V. (2015). Research methods in human resource management-Investigating a business issue. *CIPD Publishing*.

Andriopoulou, P., & Prowse, A. (2020). Towards an effective supervisory relationship in research degree supervision: insights from attachment theory. *Teaching in Higher Education*, *25*(5), 648-661.

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: why HR is set to fail the big data challenge. *Human Resource Management Journal*, *26*(1), 1-11.

Ansari, S., Du, H., & Naghdy, F. (2020). Driver's Foot Trajectory Tracking for Safe Maneuverability Using New Modified reLU-BiLSTM Deep Neural Network. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 4392-4397). IEEE.

Anwar, K., & Ghafoor, C. (2017). Knowledge management and organizational performance: A study of private universities in Kurdistan. International Journal of Social Sciences & Educational Studies, 4(2), 53.

Aragao, R., & El-Diraby, T. E. (2021). Network analytics and social BIM for managing project unstructured data. *Automation in Construction*, *122*, 103512.

Argyris, C., Putman, R., & Smith, D. M. (1985). Action science (Vol. 13). *Jossey-bass*.

Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index-insights from Facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, *49*, 86-101.

Asendorpf, J. B. (2009). Personality: Traits and situations. *The Cambridge handbook of personality psychology*, 43-53.

Ashqar, B. A., & Abu-Naser, S. S. (2019). Identifying images of invasive hydrangea using pre-trained deep convolutional neural networks. *International Journal of Academic Engineering Research (IJAER)*, *3*(3), 28-36.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507-547). University of Chicago Press.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, *38*, 100311.

Azure, J. (2016). Students' perspective of effective supervision of graduate programme in Ghana. *American Journal o0f Educational Research*, *4*(2), 163.

Babak, V. P., Babak, S. V., Myslovych, M. V., Zaporozhets, A. O., Zvaritch, V. M., & Zvaritch, V. M. (2020). Methods and models for information data analysis. *Diagnostic Systems For Energy Equipments*, 23-70.

Babbie, E. (2015). *Observing ourselves: Essays in social research*. Waveland Press.

Babbie, E. R. (2020). *The practice of social research (5th ed.)*. Cengage learning.

Bagui, S., & Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, *8*(1), 1-41.

Balas, V. E., & Fodor, J. (2013). *New Concepts and Applications in Soft Computing*. A. R. Várkonyi-Kóczy (Ed.). Springer.

Banach, C. (1995). Cechy osobowościowe nauczycieli. *Nowa Szkoła*, *3*(10).

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., & Chowell, G. (2021). A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, *2*(3), 315-324.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology*, *33*(4), 445-459.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2022). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Key Topics in Psychological Methods*, 125-139.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., Frear, K. A., Berka, G., Heggestad, E. D., & Gordon, H. L. (2019). Strategic recruitment across borders: An investigation of multinational enterprises. *Journal of Management*, *45*(2), 476-509.

Barnaghi, P., Ghaffari, P., & Breslin, J. G. (2016). Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In *2016 IEEE second international conference on big data computing service and applications (BigDataService)* (pp. 52-57). IEEE.

Barnes, B. J., Williams, E. A., & Archer, S. A. (2010). Characteristics that matter most: Doctoral students' perceptions of positive and negative advisor attributes. *Nacada Journal*, *30*(1), 34-46.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, *44*(1), 1-26.

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, *94*(6), 1394.

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, *94*(6), 1394.

Bassi, L., & McMurrer, D. (2015). A Quick Overview of HR Analytics: Why, What, How, and When?. *Association for talent development*. From https://www.td.org/insights/a-quick-overview-of-hr-analytics-why-what-how-and-when [12/08/2022]

Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M. F., Schuller, B., & André, E. (2020). Explainable cooperative machine learning with NOVA. *KI-Künstliche Intelligenz*, 1-22.

Becker, B. E., Huselid, M. A., Huselid, M. A., & Ulrich, D. (2001). *The HR scorecard: Linking people, strategy, and performance*. Harvard Business Press.

Becker, B., & Gerhart, B. (1996). The impact of human resource management on organizational performance: Progress and prospects. *Academy of management journal*, *39*(4), 779-801.

Behera, B., Kumaravelan, G., & Kumar, P. (2019). Performance evaluation of deep learning algorithms in biomedical document classification. In *2019 11th International Conference on Advanced Computing (ICoAC)* (pp. 220-224). IEEE.

Belinkov, Y., Lei, T., Barzilay, R., & Globerson, A. (2014). Exploring compositional architectures and word vector representations for prepositional phrase attachment. *Transactions of the Association for Computational Linguistics*, *2*, 561-572.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, 45-59.

Berger, A., & Guda, S. (2020). Threshold optimization for F measure of macro-averaged precision and recall. *Pattern Recognition*, *102*, 107250.

Berrar, D. (2019). Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology, Volume 1, Elsevier*, pp. 542–545.

Berry, A. B., Petrin, R. A., Gravelle, M. L., & Farmer, T. W. (2011). Issues in special education teacher recruitment, retention, and professional development: Considerations in supporting rural teachers. *Rural Special Education Quarterly*, *30*(4), 3-11.

Best, A., Greenhalgh, T., Lewis, S., Saul, J. E., Carroll, S., & Bitz, J. (2012). Large-system transformation in health care: a realist review. *The Milbank Quarterly*, *90*(3), 421-456.

Bigsby, K. G., Ohlmann, J. W., & Zhao, K. (2019). Keeping it 100: social media and self-presentation in college football recruiting. *Big data*, *7*(1), 3-20.

Bigsby, K. G., Ohlmann, J. W., & Zhao, K. (2019). The turf is always greener: Predicting decommitments in college football recruiting using Twitter data. *Decision Support Systems*, *116*, 1-12.

Bikmukhametov, T., & Jäschke, J. (2020). Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Computers & Chemical Engineering*, *138*, 106834.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Bolino, M., Long, D., & Turnley, W. (2016). Impression management in organizations: Critical questions, answers, and areas for future research. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 377-406.

Bollegala, D., Alsuhaibani, M., Maehara, T., & Kawarabayashi, K. I. (2016). Joint word representation learning using a corpus and a semantic lexicon. In *Proceedings of the AAAI Conference on Artificial Intelligence* 30(1).

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications*, *162*, 113746.

Bourdage, J. S., Roulin, N., & Tarraf, R. (2018). "I (might be) just that good": Honest and deceptive impression management in employment interviews. *Personnel Psychology*, *71*(4), 597-632.

Bourdage, J. S., Schmidt, J., Wiltshire, J., Nguyen, B., & Lee, K. (2020). Personality, interview performance, and the mediating role of impression management. *Journal of Occupational and Organizational Psychology*, *93*(3), 556-577.

Bowen, D. E., & Ostroff, C. (2004). Understanding HRM–firm performance linkages: The role of the "strength" of the HRM system. *Academy of management review*, *29*(2), 203-221.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679.

Boyd, R. (2016). General use: RIOT scan. https://riot.ryanb.cc/general-use/ [15/04/2022]

Bozeman, D. P., & Kacmar, K. M. (1997). A cybernetic model of impression management processes in organizations. *Organizational behavior and human decision processes*, *69*(1), 9-30.

Brink, K. E., & Costigan, R. D. (2015). Oral communication skills: Are the priorities of the workplace and AACSB-accredited business programs aligned?. *Academy of Management Learning & Education*, *14*(2), 205-221.

Brown, P., Von Daniels, C., Bocken, N. M. P., & Balkenende, A. R. (2021). A process model for collaboration in circular oriented innovation. *Journal of Cleaner Production*, *286*, 125499.

Brown, S. (2021). Machine Learning explained. [02/06/2021] https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *Journal of Big data*, *5*(1), 1-10.

Buehl, A. K., & Melchers, K. G. (2017). Individual difference variables and the occurrence and effectiveness of faking behavior in interviews. *Frontiers in Psychology*, *8*, 686.

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction*, *3*(CSCW), 1-24.

Camacho, J., Smilde, A. K., Saccenti, E., & Westerhuis, J. A. (2020). All sparse PCA models are wrong, but some are useful. Part I: computation of scores, residuals and explained variance. *Chemometrics and Intelligent Laboratory Systems*, *196*, 103907.

Cambria, E., Rajagopal, D., Olsher, D., & Das, D. (2013). Big social data analysis. *Big data computing*, *13*, 401-414.

Cambria, E., Wang, H., & White, B. (2014). Guest editorial: Big social data analysis. *knowledge–based systems, (69), 1-2.*

Cao, J., Su, Z., Yu, L., Chang, D., Li, X., & Ma, Z. (2018). Softmax cross entropy loss with unbiased decision boundary for image classification. In *2018 Chinese Automation Congress (CAC)* (pp. 2028-2032). IEEE.

Cao, S., & Lu, W. (2017). Improving word embeddings with convolutional feature learning and subword information. In *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1).

Cappelli, P., & Tavis, A. (2018). HR goes agile. *Harvard Business Review*, *96*(2), 46-52.

Cappelli, P., Tambe, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward (April 8, 2019). *Saatavilla internet-osoitteessa: http://dx. doi. org/10.2139/ssrn*, *3263878*.

Cappelli, P., Tambe, P., & Yakubovich, V. (2020). Can data science change human resources?. *The future of management in an AI world: Redefining purpose and strategy in the fourth industrial revolution*, 93-115.

CareerBuilder. (2012). Thirty- seven percent of companies use social networks to research potential job candidates. [18/11/2018] Retrieved from https://www.careerbuilder.com/share/aboutus/pressreleasesdetail.aspx?id=pr691&sd=4%2F18%2F2012&ed=4%2F18%2F2099

CareerBuilder. (2016). Number of Employers Using social media to screen candidates has increased 500 percent over the last decade. [18/06/2019] Retrieved from http://www.careerbuilder.com/share/aboutus/pressreleasesdetail.aspx?ed=12%2F31%2F2016&id=pr945&sd=4%2F28%2F2016

Carifio, M. S., & Hess, A. K. (1987). Who is the ideal supervisor?. *Professional Psychology: Research and Practice*, *18*(3), 244.

Carrillo-Larco, R. M., & Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. *Wellcome Open Research*, *5*.

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. *Physica Medica*, *83*, 9-24.

Chapman, C., & Stolee, K. T. (2016). Exploring regular expression usage and context in Python. In *Proceedings of the 25th International Symposium on Software Testing and Analysis* (pp. 282-293).

Chen, D. Y. (2017). *Pandas for everyone: Python data analysis*. Addison-Wesley Professional.

Chen, J., Cheng, C., Collins, L., Chharbria, P., & Cheong, H. (2018). The Rise of Analytics in HR: The era of talent intelligence is here. *LinkedIn Report*.

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review*, *31*(1), 100698.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

Chiresche, R. (2017). "Research Supervision: Postgraduate Students' Experiences in South Africa." Journal of Social Sciences 31 (2): 229–234. doi:10.1080/09718923.2012.11893032.

Chitra, S., & Srivaramangai, P. (2018). A Study on Analytics of Human Resource Management in Big Data. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *3*(3), 58-68.

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th workshop on biomedical natural language processing* (pp. 166-174).

Chong, E. S., Zhang, Y., Mak, W. W., & Pang, I. H. (2015). Social media as social capital of LGB individuals in Hong Kong: Its relations with group membership, stigma, and mental well-being. *American journal of community psychology*, *55*, 228-238.

Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., & Lin, C. (2012). Alzheimer's Disease Neuroimaging Initiative: does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, *60*(1), 59-70.

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, *7*, 24634-24648.

Clark, K. R., & Vealé, B. L. (2018). Strategies to enhance data collection and analysis in qualitative research. *Radiologic technology*, *89*(5), 482CT-485CT.

Colah. (2015). Understanding LSTM Networks. [26/12/2021] retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Collis, J., & Hussey, R. (2014). Writing up the Research. In *Business Research* (pp. 297-330). Palgrave, London.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Dang, Y., Zhang, Y., & Chen, H. (2009). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, *25*(4), 46-53.

Dansana, D., Kumar, R., Bhattacharjee, A., Hemanth, D. J., Gupta, D., Khanna, A., & Castillo, O. (2020). Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm. *Soft computing*, 1-9.

Dany, F., & Torchy, V. (2017). Recruitment and selection in Europe Policies, practices and methods 1. In *Policy and practice in European human resource management* (pp. 68-88). Routledge.

Das, K. G., & Das, D. (2017). Developing lexicon and classifier for personality identification in texts. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)* (pp. 362-372).

Dastin, J. (2018). Amazon scraps secret recruiting AI tool that showed against women. [18/07/2023]. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davis, D. (2020). The ideal supervisor from the candidate's perspective: what qualities do students actually want?. *Journal of Further and Higher Education*, *44*(9), 1220-1232.

Davis, J. C. (2019). Rethinking Regex engines to address ReDoS. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 1256-1258).

Davison, H. K., Maraist, C. C., Hamilton, R. H., & Bing, M. N. (2012). To screen or not to screen? Using the internet for selection decisions. *Employee Responsibilities and Rights Journal*, *24*, 1-21.

Deloitte, L. L. P. (2017). Rewriting the rules for the digital age: 2017 Deloitte global human capital trends.

Deloitte Global Human Capital Trends (2018). *The Rise of the Social Enterprise*. Available from https://www2.deloitte.com/content/dam/insights/us/articles/HCTrends2018

Deng, Q., & Ji, S. (2018). A review of design science research in information systems: concept, process, outcome, and evaluation. *Pacific Asia journal of the association for information systems*, *10*(1), 2.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168-189.

Department of Higher Education and training. (2016). International educationists collaborate to share solutions to common problems. [23/08/2028]. From https://www.dhet.gov.za/SiteAssets/Latest%20News/Independent%20Thinking%208th%20Edition/TrainingP6.pdf

Desai, R. D. (2018). Sentiment analysis of Twitter data. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 114-117). IEEE.

Desmond, M., Duesterwald, E., Brimijoin, K., Brachman, M., & Pan, Q. (2021). Semi-Automated Data Labeling. In *NeurIPS 2020 Competition and Demonstration Track* (pp. 156-169). PMLR.

Desmond, M., Muller, M., Ashktorab, Z., Dugan, C., Duesterwald, E., Brimijoin, K., ... & Pan, Q. (2021). Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In *26th International Conference on Intelligent User Interfaces* (pp. 392-401).

DeVaro, J. (2020). Internal hiring or external recruitment?. *IZA World of Labor*. From https://wol.iza.org/articles/internal-hiring-or-external-recruitment/long [14/10/2021]

Dharma, E. M., Gaol, F. L., Warnars, H. L. H. S., & Soewito, B. (2022). The Accuracy Comparison Among Word2vec, Glove, And Fasttext Towards Convolution Neural Network (CNN) Text Classification. *Journal of Theoretical and Applied Information Technology*, *100*(2).

Dia, I., Ahvar, E., & Lee, G. M. (2022). Performance evaluation of machine learning and neural network-based algorithms for predicting segment availability in AIoT-based smart parking. *Network*, *2*(2), 225-238.

Djabatey, E. N. (2012). Recruitment and selection practices of organizations: A case study of HFC Bank (GH) Ltd. *Unpublished thesis submitted to the Institute of Distance Learning, Kwame Nkrumah University of Science and Technology. Ghana: Kwame Nkrumah University of Science and Technology*.

Djabatey, E. N. (2012). Recruitment and selection practices of organizations. *A case study of HFC Bank (GH) Ltd., Degree of Master Thesis, Kwame Nkrumah University of Science and Technology, Ghana*.

Dogan, O., Martinez-Millana, A., Rojas, E., Sepúlveda, M., Munoz-Gama, J., Traver, V., & Fernandez-Llatas, C. (2019). Individual behavior modeling with sensors using process mining. *Electronics*, *8*(7), 766.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International journal of information management*, *48*, 63-71.

Duggan, J., Sherman, U., Carbery, R., & McDonnell, A. (2020). Algorithmic management and app-work in the gig economy: A research agenda for employment relations and HRM. *Human Resource Management Journal*, *30*(1), 114-132.

Dutta, H. S., Chetan, A., Joshi, B., & Chakraborty, T. (2018). Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 242-249). IEEE.

Dzisevič, R., & Šešok, D. (2019). Text classification using different feature extraction approaches. In *2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)* (pp. 1-4). IEEE.

Ebneyamini, S., & Sadeghi Moghadam, M. R. (2018). Toward developing a framework for conducting case study research. *International journal of qualitative methods*, *17*(1), DOI:10.1177/1609406918817954.

Edwards, M. R. (2018). HR metrics and analytics. In *e-HRM* (pp. 89-105). Routledge.

Edwards, M. R., & Edwards, K. (2019). *Predictive HR analytics: Mastering the HR metric*. Kogan Page Publishers.

Egger, R. (2022). Text Representations and Word Embeddings. In *Applied Data Science in Tourism* (pp. 335-361). Springer, Cham.

Ekawati, A. D. (2019). Predictive analytics in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences*, *22*(4), 387-397.

Ekwoaba, J. O., Ikeije, U. U., & Ufoma, N. (2015). The Impact of Recruitment and Selection Criteria on Organizational Performance.

El Ouirdi, M., El Ouirdi, A., Segers, J., & Pais, I. (2016). Technology adoption in employee recruitment: The case of social media in Central and Eastern Europe. *Computers in human behavior*, *57*, 240-249.

Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 122, p. 16).

Ellis, A. P., West, B. J., Ryan, A. M., & DeShon, R. P. (2002). The use of impression management tactics in structured interviews: A function of question type?. *Journal of Applied Psychology*, *87*(6), 1200.

Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. *Information Processing & Management*, *57*(1), 102121.

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An introductory review of deep learning for prediction models with big data. *Frontiers in Artificial Intelligence*, *3*, 4.

Erez, M., & Shneorson, Z. (1980). Personality types and motivational characteristics of academics versus professionals in industry in the same occupational discipline. *Journal of Vocational Behavior*, *17*(1), 95-105.

Erickson, B. J., & Kitamura, F. (2021). Magician's Corner: 9. Performance Metrics for Machine Learning Models. *Radiology: Artificial Intelligence*, *3*(3), e200126.

Esposito, M., Damiano, E., Minutolo, A., De Pietro, G., & Fujita, H. (2020). Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Information Sciences*, *514*, 88-105.

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

Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Elaziz, M. A., ... & Al-qaness, M. A. (2021). Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. *Electronics*, *10*(11), 1332.

Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM SIGKDD explorations newsletter*, *14*(2), 1-5.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, *57*(6), 74-81.

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4647-4657).

Fayyad, U. M., Piatetsky-Shapiro, G., & Uthurusamy, R. (2003). Summary from the KDD-03 panel: data mining: the next 10 years. *ACM Sigkdd Explorations Newsletter*, *5*(2), 191-196.

Fernandez, J. (2019). The ball of wax we call HR analytics. *Strategic HR Review*.

Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., & González-Castano, F. J. (2018). Creating emoji lexica from unsupervised sentiment analysis of their descriptions. *Expert Systems with Applications*, *103*, 74-91.

Ferris, G. R., Hochwarter, W. A., Buckley, M. R., Harrell-Cook, G., & Frink, D. D. (1999). Human resources management: Some new directions. *Journal of management*, *25*(3), 385-415.

Fitz-Enz, J. (1984). How to measure human resources management. McGraw-Hill. 1-237.

Fitz-Enz, J., & John Mattox, I. I. (2014). *Predictive analytics for human resources*. John Wiley & Sons.

Flach, P. (2019). Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9808-9814).

Fredriksson, T., Mattos, D. I., Bosch, J., & Olsson, H. H. (2020, November). Data labeling: an empirical investigation into industrial challenges and mitigation strategies. In *International Conference on Product-Focused Software Process Improvement* (pp. 202-216). Springer, Cham.

Fullan, M., & Scott, G. (2009). *Turnaround leadership for higher education*. John Wiley & Sons.

Gaddis, B. H., & Foster, J. L. (2015). Meta-analysis of dark side personality characteristics and critical work behaviors among leaders across the globe: Findings and implications for leadership development and executive coaching. *Applied Psychology*, *64*(1), 25-54.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

Gardner, W. L., & Martinko, M. J. (1988). Impression management in organizations. *Journal of management*, *14*(2), 321-338.

Ghani, N. A., Hamid, S., Hashem, I. A. T., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior*, *101*, 417-428.

Ghazi, A. N., Petersen, K., Reddy, S. S. V. R., & Nekkanti, H. (2018). Survey research in software engineering: Problems and mitigation strategies. *IEEE Access*, *7*, 24703-24718.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215-230.

Goffman, E. (2002). The presentation of self in everyday life. 1959. *Garden City, NY*, *259*.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing* (pp. 149-156). IEEE.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Goundar, S. (2012). Research methodology and research method. *Victoria University of Wellington*.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354-377.

Gujjar, J. P., & Kumar, H. P. (2021). Sentiment analysis: Textblob for decision making. *Int. J. Sci. Res. Eng. Trends*, *7*(2), 1097-1099.

Gull, R., Shoaib, U., Rasheed, S., Abid, W., & Zahoor, B. (2016). Pre-processing of twitter's data for opinion mining in political context. *Procedia Computer Science*, *96*, 1560-1570.

Guo, H., Nguyen, H., Vu, D. A., & Bui, X. N. (2021). Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach. *Resources Policy*, *74*, 101474.
Gupta, G., & Malhotra, S. (2015). Text document tokenization for word frequency count using rapid miner (taking resume as an example). *International Journal of Computer Applications*, *975*, 8887.

Gupta, G., & Malhotra, S. (2015). Text document tokenization for word frequency count using rapid miner (taking resume as an example). *Int. J. Comput. Appl*, *975*, 8887.

Halbert, K. (2015). Students' perceptions of a 'quality'advisory relationship. *Quality in Higher Education*, *21*(1), 26-37.

Hameed, A. A., & Anwar, K. (2018). Analyzing the Relationship between Intellectual Capital and Organizational Performance: A Study of Selected Private Banks in Kurdistan. *International Journal of Social Sciences & Educational Studies*, *4*(4), 39.

Hamilton, R. H., & Sodeman, W. A. (2020). The questions we ask: Opportunities and challenges for using big data analytics to strategically manage human capital resources. *Business Horizons*, *63*(1), 85-95.

Hamza, P. A., Othman, B. J., Gardi, B., Sorguli, S., Aziz, H. M., Ahmed, S. A., ... & Anwar, G. (2021). Recruitment and Selection: The Relationship between Recruitment and Selection with Organizational Performance. *International Journal of Engineering, Business and Management*, *5*(3), 1-13.

Han, S., Huang, H., & Tang, Y. (2020). Knowledge of words: An interpretable approach for personality recognition from social media. *Knowledge-Based Systems*, *194*, 105550.

Hao, J., & Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348-361.

Haralabopoulos, G., Anagnostopoulos, I., & McAuley, D. (2020). Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms*, *13*(4), 83.

Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), 1-16.

Harrison, J. T., & Hartwell, C. J. (2022). HR machine learning on social media data. In *Handbook of Research on Artificial Intelligence in Human Resource Management* (pp. 89-104). Edward Elgar Publishing.

Hart, C. (2018). Doing a literature review: Releasing the research imagination. 2nd Ed. *SAGE Publication Ltd.*

Haryanto, A. W., & Mawardi, E. K. (2018, September). Influence of word normalization and chi-squared feature selection on support vector machine (svm) text classification. In *2018 International seminar on application for technology of information and communication* (pp. 229-233). IEEE.

Hasan, A. (2021). Ethical considerations in the use of secondary data for built environment research. In *Secondary Research Methods in the Built Environment* (pp. 26-39). Routledge.

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11.

He, H., & Lin, J. (2016). Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 conference of the north American chapter of the Association for Computational Linguistics: human language technologies* (pp. 937-948).

He, Z. F., Yang, M., Gao, Y., Liu, H. D., & Yin, Y. (2019). Joint multi-label classification and label correlations with missing labels and feature selection. *Knowledge-Based Systems*, *163*, 145-158.

Heap, V., & Waters, J. (2019). Data collection methods. In *Mixed Methods in Criminology* (pp. 141-176). Routledge.

Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2018). Requirements for data quality metrics. *Journal of Data and Information Quality (JDIQ)*, *9*(2), 1-32.

Hermansyah, R., & Sarno, R. (2020). Sentiment Analysis about Product and Service Evaluation of PT Telekomunikasi Indonesia Tbk from Tweets Using TextBlob, Naive Bayes & K-NN Method. In *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)* (pp. 511-516). IEEE.

Herschberg, C., Benschop, Y., & Van den Brink, M. (2018). Precarious postdocs: A comparative study on recruitment and selection of early-career researchers. *Scandinavian Journal of Management*, *34*(4), 303-310.

Hevner, A. R., March, S. T., Park, J., Ram, S., & Ram, S. (2004). Research essay design science in information. *MIS Q*, *28*(1), 75-105.

Hevner, A., & Chatterjee, S. (2010). *Design research in information systems. Theory and practice*. Springer.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 1094428120971683.

Hino, A., & Fahey, R. A. (2019). Representing the Twittersphere: Archiving a representative sample of Twitter data under resource constraints. *International journal of information management*, *48*, 175-184.

Hiraoka, T., Shindo, H., & Matsumoto, Y. (2019). Stochastic tokenization with a language model for neural text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1620-1629).

Hoffmann, S., Wiben, A., Kruse, M., Jacobsen, K. K., Lembeck, M. A., & Holm, E. A. (2020). Predictive validity of PRISMA-7 as a screening instrument for frailty in a hospital setting. *BMJ open*, *10*(10), e038768.

Holland, J. L. (1966). The psychology of vocational choice: A theory of personality types and model environments.

Holsapple, C. W., Hsiao, S. H., & Pakath, R. (2018). Business social media analytics: Characterization and conceptual framework. *Decision Support Systems*, *110*, 32-45.

Hosain, S., Manzurul Arefin, A. H. M., & Hossin, M. (2020). E-recruitment: A social media perspective. *Asian Journal of Economics, Business and Accounting*, *16*(4), 51-62.

HR Outlook. (2016-2017). Views of our profession. [30/06/2021]. Available From https://www.cipd.co.uk/Images/hr-outlook_2017_tcm18-17697.pdf

Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems* (pp. 2042-2050).

Hu, Y., Ni, J., & Wen, L. (2020). A hybrid deep learning approach by integrating LSTM-ANN networks with GARCH model for copper price volatility prediction. *Physica A: Statistical Mechanics and its Applications*, *557*, 124907.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

Huang, J., Li, Y. F., & Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, *67*, 108-127.

Huang, M., Xie, H., Rao, Y., Feng, J., & Wang, F. L. (2020). Sentiment strength detection with a context-dependent lexicon-based convolutional neural network. *Information Sciences*, *520*, 389-399.

Huffcutt, A. I., Van Iddekinge, C. H., & Roth, P. L. (2011). Understanding applicant behavior in employment interviews: A theoretical model of interviewee performance. *Human Resource Management Review*, *21*(4), 353-367.

Hugo, G. (2005). Demographic trends in Australia's academic workforce. *Journal of Higher Education Policy and Management*, *27*(3), 327-343.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Imane, A., & Mohamed, B. A. (2017). Multi-label categorization of french death certificates using nlp and machine learning. In *Proceedings of the 2nd international conference on big data, cloud and applications* (pp. 1-4).

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2018). Processing social media messages in mass emergency: Survey summary. In *Companion Proceedings of the The Web Conference 2018* (pp. 507-511).

Ingold, P. V., Kleinmann, M., König, C. J., & Melchers, K. G. (2015). Shall we continue or stop disapproving of self-presentation? Evidence on impression management and faking in a selection context and their relation to job performance. *European Journal of Work and Organizational Psychology*, *24*(3), 420-432.

InsiderIntelligence. (2022). Twitter in 2022: Global user statistics, demographics and marketing trends to know https://www.insiderintelligence.com/insights/twitter-user-statistics-trends/ [09/09/2022]

Isson, J. P., & Harriott, J. S. (2016). *People analytics in the era of big data: Changing the way you attract, acquire, develop, and retain talent*. John Wiley & Sons.

Izatovich, B. I. (2021). Development of a stemming algorithm based on a linguistic approach for words of the uzbek language. In *E-Conference Globe* (pp. 195-202).

Jabir, B., Falih, N., & Rahmani, K. (2019). HR analytics a roadmap for decision making: case study. *Indonesian Journal of Electrical Engineering and Computer Science*, *15*(2), 979-990.

Jansson, P., & Liu, S. (2017). Distributed representation, LDA topic modelling and deep learning for emerging named entity recognition from social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text* (pp. 154-159).

Japkowicz, N., & Shah, M. (2015). Performance evaluation in machine learning. *Machine Learning in Radiation Oncology: Theory and Applications*, 41-56.

Javid, A. M., Das, S., Skoglund, M., & Chatterjee, S. (2021, June). A relu dense layer to improve the performance of neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2810-2814). IEEE.

Ji, D., Tao, P., Fei, H., & Ren, Y. (2020). An end-to-end joint model for evidence information extraction from court record document. *Information Processing & Management*, *57*(6), 102305.

Ji, Y., Cohn, T., Kong, L., Dyer, C., & Eisenstein, J. (2015). Document context language models. *arXiv preprint arXiv:1511.03962*.

Jia, Q., Guo, Y., Li, R., Li, Y., & Chen, Y. (2018). A conceptual artificial intelligence application framework in human resource management. In *Proceedings of the International Conference on Electronic Business* (pp. 106-114).

Jiang, A. (2019). QMUL-NLP at HASOC 2019: offensive content detection and classification in social media.

Jiang, H., Song, Y., Wang, C., Zhang, M., & Sun, Y. (2017). Semi-supervised Learning over Heterogeneous Information Networks by Ensemble of Meta-graph Guided Random Walks. In *IJCAI* (pp. 1944-1950).

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: a brief primer. *Behavior Therapy*, *51*(5), 675-687.

Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE access*, *5*, 2870-2879.

Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, *6*, 23253-23260.

Johansson, J., & Herranen, S. (2019). The application of artificial intelligence (AI) in human resource management: Current state of AI and its impact on the traditional recruitment process.

Johansson, J., & Herranen, S. (2019). The application of artificial intelligence in human resources management. *Business Administration Thesis, Jonkoping University, Sweden*.

Jones, E. E. (1964). Ingratiation : A social psychological analysis. New York: Appleton-Century-Crofts.

Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Juba, B., & Le, H. S. (2019). Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 4039-4048).

Judge T.A., Higgins C.A., Thoresen C.J., & Barrick M.R. (2006). The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology*, *52*(3), 621-652.

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology*, *52*(3), 621-652.

Jurafsky, D., & Martin, J. H. (2014). N-grams. *Speech and Language Processing*, 1-28.

Kadhim, A. I. (2018). An evaluation of preprocessing techniques for text classification. *International Journal of Computer Science and Information Security (IJCSIS)*, *16*(6), 22-32.

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273-292.

Kagalkar, A., & Raghuram, S. (2020). CORDIC based implementation of the softmax activation function. In *2020 24th International Symposium on VLSI Design and Test (VDAT)* (pp. 1-4). IEEE.

Kalra, V., & Aggarwal, R. (2017). Importance of Text Data Preprocessing & Implementation in RapidMiner. *ICITKM*, *14*, 71-75.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and electronics in agriculture, 147, 70-90.

Kassu, J. S. (2019). Research Design and Methodology, Cyberspace, Evon Abu-Taieh, Abdelkrim El Mouatasim and Issam H. Al Hadid, IntechOpen, DOI: 10.5772/intechopen. 85731.

Kathuria, R. S., Gautam, S., Singh, A., Khatri, S., & Yadav, N. (2019). Real time sentiment analysis on twitter data using deep learning (Keras). In *2019 international conference on computing, communication, and intelligent systems (ICCCIS)* (pp. 69-73). IEEE.

Kehinde, J.S. (2012). Talent management: Effect on organizational performance. *Journal of management research*, *4*(2), 76-88.

Kelleher, J. D. (2019). *Deep learning*. MIT press.

Kerey, A, B. (2021). AI in recruitment: an exploratory study into the factors that impact its pace of adoption. Master dissertation. *UPPSALA UNIVERSITET*.

Kern, M. L., McCarthy, P. X., Chakrabarty, D., & Rizoiu, M. A. (2019). Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences*, *116*(52), 26459-26464.

Ketkar, N. (2017). Introduction to keras. In *Deep learning with Python* (pp. 97-111). Apress, Berkeley, CA.

Khalaf, M., Hussain, A. J., Alafandi, O., Al-Jumeily, D., Alloghani, M., Alsaadi, M., ... & Abd, D. H. (2019). An application of using support vector machine based on classification technique for predicting medical data sets. In *International Conference on Intelligent Computing* (pp. 580-591). Springer, Cham.

Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H. (2016). Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling*, *4*, 1-19.

Kharwal, A. (2021). Mini-batch K-means Clustering in Machine Learning. The Clever Programmer, September, 10.

Kim Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, p. 1746–1751

Kim, J. W. (2018). Rumor has it: The effects of virality metrics on rumor believability and transmission on Twitter. *New Media & Society*, *20*(12), 4807-4825.

Kim, T. Y., & Cho, S. B. (2019). Predicting residential energy consumption using CNN-LSTM neural networks. *Energy*, *182*, 72-81.

King, N., Horrocks, C., & Brooks, J. (2018). *Interviews in qualitative research*. sage.

Kirsch, A., Van Amersfoort, J., & Gal, Y. (2019). Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, *32*.

Kissoonduth, K. (2017). Talent Management: Attracting and retaining academic staff at selected public higher education institutions. *Pretoria: Department of Public Administration, University of Unisa*.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*, *79*, 1-14.

Ko, H., Chung, H., Kang, W. S., Kim, K. W., Shin, Y., Kang, S. J., ... & Lee, J. (2020). COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation. *Journal of medical Internet research*, *22*(6), e19569.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational research methods*, *21*(3), 733-765.

Koch, T., Gerber, C., & De Klerk, J. J. (2018). The impact of social media on recruitment: Are you LinkedIn?. *SA Journal of Human Resource Management*, *16*(1), 1-14.

Koenig, T. W., Parrish, S. K., Terregino, C. A., Williams, J. P., Dunleavy, D. M., & Volsch, J. M. (2013). Core personal competencies important to entering students' success in medical school: what are they and how could they be assessed early in the admission process?. *Academic Medicine*, *88*(5), 603-613.

Köhn, A. (2015). What's in an embedding? Analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2067-2073).

Kooij, D. T., Zacher, H., Wang, M., & Heckhausen, J. (2020). Successful aging at work: A process model to guide future research and practice. *Industrial and Organizational Psychology*, *13*(3), 345-365.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, *110*(15), 5802-5805.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*(6130), 297-300.

Kristof-Brown, A. L., Jansen, K. J., & Colbert, A. E. (2002). A policy-capturing study of the simultaneous effects of fit with jobs, groups, and organizations. *Journal of Applied psychology*, *87*(5), 985.

Krouska, A., Troussas, C., & Virvou, M. (2016). The effect of preprocessing techniques on Twitter sentiment analysis. In *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1-5). IEEE.

Krumeich, J., Werth, D., & Loos, P. (2016). Prescriptive control of business processes: new potentials through predictive analytics of big data in the process manufacturing industry. *Business & Information Systems Engineering*, *58*, 261-280

Kryscynski, D., Reeves, C., Stice-Lusvardi, R., Ulrich, M., & Russell, G. (2017). Analytical abilities and the performance of HR professionals. *Human Resource Management*, *57*(3), 715-738.

Kumar, A., & Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, *32*(1), e5107.

Kumar, R. (2018). *Research methodology: A step-by-step guide for beginners*. Sage.

Kunte, A. V., & Panicker, S. (2019). Using textual data for personality prediction: a machine learning approach. In *2019 4th international conference on information systems and computer networks (ISCON)* (pp. 529-533). IEEE.

Kursuncu, U., Gaur, M., Lokala, U., Thirunarayan, K., Sheth, A., & Arpinar, I. B. (2019). Predictive analysis on Twitter: Techniques and applications. In *Emerging research challenges and opportunities in computational social network analysis and mining* (pp. 67-104). Springer, Cham.

Kuyumcu, B., Aksakalli, C., & Delil, S. (2019). An automated new approach in fast text classification (fastText) A case study for Turkish text classification without pre-processing. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval* (pp. 1-4).

Ladimeji, K. (2013). Things that HR predictive analytics will actually predict. *Recruiter sec. I, 1*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Landers, R. N., & Schmidt, G. B. (2016). Social media in employee selection and recruitment: An overview. *Social media in employee selection and recruitment: Theory, practice, and current challenges*, 3-11.

Langer, M., König, C. J., & Scheuss, A. I. (2019). Love the way you lie: Hiring managers' impression management in company presentation videos. *Journal of Personnel Psychology*, *18*(2), 84.

Larkin, J., & Neumann, R. (2012). Ageing academics: Workforce priorities for universities. *International Journal of Employment Studies*, *20*(1), 3-24.

Lavanya, P. M., & Sasikala, E. (2021). Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)* (pp. 603-609). IEEE.

Lawler, J. J., & Elliot, R. (1996). Artificial intelligence in HRM: An experimental study of an expert system. Journal of Management, 22(1), 85–111. Available From https://doi.org/10.1177/014920639602200104

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196). PMLR.

Leary, M. R. (2019). *Self-presentation: Impression management and interpersonal behavior*. Routledge.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin*, *107*(1), 34.

Lebedeva, T. E., Golubeva, O. V., Chaykina, Z. V., Egorov, E. E., & Romanovskaya, E. V. (2022). Tendencies and Trends in the Process of Digitalization of Personnel Selection by Heads

of Commercial Companies. In *Big Data in the GovTech System* (pp. 129-135). Cham: Springer International Publishing.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Lee, S. B., Gui, X., Manquen, M., & Hamilton, E. R. (2019). Use of training, validation, and test sets for developing automated classifiers in quantitative ethnography. In *International Conference on Quantitative Ethnography* (pp. 117-127). Springer, Cham.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. In *Business and the Ethical Implications of Technology* (pp. 71-86). Cham: Springer Nature Switzerland.

Lemenkova, P. (2020). Python libraries matplotlib, seaborn and pandas for visualization geospatial datasets generated by QGIS. *Analele stiintifice ale Universitatii" Alexandru Ioan Cuza" din Iasi-seria Geografie*, *64*(1), 13-32.

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, 57-70.

Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment*, *14*(4), 299-316.

Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: development and validation of an interview faking behavior scale. *Journal of applied psychology*, *92*(6), 1638.

Li, H. (2017). Deep learning for natural language processing: advantages and challenges. *National Science Review*.

Li, W., Qi, F., Tang, M., & Yu, Z. (2020). Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing*, *387*, 63-77.

Li, Y., Dong, M., Wang, Y., & Xu, C. (2020). Neural architecture search in a proxy validation loss landscape. In *International Conference on Machine Learning* (pp. 5853-5862). PMLR.

Likhitkar, P., & Verma, P. (2020). HR value proposition using predictive analytics: An overview. *New Paradigm in Decision Science and Management*, 165-171.

Lin, J. C. W., Shao, Y., Djenouri, Y., & Yun, U. (2021). ASRNN: a recurrent neural network with an attention model for sequence labeling. *Knowledge-Based Systems*, *212*, 106548.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Lindsay, S. (2015). What works for doctoral students in completing their thesis?. *Teaching in Higher Education*, *20*(2), 183-196.

Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, *337*, 325-338.

Liu, H., Morstatter, F., Tang, J., & Zafarani, R. (2016). The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics*, *1*, 137-143.

Liu, Q., Huang, H., Zhang, G., Gao, Y., Xuan, J., & Lu, J. (2018). Semantic structure-based word embedding by incorporating concept convergence and word divergence. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Liu, T., & Homan, C. M. (2019). Twitter Job/Employment Corpus: A Dataset of Job-Related Discourse Built with Humans in the Loop. *arXiv preprint arXiv:1901.10619*.

Liu, T., & Homan, C. M. (2019). Twitter Job/Employment Corpus: A Dataset of Job-Related Discourse Built with Humans in the Loop. *arXiv preprint arXiv:1901.10619*.

Liu, W., Luo, X., Gong, Z., Xuan, J., Kou, N. M., & Xu, Z. (2016). Discovering the core semantics of event from social media. *Future Generation Computer Systems*, *64*, 175-185.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Long, F., Zhou, K., & Ou, W. (2019). Sentiment analysis of text based on bidirectional LSTM with multi-head attention. *IEEE Access*, *7*, 141960-141969.

López-Hernández, J., Almela, A., & Valencia-García, R. (2019). Automatic spelling detection and correction in the medical domain: A systematic literature review. In *Technologies and Innovation: 5th International Conference, CITI 2019, Guayaquil, Ecuador, December 2–5, 2019, Proceedings 5* (pp. 95-108). Springer International Publishing.

Loria, S. (2018). Textblob Documentation. *Release 0.15*, *2*, 269.

Lowe, B., & Laffey, D. (2011). Is Twitter for the birds? Using Twitter to enhance student learning in a marketing course. *Journal of Marketing Education*, *33*(2), 183-192.

Luo, X. (2021). Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal*, *60*(3), 3401-3409.

Luque, A., Carrasco, A., Martín, A., & de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216-231.

Ma, Q. (2020). Comparison of ARIMA, ANN and LSTM for stock price prediction. In *E3S Web of Conferences* (Vol. 218, p. 01026). EDP Sciences.

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Mac Naughton, G. (2020). Action research. In *Doing early childhood research* (pp. 208-223). Routledge.

Machi, L. A., & McEvoy, B. T. (2016). The literature review: Six steps to success. *Corwin Press*,1-192.

Makrehchi, M., & Kamel, M. S. (2017). Extracting domain-specific stop words for text classifiers. *Intelligent Data Analysis*, *21*(1), 39-62.

Malhotra, A., & Jindal, R. (2020). Multimodal deep learning-based framework for detecting depression and suicidal behaviour by affective analysis of social media posts. *EAI Endorsed Transactions on Pervasive Health and Technology*, *6*(21).

Malik, P., Aggrawal, A., & Vishwakarma, D. K. (2021). Toxic speech detection using traditional machine learning models and bert and fasttext embedding with deep neural networks. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1254-1259). IEEE.

Malisetty, S., Archana, R. V., & Kumari, K. V. (2017). Predictive Analytics in HR Management. *Indian Journal of Public Health Research & Development*, *8*(3).

Malone, W. T., Rus, D. & Laubacher, R. (2020). Artificial Intelligence and the Future of Work. https://workofthefuture.mit.edu/wp-content/uploads/2020/12/2020-Research-Brief-Malone-Rus-Laubacher2.pdf [2/06/2021]

Manaswi, N. K. (2018). Understanding and working with Keras. In *Deep Learning with Applications Using Python* (pp. 31-43). Apress, Berkeley, CA.

Manguri, K. H., Ramadhan, R. N., & Amin, P. R. M. (2020). Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research*, 54-65.

Margherita, A. (2021). Human resources analytics: A systematization of research topics and directions for future research. *Human Resource Management Review*, 100795.

Marr, B. (2017). Really big data at walmart: Real-time insights from their 40+ petabyte data cloud. *Forbes. com*, *23*.

Maslen, G. (2019). Recruitment of top academic leaders is getting harder. [14/07/2021] from https://www.universityworldnews.com/post.php?story=20190709153145683

Mayfield, C., Perdue, G., & Wooten, K. (2008). Investment management and personality type. *Financial services review*, *17*(3), 219-236.

Maynooth, U. O. (2016). *Recruitment and Selection Procedures*. [02/01/2019] Retrieved from https://www.maynoothuniversity.ie/sites/default/files/assets/document/Recruitment&SelectionGuidelines_3.pdf

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

Mehanna, Y. S., & Mahmuddin, M. (2021). The Effect of Pre-processing Techniques on the Accuracy of Sentiment Analysis Using Bag-of-Concepts Text Representation. *SN Computer Science*, *2*(4), 1-13.

Meijerink, J., & Keegan, A. (2019). Conceptualizing human resource management in the gig economy: Toward a platform ecosystem perspective. *Journal of managerial psychology*, *34*(4), 214-232.

Meijerink, J., Boons, M., Keegan, A., & Marler, J. (2021). Algorithmic human resource management: Synthesizing developments and cross-disciplinary insights on digital HRM. *The InTernaTIonal Journal of human resource managemenT*, *32*(12), 2545-2562.

Melanthiou, Y., Pavlou, F., & Constantinou, E. (2015). The use of social network sites as an e-recruitment tool. *Journal of Transnational Management*, *20*(1), 31-49.

Melchers, K. G., Roulin, N., & Buehl, A. K. (2020). A review of applicant faking in selection interviews. *International Journal of Selection and Assessment*, *28*(2), 123-142.

Melchers, K. G., Roulin, N., & Buehl, A. K. (2020). A review of applicant faking in selection interviews. *International Journal of Selection and Assessment*, *28*(2), 123-142.

Menon, V. M., & Rahulnath, H. A. (2016). A novel approach to evaluate and rank candidates in a recruitment process by estimating emotional intelligence through social media data. In *2016 International Conference on Next Generation Intelligent Systems (ICNGIS)* (pp. 1-6). IEEE.

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for big data. *Future Generation Computer Systems*, *63*, 123-130.

Mestry, S., Singh, H., Chauhan, R., Bisht, V., & Tiwari, K. (2019, April). Automation in social networking comments with the help of robust fasttext and cnn. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)* (pp. 1-4). IEEE.

Michael, L. G., Donohue, J., Davis, J. C., Lee, D., & Servant, F. (2019). Regexes are hard: Decision-making, difficulties, and risks in programming regular expressions. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (pp. 415-426). IEEE.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning--based Text Classification: A Comprehensive Review. *ACM Computing Surveys (CSUR)*, *54*(3), 1-40.

Mishra, P., Roger, J. M., Marini, F., Biancolillo, A., & Rutledge, D. N. (2021). Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *212*, 104190.

Mishra, R., & Sharma, R. (2015). Big data: opportunities and challenges. *International Journal of Computer Science and Mobile Computing*, *4*(6), 27-35.

Mishra, S. B, & Alok, S. (2017). Handbook of research methodology: a compendium for scholars and researchers. *Educreation Publishing*.

Mishra, S. N., Lama, D. R., & Pal, Y. (2016). Human Resource Predictive Analytics (HRPA) for HR management in organizations. *International Journal of Scientific & Technology Research*, *5*(5), 33-35.

Mishra, S. N., Lama, D. R., & Pal, Y. (2016). Human Resource Predictive Analytics (HRPA) for HR management in organizations. *International Journal of Scientific & Technology Research*, *5*(5), 33-35.

Mohammed, S. M., Jacksi, K., & Zeebaree, S. R. (2020). Glove word embedding and DBSCAN algorithms for semantic document clustering. In *2020 International Conference on Advanced Science and Engineering (ICOASE)* (pp. 1-6). IEEE.

Molefe, M. (2014). *From data to insights: HR analytics in organisations* (Doctoral dissertation, University of Pretoria).

Moolayil, J., Moolayil, J., & John, S. (2019). *Learn Keras for Deep Neural Networks* (pp. 1-192). Birmingham: Apress.

Morra, G. (2018). Fast Python: NumPy and Cython. In *Pythonic Geodynamics* (pp. 35-60). Springer, Cham.

Mosley Jr, R. C. (2012). Social media analytics: Data mining applied to insurance Twitter posts. In *Casualty Actuarial Society E-Forum* (Vol. 2, p. 1).

Mullen, L. A., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, consistent tokenization of natural language text. *Journal of Open Source Software*, *3*(23), 655.

Mullen, T., & Malouf, R. (2006). A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In *AAAI spring symposium: computational approaches to analyzing weblogs* (pp. 159-162).

Muscalu, E. (2015). Sources of human resources recruitment organization. *Land Forces Academy Review*, *20*(3), 351.

Mustard-Research. Top 10 qualities required to be a good researcher [21/07/2021] from retrieved https://www.mustard-research.com/blog/general/top-10-qualities-required-be-good-researcher/

Nabukenya, J. (2012). Combining case study, design science and action research methods for effective collaboration engineering research efforts. In *2012 45th Hawaii International Conference on System Sciences* (pp. 343-352). IEEE.

Naseem, U., Razzak, I., & Eklund, P. W. (2021). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, *80*, 35239-35266.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77).

Neelankavil, J. P. (2015). *International business research*. Routledge.

Nelson, L. K. (2021). Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century US South. *Poetics*, *88*, 101539.

Ng, W. L., Collins, P. F., Hickling, D. F., & Bell, J. J. (2019). Evaluating the concurrent validity of body mass index (BMI) in the identification of malnutrition in older hospital inpatients. *Clinical Nutrition*, *38*(5), 2417-2422.

Nield, T. (2017). An introduction to regular expressions. From oreilly.com/An introduction to regular expressions [11/12/2021]

Nowakowski, P. T. (2013). Criteria for assessment of ethical conduct of academics. *Acta Prosperitatis*, 49.

Nowakowski, P. T. (2018). Criteria for assessment of ethical conduct of academics.

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*.

Oh, C., Sasser, S., & Almahmoud, S. (2015). Social media analytics framework: the case of Twitter and Super Bowl ads. *Journal of Information Technology Management*, *26*(1), 1-18.

Oke, G. G., Okunola, P. O., Oni, A. A., & Adetoro, J. A. (2010). The Relationship Between Vice-Chancellors' Leadership Behaviour and the Work Behaviour of Lecturers in Nigerian Universities: Implication for Leadership Training for Vice-Chancellors. *Journal of Higher Education in Africa/Revue de l'enseignement supérieur en Afrique*, *8*(1), 123-139.

Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, *33*(23), doi.org/10.1002/cpe.5909.

Orbay, A., & Akarun, L. (2020, November). Neural sign language translation by learning tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 222-228). IEEE.

Orellana, G., Arias, B., Orellana, M., Saquicela, V., Baculima, F., & Piedra, N. (2018, November). A study on the impact of pre-processing techniques in Spanish and english text classification over short and large text documents. In *2018 international conference on information systems and computer science (INCISCOS)* (pp. 277-283). IEEE.

Otoo, I. C., Assuming, J., & Agyei, P. M. (2018). Effectiveness of recruitment and selection practices in public sector higher education institutions: Evidence from Ghana. *European scientific journal*, *14*(13), 199-214.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604-624.

Palomino, M. A., & Aider, F. (2022). Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis. *Applied Sciences*, *12*(17), 8765.

Palomino, M., Grad, D., & Bedwell, J. (2021). GoldenWind at SemEval-2021 Task 5: Orthrus–An Ensemble Approach to Identify Toxicity.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(2), 1-135.

Panicheva, P., Ledovaya, Y., & Bogolyubova, O. (2016). Lexical, morphological and semantic correlates of the dark triad personality traits in russian facebook texts. In *2016 IEEE artificial intelligence and natural language conference (AINL)* (pp. 1-8). IEEE.

Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., & Elmqvist, N. (2017). ConceptVector: text visual analytics via interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics*, *24*(1), 361-370.

Park, K., Choi, Y., Choi, W. J., Ryu, H. Y., & Kim, H. (2020). LSTM-based battery remaining useful life prediction with multi-channel charging profiles. *Ieee Access*, *8*, 20786-20798.

Park, K., Lee, J., Jang, S., & Jung, D. (2020). An empirical study of tokenization strategies for various Korean NLP tasks. *arXiv preprint arXiv:2010.02534*.

Paul, A., Ahmad, A., Rathore, M. M., & Jabbar, S. (2016). Smartbuddy: defining human behaviors using big data analytics in social internet of things. *IEEE Wireless communications*, *23*(5), 68-74.

Peck, J.A., & Levashina, J. (2017). Impression management and interview and job performance ratings: A meta-analysis of research design with tactics in mind. *Frontiers in psychology*, *8*, 201.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research. *European Journal of Information Systems*, *27*(2), 129-139.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

Pei, F. Q., Li, D. B., & Tong, Y. F. (2018). Double-layered big data analytics architecture for solar cells series welding machine. *Computers in Industry*, *97*, 17-23.

Peng, S., Wang, G., Zhou, Y., Wan, C., Wang, C., Yu, S., & Niu, J. (2017). An immunization framework for social networks through big data based influence modeling. *IEEE transactions on dependable and secure computing*, *16*(6), 984-995.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pereira, P. (2020). Towards Helping Data Scientists. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 1-2). IEEE.

Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-API environment. *International Journal of Social Research Methodology*, *23*(3), 277-290.

Pirina, I., & Çöltekin, Ç. (2018). Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task* (pp. 9-12).

Pradha, S., Halgamuge, M. N., & Vinh, N. T. Q. (2019). Effective text data pre-processing technique for sentiment analysis in social media data. In *2019 11th international conference on knowledge and systems engineering (KSE)* (pp. 1-8). IEEE.

Purnamasari, K. K., & Suwardi, I. S. (2018). Rule-based part of speech tagger for Indonesian language. In *IOP Conference Series: Materials Science and Engineering* (Vol. 407, No. 1, p. 012151). IOP Publishing.

Putra, O. V., Wasmanson, F. M., Harmini, T., & Utama, S. N. (2020). Sundanese twitter dataset for emotion classification. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)* (pp. 391-395). IEEE.

Qi, C. C. (2020). Big data management in the mining industry. *International Journal of Minerals, Metallurgy and Materials*, *27*(2), 131-139.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., & Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation?. *arXiv preprint arXiv:1804.06323*.

Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications*, *212*, 118715.

Rajasekar, S., Philominathan, P., & Chinnathambi, V. (2013). Research methodology. eprint. *arXiv preprint physics/0601009*, 1-53.

Rane, A., & Kumar, A. (2018). Sentiment classification system of Twitter data for US airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 1, pp. 769-773). IEEE.

Ranganathan, J., Hedge, N., Irudayaraj, A. S., & Tzacheva, A. A. (2018). Automatic detection of emotions in Twitter data: a scalable decision tree classification method. In *Proceedings of the Workshop on Opinion Mining, Summarization and Diversification* (pp. 1-10).

Rapson, C. J., Seet, B. C., Naeem, M. A., Lee, J. E., Al-Sarayreh, M., & Klette, R. (2018). Reducing the pain: A novel tool for efficient ground-truth labelling in images. In *2018 international conference on image and vision computing New Zealand (IVCNZ)* (pp. 1-9). IEEE.

Raschka, S., & Mirjalili, V. (2017). Python Machine Learning: Machine Learning and Deep Learning with Python. *Scikit-Learn, and TensorFlow. Second edition ed*.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems, 89*, 14-46.

Raza, M., & Gulwani, S. (2020). Web data extraction using hybrid program synthesis: A combination of top-down and bottom-up inference. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (pp. 1967-1978).

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, *8*, 54776-54788.

Ren, Z., Shen, Q., Diao, X., & Xu, H. (2021). A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, *58*(3), 102532.

Reshi, A. A., Rustam, F., Aljedaani, W., Shafi, S., Alhossan, A., Alrabiah, Z., & Ashraf, I. (2022). COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset. In *Healthcare* (Vol. 10, No. 3, p. 411). MDPI.

Resnik, D. B. (2015). What is ethics in research & why is it important. [21/11/2022] From https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm?links=falseWijaya

Reuter, T., & Cimiano, P. (2012). A systematic investigation of blocking strategies for real-time classification of social media content into events. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 6, No. 3, pp. 8-15).

Reuter, T., & Cimiano, P. (2012). Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (pp. 1-8).

Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, *11*(1), 3923.

Ridzuan, F., & Zainon, W. M. N. W. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, *161*, 731-738.

Roach, A., Christensen, B. K., & Rieger, E. (2019). The essential ingredients of research supervision: A discrete-choice experiment. *Journal of Educational Psychology*, *111*(7), 1243.

Robie, C., Christiansen, N. D., Bourdage, J. S., Powell, D. M., & Roulin, N. (2020). Nonlinearity in the relationship between impression management tactics and interview performance. *International Journal of Selection and Assessment*, *28*(4), 522-530.

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Roth, L., Klehe, U. C., & Willhardt, G. (2021). Liar, liar, pants on fire: How verbal deception cues signal deceptive versus honest impression management and influence interview ratings. *Personnel Assessment and Decisions*, *7*(1), 7.

Roulin, N., Bangerter, A., & Levashina, J. (2015). Honest and deceptive impression management in the employment interview: Can it be detected and how does it impact evaluations?. *Personnel Psychology*, *68*(2), 395-444.

Rozario, S. D., Venkatraman, S., & Abbas, A. (2019). Challenges in recruitment and selection process: An empirical study. *Challenges*, *10*(2), 35.

Ruby, U., & Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng*, *9*(10).

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., & Choi, G. S. (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE access*, *8*, 101489-101499.

Ryazanov, I., Nylund, A. T., Basu, D., Hassellöv, I. M., & Schliep, A. (2021). Deep Learning for Deep Waters: An Expert-in-the-Loop Machine Learning Framework for Marine Sciences. *Journal of Marine Science and Engineering*, *9*(2), 169.

Sager, C., Janiesch, C., & Zschech, P. (2021). A survey of image labelling for computer vision applications. *Journal of Business Analytics*, *4*(2), 91-110.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. In: LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings., pp. 810–817.

Saini, S., Punhani, R., Bathla, R., & Shukla, V. K. (2019, April). Sentiment analysis on twitter data using R. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 68-72). IEEE.

Saks, A. M. (2017). The Impracticality of Recruitment Research. *The Blackwell handbook of personnel selection*, 47-72.

Sambasivam, G. A. O. G. D., & Opiyo, G. D. (2021). A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. *Egyptian informatics journal*, *22*(1), 27-34.

Sammons, M., Christodoulopoulos, C., Kordjamshidi, P., Khashabi, D., Srikumar, V., & Roth, D. (2016). Edison: Feature extraction for nlp, simplified. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4085-4092).
Samuel, A. L. (1959). Machine learning. *The Technology Review*, *62*(1), 42-45.

Santhosh Kumar, D. K., & D 'Mello, D. A. (2020). Strategies and challenges in big data: a short review. In *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2* (pp. 34-47). Springer International Publishing.

SAS. Predictive Analytics: What it is and why it matters. [11/05/2020] Retrieved from https://www.sas.com/en_za/insights/analytics/predictive-analytics.html

Satapathy, R., Guerreiro, C., Chaturvedi, I., & Cambria, E. (2017). Phonetic-based microtext normalization for twitter sentiment analysis. In *2017 IEEE international conference on data mining workshops (ICDMW)* (pp. 407-413). IEEE.

Schlenker, B. R. (1980). *Impression management* (Vol. 526). Monterey, CA: Brooks/Cole.

Schmidhuber, J. (2015). "Deep learning in neural networks: An overview," Neural Netw., vol. 61, pp. 85–117.

Schraeder, M., & Jordan, M. (2011). Managing performance: A practical perspective on managing employee performance. *The journal for quality and participation*, *34*(2), 4.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Seetha, J., & Raja, S. S. (2018). Brain tumor classification using convolutional neural networks. *Biomedical & Pharmacology Journal*, *11*(3), 1457.

Sehgal, G., Gupta, B., Paneri, K., Singh, K., Sharma, G., & Shroff, G. (2017). Crop planning using stochastic visual optimization. In *2017 IEEE Visualization in Data Science (VDS)* (pp. 47-51). IEEE.

Sharma, S., & Sharma, S. (2017). Activation functions in neural networks. *Towards Data Science*, *6*(12), 310-316.

Sharma, S., Dashora, J., & Saxena, K. (2021). Application of Business Intelligence Solutions for Human Resource Analytics in the Context of Industry 4.0. *Weser Edited book (Germany) Titled" Recent Advances in the Field of Accounting, Finance & Management*.

Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., & Zhang, C. (2018). Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Shen, Y., Lin, Z., Jacob, A. P., Sordoni, A., Courville, A., & Bengio, Y. (2018). Straight to the tree: Constituency parsing with neural syntactic distance. *arXiv preprint arXiv:1806.04168*.

Shenfield, A., Day, D., & Ayesh, A. (2018). Intelligent intrusion detection systems using artificial neural networks. *ICT Express*, *4*(2), 95-99.

Shirsat, V. S., Jagdale, R. S., & Deshmukh, S. N. (2017, August). Document level sentiment analysis from news articles. In *2017 international conference on computing, Communication, Control and Automation (ICCUBEA)* (pp. 1-4). IEEE.

Siamian, H., Bala Ghafari, A., Aligolbandi, K., Seyyede Fereshteh Reza Nezhad, S. F., Sharifi Nick, M., Shahrabi, A., & Ghazi Zadeh, Z. (2013). Characteristics of a good university lecturer according to students. *Journal of Mazandaran University of Medical Sciences*, *22*(96), 106-113.

Silander, C., & Stigmar, M. (2019). Individual growth or institutional development? Ideological perspectives on motives behind Swedish higher education teacher training. *Higher Education*, *77*, 265-281.

Singh, A. K., & Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. *Int. J. Adv. Comput. Sci. Appl*, *10*(7).

Singh, J., & Gupta, V. (2016). Text stemming: Approaches, applications, and challenges. *ACM Computing Surveys (CSUR)*, *49*(3), 1-46.

Singh, R. K., & Verma, H. K. (2022). Effective parallel processing social media analytics framework. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 2860-2870.

Singh, T., & Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, *89*, 549-554.

Sinha, V., & Thaly, P. (2013). A review on changing trend of recruitment practice to enhance the quality of hiring in global organizations. *Management: journal of contemporary management issues*, *18*(2), 141-156.

Sivarajah, U., Irani, Z., Gupta, S., & Mahroof, K. (2020). Role of big data and social media analytics for business to business sustainability: A participatory web context. *Industrial Marketing Management*, *86*, 163-179.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263-286.

Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. *Cambridge University, UK*, *32*(34), 2008.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631-1642).

Song, H., Kim, M., Park, D., & Lee, J. G. (2019). How does early stopping help generalization against label noise?. *arXiv preprint arXiv:1911.08059*.

Srivastava, H. (2017). What is K-Fold Cross Validation? [11/05/2021] Retrieved from https://magoosh.com/data-science/k-fold-cross-validation/

Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, *471*, 216-232.

Stevens, C. K., & Kristof, A. L. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of applied psychology*, *80*(5), 587.

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering*, *6*, 89-96.

Strohmeier, S., & Piazza, F. (2015). Artificial intelligence techniques in human resource management—a conceptual exploration. In *Intelligent techniques in engineering management* (pp. 149-172). Springer, Cham.

Study Lecture Notes. Qualities of a good researcher. [21/07/2021]. Available from http://studylecturenotes.com/qualities-of-a-good-researcher/

Sullivan, J. (2013). How Google is using people analytics to completely reinvent HR. *TLNT: The Business of HR*, *26*, 1-18.

Swanepoel, B., Erasmus B.J., Schenk H.W. and Tshilongamulenzhe T. (2014). South African Human Resource management: Theory and practice,4th ed., Juta, Cape town.

Swanepoel, B., Erasmus, B., & Schenk, H. (2008). South African human resource management: Theory & practice. *Juta and Company Ltd*.

Swearingen, T., Drevo, W., Cyphers, B., Cuesta-Infante, A., Ross, A., & Veeramachaneni, K. (2017). ATM: A distributed, collaborative, scalable system for automated machine learning. In *2017 IEEE international conference on big data (big data)* (pp. 151-162). IEEE.

Swider, B. W., Barrick, M. R., Harris, T. B., & Stoverink, A. C. (2011). Managing and creating an image in the interview: The role of interviewee initial impressions. *Journal of Applied Psychology*, *96*(6), 1275.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications, 110*, 298-310.

Taboada, M., Anthony, C., & Voll, K. D. (2006). Methods for Creating Semantic Orientation Dictionaries. In *LREC* (pp. 427-432).

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2018). Personality predictions based on user behavior on the facebook social media platform. *IEEE Access*, *6*, 61959-61969.

Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. (2019, June). Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning* (pp. 1-4).

Taggart, C. (2015). *New words for old: Recycling our language for the modern world*. Michael O'Mara Books.

Taherdoost, H. (2016). Sampling methods in research methodology; how to choose a sampling technique for research. *How to choose a sampling technique for research (April 10, 2016)*.

Tait, A. (2016). Why Are Online Jokes Funnier without Punctuation and Capital Letters? The New Statesman. Available online: https://www.newstatesman.com/science-tech/2016/10/why-are-online-jokes-funnier-without-punctuation-and-capital [29/07/ 2022].

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61*(4), 15-42.

Tamizharasi, K., & Rani, U. (2014). Employee turnover analysis with application of data mining methods. *International Journal of Computer Science and Information Technologies*, *5*(1), 562-566.

Tellex, S., Katz, B., Lin, J., Fernandes, A., & Marton, G. (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 41-47).

TensorFlow. Tf.keras.preprocessing.text.Tokenizer. [18/02/2022] available from https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer

TextBlob. (2020). TextBlob: Simplified Text Processing. [22/10/2021] available from https://textblob.readthedocs.io/en/dev/index.html

Tham, T. L., & Holland, P. (2018). What do business school academics want? Reflections from the national survey on workplace climate and well-being: Australia and New Zealand. *Journal of Management & Organization*, *24*(4), 492-499.

Thunnissen, M. (2016). Talent management: For what, how and how well? An empirical exploration of talent management in practice. *Employee Relations*.

Thunnissen, M., & Van Arensbergen, P. (2015). A multi-dimensional approach to talent: An empirical analysis of the definition of talent in Dutch academia. *Personnel Review*.

Toledo-Pereyra, L. H. (2012). Ten qualities of a good researcher. *Journal of Investigative Surgery*, *25*(4), 201-202.

Tommasel, A., Rodriguez, J. M., & Godoy, D. (2018). Textual aggression detection through deep learning. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)* (pp. 177-187).

Top, C., & Ali, B. J. (2021). Customer satisfaction in online meeting platforms: Impact of efficiency, fulfillment, system availability, and privacy. *Amazonia Investiga*, *10*(38), 70-81.

Troussas, C., Virvou, M., & Espinosa, K. J. (2015). Using Visualization Algorithms for Discovering Patterns in Groups of Users for Tutoring Multiple Languages through Social Networking. *J. Networks*, *10*(12), 668-674.

Troussas, C., Virvou, M., & Mesaretzidis, S. (2015). Comparative analysis of algorithms for student characteristics classification using a methodological framework. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-5). IEEE.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, *49*(11), 1225-1231.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth international AAAI conference on weblogs and social media*.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384-394).

Turk, M., & Ledić, J. (2016). Between teaching and research: Challenges of the academic profession in Croatia. *Center for Educational Policy Studies Journal*, *6*(1), 95-111.

Uddin, A. H., Bapery, D., & Arif, A. S. M. (2019). Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.

Udjo, E. O., & Erasmus, B. (2014). Impact of Retirement Age Policy on the Workforce of a Higher Education Institution in South Africa. *Politics & Policy*, *42*(5), 744-768.

Umer, M., Ashraf, I., Mehmood, A., Kumari, S., Ullah, S., & Sang Choi, G. (2021). Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model. *Computational Intelligence*, *37*(1), 409-434.

Uthayasankar, S., Mustafa M. K., Zahir, I., & Vishanth, W. (2017). Critical analysis of Big Data challenges and analytical methods, Journal of Business Research, vol. 70, pp. 263- 286, 2017.

Vallance, C. (2022). AI tools fail to reduce bias. [18/07/2023]. Retrieved from https://www.bbc.com/news/technology-63228466

Valenzuela, A. (2019). Recruitment and Selection Process of Faculty in Higher Education Institutions in the Philippines. *Available at SSRN 3445566*.

Van den Brink, M. (2010). *Behind the scenes of science: Gender practices in the recruitment and selection of professors in the Netherlands*. Amsterdam University Press.

van der Togt, J., & Rasmussen, T. H. (2017). Toward evidence-based HR. *Journal of Organizational Effectiveness: People and Performance*, *4*(2), 127-132.

Van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215-222.

Vanaja, S., & Belwal, M. (2018). Aspect-level sentiment analysis on e-commerce data. In *2018 International conference on inventive research in computing applications (ICIRCA)* (pp. 1275-1279). IEEE.

Vega, D. (2018). Case studies in humanitarian logistics research. *Journal of Humanitarian Logistics and Supply Chain Management*, *8*(2), 134-152.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. In *Design Science Research in Information Systems. Advances in Theory and Practice: 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings 7* (pp. 423-438). Springer Berlin Heidelberg.

Verma, R. (2022).SMS Slang Translator; GitHub, Inc.: San Francisco, CA, USA.

Verma, R., & Bandi, S. (2019). Artificial intelligence & Human resource management in Indian IT sector. In *Proceedings of 10th International Conference on Digital Strategies for Organizational Success*. 8(4), 962-967.

Verma, S., Singh, V., & Bhattacharyya, S. S. (2020). Do big data-driven HR practices improve HR service quality and innovation competency of SMEs. *International Journal of Organizational Analysis*, *29*(4), 950-973.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, *325*(5939), 425-428.

Villeda, M., McCamey, R., Essien, E., & Amadi, C. (2019). Use of social networking sites for recruiting and selecting in the hiring process. *International business research*, *12*(3), 66-78.

Vinayakumar, R., Alazab, M., Jolfaei, A., Soman, K. P., & Poornachandran, P. (2019). Ransomware triage using deep learning: twitter as a case study. In *2019 Cybersecurity and Cyberforensics Conference (CCC)* (pp. 67-73). IEEE.

Virmani, C., Juneja, D., & Pillai, A. (2018). Design of query processing system to retrieve information from social network using NLP. *KSII Transactions on Internet and Information Systems (TIIS)*, *12*(3), 1168-1188.

Vom Brocke, J., Winter, R., Hevner, A., & Maedche, A. (2020). Special issue editorial–accumulation and evolution of design knowledge in design science research: a journey through time and space. *Journal of the Association for Information Systems*, *21*(3), 9.

Vyas, V., & Uma, V. (2018). An extensive study of sentiment analysis tools and binary classification of tweets using rapid miner. *Procedia Computer Science*, *125*, 329-335.

Wagh, R., & Punde, P. (2018). Survey on sentiment analysis using twitter dataset. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 208-211). IEEE.

Waheeb, S. A., Ahmed Khan, N., Chen, B., & Shang, X. (2020). Machine learning based sentiment text classification for evaluating treatment quality of discharge summary. *Information*, *11*(5), 281.

Wang, D., Su, J., & Yu, H. (2020). Feature extraction and analysis of natural language processing for deep learning english language. *IEEE Access*, *8*, 46335-46345.

Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., ... & Gray, A. (2019). Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-24.

Wang, F., Xuan, Z., Zhen, Z., Li, K., Wang, T., & Shi, M. (2020). A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework. *Energy Conversion and Management*, *212*, 112766.

Wang, L., Niu, J., & Yu, S. (2019). Sentidiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *32*(10), 2026-2039.

Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 189-198).

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., & Liu, H. (2018). Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, *77*, 34-49.

Wang, Y., Wang, L., Yang, F., Di, W., & Chang, Q. (2021). Advantages of direct input-to-output connections in neural networks: The Elman network for stock index forecasting. *Information Sciences*, *547*, 1066-1079.

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., & Lestantyo, P. (2019). Cross-validation metrics for evaluating classification performance on imbalanced data. In *2019*

international conference on computer, control, informatics and its applications (ic3ina) (pp. 14-18). IEEE.

Warfield, D. (2010). IS/IT Research: A research methodologies review. *Journal of theoretical & applied information technology*, *13*.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wei, C. P., & Lee, Y. H. (2004). Event detection from online news documents for supporting environmental scanning. *Decision Support Systems*, *36*(4), 385-401.

Winter, R. P. (2017). *Managing academics: A question of perspective*. Edward Elgar Publishing.

Wolf, C., & Blomberg, J. (2019). Evaluating the promise of human-algorithm collaborations in everyday work practices. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-23.

Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., ... & Liu, T. Y. (2021). R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, *34*.

Wu, Y., Xu, J., Jiang, M., Zhang, Y., & Xu, H. (2015). A study of neural word embeddings for named entity recognition in clinical text. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 1326). American Medical Informatics Association.

Xu, D., Zhang, S., Zhang, H., & Mandic, D. P. (2021). Convergence of the RMSProp deep learning method with penalty for nonconvex optimization. *Neural Networks*, *139*, 17-23.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, *2*(3), 249-262.

Xue, D., Wu, L., Hong, Z., Guo, S., Gao, L., Wu, Z., Zhong, X., & Sun, J. (2018). Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, *48*(11), 4232-4246.

Yamak, Z. R. (2018). *Multiple identities detection in online social media* (Doctoral dissertation, Normandie Université).

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, *44*(3), 363-373.

Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.

Yin, R. K. (1994). Discovering the future of the case study. Method in evaluation research. *Evaluation practice*, *15*(3), 283-290.

Yin, R. K. (1994). Discovering the future of the case study. Method in evaluation research. *Evaluation practice*, *15*(3), 283-290.

Yogish, D., Manjunath, T. N., & Hegadi, R. S. (2018). Review on natural language processing trends and techniques using nltk. In *International Conference on Recent Trends in Image Processing and Pattern Recognition* (pp. 589-606). Springer, Singapore.

Younas, M. (2019). Research challenges of big data. *Service Oriented Computing and Applications*, *13*, 105-107.

Yu, Y., Zhang, L., Chen, L., & Qin, Z. (2021). Adversarial Samples Generation Based on RMSProp. In *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)* (pp. 1134-1138). IEEE.

Yuan, X., Li, L., Shardt, Y. A., Wang, Y., & Yang, C. (2020). Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development. *IEEE Transactions on Industrial Electronics*, *68*(5), 4404-4414.

Zabkar, V., Arslanagic-Kalajdzic, M., Diamantopoulos, A., & Florack, A. (2017). Brothers in blood, yet strangers to global brand purchase: A four-country study of the role of consumer personality. *Journal of Business Research*, *80*, 228-235.

Zad, S., Heidari, M., Jones, J. H., & Uzuner, O. (2021, May). A survey on concept-level sentiment analysis techniques of textual data. In *2021 IEEE World AI IoT Congress (AIIoT)* (pp. 0285-0291). IEEE.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56-70.

Zelaya, C. V. G. (2019). Towards explaining the effects of data pre-processing on machine learning. In *2019 IEEE 35th international conference on data engineering (ICDE)* (pp. 2086-2090). IEEE.

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics-Theory and Methods*, *49*(9), 2080-2093.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems*, *25*(6), 13-16.

Zhai, C., & Massung, S. (2016). *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool.

Zhang, Y. D., Pan, C., Sun, J., & Tang, C. (2018). Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU. *Journal of computational science*, *28*, 1-10.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, *6*(1), 1-9.

Zhang, Y., Li, Z., & Zhang, M. (2020). Efficient second-order TreeCRF for neural dependency parsing. *arXiv preprint arXiv:2005.00975*.

Zhang, Y., Wang, J., & Zhang, X. (2018). Ynu-hpcc at semeval-2018 task 1: Bilstm with attention based sentiment analysis for affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 273-278).

Zhang, Z. (2018). Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)* (pp. 1-2). IEEE.

Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, *31*.

Zhang, Z., Liu, S., Li, M., Zhou, M., & Chen, E. (2017). Stack-based multi-layer attention for transition-based dependency parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1677-1682).

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312-323.

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, *9*(2), 1-29.

Zou, F., Shen, L., Jie, Z., Zhang, W., & Liu, W. (2019). A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11127-11135).

10. APPENDICES

A. Letter from Twitter

From: developer-accounts@twitter.com <developer-accounts@twitter.com>
Sent: Friday, August 31, 2018 4:14:19 PM
To: Junior Vela Vela (210535115) <210535115@stuukznac.onmicrosoft.com>
Subject: Twitter developer account application [ref:_00DA0K0A8._5004A1Sw5SD:ref]

Your Twitter developer account application has been approved!

Thanks for applying for access. We've completed our review of your application, and are excited to share that your request has been approved.

Sign in to your developer account to get started.

Thanks for building on Twitter!

Help

Twitter, Inc. 1355 Market Street, Suite 900 San Francisco, CA 94103

B. Ethical Clearance

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

22 October 2019

Mr Junior Vela Vela (210535115)
School Of Man Info Tech &Gov
Pietermaritzburg

Dear Mr Vela Vela,

Protocol reference number: HSSREC/00000628/2019
Project title: Developing a Predictive Model Using Twitter Dataset For Recruiting Job-fit Candidates in Higher Education Institutions

Full Approval – Expedited Application

This letter serves to notify you that your application received on 08 August 2019 in connection with the above, was reviewed by the Humanities and Social Sciences Research Ethics Committee (HSSREC) and the protocol has been granted **FULL APPROVAL**

Any alteration/s to the approved research protocol i.e. Questionnaire/Interview Schedule, Informed Consent Form, Title of the Project, Location of the Study, Research Approach and Methods must be reviewed and approved through the amendment/modification prior to its implementation. In case you have further queries, please quote the above reference number. PLEASE NOTE: Research data should be securely stored in the discipline/department for a period of 5 years.

This approval is valid for one year from 22 October 2019.
To ensure uninterrupted approval of this study beyond the approval expiry date, a progress report must be submitted to the Research Office on the appropriate form 2 - 3 months before the expiry date. A close-out report to be submitted when study is finished.

Yours sincerely,

[redacted]

Dr Rosemary Sibanda (Chair)

/dd