A Data Science Analysis of the South African COVID-19 Infodemic on Twitter

Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy in Information Technology
in the Faculty of Accounting and Informatics at
Durban University of Technology

By

Yaseen Khan

(21855590)

March 2025

Declaration

I, Yaseen Khan, hereby declare that the content within this thesis is my own work. All sources that I have used or quoted have been acknowledged in the text by the means of completed references. This study has not been previously submitted in any form to the Durban University of Technology or to any other institution for assessment or for any other purpose.

Student: _ _____ 20 March 2025 _____

Yaseen Khan Date

Approved for final submission

Superviso 20 March 2025 _____

Prof. Surendra Thakur (DTech. ICT) Date

Disclaimer

The research presented in this thesis is focused on an information and communications technology (ICT) approach to analyse an infodemic and thereby mitigating its spread. The researcher does not claim to be a medical expert on COVID-19 or to make definitive judgments about what constitutes ‘Fake News’ concerning the pandemic.

The views and findings expressed in this thesis are purely academic in nature and should not be interpreted as the researcher’s personal agreement or disagreement with any particular information or narratives surrounding COVID-19. The sole purpose of this work is to explore ICT-related approaches to deal with an infodemic. The researcher focuses on techniques and approaches rather than endorsing or refuting any specific claim or counter claim about the virus or its impacts.

Table of Contents

List of Figures

List of Tables

List of Equations

Dedication

Allah ﷻ, The Most Compassionate, The Most Merciful, in whose name I begin and continuous Blessings and Peace be upon The Beloved of Allah, The Chosen, Muhammad ﷺ. All praises directed towards me are for Allah and all faults are my own.

I hereby dedicate this thesis to my loving parents, brothers, nieces, nephews, cousins, extended family and friends. Special appreciation goes to colleagues, professionals and industry experts. The network of support offered from all these personalities motivates me to continuously strive for excellence.

Acknowledgements

I hereby acknowledge Prof. Surendra Thakur who has been a wonderful supervisor and motivating mentor for this thesis. His interactive feedback and willingness to sacrifice personal time is an appreciated trademark. Further, the conducive environment and *ad hoc* assistance provided by DUT's Enterprise Development team are acknowledged.

Financial Acknowledgement

The assistance of InSETA and NEMISA towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and should not necessarily be attributed to InSETA or NEMISA.

Abstract

The rapid dissemination of information on Twitter (X), particularly during COVID-19, has exacerbated infodemics, marked by the proliferation of both accurate and false information. Users were inundated with Fake News, encompassing misinformation, disinformation and malinformation. Misinformation entails the unintentional dissemination of false information, whereas disinformation involves deliberate deception. Malinformation, although grounded in truth, is manipulated to inflict harm. Traditional human-led verification systems have proven inadequate, as seen in prior infodemics such as the 2016 U.S. elections and South Africa's #FeesMustFall. Legislative measures in South Africa aimed at curbing Fake News during the pandemic were insufficient because of the vast volume of tweets, necessitating computational approaches.

Despite global focus on COVID-19 Fake News, South African research on Twitter infodemic analysis remains limited, particularly in the areas of Big Data, longitudinal analysis, fake news detection, sentiment analysis and social bot detection. This study addresses these gaps through advanced data science methods, including Natural Language Processing (NLP), Machine Learning (ML) and Change Point Analysis (CPA), to analyse the South African COVID-19 Twitter infodemic.

A longitudinal South African COVID-19 dataset (SAcovid19dataset), comprising 976 086 tweets from 8 November 2019 to 19 July 2021, was curated. Additionally, a labelled dataset (C19MLdataset) of 30 193 tweets was created for Fake News detection, containing 17 069 'Fake News' and 13 124 'Not Fake News' tweets. The study focused on textual analysis, as audio, video and image analytics were beyond its scope.

This research uniquely employs an exhaustive approach to compare a wide range of models for COVID-19 Fake News detection prioritising balanced accuracy and execution time performance metrics. Using the C19MLdataset, twenty-seven (27) shallow, five (5) deep learning (DL) and seven (7) transformer models were systematically evaluated. ExtraTreesClassifier, RandomForestClassifier and LightGBM emerged as the top-performing shallow models. RoBERTa was the top performing transformer model and Bi-LSTM outperformed other DL models.

LightGBM was identified as the most efficient model because of its speed and low computational demands. After optimization, it achieved a balanced accuracy of 88.76% and detected 262 508 (26.89%) Fake News tweets from the full SAcovid19dataset.

Sentiment analysis, performed using VADER and CPA, revealed 16 significant shifts in sentiment due to real-world events. Approximately 56% were related to lockdown announcements and restrictions. For instance, the South African national state of emergency on 15 March 2020 led to a shift from neutral to positive sentiment. Contrastingly, the 26 February 2021 South African state of the nation address saw sentiment shift from positive to negative.

Social bot activity was examined using three novel algorithms that analysed tweet timestamps, content duplication and sources. Results showed that three of the Top 10 Fake News accounts exhibited bot-like behaviour, confirming the presence of automated accounts in the spread of false information.

This study significantly contributes to the understanding of South Africa's COVID-19 infodemic by providing a robust Fake News detection model and linking real-world events to shifts in public sentiment. The development of social bot detection algorithms further illuminates the role of automated accounts in the dissemination of Fake News. These findings have practical implications for policymakers and researchers aiming to combat infodemics using computational tools.

Keywords

#COVID-19, Opinion Mining, Fake News, Natural Language Processing, Infodemic, Data Science, Machine Learning, Sentiment Analysis, Social Bot Detection.

List of Acronyms

BERT	Bidirectional Encoder Representations from Transformers
C19MLdataset	COVID-19 Labelled Dataset
COVID-19	Coronavirus Disease of 2019
CPA	Change Point Analysis
DL	Deep Learning
DUT	Durban University of Technology
IEEE	Institute of Electronic and Electrical Engineers
JSON	JavaScript Object Notation
LSTM	Long Short-Term Memory
ML	Machine Learning
NEMISA	National Electronic Media Institute of South Africa
SAcovid19dataset	South African COVID-19-related Twitter data
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary and sEntiment Reasoner
WHO	World Health Organization
XML	Extensible Markup Language

Output

Journals and Conference Publications

Khan, Y. and Thakur, S., 2022, August. Fake News detection of South African COVID-19 related tweets using machine learning. In: *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. IEEE, pp. 1-5.

Khan, Y., Thakur, S., Obiyemi, O. and Adetiba, E., 2022. Exploring links between online activism and real-world events: A case study of the# FeesMustFall. *Scientific Programming*, 13, pp. 219-237. https://doi.org/10.1155/2022/1562592

Khan, Y., Thakur, S., Obiyemi, O. and Adetiba, E., 2022, March. Identification of bots and cyborgs in the# FeesMustFall campaign. *Informatics*, 9(1), p. 21. MDPI. https://doi.org/10.3390/informatics9010021

Khan, Y. and Thakur, S. 2018. The presence of Twitter bots and cyborgs in the #FeesMustFall campaign. In: *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*. Plaine Magnien, Mauritius, 6-7 December 2018. IEEE, pp. 543-547.

Presentations Arising from this Study

AI and its applications in learning, teaching and assessment – 25 July 2023, CELT, DUT.

Foundations of Data Science – 29 May 2023, INSETA, MS Teams.

NLP and Sentiment Analysis – 15 February 2023, NEMISA, Coastlands Umhlanga.

AI Session – 5 November 2022, NEMISA, MS Teams.

Social Media and Data Science – 30 September 2022, BANKSETA, DUT.

Research methods for Unstructured Data – 8 July 2021, MICTSETA, DUT.

Media Engagement and Recognition Arising from this Study

Thakur, S., Khan, Y., Govender, M. and Moyo, S., 2022. Data science usage in government and data science skills survey. NEMISA, Parktown, Johannesburg.

Khan, Y. and Thakur, C., 2021. Fake news and the search for a ‘news vaccine algorithm’. *Sunday Tribune* (online). Available: https://www.pressreader.com/south-africa/sunday-tribune-south-africa/20210627/281943135855356 (Accessed 22 October 2024).

Agenda Setting and Misinformation - Hindvani Community Radio, 15 September 2020, Durban.

Conflict of Interest

There was no conflict of interest.

Chapter 1: Introduction

1.1 Background of the Study

This research employs data science techniques to computationally analyse Twitter content (tweets) for their veracity, sentiment and social bot involvement during the coronavirus disease 2019 (COVID-19) infodemic within a South African context. An infodemic is an extensive and rapid spread of information on digital and physical environments related to a disease outbreak, event, crisis or controversial issue, characterized by a confusing blend of facts, misinformation, rumours and opinions (Oxford English Dictionary, 2023).

The COVID-19 pandemic represented an unprecedented global health crisis (Ghebreyesus, 2020), occurring within an era of advanced digital technology, which signifies the importance of research and analysis, particularly in the realm of social media discourse. Twitter was selected for analysis in this study because of its feasible features, such as hashtag tracking, access to large volumes of publicly available data and its role in amplifying public discourse throughout the pandemic.

The rise of social media platforms like Twitter has led to the generation and rapid dissemination of vast amounts of information at unprecedented speeds (Leung, Sun and Bai, 2019), contributing to the emergence of unstructured big data, which presents a range of complex challenges (Chen, Mao and Liu, 2014). This environment inadvertently facilitates the possible occurrences of numerous infodemics, making research in this area both significant and essential for understanding and mitigating their effects.

The spread of Fake News on Twitter, whether intentional or accidental, negatively impacts both novice and experienced users (Visentin, Pizzi and Pichierri, 2019). As a core component of infodemics, Fake News consists of misleading or deceptive news-related content, frequently observed in contexts such as election campaigns and pandemics (Allcott and Gentzkow, 2017; Gelfert, 2018; Lazer *et al.*, 2018; Sharma *et al.*, 2019). Social media platforms like Facebook and Twitter have faced criticism for their inadequate responses to the proliferation of Fake News and incitement (BBC News, 2020; Euronews, 2021; Wakefield, 2021), prompting them to revise their

policies (Paul, 2020) and invest in research efforts to combat the issue (Facebook, 2020).

Social media data exists in unstructured forms, including text, audio, images and videos. The complexity of efficiently analysing this unstructured data necessitates the use of specialized methodologies found within the interdisciplinary field of Data Science. Data science provides valuable insights by addressing the challenges associated with big data and social media interactions (Asur and Huberman, 2010; Provost and Fawcett, 2013; Oracle, 2020; Cinelli *et al.*, 2020).

The purpose of this study is to analyse the COVID-19 infodemic on Twitter within a South African context by leveraging data science techniques. The analysis includes the detection of COVID-19-related Fake News, tracking of sentiment trends and investigation of social bot involvement, providing crucial insights into how misinformation propagates during global crises. Preliminary research revealed that while international literature on this topic is growing, there is a significant gap in studies focused on South African COVID-19 Fake News on Twitter, presenting an opportunity to address this research need.

1.2 Research Problem

South Africa's digital transformation, accelerated by the Fourth Industrial Revolution (4IR), has positioned social media platforms like Twitter at the forefront of rapid information dissemination. However, the COVID-19 pandemic revealed Twitter's dual role as both a vital source for public health information and a breeding ground for misinformation. Unfounded claims, such as loan relief promises (South Africa, 2020a) and rumours of tainted COVID-19 test kits (Madiba, 2020), signifies the urgency of addressing misinformation on the platform.

In response, the South African government launched a specialized portal to combat COVID-19-related Fake News (South Africa, 2020a, 2020b). Similarly, a human processing Real411 Platform[1] has been formed as a result of disinformation circulating across digital media. However, these efforts are predominantly human-

[1] https://real411.org.za/

driven, making them slow, highlighting the need for an automated process to effectively keep pace with the rapid spread of Fake News during an infodemic (Bird and Smith, 2020).

The spread of COVID-19 Fake News on Twitter poses significant risks, including undermining public health efforts, exacerbating societal tensions and misleading the South African populace during a critical period. Addressing this challenge requires not only distinguishing between true and false information but also involves understanding public sentiment around the pandemic and identifying social bots that may amplify Fake News. Current local strategies to mitigate these risks remain limited, reactive and insufficiently sophisticated to handle the dynamic nature of social media's information spread (South Africa, 2020a). The manual identification of Fake News spreaders and analysis of public sentiment are impractical given the vast amounts of data generated on Twitter, thereby necessitating an automated, computational solution (An *et al.*, 2013).

The field of infodemiology, which examines the distribution and determinants of information in digital platforms, particularly within the context of public health (Eysenbach, 2011), signifies the importance of interdisciplinary research to address the complexities of digital misinformation. Accordingly, this study seeks to answer the following research question:

> ***How can the COVID-19 infodemic on Twitter be computationally analysed within a South African context for Fake News?***

Because of the high volume of tweets, this study aims to computationally detect and analyse COVID-19-related Fake News in South African Twitter data. This involves not only detecting Fake News but also examining sentiment trends and determining the presence of social bots responsible for spreading Fake News. In doing so, this research contributes to the broader understanding of how Fake News propagates during global crises, offering insights for improving information integrity during public health emergencies. Additionally, by leveraging data science techniques such as Machine Learning (ML) and Natural Language Processing (NLP), the study introduces computational models capable of addressing the key challenges of Fake News, sentiment analysis and social bot detection. Given the scarcity of data science expertise in South Africa (Osanga, 2020), this study therefore enriches the local

knowledge base and introduces a toolset that can be applied to academic, public health and industrial contexts.

1.3 Definitions

The following provides a working definition of key terms used in this research:

Click Bait: A strategy employed by content creators to craft headline or thumbnail links that entice users to click through to the content. This tactic leverages curiosity or sensationalism to drive web traffic (Blom and Hansen, 2015).

Change Point Analysis (CPA): A statistical technique that identifies points within a time series where the statistical properties of the data change significantly. CPA is particularly useful in analysing extensive datasets to detect shifts in trends, behaviours, or conditions (Aminikhanghahi and Cook, 2017).

Infodemic: An extensive and rapid spread of information on digital and physical environments related to a disease outbreak, event, crisis or controversial issue, characterized by a confusing blend of facts, misinformation, rumours and opinions (Oxford English Dictionary, 2023).

Fake News: Articles or messages disseminated through various media that contain false information, which can manifest as misinformation, disinformation or malinformation, regardless of the methods or motives involved (Sharma *et al.*, 2019).

Hashtag: A metadata tag prefixed by the hash symbol (#), used on social media for indexing and categorizing content related to specific themes or topics, facilitating content discovery (Small, 2011).

Natural Language Processing (NLP): The field of study that focuses on the interaction between computers and humans through natural language, aiming to enable computers to understand, interpret and generate human languages (Jurafsky and Martin, 2024).

Python: A high-level, general-purpose programming language known for its clear syntax and readability (Van Rossum, 2007).

Sentiment Analysis: An NLP technique that systematically identifies, extracts, quantifies and studies affective states and subjective information from text sources,

often used to gauge public opinion, consumer sentiment, or social media discourse (Liu, 2012).

Social Bot: An algorithmic software entity that autonomously generates content and interacts with users on social media platforms, sometimes used to disseminate information or influence social discourse (Ferrara *et al.*, 2016).

Tweet: A posting made on the Twitter platform, limited to 280 characters, which can include text, links, images and videos (Kwak *et al.*, 2010).

Tweeter: An individual or entity that creates and posts tweets on the Twitter platform.

Twitter: A microblogging and social networking service where users post and interact through messages known as tweets (Kwak *et al.*, 2010).

Twitterbots: Automated software that interfaces with the Twitter API to perform actions such as tweeting, retweeting, liking and following other accounts, often used for information dissemination or interaction automation (Bessi and Ferrara, 2016).

VADER (Valence Aware Dictionary and sEntiment Reasoner): A lexicon and rule-based sentiment analysis tool specifically designed for capturing sentiments expressed in social media contexts (Hutto and Gilbert, 2014).

X (formerly Twitter): Refers to the social media platform's name after its acquisition. The term "Twitter" is retained in this study to reflect the platform's identity during the research period (Mehta, 2023).

1.4 Research Objectives

To fulfil the primary and secondary goals of this study, the following research objectives have been established:

Main Objective: To computationally analyse South African-related COVID-19 Twitter data for Fake News using data science.

Secondary Objective 1: To examine the South African-related COVID-19 Twitter data for changes in average sentiment.

Secondary Objective 2: To determine the presence of social bots within the South African COVID-19-related Twitter data.

1.5 Research Questions

To address each of the established objectives, the subsequent research questions have been formulated:

Research Question 1: How can the COVID-19 infodemic on Twitter be computationally analysed for Fake News within the South African context?

Research Question 2: Were there significant changes in the average sentiment from the South African COVID-19-related Twitter data?

Research Question 3: Were social bots leveraged to produce content around South African-related COVID-19 Twitter content?

1.6 Significance of the Study

In the digital age, technology enables widespread dissemination of information, but it also amplifies the spread of Fake News (Chadwick, Vaccari and Kaiser, 2022). Eysenbach (2020) suggests that the same level of effort used to combat a pandemic should be applied to detecting Fake News and other counter measures. The COVID-19 infodemic highlights the need for effective detection and analytical methods. The significance of this study is, therefore, multifaceted:

1. **Supporting Computational Approaches for Fake News Detection**: The study customizes ML and NLP techniques for identifying Fake News in unstructured Twitter data, refining tools for distinguishing between factual and misleading content.
2. **Enhancing Public Health Information Integrity**: By systematically identifying COVID-19 related Fake News, the research supports the integrity of public health communications, helping to protect the South African populace.
3. **Contributing to Infodemiology**: This research offers insights into information flow during health crises on Twitter, providing evidence-based strategies for managing misinformation in future pandemics.
4. **Building Data Science Capacity in South Africa**: Through the creation of an open-source Twitter analysis toolkit, the study fosters local data science expertise, contributing to long-term digital development.

5. **Interdisciplinary Impact**: The research bridges data science, public health and social media, encouraging cross-disciplinary collaboration to tackle societal challenges.
6. **Empowering Policy Makers and Stakeholders**: By identifying Fake News patterns and sources, the study equips stakeholders with actionable insights to manage misinformation and improve public communication.

The significance of this study extends beyond the academic realm, offering practical solutions to online infodemics, enhancing societal resilience against Fake News and fostering an informed and health-conscious public discourse.

1.7 Contributions of the Study

The study makes several key contributions: the development of a fast-processing COVID-19 Fake News detection model, the creation of a nearly 1 million (976 086) South African COVID-19 Twitter dataset, an extensive collection of 30 193 labelled dataset for Fake News modelling, the identification of plausible links between online sentiment and real-world events, and the detection of social bot activity using novel algorithms. These collectively enhance the tools and insights available for analysing the COVID-19 infodemic, which can assist mitigation strategies.

These contributions are perceived to affect the following:

1. **Public Health Communication**: By analysing South African COVID-19 related Fake News in South Africa, the study enhances public health communication strategies. It helps health authorities identify prevailing false information, enabling more targeted and effective responses.
2. **Information Integrity**: The research supports information integrity on social media platforms by uncovering patterns and sources of Fake News. These insights can inform policymakers, social media platforms and researchers in developing stronger verification mechanisms.
3. **Social Media Literacy**: The findings can inform educational campaigns aimed at improving public social media literacy, helping individuals critically assess the content they encounter online and better understand Fake News dynamics.

4. **Policy and Regulation**: The insights generated could influence policy-making regarding digital communication and public health, highlighting the need for regulations that balance free expression with the prevention of harmful Fake News.
5. **Academic Contribution**: This research contributes to fields such as data science, public health informatics and social media analytics, bridging computational methods with social science research to provide a novel approach to studying infodemics.

1.8 Limitations and Delimitations

This research analyses the COVID-19 infodemic on Twitter within the context of South Africa, specifically focusing on Fake News detection, sentiment analysis and social bot detection. The research is regionally constrained and employs a finite set of analysis tools, which could be expanded in future studies to enhance the depth of analyses. While the study identifies Fake News, it does not examine mitigation strategies, which is an important area for future research.

The dataset used for developing a COVID-19 Fake News detection model predominantly consists of English-language tweets. Future work could incorporate other South African languages, such as Zulu and Xhosa, to provide a more comprehensive view of local COVID-19 discourse. Additionally, the study restricted its analysis to specific COVID-19 keywords. Expanding this keyword list in future research would provide a broader understanding of the pandemic's online discourse.

The study exclusively considers textual tweets from Twitter, without analysing embedded videos, audio or images. Multimedia analysis requires specialized techniques beyond the scope of this research. The data collection period, from 08 November 2019 to 19 July 2021, focuses on content during the pandemic's peak but restricts the study to a defined historical window.

Twitter was selected because of its accessibility and widespread use, though it is acknowledged that other platforms, such as Facebook and WhatsApp, may have hosted relevant discussions. These platforms were excluded because of technical and privacy challenges in data access. Further research could explore Fake News propagation on these alternative platforms. Additionally, not every Twitter user

affected by the COVID-19 discourse actively tweeted about it, limiting the study's capture of public sentiment.

Finally, the study excludes broader societal and economic impacts, such as supply chain disruptions and corruption, which fall outside the scope of this analysis. These aspects, while significant, are recommended for future exploration.

1.9 Ethical Considerations

Only data publicly available on social media platforms was sourced and processed in accordance with the South African Protection of Personal Information Act, 4 of 2013 (POPIA) (South Africa, 2013).

1.10 Chapter Synopsis

The thesis is organised as follows:

Chapter 1 introduces the background, aims and objectives of the study, while highlighting the significance and contributions of the research.

Chapter 2 critically reviews the literature on South African-related COVID-19 Fake News on Twitter, examining key concepts and analytical frameworks. It identifies contributions and limitations of prior research, establishes the background for analysis, and delineates research gaps that justify the study's relevance and contributions.

Chapter 3 discusses the relevant theoretical and analysis frameworks guiding the study. It contains literature on many ML algorithms utilized in the development of a COVID-19 Fake News detection model. The complexity and technical depth of this chapter warrants its separation from the methodology chapter.

Chapter 4 provides a detailed exposition of the research methodology and design in analysing the South African COVID-19 infodemic on Twitter. It pedantically describes the data sources and procedures for data handling within the study. Moreover, it details the nine steps of the data analytics lifecycle applied specifically to this research, offering insight into the systematic approach adopted for data processing and analysis.

Chapter 5 presents the findings and results of the study, addressing each research question. It serves as the empirical foundation for the study's conclusions.

Chapter 6 discusses the findings in relation to each objective and research question, contextualizing them within the relevant literature.

Chapter 7 concludes the thesis, summarizing the key research achievements and proposing future avenues for investigation.

Chapter 2: Literature Review

2.1 Introduction

This chapter undertakes a critical examination of the existing literature pertinent to the COVID-19 Fake News on Twitter within the South African context, through an exploration of the theoretical underpinnings, key concepts and terminologies integral to the study. It contextualizes the investigation within the broader academic discourse, identifying the contributions and limitations of prior studies. The chapter establishes a comprehensive background and setting for the subsequent analysis and discussions that underpin the study's objectives. The literature review highlights the current state of knowledge and delineates the research gaps that this study seeks to address, thereby justifying its significance and potential contribution to the field.

2.2 COVID-19

COVID-19 is acknowledged as a deadly pandemic associated with millions of deaths since its outbreak and worldwide spread within a short duration, prompting drastic actions by governments and organisations (Ghebreyesus, 2020). COVID-19 is a disease caused by a new type of coronavirus known as SARS-CoV-2. The alleged origin of this pandemic can be traced to the outbreak that occurred in China in December 2019, as stated by Guan *et al.* (2020), Wu and McGoogan (2020), Zhu *et al.* (2020) and Zu *et al.* (2020). However, it is important to note that China denies this claim (Jaworsky and Qiaoan, 2021). The focus of this study pertains to the dissemination of information regarding the disease, rather than its point of origin.

The respiratory infection caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) can manifest in either the upper respiratory tract, including the sinuses, nose and throat, or the lower respiratory tract, encompassing the windpipe and lungs. The World Health Organisation (WHO) officially declared COVID-19 a pandemic on 11 March 2020 (Cucinotta and Vanelli, 2020; Ghebreyesus, 2020) and according to definitions provided by the South African Government (2020a), a *medical outbreak* is a sudden rise in cases of a disease in a particular place, an *epidemic* is a large outbreak, while a *pandemic* means a global epidemic. Horton (2020) goes further, arguing that COVID-19 is not a pandemic but rather a *syndemic*. A syndemic, as described by Singer and Clair (2003), refers to the co-

occurrence and interaction of two or more concurrent epidemics or disease clusters within a population. This synergistic phenomenon heightens the prognosis and burden of disease, implicating biological interplay as a significant factor in the progression and impact of health issues within the affected community. While there remain varying opinions on COVID-19, it does not detract from the global impact it has had on numerous aspects such as health, the economy and politics.

The COVID-19 infodemic refers to the overwhelming surge of information—both accurate and misleading—surrounding the pandemic, which has significantly impacted public health responses globally. The World Health Organization (WHO) utilized the term “infodemic” to describe this phenomenon, highlighting the challenges posed by the rapid spread of misinformation and disinformation through social media and other communication channels (Golos *et al.*, 2023; Choi, 2024). This infodemic caused confusion and complicated efforts to manage the pandemic effectively, which contributed to the adoption of alternate and sometimes harmful behaviours (Chen, Lee and Lin, 2022; White *et al.*, 2022; Golos *et al.*, 2023). As a result, addressing the infodemic has become a critical component of public health strategies, necessitating the implementation of effective communication practices and the promotion of health literacy to ensure that transparent and accurate information prevails over misinformation (Eysenbach, 2020; Lohiniva *et al.*, 2022). This study contributes to these efforts by analysing the COVID-19 infodemic utilizing data science, which offers insights and approaches that can assist public health in their communication and mitigation strategies.

Given the unprecedented nature of COVID-19, it is essential to explore the consequential social dialogue within communities. The pandemic generated billions of online and offline conversations among experts and concerned citizens. This study includes the detection of Fake News, recognizing its potential to cause significant economic, social, medical and political harm through the spread of misleading and false information. Twitter (now X), a widely utilized social media platform, emerged as a valuable resource for examining discussions on COVID-19-related Fake News. Its significance is highlighted by its frequent application in various studies such as sentiment analysis (Saif, 2015; Khan, 2019), hate speech (Waseem and Hovy, 2016; Watanabe, Bouazizi and Ohtsuki, 2018) and rumour (Bondielli and Marcelloni, 2019;

Tian *et al.*, 2020) detection. This research specifically analysed COVID-19 discussions within a South African context.

2.3 Infodemiology and Infoveillance

Infodemiology, a crucial aspect of public health informatics, is devoted to the analysis of information distribution and its determinants across electronic platforms, especially the internet, aiming to enhance public health and inform policy decisions (Eysenbach, 2009). This discipline's capability for real-time data acquisition and analysis marks a significant development from traditional epidemiological approaches (Mavragani, 2020; Klimiuk and Balwicki, 2024). It encompasses a wide range of applications, from analysing Google search queries for early disease outbreak detection to employing Twitter for syndromic surveillance (Chew and Eysenbach, 2010; Alshahrani and Babour, 2021; Bacsu *et al.*, 2022). Moreover, it includes monitoring online public health trends and utilizing automated tools to evaluate the dissemination of information and the effectiveness of knowledge translation strategies (Zeraatkar and Ahmadi, 2018; Apolinardo-Arzube *et al.*, 2019; Zhao *et al.*, 2024). The flexibility and promptness of infodemiology present substantial benefits in tackling public health issues by offering deeper insights into the dynamics of information flow and its influence on health behaviours and outcomes (Eysenbach, 2009; Mavragani, 2020).

Infoveillance, a term that originated within the infodemiology domain, refers to the surveillance of health-related information on the internet through automated, continuous analysis of unstructured text data (Eysenbach, 2009). This approach enables health professionals to track health-related discussions and misinformation trends online, allowing for timely targeted public health interventions and campaigns to mitigate misinformation. Through infoveillance, public health entities proactively engage through health marketing and informational campaigns to counteract misinformation and guide public perception and behaviour in more informed directions.

Abd-Alrazaq *et al.* (2020) applied an infoveillance strategy to identify prominent topics on Twitter amidst the COVID-19 pandemic. Their analysis identified 12 topics themed into four categories, namely, origin of the virus; information sources; impact on people, countries and economies; and techniques to mitigate effects of the virus. The authors studied tweets within a timeline of about two months, which limited the

number of topics identified. This current research studied tweet content across a timeline of more than 12 months, which provided more evidence to the results yielded.

2.4 Twitter, Tweets and Taxonomy

Twitter, now known as X after its purchase by the X consortium, is a popular social networking and microblogging tool released in 2006. The use of Twitter and www.twitter.com was mainly utilized and referenced in this study to maintain consistency with the period at which the research commenced, as X and its www.x.com domain was later introduced officially in July 2023 (Mehta, 2023). In terms of content usage growth on the platform, the tweet-rate of Twitter rose from 5 000 tweets per day in 2007 (Weil, 2010) to 500 million tweets per day in 2020 (Sayce, 2021), representing a one-hundred-thousand-fold increase. This exponential increase underlies Twitter's growing popularity.

Twitter offers a wealth of information about online communities, insights that are often challenging to obtain through traditional research methods such as interviews, surveys and questionnaires (McKenna, Myers and Newman, 2017). This rich repository of data, coupled with other advantageous attributes of Twitter, significantly enhances its utility as a credible data source for research. These benefits, when weighed against its limitations, render Twitter an effective and valuable tool for academic inquiry (Ahmed, 2018).

Twitter facilitates user interaction through the posting of concise messages known as tweets, which can include text, audios, images and videos. Tweets were originally limited to 140 characters and was later doubled to 280 characters to accommodate more expressive content (Rosen, 2017; Perez, 2018). For Twitter Blue subscribers, this character limit extends to 10 000 (Weatherbed, 2023). A distinctive feature of Twitter is the utilization of hashtags, denoted by the symbol #, which serve as dynamic, user-generated metadata tags. These hashtags enable users to categorize their messages or participate in broader conversations on specific themes, thereby enhancing the discoverability of tweets on particular topics (Perez, 2018).

Twitter has notably enhanced the utilization of various forms of expression, including abbreviations, emoticons, wit, irony, humour, slang, colloquialisms and the strategic use of uppercase and lowercase letters, alongside the employment of multiple hashtags and languages (Sykora, Elayan and Jackson, 2020). This eclectic mixture

of expressive tools leads to the categorization of tweet content from Twitter as “unstructured data”. The heterogeneity and complexity inherent in such unstructured data pose a significant analytical challenge, necessitating sophisticated methodologies and tools to derive meaningful insights. The intricate nature of this data reflects the diverse and dynamic ways in which individuals communicate on digital platforms, underscoring the challenges faced by researchers in parsing, understanding and analysing digital discourse in its myriad forms.

On Twitter, the utilization of multiple hashtags within a single tweet enables the convergence or aggregation of conversations on related topics or even ambushing topics. For instance, a tweet tagged with #sport that also includes information about football, marked with #football, will capture the attention of both sport enthusiasts and football fans specifically. It is important to note that hashtags are not case-sensitive, meaning #COVID-19, #covid-19, #Covid-19 and #CoVID-19 are all interpreted as the same tag by Twitter’s system. Among these variations, #CoVID-19 might be considered the most syntactically clear, which could contribute to its widespread use.

Twitter users are identified by unique usernames, prefixed with the @ symbol, known as handles. This convention allows for direct engagement and mentions within the platform. For example, Barack Obama’s Twitter username could be Barack, but his public handle is *@BarackObama*, facilitating easy identification and interaction. The taxonomy of a tweet, encompassing its structure and the various metadata associated with it, is further delineated in Figure 2.1 and Table 2.1. These illustrative tools provide a comprehensive breakdown of the components constituting a tweet, elucidating the framework for understanding tweet dynamics and user interactions on Twitter.

Twitter establishes two principal forms of directed relationships among its users: “friend” and “follower”. When User A elects to add User B as a friend, this action automatically designates User A as a follower of User B, thus classifying User B as a friend of User A. The relationship can be reciprocated if User B adds User A as a friend, a process colloquially referred to as “following back” or “returning the follow”, although such reciprocation is not mandatory for the relationship to exist.

Twitter information flow is straightforward: content, or tweets, are disseminated from the originator to their followers and others. Specifically, a user posts a tweet, and the content becomes accessible on their own profile page as well as on the timeline of

their followers, facilitating a broad distribution of information. This dynamic allows for the rapid and widespread sharing of content across the network, underscoring Twitter's utility as a powerful platform for information dissemination (Chu *et al.*, 2012).

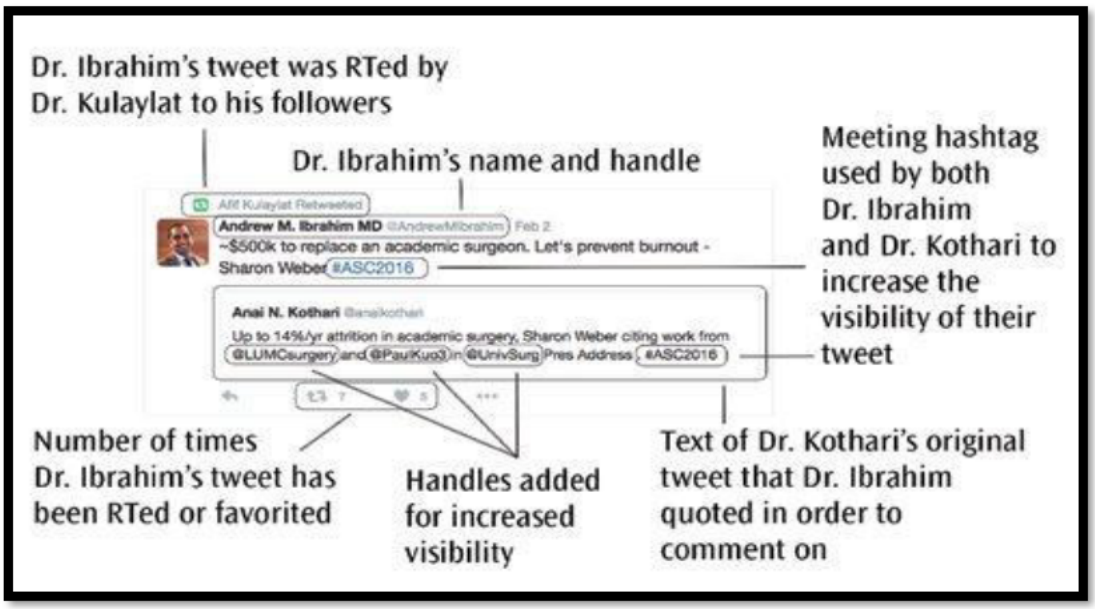

Figure 2.1: Annotated Taxonomy of a Tweet (Ferrada *et al.*, 2016)

Table 2.1 explains the common metadata of a tweet, providing a structured framework for understanding the specific attributes associated with Twitter communications.

Table 2.1: Common Metadata of a Tweet

Data	Description
Tweet	A message, up to 280 characters, may also include emoticons and symbols, allowing for succinct focused communication.
Time Stamp	The specific date and time the tweet was posted, recorded according to the Gregorian calendar, offering a temporal context of the message.
Username/ handle	The distinct identifier for the user who posted the tweet, facilitating user recognition and interaction within the platform.
Source of Tweet	Identifies the device or application used to post the tweet, providing insight into the user's access method.
Favourite	Marks a tweet that a user has tagged as a favourite, indicating appreciation or interest.
Retweeted	Denotes a tweet that has been shared or forwarded by another user, signifying its wider dissemination beyond the original audience.
Hashtag	An optional tag that users incorporate to categorize their tweet or connect it to a broader conversation on a specific topic, enhancing discoverability and engagement.

The social media landscape is vast and diverse, extending beyond microblogging platforms like Twitter and Tumblr. This includes online networks such as Facebook, WhatsApp and LinkedIn, which facilitate personal and professional connections; blogging platforms like WordPress, offering spaces for extended discourse; and social news platforms such as Reddit, where community-driven content curation prevails. Others include social bookmarking services like Pinterest and StumbleUpon, sharing economy services such as Uber and Airbnb, and media sharing sites such as Instagram, Flickr and YouTube, all of which enrich the ecosystem by enabling content discovery, sharing and service exchange. The landscape is further diversified by collaborative collective knowledge building networks like Wikipedia, interactive question-and-answer sites such as Quora for knowledge exchange, and consumer review platforms like TripAdvisor that guide consumer decisions (Barbier and Liu, 2011; Gundecha and Liu, 2012; Gandomi and Haider, 2015: 142; Sriram, Li and Hadaegh, 2021).

These platforms illustrate the expansive and multifaceted nature of social media, underscoring its role as an integral and comprehensive network of platforms facilitating a wide spectrum of digital social interactions.

2.5 Review of Relevant Twitter Hashtags

Twitter has been acknowledged for its dual impact—both positive and negative—on society. Numerous studies have delved into the influence of Twitter and its hashtags across various domains. In this study, they include its role in disaster management and terrorism (section 2.5.1), its application in infodemiology, particularly during disease outbreaks (section 2.5.2), and its involvement in anti-government protests (section 2.5.3). The intricate relationship between Twitter and electoral processes, especially concerning the use of bots, is explored (section 2.11).

2.5.1 Twitter use in Disaster Management and Terrorism

Twitter's immediacy and ease-of-use render it a pervasive, effective tool for disaster management. Its real-time capabilities allow both victims and rescuers to share their predicaments and solutions on a centralized platform. The platform's character limit encourages concise, factual communication, thereby reducing emotional content. Additionally, Twitter's geolocation feature can be instrumental in locating missing

individuals. The platform's current framework is adaptable to natural disasters such as floods, typhoons, earthquakes, tsunamis and human-induced crises like riots and terrorist attacks (Singh *et al.*, 2019). Japan, for example, experienced an earthquake (Cho, Jung and Park, 2013) followed by a tsunami (Acar and Muraki, 2011), leading to a breakdown in traditional communication infrastructures. However, Twitter and Facebook remained functional, demonstrating their utility in crisis communication (Bruns and Liang, 2012). Furthermore, Twitter data has been employed for earthquake detection, utilizing semantic analysis of search query data with keywords like *'earthquake'* and *'shaking'* (Sakaki, Okazaki and Matsuo, 2010).

Twitter data has been employed in the study and analysis of terrorism and social unrest (Compton *et al.*, 2014). Notably, during the 2011 London Riots, perpetrators exploited social media for covert communication. Rioters used the encrypted messaging service of Blackberry Messenger to orchestrate acts of violence and looting, leading to the term "Blackberry mob" and earning the moniker "riot phone" (Millham and Thakur, 2016). The 2017 Barcelona terror attack provoked Twitter users to post, en masse, images of 'cute cats'[2] to obstruct the spread of graphic content and subdue the impact of violent imagery (Woods, 2017).

In South Africa, Twitter proved instrumental in a remarkable rescue operation. A victim of carjacking[3] managed to send a tweet from the trunk of his hijacked car to his girlfriend who then retweeted this plea for help. The message quickly went viral, leading to the rescue of the victim and his vehicle within just three hours, showcasing the platform's potent role in emergency situations (Millham and Thakur, 2016).

2.5.2 Twitter and Disease Outbreaks

Ahmed (2018) highlights that research conducted on Twitter has produced socially beneficial insights within the healthcare sector, revealing a broad array of public opinions that span both positive and negative sentiments towards a variety of health-

[2] It is important to differentiate this from the concept of 'deadcatting', which is a strategic diversion tactic. Deadcatting involves introducing a dramatic, shocking or sensationalist topic to shift the discourse away from a topic that could be more harmful or damaging.

[3] In South Africa, the term 'carjacked' specifically refers to a vehicle being hijacked or forcibly taken under threatening conditions.

related topics. This platform has served as a fertile ground for examining diverse viewpoints, ranging from attitudes towards vaccination, as explored by Salathé and Khandelwal (2011), to discussions on dementia highlighted by Robillard *et al.* (2013). Twitter also facilitated discourse on sexual risk behaviours (Young, Rivers and Lewis, 2014) and public perceptions of marijuana usage (Cavazos-Rehg *et al.*, 2015).

Ahmed, Demartini and Bath (2017), Ahmed (2018), and Ahmed *et al.* (2019) underscored the importance of leveraging social media platforms such as Twitter to monitor public sentiments and opinions regarding various disease outbreaks, including swine flu, Ebola and the Zika virus. This capability of social media to act as a public sentiment barometer is crucial for health authorities in pinpointing misinformation, understanding public concerns, and responding to unresolved inquiries. Ahmed (2018) utilized a qualitative methodology for the analysis, engaging in the meticulous manual coding of tweets—a process that proved to be both time-intensive and demanding in terms of human labour (Alotaibi *et al.* 2022). The contributions of the authors to the health sector are notable, especially given the persistent global threat posed by viruses like Zika and Ebola, which continue to claim lives in the absence of comprehensive cures. The narrative sharing and exchange of experiences by those affected on Twitter become particularly relevant. A case in point is South Africa's response to a lethal Listeria outbreak, where the health authorities' active search for its origin was significantly aided by the #Listeria trend on Twitter, leading to the identification of a factory in Polokwane as the source (Van Dyk and Malan, 2018). This exemplifies the power of Twitter in aiding public health investigations and engaging communities in critical health discourse.

Social media posts detailing experiences during epidemic outbreaks contribute to the field of infodemiology. This discipline enhances understanding and knowledge about diseases. The application of infodemiology data for surveillance, known as infoveillance, enables analysts to track pandemics (Ahmed, 2018), thereby supporting their containment or eradication. While threats like Ebola and Zika are perceived as distant yet plausible, other risks such as Listeria present immediate and tangible dangers (Van Dyk and Malan, 2018). Infodemiology could accelerate the identification of disease sources. This highlights the potential societal benefits of analysing Twitter discourse concerning food and health risks, suggesting a valuable contribution to

public health management and societal well-being. This study explores Twitter conversations for COVID-19-related misinformation, which is a sources of public fear.

2.5.3 Twitter and Anti-government Protests

Twitter serves as a pervasive simple platform for protesters seeking to amplify awareness about various issues. This is highlighted by Ahmed (2018: 79), who points out:

> *"Twitter messages are easy to write and easy to read, are by default public and messages can be found and read in a variety of ways. Unlike email, unlike texting, unlike messages on Facebook, these messages are in the public domain."*

The connection between social media and activism was distinctly illustrated by Alsaedi, Burnap and Rana (2017), who utilized Twitter to research, track and predict riots and protests. Anecdotal and scholarly research suggests that Twitter acts as a catalyst for activism, as evidenced by Meijer (2017) and van der Vyver (2017), noting its potential for both positive outcomes (Bosch, 2017) and negative repercussions (Rotman *et al.*, 2011; Woode-Smith, 2016; Taghavi, 2017). A notable instance of this was the Arab Spring, a series of anti-government protests and uprisings that erupted across the Middle East in late 2010, initially sparked in Tunisia as a reaction to repressive regimes and poor living standards. This movement led to the unprecedented overthrow of four of the region's longest-serving leaders since 1927, a historical shift documented by Comninos (2011), González-Bailón *et al.* (2011) and Howard and Hussain (2013).

Howard and Hussain (2013) provide a threefold rationale for the Arab Spring's effectiveness. First, they emphasize the pivotal role of social media in framing the political discourse of the movement. Second, they note that surges in online revolutionary dialogues were typically preceded by significant on-the-ground events. In this context, this study similarly conjectures what on-the-ground event, if any, caused an increase in COVID-19 conversations in South Africa. Third, social media platforms were instrumental in disseminating democratic ideals beyond national boundaries. Importantly, the most compelling testament to the significance of digital media during the Arab Spring comes directly from the activists involved. Characterized as a grassroots movement, the Arab Spring notably circumvented

traditional political structures, including opposition leaders and other political figures not in power, a trait it shares with movements like #FeesMustFall (Khan, 2019; Khan, Thakur, Obiyemi and Adetiba, 2022a). The resulting regime changes marked the inception of a new era for these nations, with unforeseen political outcomes (Howard and Hussain, 2013). In essence, Twitter hashtags proved to be effective tools for mobilizing public opinion in situations ranging from natural disasters and health crises to anti-government protests.

The work of Ahmed, Demartini and Bath (2017), Ahmed (2018) and Ahmed *et al.* (2019) has significantly illuminated the crucial role of Twitter in monitoring public sentiments and perspectives on outbreaks of diseases such as swine flu, Ebola and the Zika virus. This monitoring is instrumental for health authorities in pinpointing misinformation, understanding public apprehensions and responding to unresolved inquiries. Reflecting on these insights, the present research corroborates the utility of Twitter in aiding authorities to combat misinformation effectively. Building on this foundation, the current study utilizes Twitter as a strategic instrument to investigate the dynamics of Fake News pertaining to COVID-19, focusing specifically on the South African milieu. This approach underscores the significance of leveraging social media analytics to enhance public health responses and interventions in the digital age.

2.6 Fake News

Fake News has been described in various ways. Allcott and Gentzkow (2017: 213) describe Fake News as, “news articles that are intentionally and verifiably false”, while Sharma *et al.* (2019: 4) negate intention and describe Fake News as, “*a news article or message published and propagated through media, carrying false information regardless the means and motives behind it*”. In trying to understand the core components to define Fake News, Tandoc, Lim and Ling (2018) reviewed 34 academic publications between 2003 and 2017 that contained the term ‘*Fake News*’ and asserted that facticity and deception are the basis for defining Fake News. This study incorporates the description offered by Sharma *et al.* (2019).

Elhadad, Fun Li and Gebali (2019) explain that Fake News may be further categorised into *Misinformation, Disinformation* and *Malinformation* as established by the Council

of Europe's taxonomy of inaccurate information (Wardle and Derakhshan, 2017). These three distinct categories are defined by veracity and intent as follows:

- **Misinformation** refers to information that, despite being incorrect, is not disseminated with the intention of causing harm.
- **Disinformation** describes information that is both incorrect and purposefully propagated to inflict harm upon an individual, social group, organization, or nation.
- **Malinformation** encompasses information that, while rooted in fact, is manipulated or presented in a manner intended to cause harm to an individual, organization, or country.

This categorisation facilitates a nuanced analysis of the phenomena of Fake News, allowing researchers to examine the relationship between the veracity of information and the motivations for its circulation within the public sphere. This study acknowledges this and delimits this opportunity for future research**.**

Fake News is inherently subjective, as its definition often depends on the interpretation of truth. Political bias, for instance, can influence what is classified as false information. Additionally, entities with substantial financial or social influence may label content as Fake News to protect their image and maintain the status quo. Addressing such biases is beyond the scope of this study.

2.6.1 COVID-19 Fake News Mitigation

To mitigate the spread of COVID-19 misinformation, several strategies and interventions have been proposed in the literature that emphasized the need for large-scale empirical investigations of the diffusion of misinformation and its social origins. (Vosoughi, Roy and Aral, 2018). Bin Naeem and Kamel Boulos (2021), in particular, suggest strategies such as machine learning-based approaches, health literacy guidelines, checklists, myth busters and fact-checkers to detect and prevent the spread of COVID-19 misinformation. Kumar *et al.* (2022) also assisted by planning a scoping review to compile evidence for designing effective interventions for COVID-19 misinformation. Chong *et al.* (2022) reviewed the segmented dissemination of COVID-19 misinformation on social media platforms to Asian Americans. On the other hand, Sule *et al.* (2023) characterized the types of COVID-19 misinformation

propagated by US physicians and the online platforms used for spreading misinformation. Ahmed and Rasul (2022) emphasized the importance of understanding social media news consumption behaviour in relation to COVID-19 misinformation.

Furthermore, Joseph *et al.* (2022) conducted an extensive scoping review to elucidate the genesis and consequences of misinformation concerning COVID-19 circulating social media, along with exploring viable approaches to mitigate its dissemination. Hadlington *et al.* (2023) highlighted the rapid spread of Fake News and misinformation during the COVID-19 pandemic. Concurrently, Ng *et al.* (2023) synthesized evidence surrounding the spread of complementary, alternative and integrative medicine-specific COVID-19 misinformation on social media. Ugarte and Young (2023) suggested that peer-led online community groups may help reduce the spread of COVID-19 misinformation among essential workers. Rao, Morstatter and Lerman (2022) asserted the urgency of addressing politically polarized content and misinformation about the COVID-19 pandemic. Ngai, Singh and Yao (2022) retrieved anti-vaccine misinformation from leading Fake News databases to track and debunk COVID-19 misinformation.

Zhao *et al.* (2022) emphasised the need for a comprehensive understanding of the virality of COVID-19-related misinformation on social media. Vivion *et al.* (2022) proposed pre-bunking messaging as an effective strategy to 'inoculate' against COVID-19 vaccine misinformation. Osborne *et al.* (2022) stressed the importance of understanding how different online communities engage with COVID-19 misinformation for an effective public health response. Pomeranz and Schwid (2021) emphasized the need for governmental actions to address COVID-19 misinformation, including disseminating factual information and supporting an independent media environment. Marwitz (2021) highlighted the active role of pharmacists in combating COVID-19 medication misinformation.

This study makes a distinct contribution by focusing specifically on the South African context, employing advanced data science techniques, such as ML, for the detection of COVID-19-related Fake News on Twitter. This localized approach allows for a deeper understanding of how misinformation spreads online within South Africa's socio-political and cultural dynamics.

Fake News is exploited by certain users to disseminate deceitful, manipulative and false information on social media platforms (Aïmeur, Amri and Brassard, 2023). Some other methods to mitigate Fake News include:

- *Geographical internet communication restrictions:* This has precedence in dire circumstances such as civil unrest and war. Throughout 2020, internet connectivity was purposely shutdown or restricted at least 155 times by about 29 countries (Duggal, 2021). However, a perceived tough approach such an outright communication ban is a violation of freedom of speech and expression which not only prevents communication, but it also stopped businesses thus destroying livelihoods.
- *Legislate the non-publication of COVID-19 information by non-medical authorities on public platforms:* This, however, will drive traffic into encrypted messaging services such as WhatsApp, Signal, Telegram and worse Dark Net. The impact of this will be that infodemiologists will be unaware of the nature, extent and range of Fake News.
- *Active moderation:* Employs a large number of human moderators who are tasked to mitigate Fake News in Real-time. This has proven not to work; Facebook employs thousands of moderators with minimal success. Active moderation has consequences such as human fatigue, bias and error.

2.6.2 Anti-vaccine and COVID-19

A vaccine denier or anti-vaxxer is described as an individual who asserts that all vaccines do not work and refuse vaccines for themselves (Benoit and Mauldin, 2021). Such individuals are known to disseminate misinformation through social media platforms, leading to a decline in vaccination rates in countries like the US and the UK. This has had tangible consequences, such as the resurgence of measles—a disease that had been nearly eradicated in the US—underscoring the lethal impact of preventable diseases and highlighting the critical need for research into social media strategies to combat Fake News (ibid).

The proliferation of coronavirus Fake News has diverse origins which may include the modern anti-vax movement that emerged in the 1990s which focused particularly (then) on childhood vaccinations against diseases like measles and mumps. This movement gained momentum from a now discredited study published in *The Lancet*,

which linked vaccines to autism, causing widespread public concern. The study has been retracted with the lead author, Andrew Wakefield, struck off the United Kingdom medical register (Pan, 2020).

Furthermore, it is noteworthy to point out that vaccine hesitancy and anti-vaccine attitudes are distinct concepts. Vaccine hesitancy refers to a delay in accepting or refusal of vaccines despite their availability, influenced by various factors such as contextual, individual and vaccine-specific issues (Wilson and Wiysonge, 2020; Hasanzad, Namazi and Larijani, 2023). On the other hand, anti-vaccine attitudes involve the rejection of vaccines based on various factors including misinformation, conspiracy theories and distrust in the pharmaceutical industry (Holford *et al.*, 2023). These two concepts should not be obfuscated into censoring legitimate critique on vaccine development, its roll-outs, associated political manipulation, conflict of interests and forced implementation.

In terms of COVID-19 vaccines, the spread of misinformation on social media has had a significant impact on vaccine intent and uptake. Studies have shown that exposure to COVID-19 vaccine misinformation on social media is associated with decreased vaccine intent (Loomba *et al.*, 2021; Skafle *et al.*, 2022; Lieneck *et al.*, 2022; Vivion *et al.*, 2022).

The dissemination of COVID-19 vaccine misinformation on social media has been facilitated by the use of various strategies, including echo-chamber structures, partisan ideology and conspiracy groups, which have contributed to the polarization of vaccine-related discussions (Sharma, Zhang and Liu, 2022).

Efforts to combat COVID-19 vaccine misinformation on social media included the development of interventions such as attitudinal inoculation videos and pre-bunking messaging to counteract false information and enhance vaccine acceptance (Piltch-Loeb *et al.*, 2022; Vivion *et al.*, 2022). Furthermore, the automatic detection of misinformation about COVID-19 vaccines on social media has been identified as a crucial component to mitigating Fake News and promoting accurate information dissemination (Weinzierl and Harabagiu, 2021).

Unlike studies that concentrate solely on anti-vaccine narratives, this research analyses a broader spectrum of misinformation, encompassing various aspects of the COVID-19 discourse.

2.6.3 Sources of COVID-19 Fake News

A report from the Centre for Countering Digital Hate (CCDH) in 2021 stated that twelve individuals, infamously dubbed the 'disinformation dozen,' were behind 65% of the anti-vaccine content circulating online, These influencers, commanding a vast audience of 59 million followers combined, were identified as the source of 73% of the content opposing vaccinations on Facebook and 65% of such messages across other significant platforms, including Twitter, Instagram and YouTube. The CCDH (2021: 5) asserts,

> "*Despite repeatedly violating Facebook, Instagram and Twitter's terms of service agreements, nine of the Disinformation Dozen remain on all three platforms, while just three have been comprehensively removed from just one platform*."

CCDH has proposed using backward tracing algorithms to track and trace accounts spreading Fake News. Remedies include restricting, removing, or de-platforming these accounts, potentially saving lives (CCDH, 2021). The efficacy of these solutions is recommended for future work. Additionally, certain hashtags, such as #IndiansMustFall, have a negative impact, contributing to localized harm. In response, platforms like Facebook have blocked hashtags like #VaccineKills, deeming them harmful although critics argue that these actions were delayed (Shah, 2021). While social media platforms possess the capability to implement CCDH's recommendations, there remains a need to strike a balance between preserving freedom of expression and avoiding censorship, which could drive users to competing platforms.

2.6.4 The KwaZulu-Natal (KZN) Riots Disinformation Dozen

A similar number of influencers, specifically 12 individuals, incited a parallel sequence of adverse violent occurrences within the KZN province of South Africa during the period spanning from 8th to 12th July 2021. The events commonly referred to as the "*KZN riots*" led to significant financial losses in the billions of South African rands and hundreds of individuals losing their lives. The riots were purportedly coordinated with the intention of expressing opposition to the incarceration of ex-President Jacob Zuma (Africa, Sokupa and Gumbi, 2021). The Centre for Analytics and Behavioural Change (CABC) portrays itself as a pivotal initiative dedicated to the identification and

mitigation of mis- and disinformation, Fake News and online rhetoric that aims to divide and polarize, thereby threatening social cohesion, the integrity of democracies and the stability of nation-states. The establishment and operations of the CABC underscore the critical need and relevance of open research in these domains.

In its analysis, the CABC discovered a significant pattern within the discourse surrounding incitement to violence, particularly noting that the majority of content disseminated by the top 12 accounts engaged in this conversation lacked originality (Mokoka, 2021). Specifically, it was observed that over 90% of the tweets from these accounts were retweets from other users who expressed support for former President Zuma, indicating a high level of echo chamber activity rather than the creation of new content. Furthermore, the hashtags associated with this discourse generated 1.29 million mentions since July 2021 as depicted in Figure 2.2, accumulating more than 1.09 million retweets and achieving a substantial reach of about 2 billion (Mokoka, 2021). This data not only highlights the scale and impact of the conversation but also the role of strategic repetition in amplifying specific narratives within the digital public sphere.

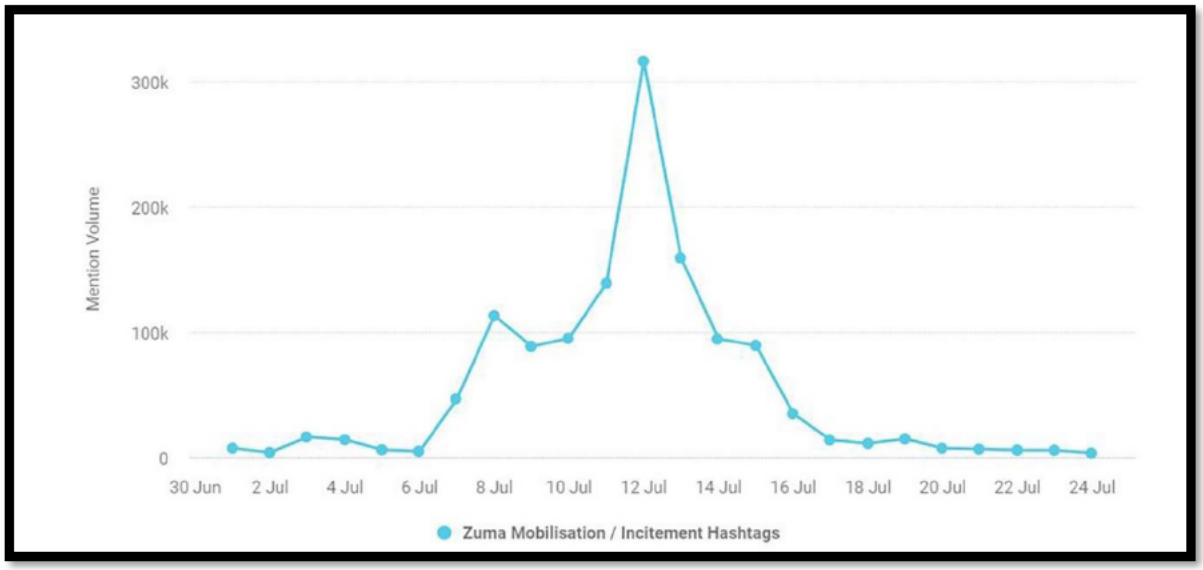

Figure 2.2: Tweet Rates by 12 Influences, KZN Riots 9–18 July 2023 (Mokoka, 2021)

It is therefore imperative to contemplate the potential impact of a larger, malevolent entity when observing the significant influence wielded by merely 12 individuals through a platform as ubiquitous as Twitter. These individuals have demonstrated the ability to orchestrate nationwide riots and propagate international anti-vaccine campaigns, showcasing the platform's power in mobilizing and influencing public

sentiment and action on a grand scale. This observation raises critical questions about the extent of what could be achieved if such tactics were employed by an organization with more extensive resources and a broader agenda. The utilization of social media for such purposes underscores the dual nature of these platforms as tools for both social engagement and potential vehicles for widespread disruption. As a result, it becomes essential to scrutinize the mechanisms through which social media can be used to amplify divisive narratives and the measures that can be taken to mitigate these risks, ensuring that these digital spaces contribute positively to public discourse and societal well-being.

2.7 Social Media, Big Data and Fake News

This section examines the role of Fake News on social media platforms, integral components of Web 2.0 technologies designed to facilitate user-generated content and foster online social networks (Obar and Wildman, 2015). The surge in social media popularity is evident, with user numbers increasing from 2.86 billion in 2017 to 4.2 billion in 2021 and an average daily usage time of 145 minutes per user (Kemp, 2021). This widespread adoption raises significant concerns about social media's potential role in disseminating and amplifying misinformation. Notably, during the 2016 United States elections, Fake News reportedly influenced voter behaviour and skewed electoral outcomes (Allcott and Gentzkow, 2017). Similarly, in South Africa, the government responded to the impact of social media by establishing platforms to counteract COVID-19-related Fake News, addressing issues such as misleading information about debt relief on Facebook and misconceptions regarding social distancing in school gatherings (South Africa, 2020a).

Social media significantly contributes to big data, encompassing diverse data formats like text, audio, images and videos. Analysing such multifaceted data poses considerable challenges, yet the field of Data Science offers methodologies and tools to effectively address many of these complexities. The academic focus on detecting and combating Fake News on social media has become prominent. Shahzad *et al.* (2022) and Al-Asadi and Tasdemir (2022) highlight the application of big data analytics and artificial intelligence tools in identifying Fake News on digital media. These studies underscore the significance of advanced technologies in combating misinformation. Shu *et al.* (2017) provide a comprehensive review of social media

Fake News and its characterization, detection algorithms, evaluation metrics and datasets, thus highlighting the interdisciplinary nature of this research field.

Furthermore, research by Pennycook and Rand (2021) illuminates factors influencing individuals' vulnerability to Fake News, such as cognitive styles and familiarity. Understanding these factors is pivotal in devising effective strategies to mitigate the impact of Fake News. Studies by Singh *et al.* (2024) and Zafarani *et al.* (2021) delve into computational models and recent trends in Fake News research, emphasizing the necessity for ongoing innovation in this domain. Soetekouw and Angelopoulos (2024) underscore the role of data mining in Fake News detection, emphasizing the importance of technological interventions. Aoun Barakat, Dabbous and Tarhini (2021) contribute to understanding users' identification of Fake News on social media, providing insights into the behavioural aspects of recognizing Fake News.

Beyond the propagation of misinformation, social media has emerged as a powerful tool for facilitating positive societal transformations, giving rise to a phenomenon known as hashtag activism (Dadas, 2017). Movements such as #FeesMustFall, the Arab Spring, the Umbrella Revolution in Hong Kong and various social justice movements worldwide, including BlackLivesMatter, the MeToo Movement, the Yellow Vests movement in France and youth-led climate actions in the US and Europe, all highlight the critical role of social media in modern activism. These movements, characterized by their widespread influence and global reach, suggest that their emergence and success would have been markedly different, if not impossible, without the amplifying power of social media platforms.

2.8 Data Science

Data Science is a multidisciplinary field that integrates computer science, mathematics and statistics to derive insights from both structured and unstructured data, particularly in the context of big data (Oracle, 2020). It emerged from the mathematics and statistics community in 1962 (Cao, 2017). The Cross Industry Standard Process for Data Mining (CRISP-DM) is a commonly applied analytical framework within data science (Wirth and Hipp, 2000). Consequently, the data science lifecycle seeks to answer the following (Martinez, Viles and Olaizola, 2021):

i. **Business understanding –** What are the specific needs and goals of the business?

ii. **Data understanding –** What data is available and is it of sufficient quality for analysis?

iii. **Data preparation –** How should the data be organized and prepared for modelling?

iv. **Modelling –** What modelling techniques are most suitable for addressing the business requirements?

v. **Evaluation –** Which model most effectively meets the predefined business objective?

vi. **Deployment –** How can the results be made accessible and understandable to the stakeholders?

This study adopts The Big Data Analytic Lifecycle (Erl, Khattak and Buhler, 2016) which incorporates the data science lifecycle to analyse the COVID-19 infodemic. In the contemporary landscape of digital communication, data science emerged as a cornerstone for the analysis of voluminous textual data, including tweets and misinformation. The evolution of this field, marked by the integration of advanced computational models and NLP techniques, has facilitated a nuanced approach to understanding and mitigating the dissemination of Fake News across social media platforms (Zhao *et al.*, 2020; Choudhary *et al.*, 2021; Rohera *et al.*, 2022).

Recent advancements in deep learning (DL), epitomized by models such as Bidirectional Encoder Representations from Transformers (BERT), have significantly propelled the capabilities for textual data processing. Devlin *et al.* (2018) underscored the transformative impact of these models, showcasing their proficiency in tasks ranging from sentiment analysis to Fake News detection. This era of ML innovation, especially during pivotal moments such as the COVID-19 pandemic, underscores the critical role of data science in navigating the challenges posed by misinformation. Raza (2021) and Montesi (2021) delineate the application of transformer-based models in discerning factual veracity amidst the informational deluge of significant events like the 2020 U.S. Election.

Moreover, the field has seen a surge in employing ML for automatic deception detection. Righetti (2021) illuminates the forefront of Fake News research, incorporating ML and network analysis to devise robust countermeasures against misinformation. Similarly, Mouratidis, Nikiforos and Kermanidis (2021) advocate for

the utilization of sophisticated ML algorithms within social networks to identify and curtail the spread of deceptive information. These interdisciplinary research efforts underscore the amalgamation of computational techniques with behavioural sciences to foster a holistic understanding of misinformation dynamics.

The intersectionality of data science with cognitive and behavioural studies further enriches our comprehension of the mechanisms underpinning the viral spread of Fake News. For instance, Petit *et al.* (2021) explored psychological interventions capable of diminishing the propagation of misinformation. Such endeavours highlight the symbiotic relationship between technological advancements and cognitive behavioural insights in crafting effective strategies against the infodemic.

The amalgamation of data science methodologies, encompassing ML models and interdisciplinary research, is indispensable in the exploration of large textual datasets, tweets and Fake News. The confluence of artificial intelligence, NLP and cognitive behavioural studies equips researchers with a robust set of tools to dissect the intricacies of misinformation, fostering a data-driven discourse on combating Fake News. This synthesis not only elevates the analytical capabilities but also propels forward the collective efforts in understanding and addressing the multifaceted challenges presented by misinformation in the digital age.

2.8.1 Fake News and Data Engineering

In order to achieve a model that can classify Fake News using data science techniques such as ML and deep leaning, there needs to be appropriately collected, formatted and prepared data. This is often considered the data engineering phase, which strategizes and develops processes that collect information and store them into databases or datasets in a format that is clean and prepared for operational analysis (Gillis, n.d.; Layton, 2020; Reis and Housley, 2022).

Data Engineering is essential in the field of Data Science, especially in tasks such as classification using Twitter data. The utilization of large datasets from platforms like Twitter has facilitated the development of open-source data-driven approaches such as SNScrape and Twint for classification tasks on Twitter data. SNScrape, a Python library, has been utilized in studies related to sentiment analysis and thematic analysis on Twitter (Yashpal *et al.*, 2022). Twint, another Python library, has been used for gathering and analysing tweets in research focusing on sentiment analysis and

classification tasks (Asare *et al.*, 2022; Mohamed *et al.*, 2023). Researchers can utilize these tools to efficiently collect tweets, provided they are still operational.

Social media is 'noisy' and highly unstructured, and information posted on it must be collected and transformed in a format that is suitable to research and analysis. This research collected and transformed South African COVID-19-related Twitter data into suitable formats by using Python and other related techniques.

2.8.2 Unstructured vs. Structured vs. Semi-structured Data

In order to effectively implement a data science framework, it is important to distinguish between the different types of data structures. Data primarily exists in two extreme forms: structured and unstructured, with semi-structured data representing a hybrid of the two (Erl, Khattak and Buhler, 2016: 19). Structured data is characterized by well-defined data types and patterns that facilitate easy searchability, typically found in databases (Mayer-Schönberger and Cukier, 2013). On the other hand, unstructured data, which includes text files, images, audio, video and social media content (Rizkallah, 2017; Shacklett, 2023), often lacks the systematic organization necessary for straightforward machine analysis. This category of data encompasses both textual elements, like social media posts and emails and non-textual forms, such as video content and can be generated either by humans or machines.

Semi-structured data primarily falls into the category of unstructured data but includes certain structured elements within its metadata, enabling analysis through some conventional structured methods (Erl, Khattak and Buhler, 2016). Formats like JSON and XML are typical examples of semi-structured data. Contrarily, unstructured data possesses an inherent internal structure but lacks the predefined data models or schemas found in structured data (Gandomi and Haider, 2015; Shacklett, 2023). This type of data represents a significant, yet largely unresolved, challenge in the field of Information Technology (Blumberg and Atre, 2003), despite ongoing research efforts (Madhoushi, Hamdan and Zainudin, 2015). Accounting for about 95% of all data (Mayer-Schönberger and Cukier, 2013; Gandomi and Haider, 2015), it encompasses a wide array of formats and sources. In this study, the focus is on Twitter data, which is classified as unstructured, highlighting the complexities and nuances associated with its analysis.

2.8.3 Machine Learning

Machine Learning (ML) stands as a cornerstone technology in tackling a myriad of complex challenges by utilizing specific algorithms across a broad spectrum of applications (Zhai and Massung, 2016). ML techniques are primarily classified into three categories: Supervised Learning, Unsupervised Learning and a combination of both known as Semi-supervised Learning. Supervised Learning involves training a computer to derive a function $\hat{y} = f(x)$, where x is the input, supplied by human-provided examples known as the training data and y is the corresponding expected outcome. The derived function (x) is then adept at recognizing patterns and generating outcomes y for novel, unseen inputs x, epitomizing the essence of ML. This approach is termed 'supervised' owing to the model learning from the input-output pairs predefined by humans, thus establishing a supervised learning scenario. On the contrary, Unsupervised Learning focuses on analysing unlabelled data, employing algorithms such as k-means, k-medoids, hierarchical clustering, hidden Markov models and certain types of unsupervised neural networks like self-organizing maps to discover inherent structures within the data (Wade and MarzoSerugendo, 2017).

In the realm of social media studies, ML has been prevalently used as a powerful tool for classification tasks including Sentiment Analysis, Hate Speech Detection, Emotion Analysis and Fake News Detection, showcasing its versatility and efficacy (Yang and Chen, 2017; Aulia and Budi, 2019; Zhang *et al.*, 2020). While DL, a subset of ML, is frequently employed for its advanced capabilities, it is often hindered by significant cost implications due to its demand for substantial computational resources and sophisticated hardware, unlike more conventional ML models.

Given the extensive body of knowledge surrounding ML algorithms and their critical role in this study, Chapter 3 is dedicated to a description and exploration of these algorithms.

2.9 Fake News Detectors

The challenge of detecting Fake News content has prompted numerous studies, leading to a substantial body of literature that explores diverse data science methodologies for this purpose. In their survey, Bondielli and Marcelloni (2019)

researched various techniques employed by researchers for Fake News and rumour detection techniques. Their findings outlined two primary categories: classification approaches and other approaches as seen in Figure 2.3.

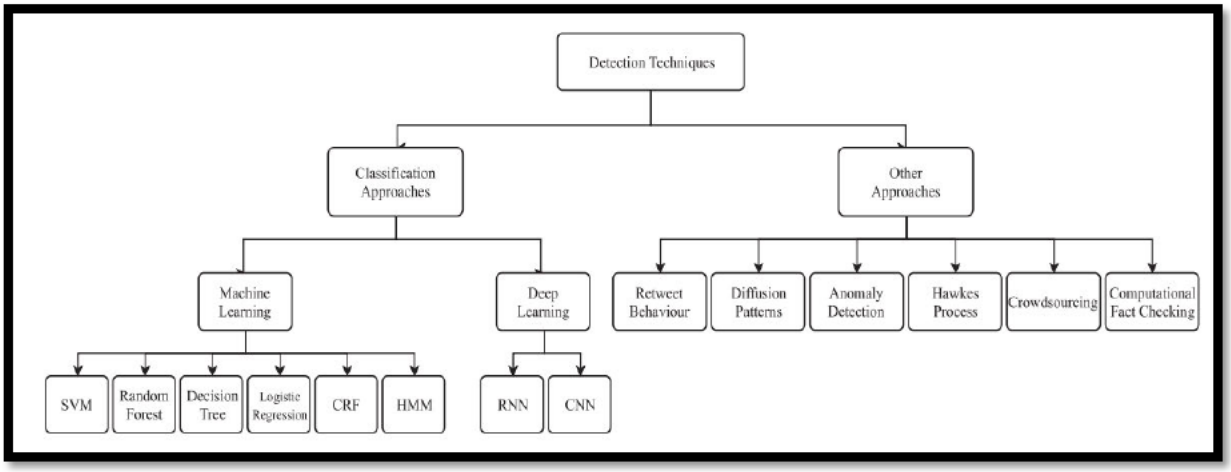

Figure 2.3: Fake News and Rumour Detection Techniques (Bondielli and Marcelloni, 2019)

Under classification approaches, two main techniques are distinguished: ML and DL. ML itself includes a variety of algorithms such as Support Vector Machines (SVM), Random Forest, Decision Trees, Logistic Regression, Conditional Random Fields and Hidden Markov Models. DL methods include Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN).

The second category, 'Other Approaches', consists of diverse strategies that do not rely on traditional classification. These include analysis of retweet behaviour, examination of diffusion patterns, anomaly detection to identify outliers or unusual patterns, the Hawkes process which models the sequence of events over time, crowdsourcing for gathering information from the public and computational fact-checking which uses databases and algorithms to verify the authenticity of information. Additionally, Wani *et al.* (2021) evaluated several supervised text classification algorithms including DL approaches such as CNN, Long Short-Term Memory (LSTM) and BERT. The models were trained on the Contraint@AAAI 2021 Covid-19 Fake News detection dataset and achieved an accuracy rate of 98.41%.

Abdelminaam *et al.* (2021) proposed CoAID-DEEP which is an optimised intelligent framework for automating the detection of COVID-19 misinformation on Twitter. Figure 2.4 represents the CoAID-DEEP framework which incorporates DL and six

regular ML techniques which yielded a significant improvement on results when compared to baseline state of the art ML models.

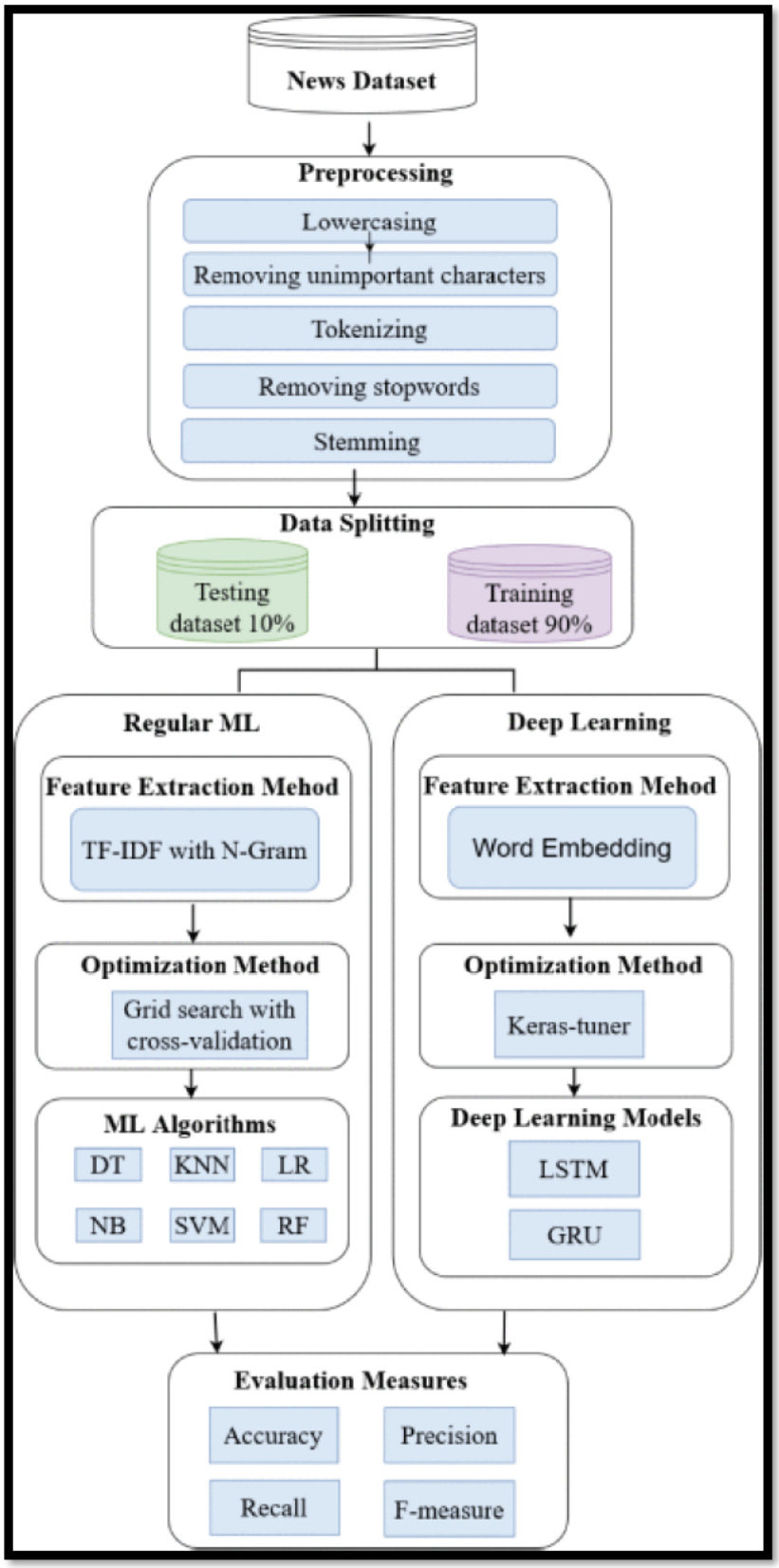

Figure 2.4: CoAID-DEEP Framework to Automatically Detect COVID-19 Misinformation (Cui and Lee, 2020)

The DL models utilised were LSTM and Gated Recurrent Units (GRU) while the six regular ML models were Decision Trees (DT), K-Nearest-Neighbours (KNN), Logistic Regression (LR), Naïve Bayes (NB), SVM and Random Forest (RF). Four Fake News labelled datasets comprising of 6 834 data points in different topics were used for this framework namely, CoAID (COVID-19 heAlthcare mIsinformation Dataset) (Cui and Lee, 2020), a Kaggle disaster dataset (Sarin and Kumar, 2020), PolitiFact and gossip cop datasets from FakeNewsNet (Shu *et al.*, 2020).

2.9.1 Fake News Characteristics

Fake News can be identified by certain patterns. The traits listed below are common to texts that are associated with Fake News:

- All text is in capital letters e.g. REVEALED: DO NOT…
- Sensationalised headlines e.g. Fountain of youth found in…
- Monetary enticement e.g. Become a billionaire in 3 months…
- Propaganda
- Unrecognisable or unknown sources

It must be noted that any text that contains elements of the above list does not necessarily mean that is fake rather it is indicative of potentially being fake and may require further investigation.

Twitter has adopted rules to assist users from viewing or sharing misinformation based on a global 6500 people survey together with advice from academic and civil society experts (Roth and Achuthan, 2020). The survey revealed that users required the following:

- Twitter should provide more information
- Different types of content should be labelled
- Removal of tweets that can potentially cause harmful
- There should be some type of enforcement actions for users who are posting harmful content

Figure 2.5 provides a guideline on how Twitter adopted these new rules:

Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **very likely** to be removed.

Figure 2.5: How Twitter Adopted Rules to Sharing Misinformation (Roth and Achuthan, 2020)

Twitter has also tested and introduced tags to its media and text posts by users which aimed to combat Fake News (Roth and Pickles, 2020). The following were some of the tags that were used:

- Manipulated
- Get the facts about COVID-19
- Misleading (Testing phase)
- Get the latest information (Testing phase)

This meant that Twitter could take action in addition to labelling based on the following three broad categories (ibid):

- Misleading
- Disputed
- Unverified claim

Figure 2.6 below is a guide on how Twitter viewed potential misinformation:

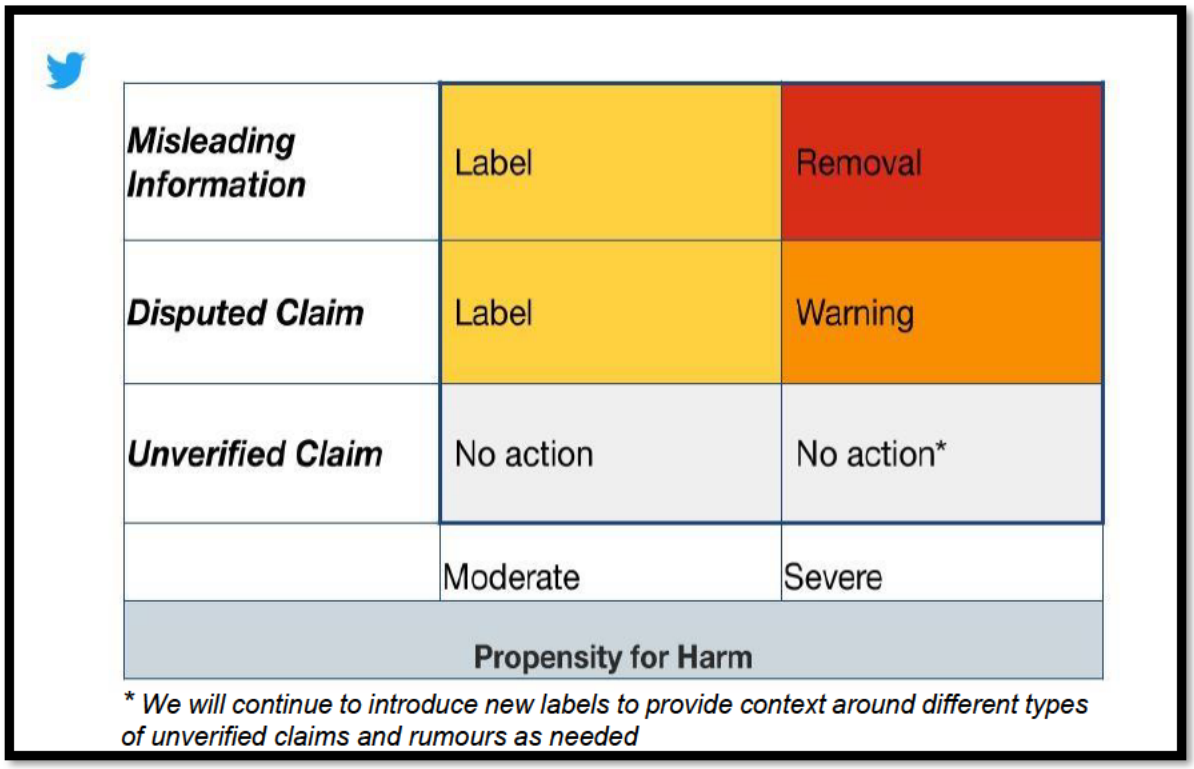

Figure 2.6: Twitter Potential Misinformation Action Guide (Roth and Pickles, 2020)

Twitter has since introduced its community-led approach to addressing misinformation, particularly through the "Community Notes" initiative. This approach emphasizes trust-building, speed and scalability in content moderation. Unlike traditional methods, Community Notes aims to leverage diverse viewpoints to ensure fairness, utilizing open-source algorithms for transparency. Twitter indicates that this initiative has shown positive results in providing context to potentially misleading content, thereby reducing its spread across the platform. Continuous improvements and community feedback are integral to its ongoing development (Coleman, 2023).

It can be argued that the reliance on community input for content moderation may introduce biases, especially if the contributor pool is not diverse enough. Additionally, the speed at which notes are added, though improved, might still lag behind the rapid spread of misinformation. The open-source nature of the algorithm and data is commendable, but it also requires continuous scrutiny to ensure it cannot be gamed or manipulated by malicious actors.

2.9.2 Fake News Datasets

There have been many diverse approaches utilised by researchers to deal with Fake News. For example, Muraya *et al.* (2021) constructed a simple mathematical model

to determine patterns for the spread of Fake News on Twitter in order to assist mitigation strategies whilst Hansrajh, Adeliyi and Wing (2021) focused on blended learning techniques to appropriately classify Fake News content.

In order to build automated classification models to detect Fake News content, there needs to be sufficient amount of labelled data available. For example, Patwa *et al.* (2021) have provided 10700 COVID-19 Fake News datasets comprised of social media posts and news articles which were manually labelled into real and fake categories. In addition, their SVM model yielded a F1-score of 93.32%, which outperformed the Decision Trees, Logistic Regression and Gradient Boost models.

Table 2.2 is reflection of various Fake News datasets provided by D'Ulizia *et al.* (2021) between 2009 and 2019.

Table 2.2: List of Fake News Datasets Provided by Researchers until 2019 (D'Ulizia *et al.* 2021)

No	Year	Dataset Name	Source
1	2019	Yelp dataset	https://www.yelp.com/dataset/download
2	2019	MisInfoText dataset	https://github.com/sfu-discourse-lab/MisInfoText
3	2019	Spanish Fake News corpus	https://github.com/jpposadas/FakeNewsCorpusSpanish
4	2019	Fake_or_real_news	https://www.kaggle.com/rchitic17/real-or-fake
5	2019	NELA-GT-2018	https://dataverse.harvard.edu/dataverse/nela
6	2019	TW_info	http://www.mdpi.com/2079-9292/8/12/1377/s1
7	2019	FCV-2018	https://mklab.iti.gr/results/fake-video-corpus/
8	2019	Verification Corpus	github.com/MKLab-ITI/image-verification-corpus
9	2019	CNN / Daily Mail summarization dataset	https://github.com/abisee/cnn-dailymail
10	2019	Rumors dataset	http://tiny.cc/p1s2qy
11	2018	BuzzFace	https://github.com/gsantia/BuzzFace
12	2018	FEVER	http://fever.ai/resources.html
13	2018	FakeNewsNet	https://github.com/KaiDMML/FakeNewsNet
14	2017	FacebookHoax	https://github.com/gabll/some-like-it-hoax
15	2017	LIAR	https://www.cs.ucsb.edu/?william/data/liar_dataset.zip
16	2017	Benjamin Political News Dataset	https://github.com/rpitrust/fakenewsdata1
17	2017	BuzzFeed News dataset	https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data
18	2017	FNC-1 dataset	https://github.com/uclmr/fakenewschallenge
19	2017	TSHP-17	https://homes.cs.washington.edu/~hrashkin/factcheck.html
20	2017	QProp	http://proppy.qcri.org/about.html

No	Year	Dataset Name	Source
21	2017	CBCNN datasets	https://github.com/chenjinyuan87/cbcnn
22	2016	PHEME dataset	https://figshare.com/articles/PHEME_rumour_scheme_dataset_journalism_use_case/2068650/2
23	2016	EMERGENT	https://github.com/willferreira/mscproject
24	2015	CREDBANK	https://github.com/compsocial/CREDBANK-data
25	2014	Fact checking dataset	https://andreasvlachos.github.io//publications/
26	2011	Deceptive Opinion Spam Corpus	https://myleott.com/op-spam.html
27	2009	Burfoot Satire News Dataset	https://github.com/rfong/satire/tree/master/corpus

All the datasets from table 2.2 contain Fake News data based on text while FCV-2018 and FakeNewsNet further contains video and image data respectively. The Verification Corpus is a standout due to it containing text, video and image data.

2.9.3 Fake News Detection Systems and Tools

The detection of Fake News on social media platforms has become increasingly crucial because of their unique characteristics and challenges. Existing methods for Fake News detection are predominantly supervised, requiring extensive time and labour to build reliably annotated datasets (Shu *et al.*, 2017; Yang *et al.*, 2019). To address language barriers, research has developed tools tailored for different languages, such as DL techniques for Vietnamese (Vo, Phan and Ninh, 2022). These advancements are essential for the broader applicability of detection systems across various linguistic contexts.

The adverse effects of Fake News have highlighted the necessity for automatic detection of social media content (Xing *et al.*, 2021). Consequently, approaches leveraging crowd signals and ensemble methods have been explored (Tschiatschek *et al.*, 2018; Reddy *et al.*, 2020). Artificial intelligence tools, including ML and DL, have been widely adopted to create robust systems for detecting Fake News across diverse fields (Al-Asadi and Tasdemir, 2022). Models such as the FAKEDETECTOR and the dEFEND system focus on explainable detection mechanisms (Cui *et al.*, 2019; Zhang, Dong and Yu, 2020), while multimedia Fake News detection utilizes information from textual, visual and social modalities (Cao *et al.*, 2020). Some innovative approaches integrate fact verification models with Fake News detection to enable zero-shot detection, thereby reducing the need for extensive training data (Li and Zhou, 2020).

Surveys and previous work have also modelled the flow of affective information for Fake News detection (Gaonkar *et al.*, 2019; Ghanem *et al.*, 2021).

Notable developments such as Hoaxy, Fakey and CoVaxxy by Indiana University's Observatory on Social Media (OSoMe) underscore the significant contributions of academic institutions in combating Fake News. Hoaxy visualizes the spread of claims and related fact-checking efforts online, enabling users to search Twitter with a seven-day historical limit on topics such as COVID-19 (Hui *et al.*, 2018). Fakey, on the other hand, is a news-oriented game designed to educate users on detecting Fake News by simulating a social media environment (Micallef *et al.*, 2021). CoVaxxy visually analyses the relationship between online misinformation and COVID-19 vaccine adoption in the United States (DeVerna *et al.*, 2021). Similarly, CoVerifi, not an OSoMe development, is a news verification system aimed at mitigating the spread of COVID-19 misinformation through a web application where users vote on news content, providing feedback that is then incorporated with ML for credibility assessment (Kolluri and Murthy, 2021).

One of the primary constraints associated with these tools is their economic accessibility, particularly for academics, researchers, non-governmental individuals, as well as small organizations and companies. Another important consideration is that vendor lock-in restricts the distribution of data or outputs that depend on proprietary software, thereby hindering the development of communities of practice.

2.10 Sentiment Analysis

Sentiment analysis, often termed opinion mining, embodies a computational methodology aimed at the discernment, extraction and classification of sentiments, emotions and opinions articulated within textual content. This domain has witnessed burgeoning interest, notably within the realm of social media analytics, where platforms such as Twitter present a fertile ground for empirical investigation (Lin and He, 2009; Kouloumpis, Wilson and Moore, 2021). The primary objective of sentiment analysis is to categorize textual expressions into positive, negative, or neutral sentiments, reflecting the underlying emotional and opinionated dimensions (Park, 2020). This entails the identification of subjective elements—ranging from attitudes and sentiments to evaluative judgments—embedded within the text (Lin and He, 2009).

The task of sentiment analysis in Twitter texts has seen scholars exploring linguistic attributes to discern sentiment orientation within tweets, a pivotal step towards comprehending user sentiments across varied subjects, ranging from events and individuals to products and organizations (Phan *et al.*, 2020; Khan, Thakur, Obiyemi and Adetiba, 2022a). The significance of sentiment analysis extends into critical applications such as the detection of Fake News, where discerning sentiment towards specific articles or entities aids in the identification of misinformation (Shu *et al.*, 2017).

The application of sentiment analysis in the context of Fake News detection on platforms like Twitter incorporates the assessment of sentiment to differentiate authentic from deceptive content. By scrutinizing the sentiment embedded in tweets, algorithms can be formulated to detect Fake News by analysing the emotional and opinionated undertones of the text (Shu *et al.*, 2017).

Sentiment analysis is instrumental in the classification of opinions, sentiments and emotions conveyed through textual mediums, particularly within the milieu of social media. The application of sentiment analysis techniques facilitates a deeper engagement with areas like Fake News detection, evaluation of customer satisfaction and analysis of public sentiment, underscoring its importance in Twitter text analytics.

2.11 Twitter Bots and Cyborgs

Within the digital Twitter ecosystem, an intricate interplay of various bot entities influences online dialogue and the proliferation of information. These automated agents are differentiated by their operational behaviours and objectives, forming a spectrum of classifications. Haustein *et al.* (2016) distinguish between bots programmed for the dissemination of bibliographic references and those engaging in interactive dialogues or providing commentary on tweet content. Furthermore, a nuanced differentiation exists among bots, categorizing them into entities such as spammers, self-promotional agents and application-based content distributors (Daniel and Millimaggi, 2020). The proliferation of such bots, especially during pivotal occurrences like COVID-19 pandemic, has been intricately linked with the dissemination of misinformation, underscoring the dual role they play in both enriching and distorting information (Himelein-Wachowiak *et al.*, 2021; Alqurashi *et al.*, 2021).

Compounding the complexity within Twitter's virtual domain is the introduction of 'cyborgs,' a term ascribed to accounts that embody attributes of both automated bots

and human operators. This hybrid categorization encompasses ‘bot-assisted humans’ and ‘human-assisted bots,’ collectively recognized as cyborgs (Matamoros-Fernández, Bartolo and Alpert, 2024). Such entities challenge conventional delineations between automated and human-driven interactions, thus complicating the task of discerning their influence on digital conversations.

In strategic communications across platforms like Twitter, cyborg accounts emerge as pivotal agents, orchestrating a blend of automation and human intervention to propagate specified narratives (Ng *et al.*, 2024). This hybridity accentuates the intricate nature of their classification and the analytical challenges they pose in understanding their roles within digital communities.

This exploration into Twitter’s bot landscape and the advent of cyborg accounts illuminates the dynamic intricacies of online engagements. It highlights the criticality of discerning the roles played by these digital actors in shaping discourse, navigating the challenges they present and unravelling the complex tapestry of human-bot interactions in the vast expanse of social media.

2.12 Conclusion

This chapter initiated a review of existing literature, focusing on the topic of Twitter and its various potential applications and drawbacks. Research has demonstrated that Twitter hashtags have proven to be a successful tool in exerting influence over the general public during times of crises and that these also may be appropriates for Fake News. The next chapter introduces some of the ML algorithms that were utilised for constructing the Fake News detection models in this study.

Chapter 3: Theoretical Analysis Frameworks

3.1 Introduction

This chapter discusses the foundational concepts that underpin our investigation into the South African COVID-19 infodemic on Twitter. The study is anchored in Information Theory as its theoretical framework, while employing ML as the primary analytical tool. Information Theory, established by Shannon (1948), offers a quantifiable lens through to understand the dynamics of information processing, transmission and reception (MacKay, 2003). Its relevance to our study stems from its ability to measure the entropy or unpredictability of information, a critical aspect when analysing the dissemination of both factual information and misinformation within the vast, complex network of Twitter.

ML, on the other hand, serves as the analytical backbone of our research, providing the computational mechanisms necessary to sift through, categorize and analyse the massive corpus of Twitter data related to COVID-19. By applying various ML models, we aim to identify patterns of misinformation, gauge sentiment and understand the spread and impact of information within the context of the pandemic. This analytical approach not only enhances the precision and depth of our analysis but also aligns with the quantitative nature of Information Theory, allowing for a synergistic examination of the infodemic.

Together, Information Theory and ML forge a comprehensive framework for this study. Information Theory's principles guide our understanding of the infodemic's nature and its implications on public discourse and perception. Concurrently, ML techniques operationalize these principles, enabling a detailed nuanced analysis of the data. This integration of theoretical insights and analytical rigor positions our study at the forefront of contemporary research on digital communication and public health crises, by offering valuable insights into the mechanisms of information spread and the mitigation of misinformation in the age of social media.

3.2 Information Theory

Information Theory is a foundational field that focuses on the quantification, storage and transmission of information (Mackay, 2003). It encompasses various aspects

such as inference, learning algorithms and the mathematical underpinnings of communication (Shannon, 1948). Key works like “Elements of Information Theory” played a crucial role in shaping the understanding of this discipline (Cover and Thomas, 2001). Information Theory has diverse applications in fields such as computer science, artificial intelligence and library science (Cornelius, 2002).

The concept of information gained importance with the advancement of information technology, where the notion of a *bit* has become central to the field (Lambrou *et al.*, 2021). Additionally, the historical origins of information theory can be traced back centuries and is a cornerstone in various scientific disciplines, offering a framework for comprehending data, communication and uncertainty (Ellerman, 2017). Its applications span from statistical mechanics to quantum information, highlighting its versatility and significance in contemporary research and now social media analytics.

The employment of Information Theory as the theoretical framework in this study is academically justified by its foundational principles that facilitate a comprehensive understanding of information’s quantification, transmission and processing (Shannon, 1948; Mackay, 2003). This framework is instrumental in the quantitative analysis of textual data within social media platforms, specifically Twitter, in the context of the COVID-19 infodemic.

Firstly, Information Theory provides a methodological basis for quantifying the information embedded in the vast amounts of textual data on Twitter. Through the measurement of entropy, it allows for the assessment of information diversity and the predictability of information content, which are crucial for identifying misinformation patterns and gauging the spread of COVID-19-related news.

Secondly, the theory’s insights into the efficiency of information transmission and the impact of noise offer a valuable perspective on the dissemination mechanisms of both accurate information and misinformation. By conceptualizing misinformation as noise within the communication channel of Twitter, the study can apply Information Theory to analyse how this noise affects the clarity and reception of intended messages about COVID-19, thereby influencing public perception and behaviour.

Moreover, Information Theory’s contributions to understanding and mitigating errors in data transmission directly translate to detecting and filtering misinformation. This aspect is particularly relevant for deploying ML techniques aimed at identifying

inaccuracies and inconsistencies in COVID-19-related tweets, enhancing the reliability of information presented to the public.

Additionally, the application of Information Theory enables the modelling of information flow within Twitter's complex network. This facilitates an exploration of the propagation dynamics of COVID-19 information, encompassing the roles of various actors within the network, such as influencers, bots and the general user base, in amplifying or counteracting the spread of misinformation.

The utilization of Information Theory as the theoretical backbone of this study not only enriches the analysis with a robust and scientifically rigorous framework but also aligns perfectly with the study's goal of employing quantitative methods and ML techniques for the analysis of textual data. This theoretical alignment underscores the study's academic rigour and contributes significantly to its aim of elucidating the nature and dynamics of the COVID-19 infodemic on Twitter within the South African context.

3.3 Analysis Framework – Machine Learning

This section provides an overview of the ML algorithms applicable to the development of the Fake News detection models. The need to implement automated methods for identifying Fake News arises from the impracticality of relying on manual techniques, given the sheer volume and velocity that tweets are generated on the Twitter platform. Moreover, the study opted for a supervised ML approach, given its objective of classifying tweets as either 'Fake' or 'Not Fake' using pre-annotated datasets.

3.4 Shallow Machine Learning Algorithms

Shallow learning models typically refer to traditional ML algorithms that do not involve neural networks with one or two hidden layers or require multiple levels of feature abstraction (Janiesch, Zschech and Heinrich, 2021). Conventional and baseline supervised ML algorithms such as Decision Tree, Random Forest, K Nearest Neighbors, Naïve Bayes, Logistic Regression and SVM are discussed in the following sections.

3.4.1 Decision Tree Models

3.4.1.1 Decision Tree

The Decision Tree algorithm stands as a robust method for addressing both regression and classification challenges within data analytics. It functions by constructing a predictive model from training data, which ascertains the class or predicts the value of a targeted variable through the derivation of straightforward decision rules. The foundation of this methodology is the root of the tree, a pivotal node that initiates the prediction process. The algorithm undertakes a comparison between the attribute value at the root and the corresponding attribute in the data record. Following this evaluation, the algorithm traverses the tree along the branch that aligns with the compared value, progressively moving to subsequent nodes until a prediction is rendered (Maimon and Rokach, 2014; Holzinger, 2015; Charbuty and Abdulazeez, 2021; Chauhan, 2022). This algorithm provides a straightforward approach to classification by creating a model based on decision rules. Its application in COVID-19 Fake News detection has been beneficial for understanding feature importance (Patwa *et al.*, 2021).

Decision trees are bifurcated into two principal categories predicated on the type of the target variable (Chauhan, 2022):

a) A Categorical Variable Decision Tree refers to a decision tree model specifically designed to handle a target variable that is categorical in nature.

b) A Continuous Variable Decision Tree refers to a type of decision tree that is utilized when the target variable is continuous in nature.

3.4.1.2 Decision Trees: Key Terminology

The following key terms, as illustrated in Figure 3.1, are essential for understanding the structure and function of a decision tree (Chauhan, 2022):

i. **Root Node:** It encapsulates the entire dataset, serving as the initial node from which the dataset is partitioned into subsets based on attribute values, leading to more uniform or homogeneous groups.

ii. **Splitting:** This refers to the act of dividing a node into multiple parts, which results in the formation of sub-nodes based on the attribute that provides the best homogeneity upon splitting.

iii. **Decision Node:** Any node that is subdivided into additional branches (sub-nodes) is a decision node, indicative of further decision points within the tree.

iv. **Leaf/Terminal Node:** Nodes at which subdivision ceases, where no further splitting occurs, are designated as leaf or terminal nodes. These nodes represent the final outcome of the decision process.

v. **Pruning:** In contrast to splitting, pruning involves the reduction of the tree's complexity by eliminating sub-nodes from a decision node, thereby streamlining the decision-making path.

vi. **Branch / Sub-Tree:** A segment of the decision tree that extends from a node to its subsequent nodes is referred to as a branch or sub-tree, representing a subset of the decision process.

vii. **Parent and Child Node:** Within the tree structure, a parent node is one that is split into sub-nodes. The subsequent nodes that originate from the division are termed child nodes.

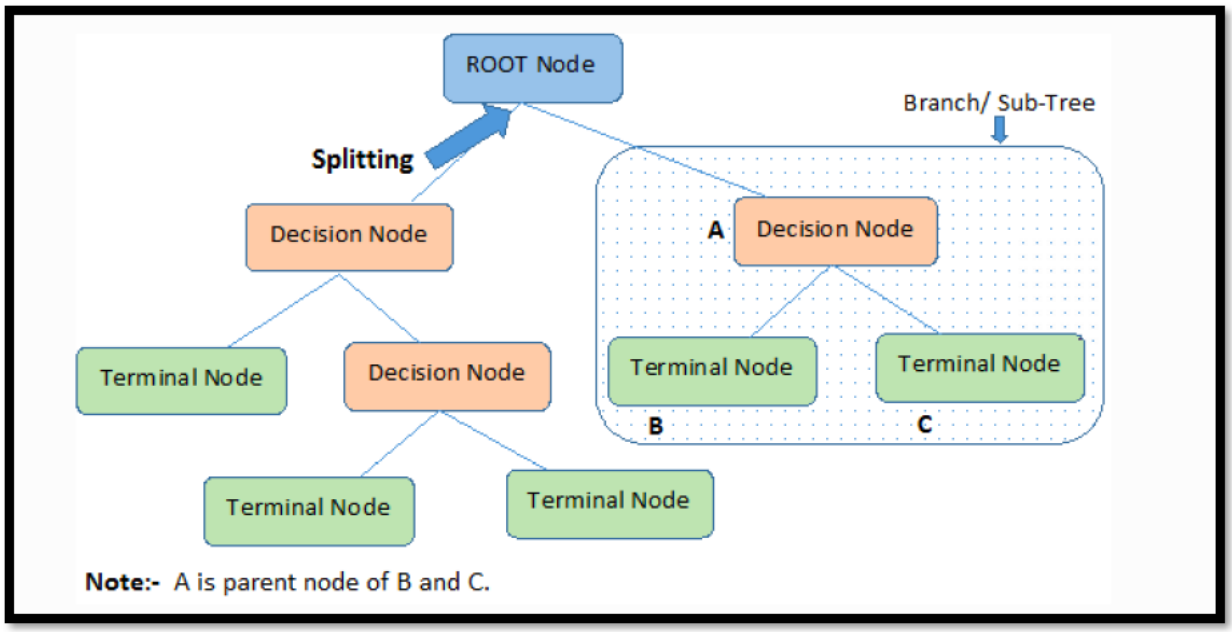

Figure 3.1: Decision Tree Nodes Relationship (Chauhan, 2022)

Decision trees operate by categorizing examples through a sorting process down from the root of the tree to a leaf node (terminal node). This leaf node ultimately

determines the classification of each example (Chauhan, 2022). In a decision tree, each node functions as a test case for a specific attribute, with every branch extending from the node representing potential responses to this test. This procedure is inherently recursive, repeating itself for each sub-tree that originates from the new node (ibid).

3.4.1.3 Understanding the Functionality of Decision Trees

The Decision tree's accuracy is significantly dependent upon the strategic selections enacted during the splitting of nodes (Kotsiantis, 2013; Chauhan, 2022). The decision criteria are multifaceted, diverging based on whether the tree's function is classification or regression. Such trees incorporate an array of algorithmic strategies to ascertain the optimal subdivision of a node into multiple sub-nodes, thus augmenting the homogeneity or purity relative to the target variable. The tree examines all accessible variables, implements divergences, and subsequently selects the split that results in the most homogeneous subsets. The algorithmic choice is further influenced by the nature of the target variable. Gupta *et al.* (2017) as well as Mienye, Sun and Wang (2019) enumerate several algorithms applied within decision tree frameworks, inclusive of:

- **Iterative Dichotomiser 3 (ID3)** – An extension of the original D3 algorithm.
- **C4.5** – The evolution of ID3, designated for enhanced performance.
- **Classification and Regression Tree (CART)** – The dual utility in classification and regression tasks.
- **Chi-square Automatic Interaction Detection (CHAID)** – Adept at executing multi-tiered splits for classification trees.
- **Multivariate Adaptive Regression Splines (MARS)** – For modelling complex, non-linear relationships between variables.

The ID3 algorithm constructs decision trees utilizing a top-down, greedy search methodology that navigates through the potential branching paths without any backtracking. In this context, a 'greedy' algorithm consistently opts for the choice that appears most optimal at the given level of the tree (Chauhan, 2022). The ID3 algorithm follows these steps in building a decision tree (ibid):

1. **Initialization**: The process starts with the original dataset S as the root node.
2. **Attribute Iteration**: In each iteration, the algorithm examines each unused attribute of the set S, calculating its Entropy (H) and Information Gain (IG).
3. **Attribute Selection**: The attribute that demonstrates the highest information gain or the lowest entropy is then selected for the dataset division.
4. **Data Splitting**: The dataset S is then divided based on the chosen attribute, creating subsets of the data.
5. **Recursion**: The algorithm recursively applies these steps to each subset, considering only those attributes that have not been used in previous splits.

3.4.1.4 Attribute Selection Measures

If the dataset comprises N attributes, the process of determining which attribute to position at the root or at various levels of the tree as internal nodes is a complex and significant task. A random selection of attributes is insufficient, as it may lead to suboptimal trees with poor predictive performance. Liu and White (1994) and Al-Ahmad *et al.* (2021) have recommended various criteria for calculating the values of each attribute in decision tree algorithms. These criteria include:

1. **Entropy**: A measure of the randomness or unpredictability in the data.
2. **Information Gain**: The reduction in entropy or unpredictability that occurs when a dataset is modified based on a particular attribute.
3. **Gini Index**: A metric that quantifies the impurity or variability of a set of elements.
4. **Gain Ratio**: A modification of information gain that accounts for the intrinsic information of an attribute and avoids bias towards attributes with many levels.
5. **Reduction in Variance**: This is used in the context of regression trees and involves selecting splits that result in subsets with lower variance.
6. **Chi-Square**: A test statistic employed to ascertain the presence of a statistically significant discrepancy between anticipated and actual frequencies within one or more categorical variables.

The attributes are evaluated based on these criteria, sorted accordingly and then placed in the decision tree. Typically, an attribute with a high value, such as high

information gain, is positioned at the root. It is important to note that when using information gain as a criterion, attributes are generally treated as categorical, while for the Gini index, attributes are considered continuous. Each of these criteria offers a unique perspective in assessing the attributes' effectiveness and suitability for inclusion at various levels within the tree (Liu and White, 1994; Al-Ahmad *et al.*, 2021).

3.4.1.5 Entropy

Entropy, as defined in the context of decision trees, is a metric for quantifying the level of randomness or disorder in the information being processed (Chauhan, 2022). The greater the entropy, the more challenging it becomes to derive clear conclusions from the given information. An example illustrating entropy is flipping a coin, where the outcome is entirely random and unpredictable.

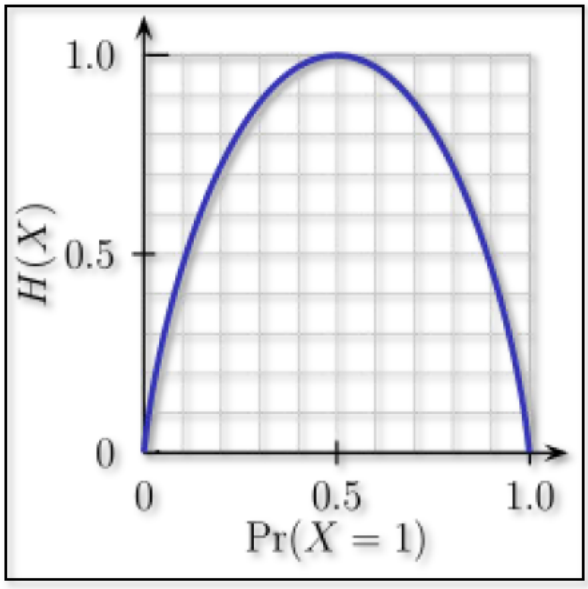

Figure 3.2: The Entropy Diagram of a Binary Information Source (Crowley, 2023)

In Figure 3.2, when the probability $P(X)$ of a particular event is 0 or 1, the entropy $H(X)$ is zero, indicating no uncertainty or surprise, which mean that the event's outcome is certain. However, when the probability is 0.5, the entropy reaches its maximum, reflecting maximum uncertainty or surprise, as the event's outcome is completely unpredictable. This scenario represents a state of perfect randomness, where there's an equal chance of the occurrence or non-occurrence of an event and thus the outcome cannot be determined with certainty.

The formula to mathematically represent entropy for one attribute is (Wang *et al.*, 2017):

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Equation 1: Entropy of One Attribute

Where:

- ***E(S)*** is the entropy of the attribute,
- ***c*** is the number of classes for the attribute, and
- $\boldsymbol{p_i}$ is the ratio of the number of elements in a specific class (i) to the total number of elements.

Figure 3.3 illustrates an example to calculate entropy for a single attribute.

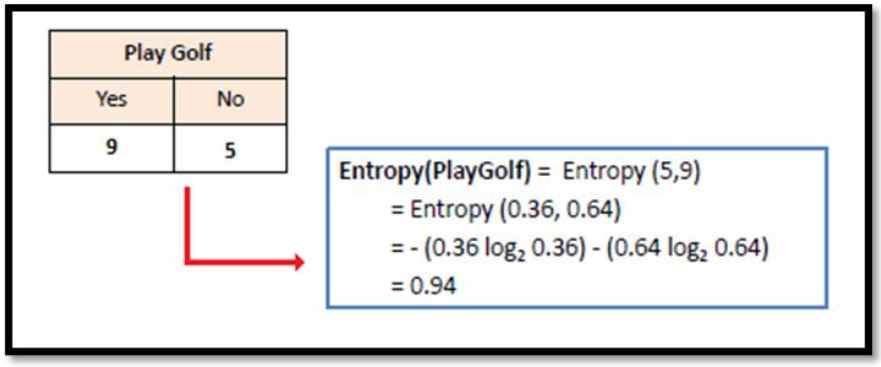

Figure 3.3: Entropy Simplified (Wang *et al.*, 2017)

The formula to mathematically represent entropy for multiple attributes is (Wang *et al.*, 2017; Chauhan, 2022):

$$E(S) = \sum_{i=1}^{n} p_i \, E(S_i)$$

Equation 2: Entropy of Multiple Attributes

Where:

- ***E(S)*** is the overall entropy of the multiple attributes,
- $\boldsymbol{p_i}$ is the probability of the (i)th attribute in the dataset, and
- $\boldsymbol{E(S_i)}$ is the entropy of the (i)th attribute.

Figure 3.4 illustrates an example to calculate entropy for multiple attributes.

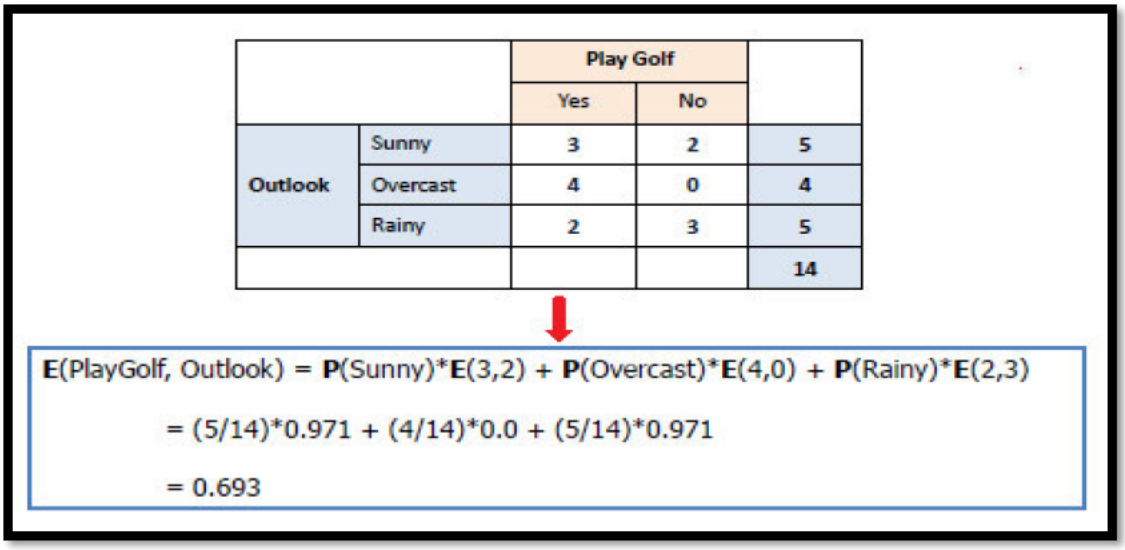

Figure 3.4: Entropy for Multiple Attributes Example (Wang *et al.*, 2017)

3.4.1.6 Information Gain

Information Gain (IG) is a statistical measure used to assess the effectiveness of an attribute in segregating the training examples into targeted classes. The construction of a decision tree is based on identifying the attribute that yields the highest IG, corresponding to the greatest reduction in entropy, thereby clarifying the classification process (Chauhan, 2022). IG is calculated by taking the difference between the initial entropy of the entire dataset and the weighted average entropy after the dataset is split according to a particular attribute. The ID3 algorithm employs IG to determine the best attribute that leads to the most informative split at each node in a decision tree (ibid).

Mathematically, IG is represented as (ibid):

$$IG = \quad Entropy(S_{before}) - \frac{1}{k}\sum_{j=1}^{k} Entropy(S_j)$$

Equation 3: Information Gain (IG)

Where:

- ***IG*** is the Information Gain,
- $\boldsymbol{Entropy}(\boldsymbol{S_{before}})$ is the entropy of the dataset before the split,
- ***k*** is the number of subsets generated by the split, and
- $\boldsymbol{Entropy}(\boldsymbol{S_j})$ is the entropy of the (j)th subset after the split.

3.4.1.7 Gini Index

The Gini index is utilized as a metric to assess the homogeneity of dataset splits. It quantifies impurity through the subtraction of the sum of the squares of class probabilities from unity. Mathematically, the Gini index is articulated as follows (Maimon and Rokach, 2014; Chauhan, 2022):

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Equation 4: Gini Index

Where:

- ***Gini***(*S*) is the Gini index of the dataset,
- ***c*** is the number of classes in the dataset, and
- $\boldsymbol{p_i}$ is the ratio of the number of elements in a specific class (i) to the total number of elements.

The Gini index quantifies dataset impurity, with a value of 0 indicating absolute purity, where all instances within the dataset are of a single class. It is leveraged to appraise the efficacy of dataset splits; it is calculated for each possible division and the partition that minimizes the weighted Gini index post-split is preferred. This minimization signifies an optimal partitioning that enhances class purity across the subsets formed. Contrasting with the Information Gain measure, which may show a preference for partitions that yield a higher number of discrete classes, the Gini index is less sensitive to such conditions, often resulting in broader, more encompassing class groupings (Kotsiantis, 2013; Maimon and Rokach, 2014; Chauhan, 2022).

The Gini index, particularly with the CART algorithm, is crucial when the target variable is categorical, manifesting in forms such as “Success” or “Failure”. CART leverages the Gini index to execute binary splits. A higher Gini index suggests greater inequality or heterogeneity within the data subsets (Chauhan, 2022). The following steps delineate the calculation of the Gini index for a particular split in the decision tree (Maimon and Rokach, 2014; Chauhan, 2022):

1. Ascertain the Gini index for each sub-node by applying the formula *Gini* = 1 − (p^2+q^2), where *p* and *q* denote the probability of success and failure, respectively, within the sub-node.

2. Compute the Gini index for the entire split by calculating a weighted average of each sub-node's Gini scores, which effectively integrates the individual heterogeneities to assess the overall quality of the split.

The Gini index, therefore, plays a pivotal role in the CART algorithm, guiding the selection of optimal split points that enhance the purity of the resultant subsets, a process integral to the construction of an effective decision tree for classification tasks.

3.4.1.8 Gain Ratio

IG demonstrates a tendency to favour attributes possessing numerous unique values, particularly for positioning as root nodes. This predisposition may result in an unwarranted preference for attributes with extensive distinct value sets. To rectify this inclination, the C4.5 algorithm, an advancement of the ID3, introduces the Gain ratio which is a refinement of IG (Maimon and Rokach, 2014; Chauhan, 2022). The Gain ratio mitigates the inherent bias by incorporating the quantity of branches that a split engenders, thus integrating the intrinsic information of a split into the selection process. This corrective mechanism enables a more equitable distribution of attributes throughout the decision tree, thereby enhancing the algorithm's ability to construct a model that reflects the underlying data more accurately (Maimon and Rokach, 2014; Chauhan, 2022).

Mathematically, Gain ratio is presented as follows (Chauhan, 2022):

$$Gain\ Ratio = \frac{Information\ Gain}{SplitInfo} = \frac{Entropy\ (before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j \log_2 w_j}$$

Equation 5: Gain Ratio

The formula in Equation 5 is presented in two parts (ibid):

1. The numerator represents the IG, which is the difference between the entropy of the original set (before the split) and the sum of the entropies of each subset resulting from the split (after), across all *K* possible values of the attribute in question.

2. The denominator, termed Split Information, is the sum across all subsets of the weighted entropy of each subset, where the weight w_j is the proportion of examples that end up in subset *j* after the split. The Split Information essentially measures the potential information generated by splitting the dataset into *K* partitions.

This ratio adjusts the IG by penalizing attributes with a larger number of partitions, thus reducing the bias towards attributes with many levels that Information Gain alone might have. It provides a more balanced approach to selecting the attribute that results in the most useful partitioning of the data for classification (ibid).

3.4.1.9 Variance Reduction

Variance reduction is an algorithmic approach predominantly applied to continuous target variables within regression analyses. This method employs the canonical variance formula as a decision-making criterion to ascertain the most propitious split within a dataset. The fundamental principle guiding this approach is the selection of a split that minimizes the variance, thus partitioning the population into more homogeneous subsets (Hastie *et al.*, 2009; Chauhan, 2022). The mathematical expression for variance is as follows (Chauhan, 2022):

$$\text{Variance} = \frac{\sum(X_i - \bar{X})^2}{n}$$

Equation 6: Variance

Where $\bar{X}$ denotes the mean of the observed values, X_i represents individual observations and *n* is the count of observations (ibid).

The procedural steps for calculating variance in the context of variance reduction are delineated below (ibid):

1. Ascertain the variance for each potential node within the tree structure.
2. Compute the variance for each prospective split by taking the weighted mean of the variances of each individual node.

This variance calculation is instrumental in the reduction of the variance technique, facilitating the identification of splits that enhance the predictive accuracy of the regression model (ibid).

3.4.1.10 Chi-Square

The CHAID analysis is one of the earliest methodologies for tree classification. It is designed to ascertain the statistical significance of the differences observed between the parent node and its sub-nodes. The measure of this significance is determined through the aggregation of the squared deviations, standardized by comparing the observed frequencies with the expected frequencies of the target variable (Chauhan, 2022).

CHAID is particularly effective when dealing with categorical target variables that can be dichotomized into outcomes such as “Success” or “Failure”. It possesses the capability to execute binary or multiple splits. A higher Chi-square value indicates a more statistically significant differentiation between the sub-node and the parent node (ibid).

The CHAID algorithm results in the generation of a decision tree that categorizes data based on the calculation of the Chi-square statistic, thus aiding in the identification of the most discriminative splits. Mathematically, the Chi-square statistic is represented as follows (ibid):

$$X^2 = \sum \left(\frac{(O - E)^2}{E} \right)$$

Equation 7: Chi-square Statistic

where:

- $\chi 2$ represents the Chi-square statistic,
- $\sum$ denotes the summation of the terms,
- O stands for the observed frequency or score, and
- E indicates the expected frequency or score under the null hypothesis.

The Chi-square statistic for a split is computed via the following steps (ibid):

1. For each node within the tree, the Chi-square statistic is computed by quantifying the deviations for both “Success” and “Failure” outcomes.
2. The Chi-square for the entire split is determined by summing the Chi-square values for the “Success” and “Failure” across all nodes affected by the split.

These calculations facilitate the examination of the associations between the categorical predictors and the target variable, thereby enabling the construction of a tree that illustrates the interactive relationships and hierarchical decision rules within the dataset (ibid).

3.4.1.11 Overfitting in Decision Trees

A commonly observed challenge with decision trees, especially when applied to datasets with a high dimensionality, is their propensity to overfit (Maimon and Rokach, 2014; Chauhan, 2022). Overfitting is characterized by the model's excessive complexity, where it appears to memorize the training data rather than learning to generalize from it. In the absence of pruning or other regularization techniques, a decision tree can ostensibly achieve an impeccable accuracy of 100% on the training data (Maimon and Rokach, 2014). This is attributed to its capacity to produce a unique leaf node for every single data point, accommodating even the most idiosyncratic patterns. However, while such a tree may fit the training data perfectly, this over-complexity compromises its ability to accurately predict unseen data, thereby diminishing its predictive performance on new samples (Chauhan, 2022).

Pruning Decision Trees and Random Forest are two ways to remove overfitting and are now presented.

3.4.1.12 Pruning Decision Trees

The procedure of tree splitting continues until a fully matured tree is developed or predefined stopping criteria are met. However, such fully developed trees are prone to overfitting the training data, which can result in suboptimal performance on new, unseen data (Maimon and Rokach, 2014; Chauhan, 2022).

To mitigate this, pruning is applied. Pruning involves the systematic removal of branches from the tree, starting from the leaves and moving towards the root, in such a manner that the overall accuracy of the model on validation data remains as high as possible. This process is facilitated by dividing the original training dataset into two distinct subsets: a training dataset *TD*, used to grow the tree, and a separate validation dataset *VD*, used to assess the impact of pruning on the model's accuracy (Maimon and Rokach, 2014; Chauhan, 2022).

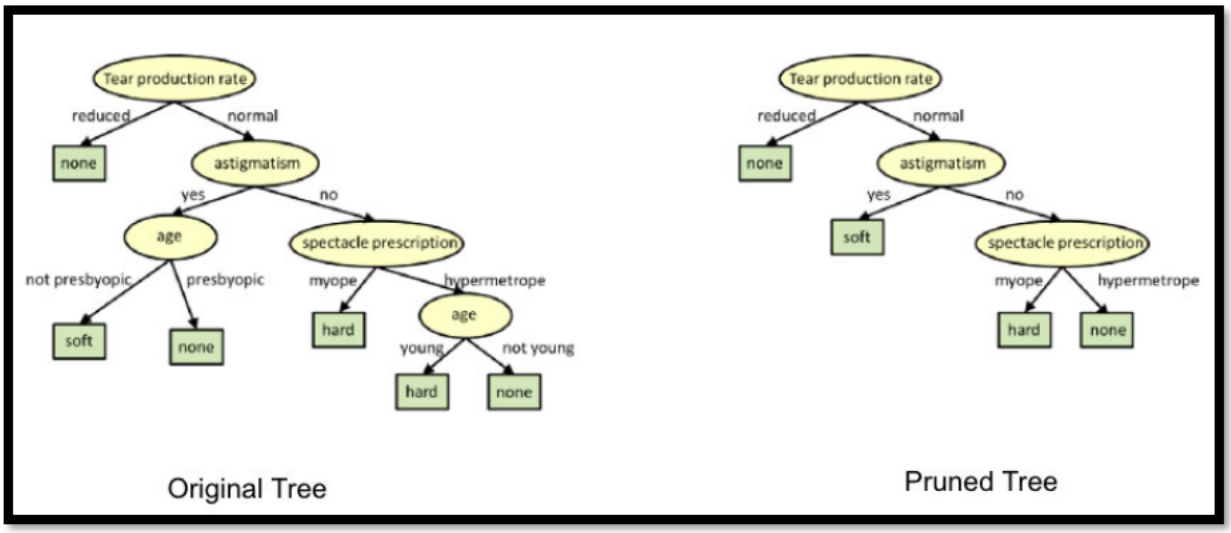

Figure 3.5: Example of Decision Tree Pruning (Chauhan, 2022)

Figure 3.5 illustrates that the 'Age' attribute (left side of the tree) may be pruned if it contributes to overfitting and if its removal increases or maintains the accuracy when predictions are validated against dataset *VD*. Pruning, in this context, is executed with the goal of enhancing the tree's generalizability to new data by eliminating branches that do not significantly contribute to the predictive power of the model (Chauhan, 2022).

3.4.2 Ensemble Learning Models

3.4.2.1 ExtraTreesClassifier

The ExtraTreesClassifier refers to the Extremely Randomized Trees, which is an ensemble learning technique based on aggregating the results of multiple decision trees. Similar to the RandomForest approach, the ExtraTreesClassifier introduces additional randomness into the model building process. It does so by randomizing certain decisions and selecting subsets of the dataset during the construction of the trees. This randomization serves as a form of regularization, intended to reduce the tendency of the model to over-learn from the training data, thereby mitigating the risk of overfitting. This method has been recognized for its effectiveness in enhancing the generalizability of the predictive model (Bhandari, 2018; Hakak *et al.*, 2021). The essence of ensemble learning lies in its capability to integrate the predictive strengths of various learners, culminating in a unified model that embodies the aggregated outputs from an array of models (Chen and Guestrin, 2016).

The ExtraTreesClassifier creates an ensemble of decision trees using bootstrapped samples of the training data and random feature sub-spacing. It builds each decision tree in the ensemble by randomly selecting a subset of the features and then selecting the best split point among the feature subset using the best-first search algorithm (Geurts, Ernst and Wehenkel, 2006). This method leads to a more diverse set of decision trees in the ensemble, which can reduce overfitting and improve the generalization performance of the model.

ExtraTreesClassifier also has several parameters that can be adjusted to control the complexity of the model and improve its performance. These include the number of decision trees in the ensemble, the maximum depth of each tree, and the minimum number of samples required to split a node (Pedregosa *et al.*, 2011). In the context of COVID-19 Fake News detection, it has been shown to outperform simpler models by leveraging multiple decision trees to improve accuracy and reduce overfitting (Olayiwola *et al.*, 2023).

3.4.2.2 Random Forest

The Random Forest algorithm exemplifies ensemble learning, a methodology that integrates the outputs of multiple ML models to achieve superior predictive accuracy. This approach capitalizes on the concept of randomness through two primary mechanisms:

1. The random selection of data samples to construct various trees, ensuring diversity in the training process.
2. The consideration of random subsets of features for node division within each tree, enhancing the robustness of the model.

Bagging, or bootstrap aggregating, plays a pivotal role in Random Forest's ensemble strategy. It involves generating multiple datasets from the original through randomized sampling with replacement, which helps to mitigate overfitting and improves the overall robustness of the model (Zhukov, Sidorov and Foley, 2017). Each dataset then serves as the foundation for building an individual tree using the same algorithm. The collective decision-making of these trees, through either voting (for classification tasks) or averaging (for regression tasks), culminates in the final prediction (Díaz-Uriarte and Andrés; 2006).

Random Forest operates effectively across both classification and regression domains by aggregating the outcomes of numerous decision trees constructed during the training phase. The algorithm assigns the most frequent output class among the trees as the final prediction for classification tasks, a process grounded in statistical mode determination (Breiman, 2001).

In constructing the decision trees, Random Forest uses a random subset of the features to choose the best split at each node. This process is repeated for a specified number of trees, and the final class or value prediction is determined by the votes of the individual trees (ibid).

Random Forest has several advantages over single decision tree models. For example, it can handle large data sets with higher dimensionality and can also be used to identify important features in the data. Additionally, the use of multiple trees can reduce overfitting and improve the generalization performance of the model (Breiman, 2001). There are also several parameters that can be adjusted to control the complexity of the model and improve its performance, such as the number of decision trees in the ensemble and the maximum depth of each tree (Pedregosa *et al.*, 2011).

3.4.3 Gradient Boosting Models

3.4.3.1 Light Gradient Boosting Machine (LightGBM)

LightGBM, a state-of-the-art gradient boosting framework that employs decision trees, significantly enhances model efficiency and minimizes memory footprint. This framework innovatively incorporates Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling methodologies, addressing the constraints present in traditional histogram-based algorithms within Gradient Boosting Decision Tree (GBDT) frameworks (Ke *et al.*, 2017). LightGBM is particularly efficient for large datasets and has been utilized in COVID-19 Fake News detection because of its speed and accuracy (Khan and Thakur, 2022; Essa, Omar and Alqahtani, 2023).

3.4.3.1.1 Mathematical Analysis for GOSS Technique: Calculation of Variance Gain at Splitting Feature j (Ke *et al.*, 2017)

The GOSS technique, a component of the LightGBM framework, offers a refined approach to Variance Gain calculation during the splitting of features within a gradient

boosting context. In a dataset composed of *n* instances, represented as $\{x_1, \cdots, x_n\}$, each instance x_i is characterized by a vector within a *s*-dimensional space X^s. Within each iteration of the gradient boosting process, the model computes the negative gradients of the loss function, denoted as $\{g_1, \cdots, g_n\}$. Employing the GOSS method, the dataset's instances are initially ranked based on the absolute values of their gradients, arranging them in descending order. The subset *A*, comprising the top $a\times$ 100% instances with the most substantial gradients, is preserved for further analysis. Conversely, from the residual set A^c, representing the lower (1 - a) × 100% gradient instances, a randomly sampled subset *B* is selected, sized at $b \times |A^c|$. The subsequent division of instances leverages the estimated variance gain at vector $V_j(d)$ across the amalgamated subset A ∪ B, mathematically represented as (Ke *et al.*, 2017):

$$\tilde{V_j}(d) = \frac{1}{n}\left(\frac{(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i)^2}{n_l^j(d)} + \frac{(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i)^2}{n_r^j(d)}\right)$$

Equation 8: Estimated Variance Gain for A ∪ B

where $A_l = \{x_i \in A : x_{ij} \leq d\}$, $A_r = \{x_i \in A : x_{ij} > d\}$, $B_l = \{x_i \in B : x_{ij} \leq d\}$, $B_r = \{x_i \in B : x_{ij} > d\}$ *and* the coefficient $(1-a)/b$ is utilized to normalize the sum of the gradients over subset B back to the size of A^c (Ahamed and Arya, 2021).

This intricate methodology underscores the efficiency of the GOSS technique in enhancing the precision of feature splits within the gradient boosting framework, thereby optimizing the model's overall predictive performance (ibid).

The paramount advantages of LightGBM are manifold, offering computational superiority compared to other GBDT frameworks by ensuring quicker training times without compromising on accuracy. Notably, it demonstrates a marked improvement in accuracy over alternative boosting algorithms and exhibits enhanced capabilities in mitigating overfitting, particularly in scenarios involving limited datasets. Additionally, LightGBM supports parallel learning, making it adept at handling datasets of varying sizes, ranging from small to large (Gilda, 2017; Bojjireddy, Chun and Geller, 2021; GeeksforGeeks, 2021; Khan and Thakur, 2022).

In our study, Khan and Thakur (2022), the LightGBM model was applied to the task of detecting Fake News related to COVID-19 within the South African context. This

exploration into its applicability highlights the model's versatility and effectiveness in addressing contemporary challenges in data analytics and ML.

3.4.3.2 XGB Classifier

XGBoost, referred to as Extreme Gradient Boosting, is an optimized version of gradient boosting and a prominent ensemble learning method designed to enhance predictive accuracy by amalgamating the outputs of multiple ML models. This method addresses the limitations inherent in depending solely on a singular ML model by leveraging the collective intelligence of an ensemble of models. Its ability to optimize for both speed and accuracy makes it a preferred choice in studies focusing on COVID-19 misinformation (Mumenin *et al.*, 2021).

The constituent models within the ensemble, referred to as base learners, may originate from identical or disparate learning algorithms. Among the ensemble techniques, bagging and boosting are notably prevalent, particularly in their application to decision trees, despite their versatility across various statistical models.

Bagging involves the parallel creation of multiple decision trees, which serve as the base learners. This technique is aimed at mitigating the high variance typically associated with decision trees. Through bagging, data is sampled with replacement and utilized to train the ensemble of decision trees, with the final output derived from averaging the predictions across all trees (Kadiyala and Kumar, 2018).

Boosting, conversely, builds trees in a sequential manner, where each tree is constructed with the objective of rectifying the errors identified in its predecessor. This process enables each successive tree to refine its understanding based on the adjusted residual errors, thereby enhancing the overall predictive capability of the ensemble. Boosting distinguishes itself by employing weak learners, which are models with minimal predictive power that marginally surpass random guessing. The aggregation of these weak learners through boosting results in a potent learner capable of reducing both bias and variance (Ghorpade, Chaudhari and Patil, 2022).

Notably, boosting prefers the utilization of trees with a limited number of splits, in contrast to the bagging approach where trees are developed to their fullest extent. The relatively shallow trees employed in boosting contribute to their interpretability. Optimal parameters, including the number of trees, learning rate and tree depth, are

determined through validation techniques like k-fold cross-validation, thereby mitigating the risk of overfitting associated with an excessive number of trees (Seiffert *et al.*, 2010).

The boosting ensemble technique unfolds in three pivotal steps:

1. An initial model F_0 is designated to predict the target variable y, introducing a residual defined as $y - F_0$.
2. A subsequent model h_1 is tailored to fit the residuals from the initial step.
3. The combination of F_0 and h_1 yields F_1, representing an enhanced version of F_0, characterized by a reduced mean squared error (MSE):

$$F_1(x) \leftarrow F_0(x) + h_1(x)$$

The performance of F1 could be improved by modelling after the residuals of F1 and to then create a new model F2:

$$F_2(x) \leftarrow F_1(x) + h_2(x)$$

A number of *'m'* iterations may be applied until residuals have reached their feasible minimum:

$$F_m(x) \leftarrow F_{m-1}(x) + h_m(x)$$

It should be noted that at this stage, the additive learners do not interfere with the previously created functions. Instead, they impart their own information to reduce the errors.

3.4.4 Linear Models

3.4.4.1 Logistic Regression

Logistic regression constitutes a quintessential algorithm within the ML spectrum, predominantly harnessed for dichotomous classification challenges. It operates on the underpinnings of probabilistic frameworks and is renowned for its predictive capabilities (Hosmer, Lemeshow and Sturdivant, 2013; Hoffman, 2021). It has also been effectively used in COVID-19 Fake News detection because of its simplicity and interpretability (Patwa *et al.*, 2021).

Distinct from its linear regression counterpart, logistic regression is underpinned by a logistic function, commonly termed the 'Sigmoid function'. This function's intrinsic architecture ensures the model's output is confined within the probabilistic bounds of 0 and 1, thereby aligning with the logistic regression hypothesis. This constraint delineates logistic regression from linear regression, where the latter's output may breach these probabilistic bounds, yielding values beyond the [0,1] interval, thereby contravening the probabilistic hypothesis inherent in logistic regression. Figure 3.6 presents a comparison between a generic linear and logistic regression graph, which highlights the aforementioned probabilistic bounds (Pant, 2019).

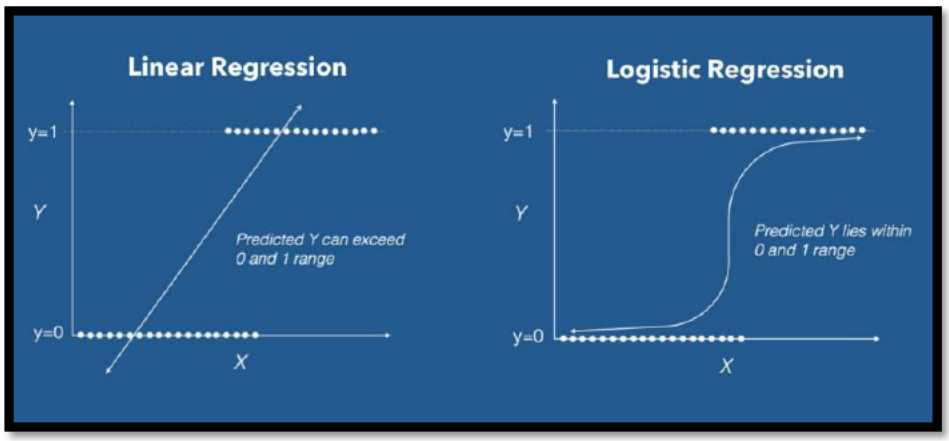

Figure 3.6: Linear Regression vs. Logistic Regression Graph (Pant, 2019)

3.4.4.1.1 Sigmoid Function

The sigmoid function is instrumental in the field of ML, particularly within the context of logistic regression. It serves as a pivotal activation function that transmutes any real-valued number into a bounded output ranging from 0 to 1, thereby rendering it eminently suitable for applications that require a probabilistic interpretation. The sigmoid function exhibits an S-shaped curve, which ensures that the output approximates the probability distribution of the binary dependent variable in logistic regression. This attribute enables the conversion of raw model outputs into interpretable probabilities, which can then be leveraged to execute classifications. The formula for the sigmoid function is presented as follows (Li *et al.*, 2022):

$$f(x) = \frac{1}{1+e^{-(x)}}$$

Equation 9: Formula of a Sigmoid Function

3.4.4.1.2 Hypothesis Representation

In logistic regression, the foundational assumption rests upon the sigmoid, or logistic function, characterized by its signature S-curve. This function is adept at transforming any real-valued number into a constrained output ranging between 0 and 1. It is noteworthy that the output asymptotically approaches these bounds, yet never reaches them precisely. The hypothesis function utilized within logistic regression is formally articulated as follows (Peng, Lee and Ingersoll, 2002; Shalizi, 2013; Pant, 2019):

$$h\theta(x) = 1/(1 + e^{-(\beta_0 + \beta_1 x)})$$

Equation 10: Logistic Regression Hypothesis Function

Where:

- $h\theta(x)$ represents the hypothesis or the estimated probability,
- β_0 and β_1 are the model parameters; β_0 is the intercept term and β_1 is the coefficient for the input feature x,
- x is the input feature value, and
- e is the base of the natural logarithm.

In this formulation, $\beta_0 + \beta_1 x$ constitutes the linear predictor component, which is then transformed by the logistic function, also known as the sigmoid function, ensuring that the output lies in the interval (0, 1). This bounded result is interpreted as a probability, providing a basis for classification: if $h\theta(x)$ is greater than or equal to a threshold value, typically 0.5, the instance is classified into one category, otherwise, it is classified into the alternative category (Peng, Lee and Ingersoll, 2002; Pant, 2019).

3.4.4.1.3 Decision Boundary

In logistic regression, the decision boundary delineates the parameter space whereby the classification algorithm assigns data points to one of two possible categories, which are typically denoted by the binary labels 0 and 1 (Pant, 2019). For instance, in a binary classification task distinguishing cats from dogs, where dogs are labelled as ‘1’ and cats as ‘0’, the decision boundary is determined by a threshold value, commonly set at 0.5. Consequently, if the predictive function yields a value exceeding the threshold, the observation is classified into Class 1 (dog), whereas values below

the threshold categorize the observation into Class 2 (cat). This is depicted in Figure 3.7.

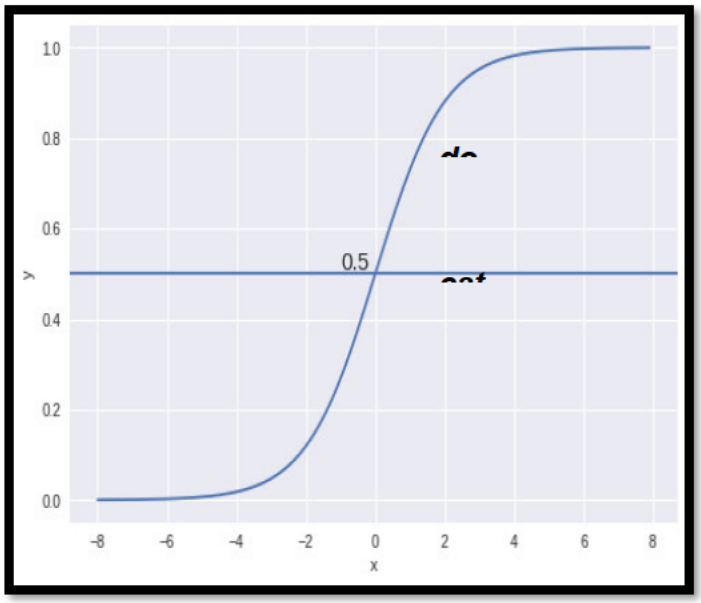

Figure 3.7: Example of Logistic Regression for Binary Classification (Pant, 2019)

3.4.4.1.4 Cost Function

The cost function, denoted as $J(\theta)$ in linear regression, serves as a cornerstone for model optimization. This function quantifies the discrepancy between the predicted values, as determined by the model and the actual values from the training data. The essence of the cost function is to encapsulate this difference in a single measure of error, facilitating the objective of minimizing this error through optimization techniques. By adjusting the model parameters θ, the goal is to find the parameter values that yield the minimum cost, thereby ensuring that the model's predictions closely align with the observed outcomes. This minimization strategy is pivotal in honing the model's accuracy, allowing it to make reliable predictions on new, unseen data. The process of optimizing the cost function is fundamental to achieving a model that not only captures the underlying pattern within the training data but also generalizes well to external data sets. Mathematically, the cost function for linear regression, also known as Mean Squared Error (MSE), is expressed as (Shalizi, 2013; Pant, 2019; Jurafsky and Martin, 2024):

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Equation 11: Cost Function for Linear Regression

Where:

- m is the number of training examples,
- $h_\theta(x^{(i)})$ is the hypothesis function, evaluated for the input features $x^{(i)}$ of the i^{th} training example, with θ representing the parameters (weights) of the model,
- $y^{(i)}$ is the actual output for the i^{th} training example, and
- the sum is taken over all m training examples.

The cost function utilized in linear regression, when applied to logistic regression, results in a non-convex function characterized by numerous local minima. This complexity significantly hampers the optimization process, making it challenging to ascertain the global minimum effectively. The non-convex nature of the function derived from applying linear regression's cost function directly to logistic regression scenarios is attributed to the sigmoid function's output, which maps predictions to a probability between 0 and 1. This mapping introduces non-linearity, rendering the cost surface non-convex and complicating the minimization of the cost function through gradient descent methods commonly employed in linear regression contexts (Jurafsky and Martin, 2024).

To adapt the cost function for logistic regression, thereby addressing the issue of the non-convex cost function encountered in linear regression when applied to logistic models, the logistic regression cost function is specifically designed to be convex. This modification ensures the efficient identification of a global minimum using optimization algorithms like gradient descent. The cost function for logistic regression is formulated as follows (Shalizi, 2013; Pant, 2019; Jurafsky and Martin, 2024):

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

Equation 12: Cost Function for Logistic Regression

Where:

- m is the number of training examples,
- $y^{(i)}$ is the actual label of the i^{th} training example,
- $h_\theta(x^{(i)})$ is the hypothesis function, given by the sigmoid function $\sigma(\theta^T x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}}$,

- θ represents the parameter vector, and
- $x^{(i)}$ represents the feature vector of the i^{th} training example.

This cost function, known as the log loss or binary cross-entropy loss, is convex, ensuring that the optimization algorithms can find the global minimum efficiently. It penalizes incorrect predictions with an increasingly larger cost as the predicted probability diverges from the actual label. For $y = 1$, the cost increases sharply as the hypothesis approaches 0 and similarly, for $y = 0$, the cost rises significantly as the hypothesis nears 1. This characteristic ensures that the optimization algorithm's goal aligns with accurately classifying the training examples, thereby effectively training the logistic regression model (Jurafsky and Martin, 2024).

3.4.4.1.5 Gradient Descent

The principal objective of the gradient descent algorithm is the minimization of the cost function $J(\theta)$, symbolizing the iterative optimization of the model parameters to reduce the discrepancy between predicted and actual outcomes. This is concisely represented by the formulation: minimize $J(\theta)$. The execution of gradient descent involves iteratively adjusting each parameter θ in the direction that most steeply decreases the cost function. This process is mathematically summarized as follows (Pant, 2019; Jurafsky and Martin, 2024):

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

$\}$

Equation 13: Execution of Gradient Descent

Where:

- θ_j is the j^{th} parameter in the parameter vector θ,
- α denotes the learning rate, a positive scalar determining the step size at each iteration, and
- $\frac{\partial}{\partial \theta_j} J(\theta)$ is the partial derivative of the cost function $J(\theta)$ with respect to θ_j.

This iterative adjustment is performed simultaneously for all parameters in θ until convergence is achieved; that is, until the reduction in the cost function $J(\theta)$ becomes negligible with successive iterations. This methodology ensures the efficient identification of the parameter set that minimizes the cost function, thereby optimizing the model's predictive accuracy (Jurafsky and Martin, 2024).

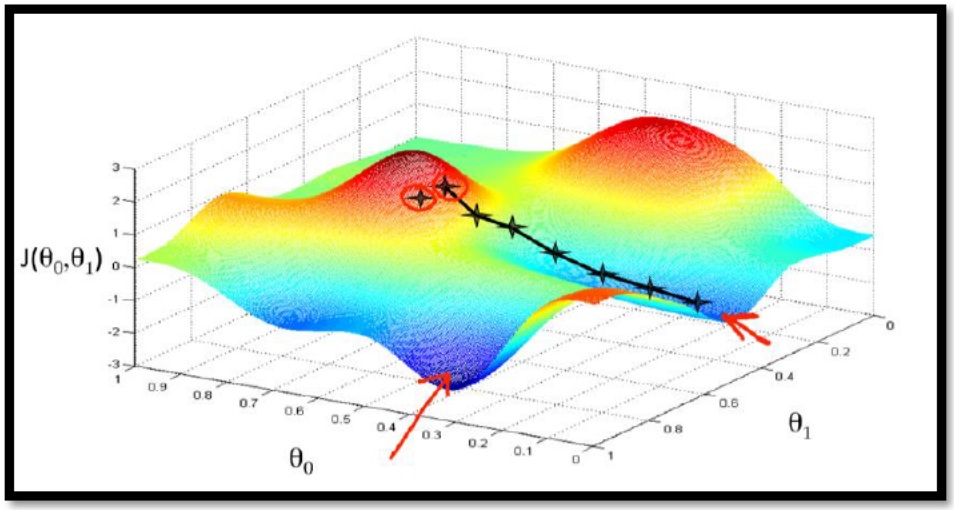

Figure 3.8: Gradient Descent Analogy (Bansal, 2019)

Figure 3.8 depicts a 3D plot with a multi-coloured surface representing a cost function $J(\theta_0, \theta_1)$ over two parameters θ_0 and θ_1. The colour gradient represents the value of the cost function, with orange and red colours indicating higher values and light and dark blue colours indicating lower values. The surface contains two local minima, indicated by the red basins (Bansal, 2019).

The path marked with black arrows and crosses on the surface represents the gradient descent algorithm's progress as it moves towards a local minimum. The arrows suggest the direction of each iterative step taken by the algorithm to minimize the cost function. The starting point is marked by a red circle with a cross in it and the path follows a curved trajectory, showing how gradient descent navigates the topography of the cost function (ibid).

There red arrows along the path indicate the direction and magnitude of the gradient at various points along the optimization path. The axes are labelled θ_0 and θ_1 and the $J(\theta_0, \theta_1)$ indicates that the surface plot is a visualization of the cost function for two variables. This type of visualization is commonly used in ML to illustrate the concept of optimizing parameters to minimize a cost function (ibid).

3.4.4.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) stands as a methodology with dual purposes of classification and dimensionality reduction within the domain of supervised ML. Devised initially by Fisher (1936) for binary classification and later extended by Rao (1948) to encompass multi-class scenarios, LDA operates by projecting high-dimensional data onto a lower-dimensional space (Xanthopoulos, Pardalos and Trafalis, 2013). The intent is to maximize the variance between distinct categories while concurrently minimizing the variance within individual categories, thus facilitating better separation of classes. In terms of Fake News classification, LDA has been leveraged as part of a blended ML model to detect misinformation on various topics (Hansrajh, Adeliyi and Wing, 2021).

The procedure of LDA unfolds through the following methodical steps (Hastie *et al.*, 2009):

1. Calculation of the mean vectors for each class within the multi-dimensional feature space, reflecting the central tendency of each class.
2. Determination of scatter matrices, encompassing both the within-class and between-class scatter, to evaluate the spread of class distributions.
3. Derivation of eigenvectors and their associated eigenvalues from the scatter matrices to ascertain the principal components.
4. Prioritization of eigenvectors based on their corresponding eigenvalues in descending order, followed by the selection of the ‘k’ most significant eigenvectors to form a matrix of dimensions $d \times k$, each column therein representing an eigenvector.
5. Transformation of the original dataset via the $d \times k$ eigenvector matrix, effectively reducing dimensionality while retaining class-discriminatory information. Mathematically, this transformation is represented by the product of the original data matrix X of dimensions $n \times d$ and the eigenvector matrix, resulting in the new subspace of transformed data points with dimensions $n \times k$.

Distinct from other methods, LDA seeks to model the conditional probability distribution of the predictors given the class, denoted as $P(X = x \mid Y = k)$, thereby enabling a probabilistic foundation for classification decisions (ibid).

LDA assumes that the probability of a feature vector $\mathbf{x}$ belonging to a class k follows a multivariate normal distribution with a class-specific mean vector $\boldsymbol{\mu_k}$ and a common covariance matrix Σ for all classes. The formula for the probability density function of class k is given by:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

Equation 14: Probability Density Function of Class k

where p is the number of features (ibid).

The posterior probability $P(Y = k \mid X = \mathbf{x})$, which is the probability that a feature vector $\mathbf{x}$ belongs to class k, can be computed using Bayes' theorem, which relates the likelihood of $\mathbf{x}$ under class k to the prior probability of class k, denoted by π_k (ibid).

The decision boundary between two classes is derived by setting the log of the posterior probabilities of the two classes equal to each other. For two classes j and l, the decision rule $\delta_k(\mathbf{x})$ can be written as (ibid):

$$\delta_k(\mathbf{x}) = \log \pi_k - \frac{1}{2}\boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_k$$

Equation 15: Decision Rule for Two Classes

For two-class problems, the decision boundary is where the decision functions for both classes are equal. This results in a linear equation in $\mathbf{x}$, hence the term 'linear' in LDA. This linear function is given as (ibid):

$$\{x : \delta_k(x) = \delta_\ell(x)\}, 1 \leq j, \ell \leq K$$

Equation 16: Linear Function for Two Classes

3.4.5 Naïve Bayes Models

3.4.5.1 Naïve Bayes

The Naïve Bayes classifier is predicated on the application of Bayes' Theorem, which provides a principled approach for calculating the conditional probability of a

hypothesis given observed evidence (Rish, 2001; Gandhi, 2018a). Bayes' Theorem expressed mathematically is as follows:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Equation 17: Bayes Theorem

In this formula, $P(A \mid B)$ denotes the probability of hypothesis A given the evidence B, $P(B \mid A)$ represents the likelihood of the evidence given the hypothesis, $P(A)$ is the prior probability of the hypothesis and *P*(*B*) is the prior probability of the evidence.

The Naïve Bayes classifier operates under the assumption that the presence (or absence) of any feature in a class is unrelated to the presence (or absence) of any other feature. This assumption of feature independence is referred to as "Naïve", hence the moniker Naïve Bayes (Jurafsky and Martin, 2024). This algorithm has also demonstrated its usage in COVID-19 Fake News classification (Malhotra, Mahur and Achint, 2022).

The classifier can be categorized into several types based on the nature of the feature distributions (Gandhi, 2018a; Jurafsky and Martin, 2024):

1. **Multinomial Naïve Bayes**: Often chosen for text classification problems, where the features are the frequencies of the words or tokens in the documents.
2. **Bernoulli Naïve Bayes**: Utilized when features are binary (Boolean). Here, the parameters used for prediction are typically indicators representing the presence or absence of a particular feature.
3. **Gaussian Naïve Bayes**: When the features have continuous values and an underlying Gaussian (normal) distribution is assumed for these values, the Gaussian Naïve Bayes is the algorithm of choice.

Despite the simplifying assumptions, the Naïve Bayes classifier is renowned for its computational efficiency and demonstrable robustness across extensive datasets and multidimensional feature spaces, underpinning its widespread utilization in numerous classification endeavours across diverse domains of application (Gandhi, 2018a). In this study, the Bernoulli and Gaussian Naïve Bayes algorithms were used.

3.4.6 Nearest Neighbours Models

3.4.6.1 K Nearest Neighbours

The k-Nearest-Neighbours (KNN) algorithm operationalizes classification by locating the proximate data instances within the training dataset and prognosticates based on their categorical affiliations. Its applicability spans across myriad sectors, notably in constructing recommendation systems, enhancing semantic search and pinpointing data anomalies (Hastie *et al.*, 2009; Soni, 2018). KNN has also been leveraged in COVID-19 Fake News classification (Malhotra, Mahur and Achint, 2022).

Representation of data points as feature vectors is central to KNN, where a vector encapsulates a numerical embodiment of the data. Since datasets may include non-quantitative elements, pre-processing and feature engineering are often prerequisites to vector creation. With a dataset comprising of **N** distinctive attributes, a feature vector emerges as an **N**-dimensional vector, with the i(th) element reflecting the value for the ith attribute. Each feature vector thus symbolizes a coordinate in the **N**-dimensional Euclidean space (Soni, 2018).

KNN is characterized as a type of instance-based or lazy learning algorithm, implying the absence of a definitive training phase prior to classification. It necessitates retaining the entire training dataset in memory, barring any dimensionality or dataset reduction, which predicates that the classification process is computationally intensive, as the algorithm necessitates scrutinizing every dataset instance for each classification. Consequently, KNN is optimally efficacious with datasets of modest size and limited features (ibid). The KNN algorithm's workflow, for each classification query, involves (ibid):

1. Computing the distance between the query instance and each instance in the training dataset.
2. Selecting the KNN based on these distances.
3. Determining the final classification through a majority vote within this neighbourhood.

Key decisions in KNN pertain to the choice of k and the distance metric. While k can be arbitrated or optimized through cross-validation, the distance metric is contingent upon the data's nature and the classification objective. Common metrics include the

Euclidean distance, which measures the vector magnitude resulting from the vector subtraction between the query point and a training point and Cosine similarity, which assesses the cosine of the angle between two vectors and is particularly pertinent in text analysis where the orientation of the word vectors holds more significance than their magnitude (ibid).

Euclidean distance and Cosine similarity formulae are expressed mathematically as follows (ibid):

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

Equation 18: General Formula for Euclidean Distance

where, $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ are n-dimensional vectors

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Equation 19: General Formula for Cosine Similarity

where A and B are n-dimensional vectors and θ is the angle between vectors A and B.

The metric selection is critical and can be refined through cross-validation. In practice, the KNN algorithm demarcates the feature space into regions corresponding to different classifications. The decision boundaries that delineate these regions are not necessarily drawn through the training instances but are calculated using the selected distance metric and the distribution of training points (ibid).

Figure 3.9 depicts a visual representation of a classification problem solved using the KNN algorithm. The decision boundaries illustrated are not confined to the exact locations of the training examples but are instead inferred by applying a chosen distance metric to the available data points. Each distinct region is color-coded to correspond with a particular class, delineating areas within the feature space where hypothetical data points would be predicted to belong to one class or another. The resulting partition of the feature space into coloured zones exemplifies the algorithm's classification mechanism, highlighting the non-linear complexity that KNN can capture despite its underlying simplicity (Soni, 2018).

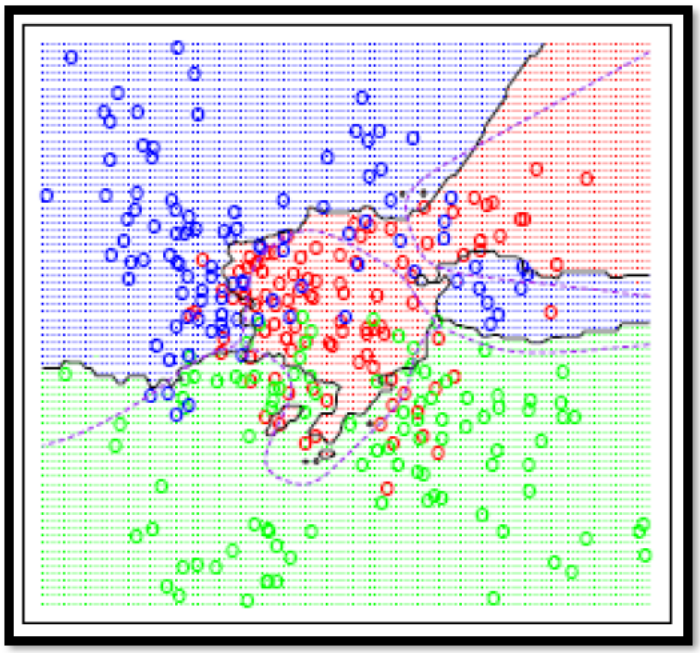

Figure 3.9: Example of KNN Utilized for Classification (Soni, 2018)

The inherent flexibility of KNN allows for enhancements through weighting schemes and pre-processing steps aimed at minimizing computational demands and mitigating noise. These may encompass feature extraction and dimensionality reduction techniques. While predominantly utilized for classification, KNN can also be adapted for regression by averaging the response variable values (ibid).

3.4.7 Support Vector Machines (SVM)

The SVM algorithm seeks to establish an optimal hyperplane in an N-dimensional feature space, where 'N' denotes the number of attributes that effectively segregates the data points into distinct classes. This N-dimensional hyperplane functions as a decision boundary, enabling the classification of new data points based on which side of the hyperplane they reside (Gandhi, 2018b). This algorithm has also demonstrated its usage in COVID-19 Fake News classification due its ability to find optimal hyperplanes (Patwa *et al.*, 2021; Malhotra, Mahur and Achint, 2022).

The SVM algorithm operates by identifying the hyperplane that maximizes the margin between the closest data points of separate classes, which are known as support vectors. This method not only aims for a clear separation of classes but also strives to position the hyperplane in a way that ensures the greatest distance from the nearest data points of any class, thereby providing a buffer zone that affords a measure of classification certainty. Figure 3.10 shows two scatter plots side by side,

depicting a binary classification problem and the concept of a SVM model (Emadi and Tanha, 2020).

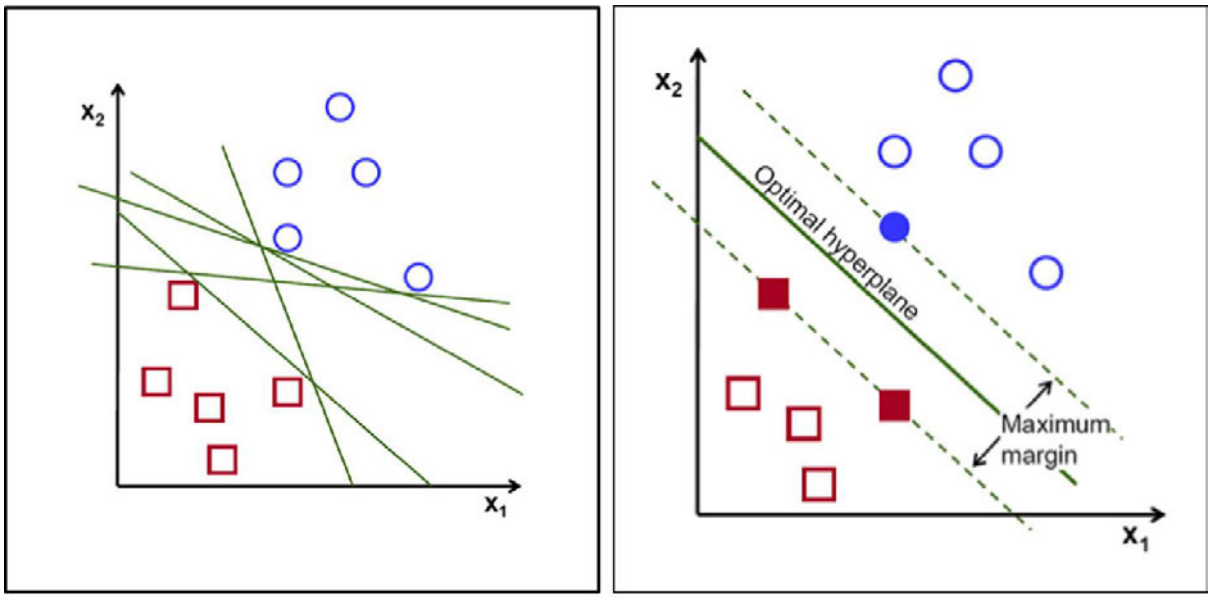

Figure 3.10: Scatterplots Highlighting the Concept of SVM for Binary Classification (Emadi and Tanha, 2020)

On the left scatter plot, there are two types of symbols: blue circles and red squares, each representing a different class. Several green lines are drawn, representing possible hyperplanes that could be used to separate the two classes. These lines, however, are not optimal as they are either too close to one of the classes or not clearly separating the classes.

On the right scatter plot, there is an optimal hyperplane indicated by a solid green line. This hyperplane has the maximum margin, which is the distance between the closest points of the two classes, referred to as support vectors. The support vectors for each class are denoted by a filled blue circle and a filled red square. Two dashed green lines parallel to the optimal hyperplane mark the boundaries of the margin.

The axes x_1 and x_2 represent features of the data set used for classification. The goal of SVM is to find the hyperplane that maximizes the margin between the two classes, thus providing the best generalization for the classification model. This image is a typical illustration used in ML to explain the concept of SVM and the importance of margin in classification problems.

The SVM algorithm is utilized to discriminate between two classes of data points by identifying a hyperplane that ensures the maximal margin. This margin is the largest possible distance between the data points of the two classes. By optimizing for the widest margin, the SVM enhances the model's robustness, affording a higher degree of confidence when it classifies new, unseen data points. The selection of such a hyperplane is strategic, as it not only segregates the classes but also minimizes classification error and enhances the algorithm's predictive accuracy (Gandhi, 2018b).

3.4.7.1 Hyperplanes and Support Vectors

Hyperplanes serve as the decision boundaries that segregate data points into distinct classes based on their feature values. The dimensionality of the hyperplane is intrinsically linked to the number of input features—with two features, the hyperplane is a line in a two-dimensional space and with three features, it manifests as a plane in a three-dimensional space (Figure 3.11). The conceptualization of hyperplanes in spaces with more than three dimensions, while mathematically feasible, transcends intuitive visualization (Gandhi, 2018b).

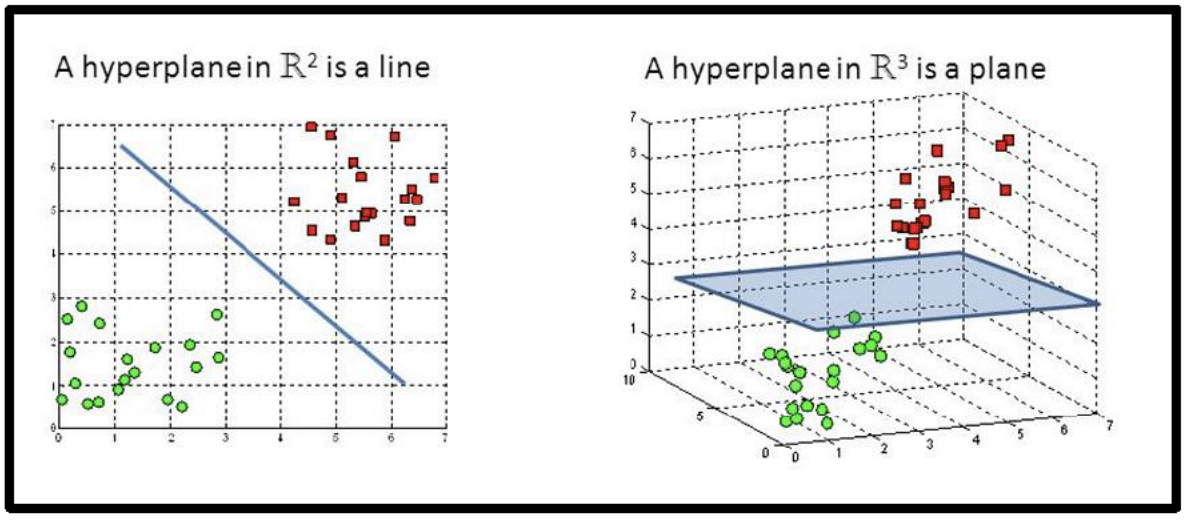

Figure 3.11: Hyperplanes in 2D and 3D Feature Space (Singh and Agarwal, 2022)

Support vectors are pivotal elements that define the optimal hyperplane. They are the data points that lie closest to the decision boundary and are crucial for determining the hyperplane's position and orientation. The classification margin, defined as the distance between the hyperplane and the nearest data points from each class, is maximized using these support vectors (Figure 3.12). The robustness of the SVM

classifier hinges on the integrity of the support vectors; any alteration or removal of these vectors would result in a recalibration of the hyperplane's position. Thus, support vectors are instrumental in constructing the SVM model, playing a decisive role in the model's structure and its subsequent classification capability (Gandhi, 2018b).

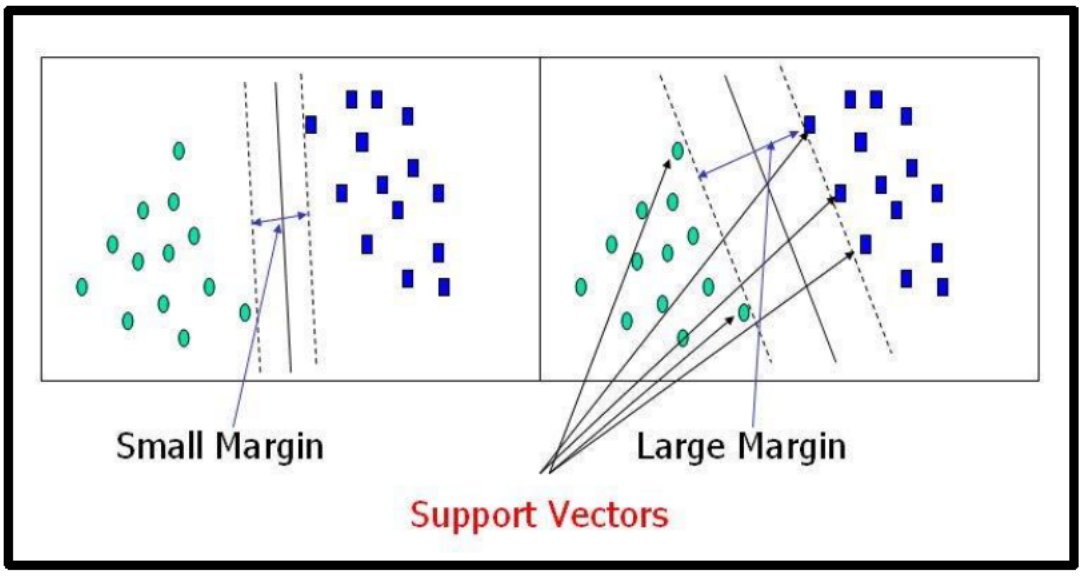

Figure 3.12: Comparison of SVM Margins (Manhas and Kotwal, 2021)

3.4.7.2 Large Margin Intuition

SVMs differentiate themselves from logistic regression in their operational mechanism. Logistic regression employs the sigmoid function to constrain the output of a linear function within a probabilistic range of [0, 1]. A value above a predefined threshold, typically 0.5, is assigned to one class, whereas values below are ascribed to another. SVMs, in contrast, adjust this threshold to values of 1 and -1, establishing a reinforced margin range of [-1, 1]. This margin serves as a buffer zone which bolsters the model's confidence in classifying future data points (Gandhi, 2018b).

3.4.7.3 Cost Function and Gradient Updates

In the context of SVM algorithms, the primary objective is to amplify the margin, or the spatial buffer, between the data points and the defining hyperplane. This margin is pivotal as it aids in determining the hyperplane's ability to classify data points with a greater degree of confidence.

The optimization of this margin is facilitated through the employment of a loss function known as hinge loss. The hinge loss function is particularly designed for SVM classification and is formulated to penalize data points that either fall within the margin or are misclassified. In mathematical notation, the loss function $c(x, y, f(x))$ for SVM is represented as follows (Hastie *et al.*, 2009; Gandhi, 2018b):

$$c(x, y, f(x)) = \begin{cases} 0 & \text{if } y \cdot f(x) \geq 1 \\ 1 - y \cdot f(x) & \text{else} \end{cases}$$

Equation 20: The Loss Function for SVM

Where y represents the true label of the data point, $f(x)$ signifies the predicted value from the SVM algorithm and x is the feature vector of the data point.

When the predicted value and the true label are in agreement and the prediction is beyond the margin (i.e., $y \cdot f(x) \geq 1$), the hinge loss is zero, indicating no penalty. Conversely, predictions that are incorrect or within the margin incur a penalty proportional to the distance from the margin. The cost increases linearly with the distance from the margin boundary (when $y \cdot f(x) < 1$).

The implementation of a regularization parameter within the hinge loss function is a crucial step that harmonizes the dual objectives of loss minimization and margin maximization. This regularization parameter introduces a penalty for increased model complexity, thereby mitigating the risk of overfitting by promoting simpler models with wider margins. With the incorporation of regularization, the augmented cost function for an SVM can be articulated as (Hastie *et al.*, 2009; Gandhi, 2018b):

$$\min_{w,b} \ \lambda \|w\|^2 + \sum_{i=1}^{n} (1 - y_i(w \cdot x_i + b))_+$$

Equation 21: Augmented Cost Function for SVM

Here, $|| w ||^2$ denotes the squared norm of the weight vector, serving as a regularization term controlled by λ, b represents the bias term, n is the number of data points and $(\cdot)_+$ denotes the positive part of the argument, ensuring that the function returns a value of zero when the prediction is correct. This cost function strategically optimizes the SVM by determining the hyperplane that best segregates the classes with an optimal margin, while keeping the model as simple and generalizable as possible (Gandhi, 2018b).

Gradient descent is then applied to iteratively update the weights w by computing the partial derivatives of the loss function with respect to the weights. The updating procedure can be summarized with the following formulas, where α is the learning rate (Gandhi, 2018b):

For correct classifications $(y_i(w \cdot x_i + b) \geq 1)$:

$$w := w - \alpha \cdot (2\lambda w)$$

Equation 22: Weight Update for Correct Classifications

For misclassifications $(y_i(w \cdot x_i + b) < 1)$:

$$w := w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Equation 23: Weight Update for Misclassifications

These updates are performed iteratively until convergence, leading to the formulation of a decision boundary with maximized margins, effectively separating the classes in the feature space.

3.4.8 Other Specialized Models

3.4.8.1 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is a statistical technique employed for pattern recognition and ML that facilitates the classification of observations into predefined classes. It is similar to LDA in that it presumes the observations from each class are drawn from a Gaussian distribution. However, unlike LDA, which assumes homogeneity of variances across classes, QDA allows for each class to have its own covariance structure (Hastie *et al.*, 2009). This algorithm allows for quadratic decision boundaries, which has been utilized in complex classification tasks like COVID-19 Fake News detection (Hauschild and Eskridge, 2024).

The classification decision in QDA is made based on which class has the highest value of the discriminant function for a given observation. This discriminant function is inherently quadratic with respect to the predictors, hence the name Quadratic Discriminant Analysis. Mathematically, the discriminant function for class y in QDA is defined as (Hastie *et al.*, 2009):

$$\delta_y(x) = -\frac{1}{2}\log|\Sigma_y| - \frac{1}{2}(x-\mu_y)^T\Sigma_y^{-1}(x-\mu_y) + \log P(Y=y)$$

Equation 24: Discriminant Function in QDA for Class y

Where:

- x is the vector of predictor variables,
- Σ_y is the covariance matrix for class y,
- μ_y is the mean vector for class y,
- $P(Y=y)$ is the prior probability of class y, also denoted as π_y
- $|\Sigma_y|$ is the determinant of the covariance matrix Σ_y, and
- ${\Sigma_y}^{-1}$ is the inverse of the covariance matrix Σ_y.

The quadratic term in the discriminant function arises from the inclusion of the term $(x-\mu_y)^T{\Sigma_y}^{-1}(x-\mu_y)$, which describes the squared Mahalanobis distance of the observation x from the mean of class y. This distance is scaled by the inverse of the covariance matrix, which accounts for the spread and orientation of each class in the feature space. For classification, the predicted class for a new observation x is the class y that maximizes the discriminant function $\delta_y(x)$ (ibid).

QDA is particularly useful when classes exhibit different variance-covariance structures and when the decision boundary between classes is not linear. While more flexible than LDA, QDA can be more prone to overfitting, especially in cases where the number of predictors is high relative to the number of observations (Rhys, 2020).

3.5 Deep Learning Models

DL models utilize neural networks, which at their core, are composed of layers of interconnected nodes, or “neurons”, each of which processes input data, applies a set of weights that signify the importance of this input, and then passes on an output to subsequent layers (Goodfellow, Bengio and Courville, 2016).

Deep neural networks differ from shallow learning models in that they can learn to represent data in a hierarchical fashion. Shallow models, like those mentioned earlier, typically involve a single layer where feature selection and classification are directly intertwined, and they might not be able to capture complex patterns in data without extensive feature engineering. In contrast, DL models can automatically discover the

representations needed for feature detection or classification from raw data (Goldberg, 2016). This automatic feature extraction makes DL particularly powerful for tasks where the input data has high dimensionality, such as text, images or sound. Figure 3.13 provides an illustration of a Neural Network Framework and a visual depiction of the general components in a neural network:

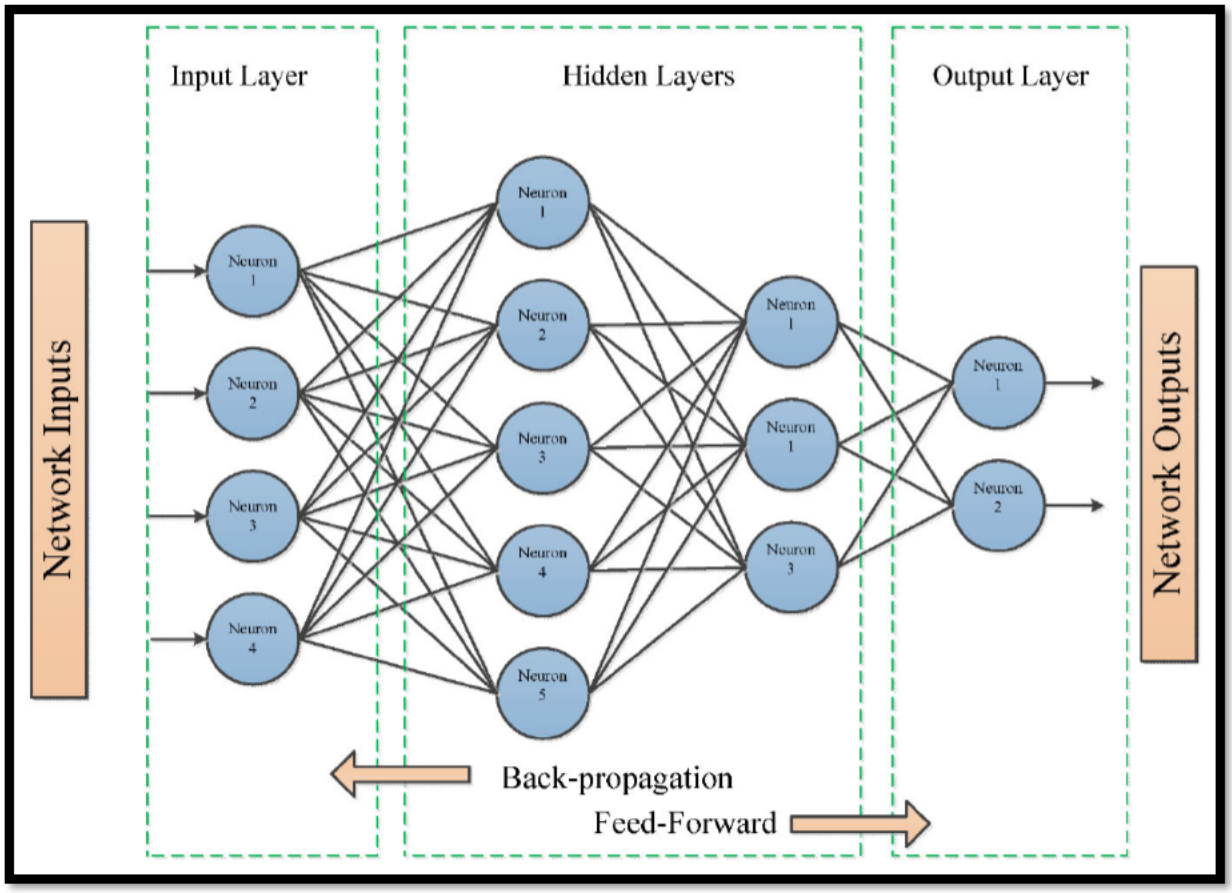

Figure 3.13: Architecture of an Artificial Neural Network with Feed-forward and Backpropagation Algorithms (Abdolrasol *et al.*, 2021)

Applying a simplified representation of what a neural network for text classification might look like within this framework is provided next. However, actual implementations can vary widely in their specifics, such as the number of layers, the types of layers (e.g., convolutional, recurrent) and the complexity of the architecture.

Input Layer: The starting point where text data is inputted into the network. Each neuron in this layer represents a unique word or character from the input text. It should be noted that an Embedding layer, although not depicted in Figure 3.13, lies between the Input and Hidden Layers and maps the words represented by the input neurons to vectors of real numbers, creating a dense representation of the words that captures

semantic meaning. Embedding layers are particularly useful in dealing with the large vocabulary sizes in text data.

Hidden Layers: These layers, typically composed of multiple neurons, transform the embedded word vectors through a series of non-linear transformations. Each hidden layer uses activation functions like ReLU (Rectified Linear Unit) or tanh (hyperbolic tangent) to introduce non-linearity, enabling the network to learn complex patterns.

Feed-Forward: In the feed-forward process, information flows from the input layer, through the hidden layers, to the output layer. Each neuron receives input from the previous layer, processes it and passes it to the next layer. This forward propagation of data through the network allows it to make predictions based on the input features.

Output Layer: The final layer of the network, which uses a softmax function to output a probability distribution over the classes. For text classification, each neuron in this layer corresponds to a potential class and the softmax function ensures that the output probabilities sum to one.

Backpropagation: This is the training algorithm used to adjust the weights of the network. During backpropagation, the network calculates the gradient of the loss function (which measures the difference between the actual and predicted outputs) with respect to each weight in the network and adjusts the weights in a way that minimizes the loss.

Feedback Loop: The mechanism through which the network learns from its errors. After the output is generated, the network's predictions are compared to the actual labels and the difference (error) is fed back into the network to adjust the weights and improve the model during the next iteration of training.

3.5.1 Recurrent Neural Network (RNN) and its Variants

RNNs are a cornerstone in the field of sequence data processing. Pioneered by Amari (1972), RNNs are uniquely equipped to handle sequential data, making them a natural fit for NLP tasks. However, traditional RNNs often grapple with capturing long-range dependencies within text sequences. This limitation led to the development of the LSTM model by Hochreiter and Schmidhuber (1997), designed to overcome the vanishing gradient problem and better retain information over longer sequences.

Building upon the LSTM framework, the Gated Recurrent Unit (GRU) emerged as an alternative. Introduced by Cho *et al.* (2014), the GRU simplifies the LSTM architecture while maintaining comparable performance, particularly in sequence modelling tasks. This was further validated by Chung *et al.* (2014), who empirically evaluated GRUs against LSTMs, highlighting their effectiveness in learning long sequences.

Another significant advancement in the realm of RNNs is the Bidirectional Long Short-Term Memory (Bi-LSTM). As discussed by Graves and Schmidhuber (2005), Bi-LSTMs enhance the standard LSTM by processing data in both forward and backward directions, thereby capturing a richer context. Jin *et al.* (2014) explored optimizations for training Bi-LSTMs, particularly for large-scale applications, underscoring their increasing relevance in contemporary NLP challenges.

RNNs, along with their more advanced counterparts LSTM, Bi-LSTM and GRU, have been instrumental in sequence modelling for Fake News detection. These models are well-suited for analysing the context in social media posts and have been the focus of recent research in the field of Fake News detection. For example, Jaybhaye *et al.* (2023) proposed a DL methodology that utilizes LSTM neural networks for identifying false news, achieving an accuracy of 94% in detecting Fake News. Additionally, Mengji *et al.* (2021) presented a model for identifying false news built on a Bi-LSTM, which assessed the model's performance using two datasets of unstructured news articles, showing that the Bi-LSTM model outperformed other approaches for detecting false news in terms of accuracy. Ali *et al.* (2023) utilized Bi-LSTM for the detection of rumours from microblogs, showing the potential of RNNs in identifying misinformation while Chen, Lai and Lian (2023) also experimented with Bi-LSTM and achieved approximately 99% accuracy for COVID-19 Fake News English text detection trained on 7 686 texts which were collected and labelled from fact-checking sites such as Poynter and Snopes. This research differs from these studies by utilizing a significantly improved labelled and larger dataset of 36 254.

The aforementioned studies collectively demonstrate the growing interest and potential of RNNs, including LSTM, Bi-LSTM and GRU, in Fake News detection, highlighting their effectiveness in discerning misinformation from factual content.

This study leveraged such neural networks to train and develop models that contributes to the ongoing efforts in combatting the spread of misinformation and highlights the potential of DL approaches in detecting Fake News.

3.5.2 Convolutional Neural Network (CNN)

CNNs, primarily known for their dominance in image processing, have been adeptly repurposed for NLP. Kim (2014) was among the first to adapt CNN architectures for sentence classification, harnessing their ability to capture spatial hierarchies in data. Further advancements in CNN applications for NLP were shown by Gehring *et al.* (2017), who introduced a novel approach for sequence transduction, exemplifying the versatility and ongoing evolution of CNNs in handling textual data.

Nasir, Khan and Varlamis (2021) also proposed a model for identifying Fake News built on a CNN, demonstrating the model's effectiveness in detecting false news surrounding the Syrian war based on its content. The study assessed the model's performance using a dataset of news articles, showing that the CNN model outperformed other approaches for detecting false news in terms of accuracy. Moreover, research has also explored the use of CNNs in the classification of social media posts. Kaliyar *et al.* (2020) proposed a CNN-based method for classifying social media posts as either true or false, demonstrating the effectiveness of CNNs in identifying patterns in textual data reaching an accuracy of 98.36%. The study utilized a dataset of social media posts related to the 2016 United States of America (USA) elections, showing the potential of CNNs in detecting misinformation in real-time depending on available computational resources and access to live Twitter data.

These studies collectively demonstrate the growing interest and potential of CNNs in Fake News detection, particularly in analysing textual data from social media posts. CNNs have shown effectiveness in identifying patterns in textual data that are indicative of misinformation, highlighting their potential in detecting Fake News.

This current study utilized CNN to supplement its DL training models because of its versatility and researched effectiveness on social media textual classification.

3.6 Transformer Models

Transformer models represent a class of architecture that is particularly well-suited for handling sequential data, like text, for tasks including classification, translation and

summarization (Wolf *et al.*, 2020). Unlike prior sequence-to-sequence models that used recurrent or convolutional layers, transformers rely entirely on attention mechanisms to draw global dependencies between input and output. Figure 3.14 below depicts a typical architecture of a transformer model:

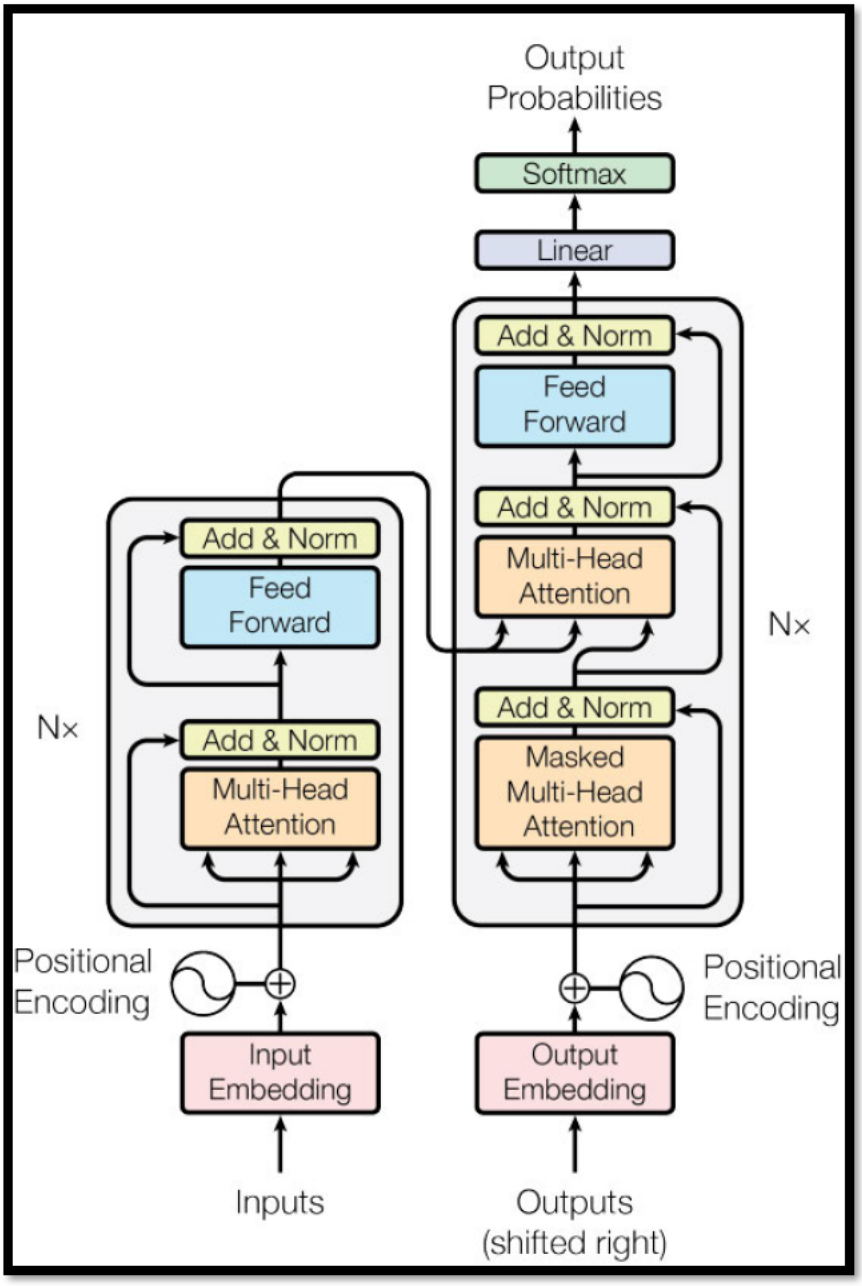

Figure 3.14: Architecture of a Transformer Model (Vaswani *et al.*, 2017)

The beginning of a transformer model for text classification contains the following:

Input Layer: This is where tokenized text data is fed into the model. Each token corresponds to a word or sub-word in the input sequence.

Positional Encoding: Since transformers do not have a recurrent structure, they use positional encodings to give the model information about the order of the tokens.

Multi-Head Self-Attention Mechanism: This is the core of the transformer model. It allows the model to weigh the influence of different parts of the input sequence differently and is executed multiple times in parallel to provide a “multi-headed” attention that captures various aspects of semantic and syntactic relationships between words.

Normalisation Layers: After each attention and feed-forward operation, a normalization step is applied to stabilize the learning process.

Feed-Forward Neural Networks: These networks are applied after the attention mechanism and are responsible for processing the output of the attention layer independently for each position.

Output Layer with Classification Head: At the top of the model, after processing through multiple attention and feed-forward layers, there is an output layer specifically tailored for classification. This usually involves a softmax function that turns the transformer’s output into probabilities over the target classes.

Residual Connections: Also known as “skip connections”, these allow the input of each sub-layer (like attention or feed-forward layers) to be added to its output, which helps mitigate the vanishing gradient problem in deep networks.

Classification Layer: This layer takes the output from the transformer model and processes it to produce the final class prediction for the input text.

The diagram provided visualizes the complex interplay of components in a transformer model. Each block and connection signify a part of the process where transforms input text into a classified output, encapsulating the power of transformers to handle long-range dependencies and various nuances in text.

The introduction of BERT (Devlin *et al.*, 2018) marked a paradigm shift in NLP. BERT’s bidirectional training approach allows it to understand the context of a word based on all surrounding words, unlike previous models that processed text in a unidirectional manner. However, the effective fine-tuning of BERT for specific tasks remains a challenge (Su and Vijay-Shankar, 2022).

In response to BERT's computational demands, two streamlined variants emerged: RoBERTa and DistilBERT. RoBERTa, an optimized version of BERT, achieving superior performance through robust training methods (Liu *et al.*, 2019). DistilBERT goes a step further, offering a more efficient yet powerful alternative, retaining much of BERT's capabilities while being significantly smaller and faster (Sanh *et al.*, 2019).

XLM-RoBERTa extends the capabilities of RoBERTa to a multilingual domain, enabling cross-lingual understanding and transfer learning (Conneau *et al.*, 2019). This advancement was further explored through unsupervised learning techniques (Artetxe and Schwenk, 2019).

XLNet, a creation of Adwaith *et al.* (2022), represents another evolution in transformer-based models. XLNet combines the best aspects of BERT with an ability to consider all permutations of the input sequence, enhancing its applicability in complex tasks like tweet analysis.

Transformer-based models like BERT, RoBERTa, DistilBERT, XLM-RoBERTa and XLNet have revolutionized the approach to NLP tasks, including Fake News detection. Their ability to process large amounts of unstructured text with deep contextual understanding makes them particularly suitable for identifying COVID-19-related misinformation. Mehta *et al.* (2021) demonstrated how BERT models could be fine-tuned to effectively identify false claims and misinformation. Similarly, RoBERTa (Rajendran *et al.* 2022), DistilBERT (Al-Garadi *et al.*, 2021), XLM-RoBERTa (Wiciaputra, Young and Rusli, 2021) and XLNet (Wang *et al.*, 2021) have been employed to enhance the accuracy of social media classification detection systems, offering improved performance over traditional models.

3.7 Conclusion

This chapter introduced and discussed some essential concepts and provided the background to some algorithms used in this study. The concepts include terms around ML, DL and transformers, while the algorithms backgrounds from various ML types such as Ensemble Learning Models, Gradient Boosting Models, Linear Models and SVMs have been outlined. Furthermore, their utilisation and suitability in text classification have been outlined and provides foundation for their selection in the experimentation phases of this study.

The next chapter presents the foundational research paradigm, delineating the structured design and methodologies employed. It meticulously outlines the origin and nature of the data sources utilized, explicates the analytical methods applied in data processing and analysis and discusses the ethical considerations integral to the integrity of the study.

Chapter 4: Research Methodology

4.1 Introduction

In this chapter, the study's underlying research paradigm, design methodology, data sourcing, analytical methods and ethical considerations are presented. The data processing of the Twitter data corpus addresses pertinent challenges associated with data mining on the Twitter platform. As this is a data science study, it pays particular attention to how it determines the identification, acquisition, range and pre-processing of the tweets. The pre-processing phase encompassed the exclusion of tweets that were indecipherable or identically duplicated. The processes for Fake News detection, Sentiment Analysis, Change-Point Analysis and Bot detection are described and presented.

Guiding this study are the core research questions, which were initially outlined in Chapter 1. These questions are as follows:

Research Question 1: How can the COVID-19 infodemic on Twitter be computationally analysed for Fake News within the South African context?

Research Question 2: Were there significant changes in the average sentiment from the South African COVID-19-related Twitter data?

Research Question 3: Were social bots leveraged to produce content around South African-related COVID-19 Twitter content?

4.2 Research Paradigm: Post-positivism

This research subscribes to the post-positivist philosophical paradigm, which presents a distinct departure from the principles of positivism. Positivism is characterized by its deterministic view of the universe, advocating that truth can only be discerned through the scientific method, focused on causal relationships (Hacking, 1983; Creswell, 2013). Post-positivism, in contrast, acknowledges the similarities in cognitive processes between scientific inquiry and everyday reasoning, suggesting that both follow a fundamentally analogous logical structure (Creswell and Creswell, 2017). This paradigm recognizes the susceptibility of observations and theories to

error and revision, thereby challenging the positivist stance that empirical science is the exclusive path to understanding reality. Post-positivists hold that while science persistently aims to accurately describe reality, achieving absolute knowledge through empirical methods alone is unattainable (Creswell, 2013; Creswell and Creswell, 2017).

The application of the post-positivist paradigm is particularly relevant to social media research. This relevance stems from the relative infancy of the underlying infrastructure of social media, such as Internet Web 2.0 (established in 2004) and platforms like Twitter (established in 2006). Furthermore, the tools and techniques used NLP for analysing extensive social media data are still evolving in terms of their methodology (be it ML or lexical analysis), scale (big data) and scope (including content analysis and sentiment analysis) (Mayer-Schönberger and Cukier, 2013). These considerations underscore the suitability of a post-positivist approach to this study, aligning with the dynamic and developmental nature of the field.

4.3 Research Strategies and Research Design

This section presents the study's methodology choice (section 4.3.1), research approach (section 4.3.2) and time horizon (section 4.3.3).

Quantitative methodology is adopted for the research design since this study primarily utilises data science techniques which are built upon the scientific method that makes use of measurements and statistical tests to test hypotheses. Furthermore, the study aims to computationally discern Fake News content and has no qualitative research elements.

This study adopted an experimental research strategy for the use of ML algorithms to develop classification models for Fake News (Creswell and Creswell, 2017). According to Christensen *et al.* (2011), experimental research studies aim to discover causal-relationships between variables, which is consistent with ML classification models. The ML models were developed and deployed to classify content as either being Fake News or not. The models were then ranked in terms of accuracy, prediction, recall and error rates. Information gained from a comprehensive literature review together with the Cross Industry Standard Process for Data Mining (CRISP-DM) were utilised to select the data science techniques used. CRISP-DM is a recognised process model to conduct data mining tasks (Wirth and Hipp, 2000).

The following is a brief overview of the different types of methodologies.

4.3.1 Methodological Choice: Quantitative

Research methodology constitutes a structured procedure for the collection, analysis and interpretation of information. Generally, research methodologies are categorized into qualitative and quantitative approaches, with a third, mixed methodology, combining elements of both (Creswell, 2013; Creswell and Creswell, 2017). This study specifically adopts a quantitative methodology, driven by the intent to quantitatively dissect, classify and statistically interpret content expressed in substantial volumes of social media data. The rationale is that a quantitative, empirical approach not only mitigates the extensive time and effort typically associated with manual analysis but also offers greater scalability and adaptability for wider application.

Quantitative methodology employs mathematical and statistical tools for the analysis of collected data, utilizing established statistical and computational techniques (Creswell and Creswell, 2017). This approach is in alignment with the post-positivist worldview, focusing on the acquisition of numerical data to elucidate a phenomenon or to enable generalizations across varied domains (Creswell, 2013; Creswell and Creswell, 2017).

Considering the scope of this study, manual analysis of the extensive COVID-19 longitudinal dataset is impractical because of its significantly high volume – examining approximately one million tweets is labour-intensive and not scalable. An automated, quantitative method is therefore imperative as indicated by Alotaibi *et al.* (2022).

4.3.2 Research Approach: Deductive

This research adopted a deductive analytical approach, in alignment with its quantitative research methodology as delineated in section 4.3.1. The nature of quantitative analysis inherently necessitates a deductive stance (Creswell, 2013). In a deductive framework, the analysis begins with the establishment of a theoretical structure. From this theoretical foundation, specific hypotheses are formulated. These hypotheses are subsequently tested through empirical observation, utilizing instruments designed to measure variables that have been clearly defined within the context of the established theory. This methodological approach allows for the

systematic verification or falsification of the theoretical propositions based on the observed data. This flow is depicted in Figure 4.1.

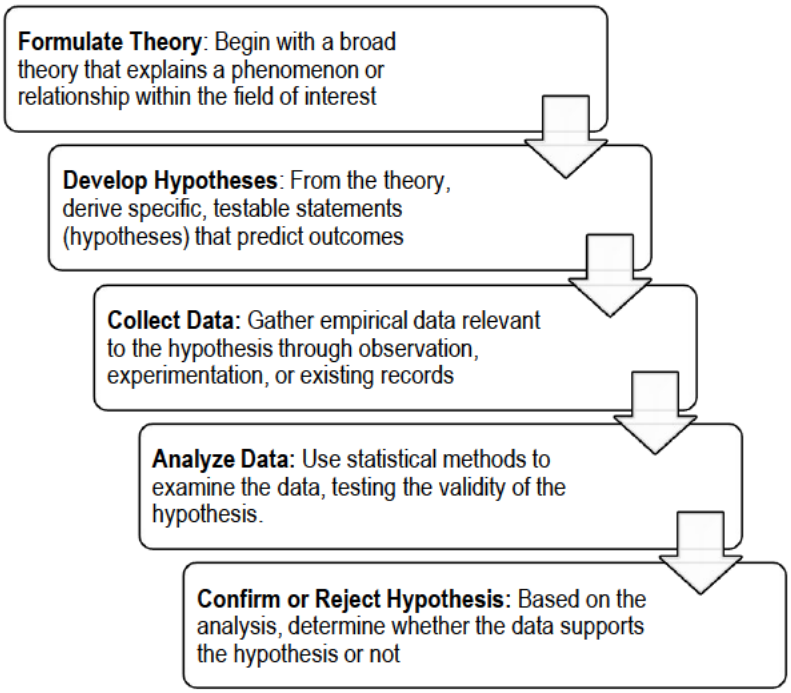

Figure 4.1: Typical Deductive Approach Utilised in Quantitative Studies (Creswell, 2013)

The deductive approach is, therefore, utilized for this study as it quantitatively analyses Twitter texts for COVID-19-related Fake News.

4.3.3 Time Horizon: Longitudinal

A longitudinal study is a type of observational research where data is gathered consistently for the same subject over a substantial period. This research method is particularly adept at tracking changes and trends over time and such studies can span several years, as highlighted by Dawber, Kannel and Lyell (1963). For his research, approximately 18 months (08 November 2019 to 19 July 2021) of data were collected focusing on the same entity hence categorizing it within the longitudinal study framework. The longitudinal study enabled a detailed analysis of the Fake News trends and public sentiment related to the pandemic throughout this period.

4.4 Data Sources and Twitter as a Data Source

This section presents a comprehensive overview of the data sources utilized in the study (section 4.4.1). It includes a rationale for selecting Twitter as the primary data source, accompanied by an exploration of the challenges inherent in this choice

(section 4.4.2). Following this, the data handling process specific to this study is thoroughly elaborated, providing insights into the methodologies employed for data collection, processing and analysis.

4.4.1 Overview of Data Sources used in the Study

The study employed a comprehensive selection of accredited electronic databases and advanced search engines to facilitate data collection across its various stages. This encompassed an array of resources, including Google Search and Google Scholar for scholarly articles; publications from relevant conferences; an assortment of media outlets encompassing both traditional and online platforms; authoritative textbooks and data extracted from Twitter. The rationale behind this eclectic choice of sources was guided by the emergent and dynamic nature of the research subject. Table 4.1 provides a structured overview of the data gathering process. The table outlines the specific data sources engaged at each distinct phase of the research endeavour.

Table 4.1: Data Sources Used During the Research

Research phase	Source
Literature Examination	Relevant articles in digital libraries, conference papers, search tools like Google and Google Scholar, digital media, literature and associated texts.
Tweet Analysis	Twitter, SNScrape, VADER sentiment analysis, Microsoft Excel, Change Point Analysis software. ML libraries in Python: Keras, LightGBM, Scikit-learn, SimpleTransformers, TensorFlow.
Data Visualisation	Python (Matplotlib), Tableau software, Microsoft Excel.

4.4.2 Twitter as a Data Source

The utilization of Twitter as a pivotal data source has garnered significant attention in a wide spectrum of research fields, encompassing finance, healthcare, disaster management, public administration and electoral studies. This multidisciplinary interest is evidenced by an extensive body of literature, including works by Sakaki, Okazaki and Matsuo (2010), Acar and Muraki (2011), Bruns and Liang (2012), Cho, Jung and Park (2013), Ahmed (2018), Singh *et al.* (2019), Valdez *et al.* (2020), Jang *et al.* (2021) and Ng *et al.* (2022). The platform's capacity to facilitate the automated

collection of public tweets has been a particularly attractive feature, empowering researchers to derive insights from Twitter data across these varied domains. The decision to choose Twitter as the data source for this study was influenced by multiple factors, notably its capability to efficiently and effectively facilitate the searching, retrieving and analysing of information (section 2.4). This choice is underpinned by Twitter's unique characteristics that enable the automated collection of public tweets, presenting a valuable resource for conducting comprehensive social media analysis.

The ascendancy of Twitter among South Africans has been notably robust, with user numbers proliferating from 2.85 million in 2022 to 3.65 million in 2023 (Kemp, 2022, 2023). Influential figures such as journalists, politicians and academics frequently utilize Twitter, leveraging its platform to engage actively with the public and disseminate information (Park, Reber and Chon, 2015; Bane, 2017; Hu and Hong, 2017; Annamalai, Chandrasekaran and Pathak, 2023; Richter *et al.*, 2024). This reinforces the rationale for selecting Twitter as a pertinent and justified data source for the present study.

It should be noted that utilizing Twitter as a data source is accompanied by specific challenges that researchers must navigate. These challenges encompass issues such as the enduring accessibility of data, ethical considerations in research conduct, the analysis and interpretation of user-generated content, and the integration of qualitative with quantitative research methodologies as well as reconciling user-centric with content-centric research strategies (Ahmed, 2018). Such complexities are not unique to Twitter; they are broadly representative of the obstacles encountered in the realm of social media research (Giglietto, Rossi and Bennato, 2012). Nonetheless, these challenges do not serve as impediments to research endeavours. Instead, they present opportunities for methodological rigor and innovation, provided they are prudently considered and integrated into the research design (Sinnenberg *et al.*, 2016).

In evaluating Twitter's validity as a source for scientific research, Bruns and Stieglitz (2014) address concerns about its representativeness and reliability. They acknowledge critiques suggesting that Twitter's user demographics and content generation may be skewed towards certain minority groups, potentially limiting its representativeness of the broader population. However, this perspective is more pertinent to the platform's initial phase of development. As Twitter's popularity and

user base have expanded significantly over time, the relevance and impact of such criticism have diminished, suggesting a broadening of the platform's user demographic and a more diverse range of content generation. Furthermore, the platform's transformational impact on research communication is notable, despite concerns about its future and implications for science (Stokel-Walker, 2023a). Indeed, O'Leary (2015) posited that Twitter data becomes increasingly valuable for analysis, prediction and understanding causality as its use in significant events grows. This viewpoint aligns with the contemporary digital era and the national impact of the COVID-19 pandemic, suggesting the potential of Twitter data in understanding large-scale societal trends and behaviours.

4.5 Data Analytics: How Data is Handled in this Study

Ahmed (2018) emphasizes the critical importance of meticulous documentation in social media research. The documentation includes recording the dates and types of data collected, as well as the source code used (Table 4.2). This section presents the specific procedures with respect to when, how, why and what data was retrieved, processed and analysed. In particular, the Data Analytics Lifecycle (Erl, Khattak and Buhler, 2016) is employed to delineate the data handling process. This particular model consists of nine distinct stages, as depicted in Figure 4.2. While alternative data analytics lifecycles exist, with Dietrich (2013) and Prajapati (2013) proposing models comprising six and five phases respectively, the Erl, Khattak and Buhler (2016) framework is selected for its detailed articulation of each stage.

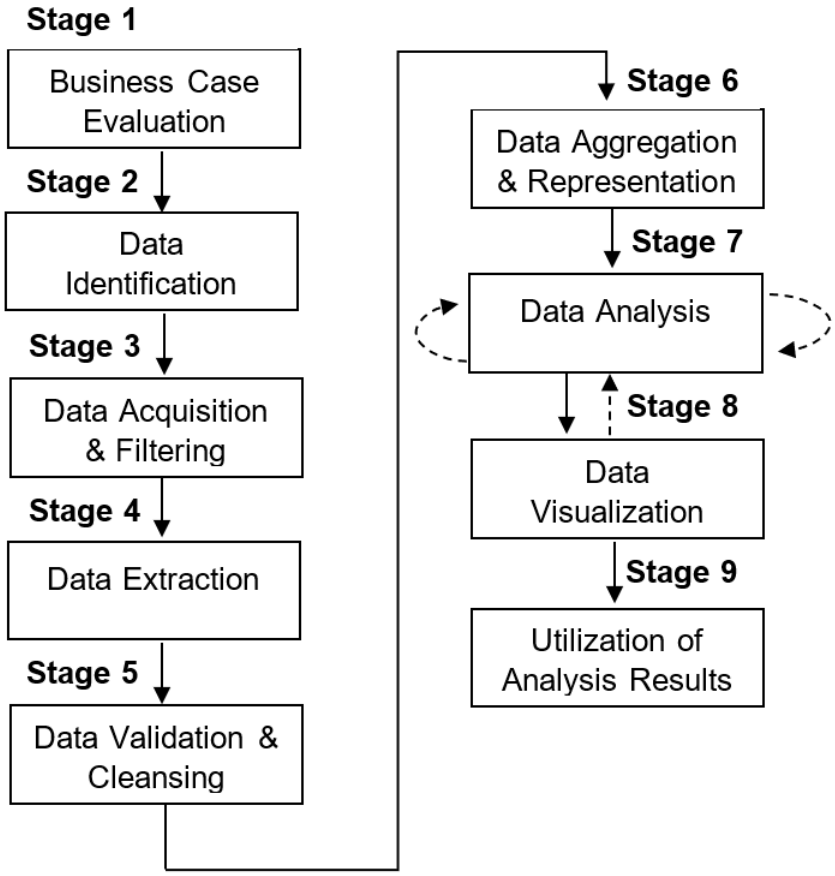

Figure 4.2: The Big Data Analytic Lifecycle (Erl, Khattak and Buhler, 2016)

Each stage is now explained.

4.5.1 Stage 1: Business Case Evaluation

Business case evaluation, the initial stage, necessitates a clear articulation of the analysis's justification, aims and objectives (Erl, Khattak and Buhler, 2016). The selection of South African COVID-19 tweets, hereby referred to as SAcovid19dataset, was the focal point for this investigation and was guided by the research problem, questions and objectives delineated in Chapter 1, supported by an appropriate examination of relevant literature in Chapter 2.

The literature review segment of a research project involves a systematic review of a diverse range of materials, including articles, abstracts, comprehensive reviews, monographs, dissertations, additional research findings and digital media (Aveyard 2023). This literature review substantiates the research study, pinpointing an unexplored niche within the research domain (Creswell and Creswell, 2017).

4.5.2 Stage 2: Data Identification

Data identification, the second stage, involves identifying the datasets and their sources necessary for the analysis (Erl, Khattak and Buhler, 2016). Twitter was chosen as the target social media platform (section 4.4.2) because of the volume and accessibility of tweets (section 2.4). The uniqueness of the South African COVID-19 experience persuaded this novel South African study. The data identification focused on two aspects: the selection of Machine Learning (ML) suitable datasets (section 4.5.2.1) and the South African COVID-19 tweets including its date range (section 4.5.2.2).

4.5.2.1 First Phase: ML Suitable Datasets (C19MLdataset)

The datasets identified to train and test the ML models (Chapter 3), consisted of labelled content, were obtained from CONSTRAINT AAAI-2021 (Patwa *et al.*, 2021), COVID-19 Rumour dataset (Cheng *et al.*, 2021), COVID-19 Fake News Infodemic Research Dataset (FNIR) (Saenz, Kalathur Gopal and Shukla, 2021), Zenodo (Banik, 2020), Google Fact Check (Google, n.d.) and PolitiFact websites (Politifact, n.d.). These datasets combined are collectively referred to as the C19MLdataset throughout this study. The rationale for this selection is to ensure academic credibility by utilizing perceived reputable databases and websites. This study focused on detection techniques rather than reviewing each labelled content to develop adaptable and broadly applicable analytical methods. Table 4.2 represents an example of labelled content extracted from the identified datasets (C19MLdataset). The labelled data consists of Text and Label metadata which facilitates ML development.

Table 4.2: Examples of C19MLdataset

Text	Label
The NZ COVID Tracker app will remain important and useful at Alert level 1. People are encouraged to download the app or a similar one or keep a record of where you’re going to be.	Real
BREAKING NEWS# The president Cryill Ramaphosa has asked all foreign nations to depart south Africa before 21 june 2020 due to increasing cases of COVID19	Fake
We are delighted that 78 high- and upper-middle countries and economies have now confirmed they will participate in the COVAX Facility and the number is growing. I urge those who have not yet joined to do so by the 18th of September-@DrTedros #COVID19.	Real

4.5.2.2 Second Phase: The South African COVID-19 Twitter Data (SAcovid19dataset)

To ensure the analysis of tweets related to South Africa, the total volume of COVID-19-related tweets had to be narrowed down. This was achieved by identifying specific phrases or keywords to filter the tweets to those pertaining to South Africa (section 4.5.3.1.2). The consolidated collection of these filtered tweets is referred to as the SAcovid19dataset, which forms the basis for the subsequent analysis and insights presented in this study. This is distinct to the C19MLdataset.

The dataset for this study consisted only of a corpus of text data (tweets) containing emojis and emoticons; audio, images and videos were not considered given distinct types of analysis required (section 1.9). These are delimited.

The collection period of tweets acquired spanned from 08 November 2019 to 19 July 2021 (section 4.5.3.1.2). The timeline for the collection of tweets was established from the first mention of South African COVID-19-related keywords (section 4.5.2.2) on 8 November 2019 to 19 July 2021. The SAcovid19dataset was not extended because of the following factors:

1. The initial proposal for this research was formulated in 2021, setting the temporal boundaries for data collection.
2. Acquiring additional Twitter datasets beyond 19 July 2021 would lead to the formation of two separate datasets for the same investigation. Merging these datasets would not equate to generating a seamless, unified dataset because of the potential for data to be deleted, accounts suspended, or tweets to be set to private in the intervening period. Such a scenario was deemed a significant research risk, necessitating a limit on the data collection period.

4.5.3 Stage 3: Data Acquisition and Filtering

The third stage is defined as the collection or gathering of all identified data sources and the filtration of irrelevant and corrupt data (Erk, Khattak and Buhler, 2016). Records containing invalid data types or missing or nonsensical values are corrupt or irrelevant data and “filtering” means removing corrupt data wherever it is detected.

4.5.3.1 C19MLdataset and SAcovid19dataset Acquisition

4.5.3.1.1 Acquiring C19MLdataset

Datasets for building ML models were downloaded and acquired from sources identified in section 4.5.2.1. The total size, unique entries and corresponding labels are described in Table 4.3 for each dataset. C19MLdataset refers to the combination of these datasets.

Table 4.3: ML Suitable Datasets for COVID-19

Name	Total Size	Unique Entries (Duplicates)	Labels (Count)
CONSTRAINT AAAI-2021	10 700	10 695 (5 duplicates)	Fake (5 098), Real (5 597)
COVID-19 Rumour dataset	7 179	7 112 (67 duplicates)	F (3 655), T (1 845), U (1 612)
Zenodo	10 201	8 965 (1 236 duplicates)	Fake (8 505), Real (460)
COVID-19 Fake News Infodemic Research Dataset (FNIR)	7 588	7 251 (337 duplicates)	Fake (3 459), Real (3 792)
Google Fact Check	3 043	3 040 (3 duplicates)	False (2 926), True (114)
PolitiFact	1 185	1 185 (0 duplicates)	False (787), True (128), Misleading (270)

4.5.3.1.2 Acquiring SAcovid19dataset

SNScrape[4] was employed to produce a dataset of tweets related to COVID-19 in South Africa (SAcovid19dataset). The acquired SAcovid19dataset contained 976 086 unique Twitter data points and was stored in the JavaScript Object Notation (JSON) file format comprising 27 metadata fields.

[4] The library is available at https://github.com/JustAnotherArchivist/snscrape (Accessed 17 April 2023)

SNScrape, a Python library, is particularly adept at scraping data from various social networking services (SNS), including Twitter. The ability of SNScrape to acquire data without the need for personal API keys, coupled with its comprehensive search capabilities, signifies its utility in academic research.

For the specific needs of this study, which required a localized understanding of the pandemic's impact, SNScrape facilitated the extraction of tweets using the following case-insensitive keywords: *covidsa; covidinsa; covid19sa; covid_19sa; covid19-sa; lockdownsa; covid-19sa; covid19insa; coronavirusSA; coronavirusinSouthAfrica; coronavirusinSA; sacoronavirus; coronaSA; coronainSA; covidSouthAfrica; covidinSouthAfrica; covid19SouthAfrica; covid_19SouthAfrica; covid-19SouthAfrica.* SNScrape's wildcard and geolocation features were leveraged to further expand the collection of tweets. These were collaborated through the researcher's frequent verification of trending tweets, related websites and an analysis of Google Trends (South Africa, 2020a; Media Update, 2021). A number of keywords were limited and chosen because of feasibility in terms of resources and research time frames.

The application of SNScrape in academic research is well-documented and justified. For instance, Ridhwan and Hargreaves (2021) employed SNScrape to gauge public sentiment concerning the COVID-19 pandemic in Singapore, showcasing the tool's effectiveness beyond traditional Twitter data collection methods. Similarly, Khan and Khan (2021) utilized SNScrape in their sentiment analysis of tweets concerning COVID-19 governmental responses in Pakistan. Furthermore, in the realm of social movements, Tong *et al.* (2022) analysed Twitter discussions on significant topics such as #BlackLivesMatter and #StopAsianHate using SNScrape to gather the necessary data.

These examples establish the applicability of SNScrape as a research tool in facilitating diverse research objectives such as sentiment analysis, topic modelling and the exploration of public opinion on critical issues.

The dichotomy between public and private data on Twitter distinguishes between openly accessible tweets and those restricted to followers. Public tweets are accessible to anyone visiting the Twitter website or utilizing an external client, whereas private tweets are viewable exclusively by users granted permission by the account holder. For this study, tweets that did not contain the keywords (section

4.5.2.2) were omitted from the scraping. This is an important decision that excluded numerous records. Time, complexity and expense influenced this decision. Thus, terms such as “Pfizer” was excluded unless they were posted together with the selected keywords (section 4.5.2.2).

4.5.3.2 Tweet Metadata

Each data point analysed in this study, represented by a tweet, encompassed several components such as the tweet’s textual content, its timestamp, the username of the tweet’s author, the device or application used to post the tweet, designations of ‘favorite’ and ‘retweet’, as well as the language settings of the user and the tweet itself (Figure 2.1 and Table 2.1).

The following list is the 27 metadata fields name and brief explanation included in each tweet prior to data extraction (section 4.5.4):

1. **url**: The URL of the tweet.
2. **date**: The date and time when the tweet was posted.
3. **content**: The actual text content of the tweet.
4. **renderedContent**: The display format of the tweet’s content.
5. **id**: The unique identifier for the tweet.
6. **user**: Information about the user who posted the tweet.
7. **replyCount**: The number of replies to the tweet.
8. **retweetCount**: The number of times the tweet was retweeted.
9. **likeCount**: The number of likes the tweet received.
10. **quoteCount**: The number of times the tweet was quoted.
11. **conversationId**: The identifier for the conversation to which the tweet belongs.
12. **lang**: The language of the tweet.
13. **source**: The tool or service used to post the tweet.
14. **sourceUrl**: The URL of the source tool or service.
15. **sourceLabel**: The label or name of the source tool or service.
16. **outlinks**: URLs included in the tweet.
17. **tcooutlinks**: T.co wrapped URLs included in the tweet.
18. **media**: Media elements (images, videos, etc.) attached to the tweet.
19. **retweetedTweet**: Information about the original tweet if it is a retweet.
20. **quotedTweet**: Information about the tweet if it is quoting another tweet.

21. **inReplyToTweetId**: The identifier of the tweet to which this tweet is a reply.
22. **inReplyToUser**: The user to whom the reply tweet is directed.
23. **mentionedUsers**: Users mentioned in the tweet.
24. **coordinates**: Geographical coordinates where the tweet was posted.
25. **place**: The place associated with the tweet.
26. **hashtags**: Hashtags included in the tweet.
27. **cashtags**: Cashtags included in the tweet.

4.5.3.3 The Acquisition of Twitter Data

The acquisition of Twitter data is usually conducted through various approaches depending on specific research and analytical requirements. Here are six different approaches that are used to acquire Twitter data:

1. **Twitter API:** The Twitter Application Programming Interface (API) is a pivotal resource in social media research. This API is extensively utilized by researchers, developers and businesses for analytical, customer service, or marketing purposes (Kim, Nordgren and Emery, 2020).
2. **Web Scraping:** Another technique employed to gather Twitter data. This process involves programmatically accessing Twitter's web interface to extract data, though it may be sensitive to website changes (Dongo *et al.*, 2021).
3. **Third-party Tools:** Various third-party tools facilitate Twitter data access, often leveraging the Twitter API to offer enhanced functionalities like advanced search, analytics and visualization (Kim, Nordgren and Emery, 2020).
4. **Twitter Data Partnerships:** Twitter engages in data partnerships with certain organizations, offering them broader data access for specific purposes such as academic research or market analysis (Susha, Janssen and Verhulst, 2017).
5. **Public Datasets:** Public datasets of Twitter data, typically curated by academic institutions or Twitter itself, are occasionally released, focusing on specific events, topics, or periods for research purposes (Banda *et al.,* 2021; Chen, Deb and Ferrara, 2022).
6. **Data Resellers:** There are companies that specialize in aggregating and reselling social media data, including that from Twitter. These entities often

combine data from multiple sources, offering it to businesses for marketing and research (Bruns, 2020).

Ethical considerations, especially in terms of user privacy and consent, are crucial in the collection and utilization of social media data (Lipschultz, 2020). This study ensured that public data was harvested that required no user consent.

4.5.4 Stage 4: Data Extraction

The fourth stage entails transforming data from its original format to a format primed for analysis (Erl, Khattak and Buhler, 2016). This transformation is essential for organizations aiming to process data further and transfer it into analytical repositories, such as data warehouses or lakes, facilitating in-depth analysis (Nambiar and Mundra, 2022).

In this study, the extracted Twitter data was, as necessitated, transformed into comma-separated values (CSV) and Microsoft Excel formats for enhanced analytical utility. The conversion process involved transforming the collected SAcovid19dataset from their original JSON format (section 4.5.3) into CSV and Microsoft Excel formats. This conversion facilitated various phases of analysis (section 4.5.7). Furthermore, specific data visualizations (section 4.5.8) required the importation of the dataset, formatted in Microsoft Excel, into Tableau. Conversely, other analyses required exporting data from Microsoft Excel for application within the Change Point Analysis (CPA) framework (section 4.5.7.4). This process ensures that the analysis is conducted efficiently across various platforms and tools.

4.5.5 Stage 5: Data Validation and Cleansing

The fifth stage is dedicated to data cleansing and validation (Erl, Khattak and Buhler, 2016). This phase involves the pedantic scrutiny and rectification of any invalid data present within the datasets. The process entails either correcting or removing such data, contingent upon the nature of the analysis to be conducted. This step is essential, as invalid data can potentially distort or invalidate the research findings. It ensures that the data is in a state suitable for comprehensive examination and interpretation (refer to section 4.5.7 for further details). This meticulous preparation of data is vital to the integrity and accuracy of the study's outcome.

This section begins with an examination of the tweets' validity (section 4.5.5.1), ensuring the data's relevance and accuracy for analysis. This is followed by a delineation of the study's population and sample (section 4.5.5.2), crucial for contextualizing the research findings. Ethical considerations are then addressed, focusing on the copyright aspects of tweets (section 4.5.5.3) and the imperative of maintaining user anonymity and privacy (section 4.5.5.4). Subsequently, the discussion shifts to the pre-processing of data (section 4.5.5.5), detailing the steps taken to ready the tweet data for analytical scrutiny. Finally, the section concludes with an exploration of data security measures (section 4.5.5.6), underscoring the efforts to protect the integrity and confidentiality of the collected data.

4.5.5.1 Validity of the Tweets

The SAcovid19dataset is considered valid as it was acquired directly from the Twitter platform itself using SNScrape. Each tweet may be traced back to the platform using the tweet identification number (Tweet ID) or its unique URL. However, a user may also delete, make private or be suspended from the platform, which may potentially make certain tweets unavailable at different periods in time.

4.5.5.2 The Population and Sample under Review

The current study analysed all accessible tweets pertaining to COVID-19 in South Africa, filtered through specific keywords outlined in section 4.5.2.2, within a defined temporal frame. It is critical to note, however, that this corpus does not encompass the totality of potential tweets on this subject within the South African context.

This limitation can be attributed to several factors, including tweets becoming inaccessible because of deletion, account suspensions by Twitter in adherence to its policy, user account deletions, privacy adjustments by users and the specificity and exclusivity of the chosen keywords—thereby excluding even those tweets with minor typographical errors such as the omission of the letter 'd' in "covi19sa" (Roth and Harvey, 2018).

In this study, all available tweets were considered to be the sample, which closely approximates the population size. This is different to traditional research methodologies, where the process involves identifying a population, determining an

appropriately sized statistical sample and collecting this sample via a designated sampling strategy, which may be purposive or random (Creswell and Creswell, 2017).

4.5.5.3 Tweet Copyright

Under the South African legal framework, copyright considerations are governed by the Copyright Act No. 98 of 1978, which has been amended to stay relevant with the digital age (South Africa: DTI, 1978). For academic and non-commercial research, the South African Copyright Act allows for the use of copyrighted material under specific conditions that align with fair use principles. This includes usage for research, review or reporting of current events, provided that such use complies with the stipulated legal requirements, including acknowledging the source of the copyrighted material.

The stipulations set forth by Twitter's terms of service delineate that tweets categorized as public are universally accessible, not contingent upon the possession of a Twitter account (Twitter, n.d.a). This transparency ensures that users are aware of the public nature of their tweets. In conducting this research, care was taken to ensure that no intentional copyright infringement occurred by adhering strictly to the guidelines set forth by both Twitter and relevant copyright laws.

4.5.5.4 User Anonymity and Privacy

Sloan *et al.* (2020: 7) recommend anonymizing Twitter usernames to protect the identity of users and to prevent the disclosure of sensitive information. The researcher acknowledges this viewpoint and publishes results with this in mind; however, Twitter does allow publicly available usernames to be published. This study only identifies the most influential tweeters. Location data of users are not published and access to the data is restricted and governed by the university regulations.

4.5.5.5 Pre-processing Data

The initial data pre-processing included the elimination of exact duplicate entries across the C19MLdataset and SAcovid19dataset. This step was crucial to ensure data integrity and relevance. However, the study adopted a nuanced approach towards "near-duplicate tweets", which are identical tweets that bear different timestamps. The rationale behind retaining such near-duplicates lies in the unique dynamics of Twitter, where influential users often retweet content to enhance its

visibility and impact. This practice, commonly observed among Twitter influencers, serves as a strategic means to augment the reach of specific messages (Bild *et al.*, 2015).

Retweets, a fundamental aspect of Twitter's interaction model, were subjected to analysis despite Bruns and Liang (2012) suggesting their exclusion due to their propensity to skew analysis towards highly retweeted content. This study includes retweets, which is justified because of their role in amplifying messages. This characteristic is pertinent to the study's focus on identifying patterns of information dissemination as well as potential bot-driven activities. The examination of retweets, therefore, provides insights into the dynamics of content amplification on Twitter, including the detection of automated accounts or bots that might employ retweeting as a mechanism to spread information or misinformation widely.

The C19MLdataset underwent further pre-processing, which included the transformation of classification labels into binary values of 1's (Fake) and 0's (Not Fake). The rationale behind this binary coding was to create a consistent framework for analysis, with the label 1 representing potential Fake News and label 0 indicating not Fake News. This transformation was necessary to standardize the labelling process and maintain consistency across the dataset. Specifically, labels 'FALSE', 'misleading', 'F', '0' and 'fake' were converted to 1, while labels 'TRUE', 'T', 'U', '1' and 'real' were converted to 0.

4.5.5.6 Data Security

The custodianship of the SAcovid19dataset was exclusively maintained by the researcher. To ensure data integrity and security, robust digital infrastructures were employed, including encryption-secured laptop storage and cloud-based backups via Google Drive. The analytical processes were solely conducted by the researcher, with data shared with academic supervisors for evaluation or administrative requirements.

4.5.6 Stage 6: Data Aggregation and Representation

The sixth stage refers to the aggregation and representation of data. The amalgamation or aggregation of disparate datasets (section 4.5.3.1) into a single unified dataset, called C19MLdataset, a process that presents challenges due to varying data structures and the semantics (Erl, Khattak and Buhler, 2016). The case

of the C19MLdataset[5] was standardised and duplicates were removed. The resulted dataset comprised of 30 193 data points, segregated into text and label columns, with 17 069 denoting ‘Fake’ (label 1) news instances and 13 124 denoting ‘Not Fake’ (label 0) news instances (section 4.5.5.5).

The SAcovid19dataset is a combination of data collected using contextualized keywords, wildcard and geolocation extraction features (section 4.5.3.1.2). It comprises of 976 086 total tweets following the removal of duplicates. Additional fields were added to classify Fake News (1 and 0) alongside sentiment polarity scores and sentiment classes (positive, neutral and negative) for each tweet.

4.5.7 Stage 7: Data Analysis

Stage 7 represents the core phase where the actual analytical process takes place (Erl, Khattak and Buhler, 2016). Typically, the analysis in this stage is iterative, involving multiple cycles of review and refinement utilising a variety of analytical techniques. This multifaceted approach allows for a comprehensive examination of the data, ensuring that insights are drawn from different methodologies.

In this study, a preliminary pilot study was first conducted to assess the retrievability and volume of the SAcovid19dataset relevant to Fake News detection (section 4.5.7.1). Subsequently, the study progressed through several analytical stages: Fake News detection (section 4.5.7.2), sentiment analysis (section 4.5.7.3), descriptive analysis (section 4.5.7.4), CPA and the CUSUM method (section 4.5.7.5) and finally, the identification of social bots (section 4.5.7.6). These stages are visually depicted in Figures 4.4 to 4.7. The final part of this analysis (section 4.5.7.7) ties these methods back to the primary research questions.

4.5.7.1 Pilot Study

In[6] (2017) emphasizes the importance of conducting a pilot study as an initial, small-scale inquiry. It is crucial for identifying possible challenges and risks inherent in the

[5] https://docs.google.com/spreadsheets/d/1wxpRFWROJNGdG5u8TIRl27mOW93Xa1j7/edit?usp=sharing&ouid=101959657091259903296&rtpof=true&sd=true

[6] Researcher’s name is Junyong In.

research process while providing practical knowledge and insights into the specific area of study. Arain *et al.* (2010) concur and add that a pilot study is instrumental in ensuring the appropriateness, comprehensiveness and effective operationalization of observational categories tailored to the study's objectives. This preliminary step is pivotal in refining the research methodology and enhancing the study's overall rigor.

Consequently, a pilot study was initiated, successfully using the Python library, SNScrape, extracting 3 200 South African COVID-19-related data points from Twitter. This specific number was chosen as it represented the maximum free download limit set by Twitter's API (Twitter, n.d.a). ML suitable datasets were also explored to develop a COVID-19-related Fake News detection model (section 4.5.2.1).

The outcomes of this preliminary investigation affirmed the feasibility and the need for Fake News detection. However, the limited number of data points suggested a more extensive dataset to analyse online discussions about COVID-19 in South Africa. This requirement steered the research towards a longitudinal study design. Given the substantial volume of data amassed, a qualitative approach to an unrestricted dataset was deemed impractical, while a manual analysis was ruled out because of its extensive time requirements. The pilot study yielded several critical conclusions, as follows:

- The incorporation of images, embedded videos and audio clips into the analysis, despite their informational value, necessitates distinct analytical methodologies that diverge from those applied to textual data.
- A vast quantity of tweets related to COVID-19 was identified; however, a focused filtration process was necessary to extract content specifically related to the South African context (section 4.5.2.2).
- The selection of SNScrape as the tool for data retrieval was determined based on its proficiency in handling substantial volumes of data efficiently.

These insights led to the implementation of specific delimitations and qualifications within the scope of this research (section 1.9).

4.5.7.2 Fake News Detection

ML models (Chapter 3) were utilised to develop and select among Fake News detection models. Figure 4.3 depicts the process to which this study developed and selected a Fake News detection model.

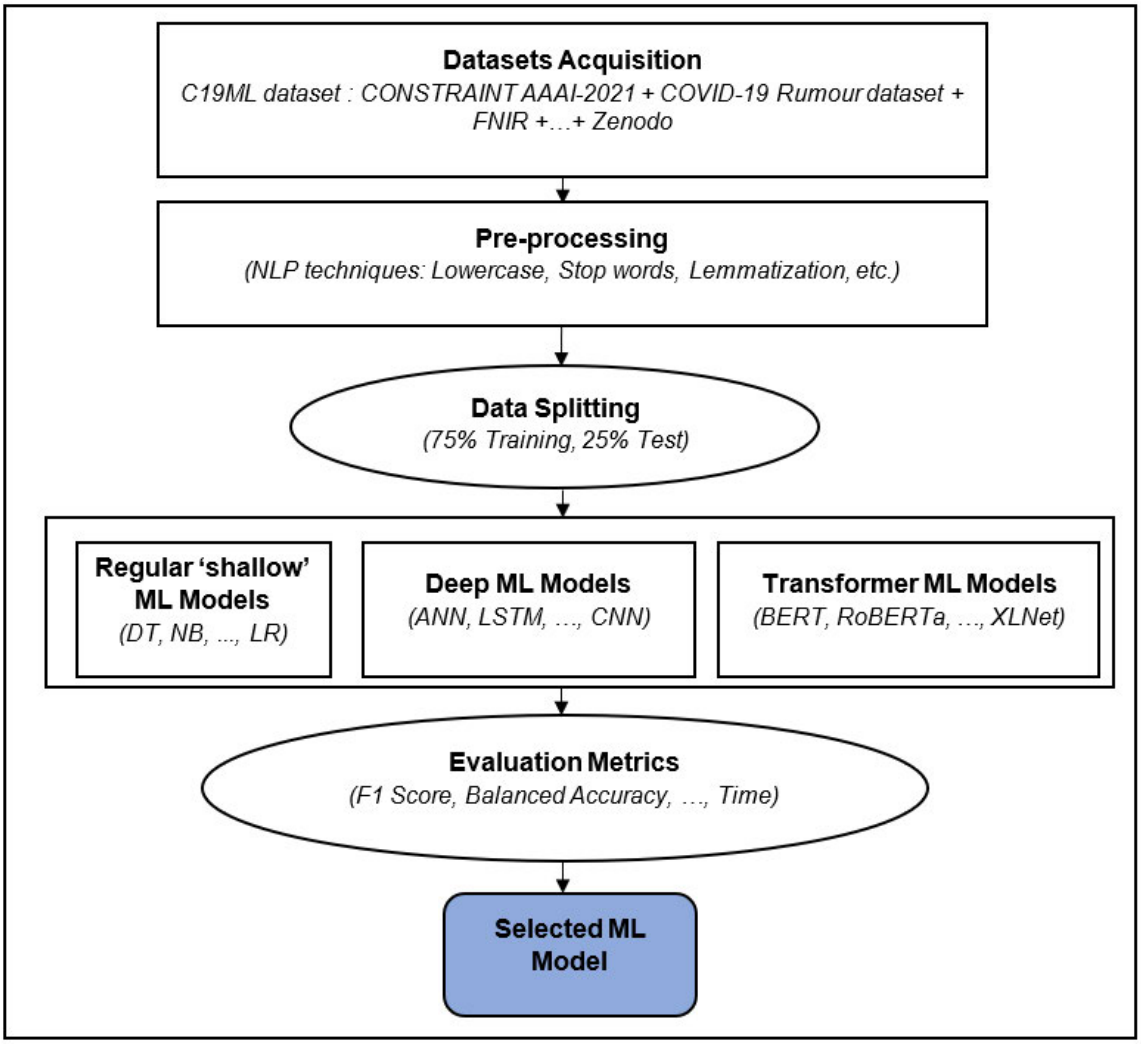

Figure 4.3: Fake News Detection ML Model Selection Framework

The C19MLdataset comprises of merged labelled datasets (section 4.5.6) and underwent comprehensive NLP pre-processing to ensure data cleanliness and readiness for subsequent analysis. This pre-processing pipeline included several critical steps: lowercasing, removing unimportant characters, tokenizing, removing stop words and lemmatization. Lowercasing was performed to maintain uniformity across the dataset by converting all text to lowercase, thus treating “Word” and “word” as identical tokens. The removal of unimportant characters, such as special symbols, was undertaken to reduce noise and focus on the meaningful content of the text.

Tokenization was employed to segment the text into individual words or tokens, transforming sentences like “Research is fun” into discrete tokens such as [“Research”, “is”, “fun”]. Stop words, which are common words like “the”, “is”, “in” and “and” that typically do not carry significant semantic weight, were removed to enhance the relevance of the dataset. Lemmatization was applied to convert words to their base or root form, for example, changing “talk”, “talking” and “talks” to “talk”, thus standardizing the lexical representations. Following these pre-processing steps, the dataset was divided into training and testing subsets, with 75% allocated for training and 25% reserved for testing. This split ensured that the model had a robust training set while retaining a substantial portion for evaluating the model’s performance.

The subsequent 75% training dataset was then leveraged to develop shallow[7] (section 3.2), deep[8] (section 3.3) and transformer models[9] (section 3.4). These ML models were evaluated using the 25% testing dataset on performance metrics, including accuracy, precision, recall and F1 scores. Accuracy, Precision, Recall and F1 Score are some of the conventional metrics used to determine the best performing classification model (Czakon, 2023). Accuracy measures how many observations, both positive and negative, were correctly classified. Precision measures the proportion of positive identifications that were actually correct. Recall measures the proportion of actual positives that were identified correctly. The general calculation for accuracy, precision, recall (sensitivity) and specificity are determined mathematically as follows:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$Precision = \frac{Number\ of\ correct\ positive\ results}{Number\ of\ positive\ results\ predicted\ by\ the\ classifier}$$

$$Recall = \frac{number\ of\ correct\ positive\ results}{number\ of\ all\ samples\ that\ should\ have\ been\ identified\ as\ positive}$$

$$Specificity = \frac{number\ of\ correct\ negative\ results}{number\ of\ all\ samples\ that\ should\ have\ been\ identified\ as\ negative}$$

[7] https://colab.research.google.com/drive/1TX3MviNbI-j9KUGKqJfWwBfZGPzNcHr4?usp=sharing
[8] https://colab.research.google.com/drive/1uBmhmrCsorBoAC5s5I8KnDgR_I_9wvbB?usp=sharing
[9] https://colab.research.google.com/drive/1dO8Q2Z0Zv7Iw8hFFzgP10Tv97EVHlKgH?usp=sharing

For binary predictions, accuracy, precision, recall and specificity can be calculated as follows:

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

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

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives and *FN* = False Negatives.

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1] or [0%, 100%] and it determines how precise and robust the classifier is (Mishra, 2018). The greater the F1 Score, the better is the performance of the model. Mathematically, it can be expressed as:

$$F1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

Balanced Accuracy is used to assess the performance of a classification model, especially when dealing with imbalanced datasets (Bej *et al.*, 2021). It is a preferred metric to use with imbalanced data as it accounts for both the positive and negative outcome classes and it does not mislead with imbalanced data. This metric was suited to this study because of the imbalanced nature of the labelled C19MLdataset (section 4.5.6).

$$Balanced\ Accuracy\ = (Sensitivity\ +\ Specificity)\ /\ 2$$

The selection of a ML model entailed assessing the aforementioned performance metrics and were compared for feasibility to facilitate this study. Priority was given to faster total training and evaluation time of a model with satisfactory metrics such as balanced accuracy and F1 Score being close to 100%.

Once the Fake News ML Detection model was decided, it was applied onto the collected SAcovid19dataset (section 4.5.3). Figure 4.4 highlights this process.

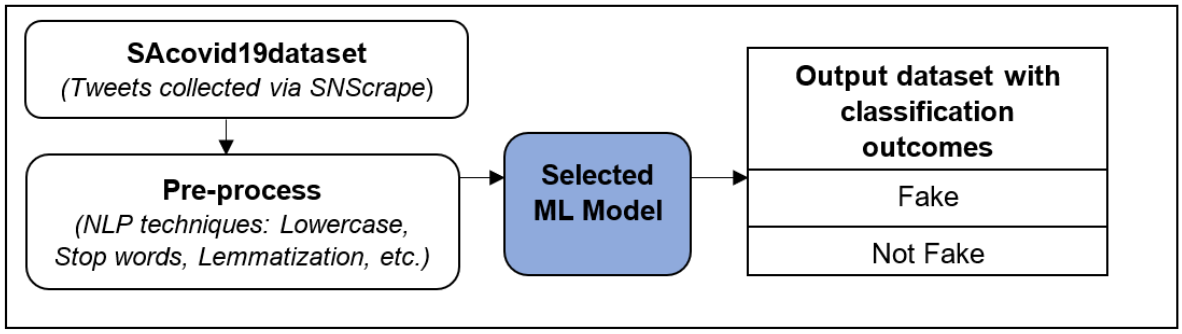

Figure 4.4: Fake News Classification of SAcovid19dataset

The SAcovid19dataset's[10] column containing tweets was pre-processed using NLP techniques consistent during the ML model development process (Figure 4.4). Thereafter, it parsed through the chosen ML model to create binary predictions for classes *Fake* or *Not Fake* and its output was placed into an added 'Prediction' column. This process to analyse COVID-19-related Fake News was demonstrated by Khan and Thakur (2022).

4.5.7.3 Sentiment Analysis

Sentiment analysis was conducted computationally, leveraging the capabilities of Python and the VADER library. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically designed for processing sentiments expressed on social media platforms. It is also effective in analysing text from various other domains (Hutto and Gilbert, 2014). This model classifies tweets into three distinct categories: positive, neutral and negative.

Tweets were categorized into different sentiment classes based on their sentiment polarity scores[11], following the guidelines established by Hutto and Gilbert (2014). The classification was as follows:

[10]https://docs.google.com/spreadsheets/d/1hIjVG2XTj7zk3SIQ-EOzhyh9OuaK4ful/edit?usp=sharing&ouid=101959657091259903296&rtpof=true&sd=true

[11] Hutto recommends a threshold of 0.05, which can be adjusted according to the desired level of sensitivity.

- Tweets with a polarity score less than -0.05 were classified as **Negative**.
- Tweets with a polarity score between -0.05 and 0.05 (inclusive) were classified as **Neutral**.
- Tweets with a polarity score greater than 0.05 were classified as **Positive.**

Figure 4.5 illustrates the sentiment analysis and classification process. Initially, Twitter data was acquired and underwent pre-processing to eliminate duplicates and corrupted data producing the SAcovid19dataset (section 4.5.6). Following this, the VADER model was applied on the SAcovid19dataset to compute sentiment polarity scores for each tweet. Based on these scores and the predefined thresholds, tweets were then stratified into one of the three sentiment classes: positive, negative, or neutral.

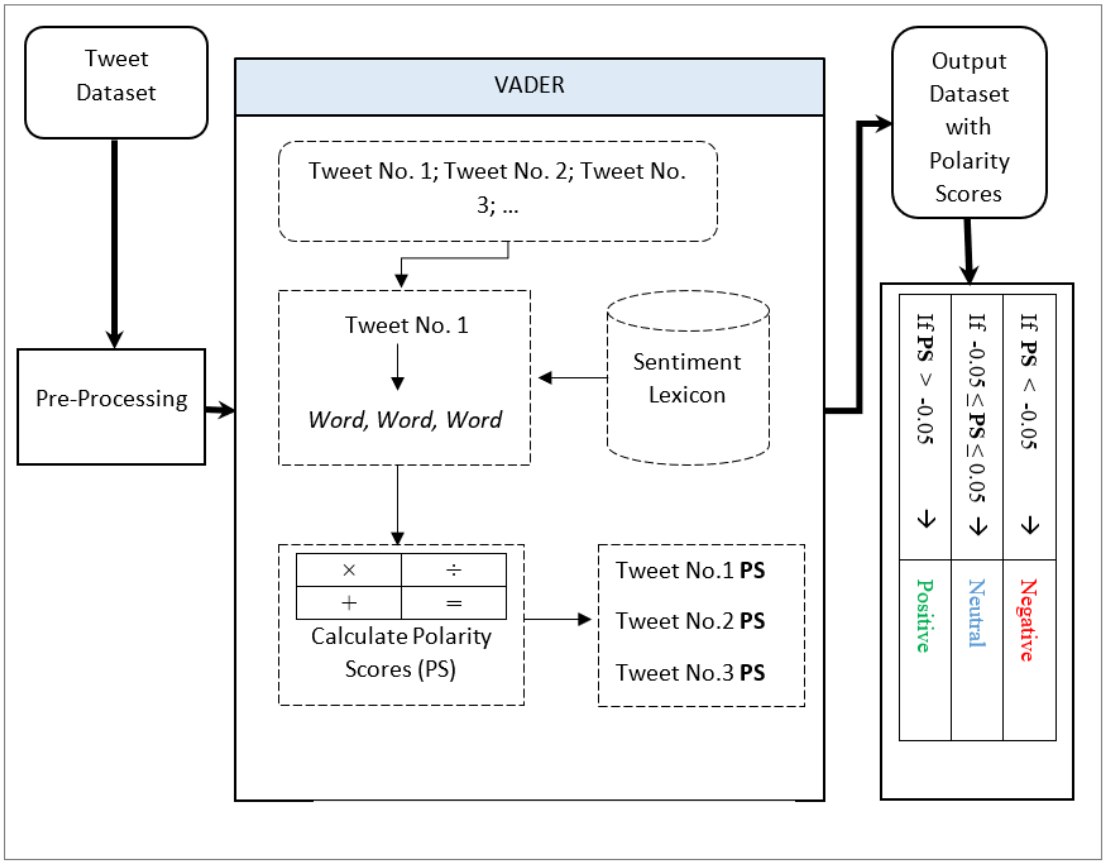

Figure 4.5: Procedure for Analysing Sentiment in Twitter Data

Consequently, additional fields were added to the dataset (SAcovid19dataset) which included the sentiment polarity scores and sentiment classifications.

4.5.7.4 Descriptive Analysis

This section describes the SAcovid19dataset following the sentiment analysis stage. It includes a comprehensive account of the corrupt and duplicate data that was identified and subsequently removed from the dataset, along with the total number of data points excluded. Additionally, the section explores the distribution of the dataset, specifically examining whether it conforms to a log-normal form, which is a common distribution pattern in social media data as suggested by Bild *et al.* (2015).

Key characteristics of the dataset are also outlined, such as the total number of data points, the linguistic composition of the tweets and the prevalence of secondary hashtags used within the dataset. The analysis extends to the examination of the volume of tweets distributed across different weekdays, providing insights into temporal patterns of Twitter engagement.

Descriptive analytics play a crucial role in transforming raw data from various sources into valuable insights about past events and trends (Bekker, 2019). In this study, descriptive analytics were employed to provide longitudinal observation insights.

Given that in the realm of Twitter-based campaigns, a log-normal distribution pattern is often observed (Bild *et al.*, 2015), the COVID-19-related data in this study was hypothesized to follow a log-normal distribution. To determine this, a log-normal test was employed. This statistical test aimed to evaluate the distribution of data over a 12-month period, from January 2020 to December 2020. This period was selected to capture a substantial amount of data, as the awareness and impact of the COVID-19 pandemic were particularly pronounced during this time. The hypothesis tests were as follows:

$$Null\ Hypothesis: \mathcal{H}_0 = The\ data\ follows\ a\ lognormal\ distribution$$

$$Alternative\ Hypothesis: \mathcal{H}_1 = The\ data\ does\ not\ follow\ a\ lognormal\ distribution$$

In addition to the existing dataset, two new fields were incorporated:

1. **Number of Hashtags**: Counting the number of hashtags used in each tweet.
2. **Number of URLs**: Tallying the number of URLs present in each tweet.

A monthly analysis of tweet frequency was conducted to identify patterns and pinpoint the months with the highest tweet volumes. Similarly, the computation of tweet

volumes by day of the week was undertaken, yielding insights into the daily tweet distribution from Monday through to Sunday.

To analyse associated tweet hashtags, tweets were standardized by making them all lowercase and thereafter filtered for hashtags, then tabulated.

Upon completion of this data augmentation and analysis, the study proceeded to examine the research questions, applying the insights gained from the statistical and descriptive analysis of the Twitter data.

4.5.7.5 Data Analysis: CPA using CUSUM Method

CPA was integral to examining Research Question 2, with its process illustrated in Figure 4.6. The CPA analysis comprised three key phases: Data Pre-processing, where the data was cleaned and organized for analysis; Data Mining, involving the extraction of patterns and identification of potential change points in the dataset; and Data Analysis, where these findings were thoroughly evaluated to interpret significant shifts or trends. This systematic approach was essential in analysing the data to uncover and understand critical changes relevant to the research question.

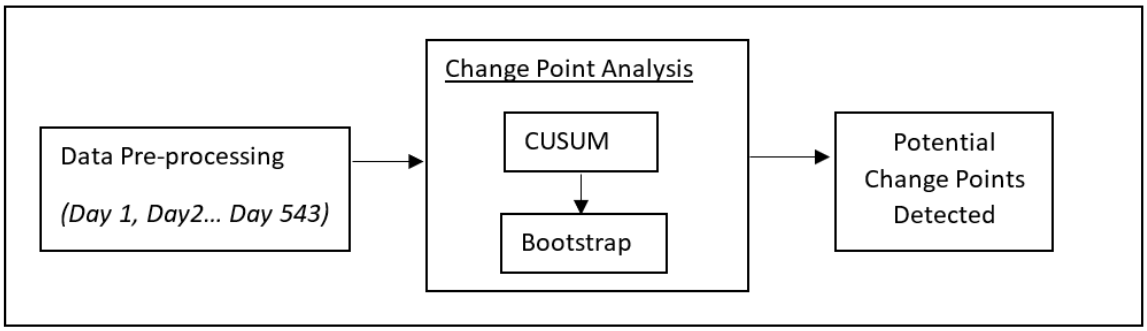

Figure 4.6: Procedure Flow for Change Point Analysis

4.5.7.5.1 Change Point Analysis (CPA)

CPA was employed to uncover subtle shifts in sentiment within the South African COVID-19 Twitter dataset. Drawing an analogy from seismology, where a foreshock cannot be identified until a subsequent larger earthquake occurs (Sykes, 1971), CPA was used to detect smaller but significant events in the pandemic data that might be overshadowed by larger occurrences. The analysis utilized the bootstrapped-based Cumulative Sum Analysis (CUSUM) method because of its ability to control the

change-wise error rate, thereby ensuring that the changes detected are likely genuine (Taylor, 2000).

CPA is adept at identifying multiple changes within large datasets and is effective when methods, such as control charts, might miss subtle shifts (Bendat and Piersol, 2011). This detection capability made CPA an ideal choice, enabling an exploration of *when* and *with what* confidence level sentiment changes in the dataset.

The detected change points were then triangulated with historic COVID-19 timelines to establish plausible correlations between online sentiment and real-world events in South Africa. Contrary to the view that CPA is akin to *ex-post-facto* research, this study contends that CPA offers valuable insights into the 'foreshocks' or early indicators preceding major events. Such analysis can provide crucial context and contribute to understanding the factors leading up to significant occurrences. While this research focuses on the national level, the application of CPA can and should be customized to suit various contexts, including individual institutions.

4.5.7.5.2 Bootstrap and Cumulative Sum of Charts (CUSUM) Algorithm

Taylor (2000) delineates a non-parametric methodology for CPA predicated on the mean-shift paradigm, positing that residuals are independent and identically distributed with a mean of zero. This methodology incorporates an iterative application of the CUSUM algorithm, augmented by bootstrapping analyses, to effectively identify shifts within time-series datasets. The distinctive attribute of this approach, facilitating the detection of multiple change points without reliance on the underlying data distribution, renders it particularly pertinent to the CPA objectives of this research.

CUSUM charts were developed by calculating and plotting cumulative sums from the dataset spanning 543 days, starting from 24 January 2020 to 19 July 2021. Initial observations identified the dates prior to 24 January 2020 as non-sequential outliers, which were subsequently excluded. Each day within this period is represented as d_i where i ranges from 1 to 543 and $i \in \mathbb{N}$. A day is delineated by the conventional 24-hour timeframe spanning from 00:00:00 to 23:59:59.

For the i^{th} day, the mean sentiment is represented as X_i and the sequence of cumulative sums $\{S_0, S_1, \dots, S_{543}\}$ is computed according to the following methodology (Taylor, 2000):

1. Calculate the overall mean sentiment, $\bar{X}$

$$\bar{X} = \frac{\sum_{i=1}^{i=N} X_i}{N} \ ,$$

   where $N = 543$, denoting the total number of days under consideration.

2. Initiate the series of cumulative sums by setting, $S_0 = 0$.

3. For each day i=1,2,…,543, compute the cumulative sums as:

$$S_i = S_{i-1} + (X_i - \bar{X})$$

   This involves adding the discrepancy between the day's average sentiment and the overall mean sentiment to the preceding cumulative sum.

Bootstrap analysis is then applied using S_{diff}, an estimator designed to quantify the extent of change detected through the cumulative sums. The estimator does not make any assumptions on how the data is distributed and is denoted as follows:

$$S_{diff} = S_{max} - S_{min} \text{ , where}$$

$$S_{max} = \max_{i=0,\dots,543} S_i \quad \text{and} \quad S_{min} = \min_{i=0,\dots,543} S_i$$

A single bootstrap involves a series of methodical steps to evaluate changes within a dataset. The process is as follows (Taylor, 2000):

1. **Generate a Bootstrap Sample**: Initially, a bootstrap sample comprising 543 units, denoted as $X_1^0, X_2^0, \dots, X_{543}^0$, is created by randomly reordering the 543 observed values. This reordering respects the original timeline of 543 days.

2. **Calculate the Bootstrap CUSUM**: Utilizing the bootstrap sample, the CUSUM for the bootstrap, represented as $S_0^0, S_1^0, \dots, S_{543}^0$, is computed.

3. **Determine Maximum, Minimum and Difference**: The maximum, S_{max}^0 and minimum, S_{min}^0, values of the bootstrap CUSUM are calculated, followed by the determination of their difference, S_{diff}^0.

4. **Comparison of Differences**: The final step involves comparing the bootstrap difference, S_{diff}^{0}, with the original difference, S_{diff}, to ascertain whether $S_{diff}^{0} < S_{diff}$

This analysis involves generating 10 000 bootstrap samples to assess the confidence level of significant changes observed. This number samples enhances the precision and reliability of the confidence intervals and p-values. By analysing the proportion of occurrences where $S_{diff}^{0} < S_{diff}$ across these samples, a 95% confidence level was applied to ascertain significant shifts in the data (Taylor, 2000). This rigorous statistical approach enhances the reliability of identifying significant changes within the dataset.

4.5.7.5.3 Accepting a Change Point Date Change

Granjon (2013) elucidates the process of identifying and accepting a change point date through a sequential change detection algorithm. The algorithm's primary objective is to methodically evaluate a time series to discern whether a change point occurs. The general structure of this algorithm is as follows:

initialisation
  if necessary
end
While *the algorithm is active* do
  Measure the current sample x[k]
  Decide between $\mathcal{H}_0$ (no change) and $\mathcal{H}_1$ (one change)
  If $\mathcal{H}_1$ *decided* **then**
    store the detection time $n_d \leftarrow k$
    estimate the change time n_c
    stop or reset the algorithm
  end
end

CPA Algorithm: General form of a sequential change detection algorithm

4.5.7.6 Data Analysis: Social Bots Identification

This section addresses Research Question 3, detailing the data analysis phase which employs three distinct methodologies for the identification of social bots, as depicted in Figure 4.7.

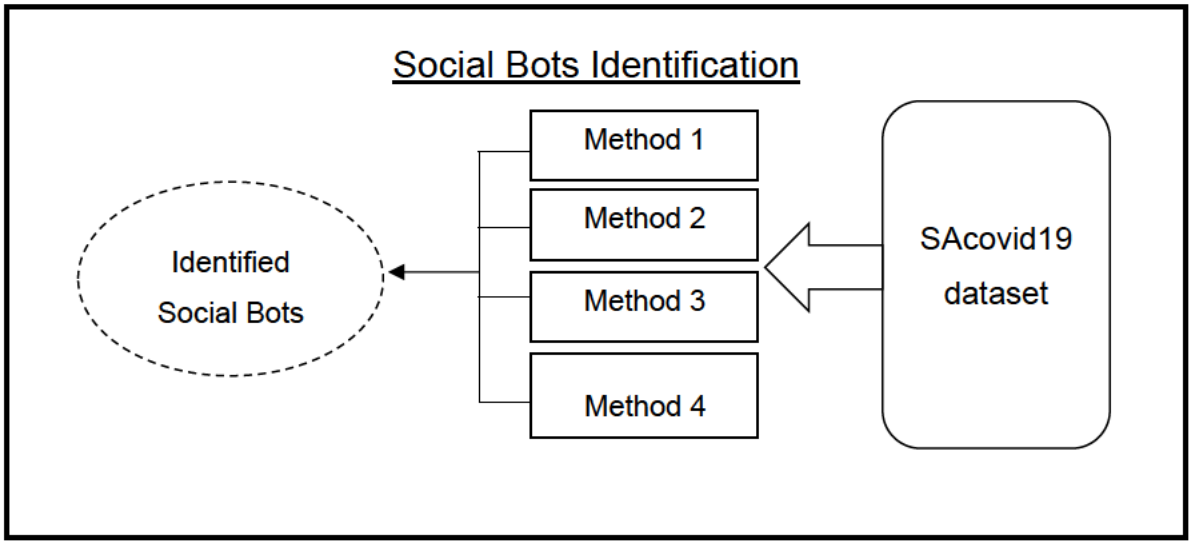

Figure 4.7: Social Bots Identification

To detect social bots such as Twitter bots and cyborgs, methods taken from Khan, Thakur, Obiyemi and Adetiba (2022b) were used and applied on the SAcovid19dataset. These were preferred as it did not require online access to detect social bots such as Botometer and DeBot which were further rendered unsustainable because of changes in Twitter's API access (Twitter, n.d.b).

Method 1: Identify tweeters with multiple postings at identical timestamps.

Identifying potential bot or cyborg accounts on Twitter relies on analysing tweeting behaviour, particularly focusing on users who posted more than once at the exact same timestamp. Given the practical limitations on how quickly a human can manually produce and post successive tweets, such behaviour suggests the possible use of automated mechanisms characteristic of bots or cyborg tweeters. The mathematical formulation for this method is delineated as follows:

Let a tweet = $(tw)_i \;\; where\; i \in \mathbb{N}$

$$\sum_i (tw)_i = Total\; No. of\; tweets$$

Let a user for a tweet = u_j , $where\ j\ \in\ \mathbb{N}\ and$

$$\sum_j u_j = Total\ No.of\ users$$

Let a Timestamp for a tweet = t_k , $where\ k\ \in\ \mathbb{N}\ \ and$

$$\sum_k t_k = Total\ No.of\ Timestamps$$

Timestamps are formatted according to the Gregorian Calendar and Coordinated Universal Time (UTC), using the structure: year/month/day hour:minute:second,

$$A\ user, u_j\ is\ assumed\ a\ bot\ or\ cyborg\ if\ at\ Timestamp, t_k:$$

$$\sum_i (tw)_{ijk} > 1$$

This method provides a quantitative criterion for flagging users who exhibit automated tweeting behaviour, contributing to the identification of non-human accounts within the dataset.

Method 2: Identify tweeters who produce multiple instances of identical tweet content.

Identifying Twitter users who exhibit a high frequency of posting duplicate tweet content, a behaviour indicative of potential bots or cyborgs. To refine the analysis, users who posted less than 30 tweets were excluded from consideration. This threshold was chosen to not only streamline the dataset for further analysis but also to leverage the Central Limit Theorem (CLT), which posits that sample means will approximate a normal distribution for sample sizes of 30 or more, irrespective of the population's distribution (Roscoe, 1975; Kwak and Kim, 2017; Islam, 2018). This principle is particularly useful for subsequent statistical analysis.

The underlying premise of Method 2 is that certain bots and cyborgs are programmed to engage in spamming behaviours, such as scheduling or automatically triggering the posting of repetitive tweet content. Thus, these accounts are expected to produce a significantly higher volume of duplicate tweets compared to human users. While human users might occasionally repost similar content, the frequency of such duplication is generally much lower.

A user is classified as a bot or cyborg under this method if the proportion of their duplicate tweet content constitutes 30% or more of their total tweets within a specified timeframe. This criterion is mathematically represented as follows:

Let each unique tweet = $(\overline{\overline{\text{tw}}})_{\text{i}}$ where $\text{i} \in \mathbb{N}$

$$\sum_{\text{i}} (\overline{\overline{\text{tw}}})_{\text{i}} = \text{Total No. of unique tweets}$$

$$\text{A user, } \text{u}_{\text{j}} \text{ is assumed to be a bot or cyborg if:}$$

$$\sum_{\text{i}} (\text{tw})_{\text{ij}} \geq \frac{10 * \sum_{\text{i}} (\overline{\overline{\text{tw}}})_{\text{ij}}}{7} \text{ and } \sum_{\text{i}} (\text{tw})_{\text{ij}} \geq 30$$

This formula establishes a quantifiable measure to distinguish users engaging in bot-like repetitive tweeting behaviour from those of typical human tweeters, thereby aiding in the identification of non-human actors within the Twitter ecosystem.

Method 3: The identification of Twitter users, particularly cyborgs and bots that predominantly utilize automation software for tweeting.

These automated behaviours include posting messages, retweeting, following users and replying to posts, often used to amplify the user's intended messages or causes. The crux of this method lies in distinguishing between users who employ automation tools for legitimate, occasional purposes (e.g., auto-replies during vacations) and those who use such tools predominantly, aligning more closely with bot or cyborg characteristics. The latter group's behaviour is marked by a significant reliance on automation, as opposed to occasional use which might be seen in more human-like interactions on the platform.

A threshold to discern users was established, stipulating a minimum criterion of 30 tweets, with a significant proportion (70% or above) stemming from recognized automated applications. This threshold aims to filter out users whose automated tweeting frequency suggests they are likely bots or cyborgs, based on the premise that genuine users would not predominantly use automation for their Twitter interactions. The sources of tweets, as captured in the dataset's metadata (section 3.6.3.2), facilitate this analysis by indicating the origin of each tweet.

The analysis focused on a set of well-known automation tools, including “IF This Then That” (IFTTT), Hootsuite, TweetDeck, TweetCaster and Buffer (IFTTT, n.d.; Hootsuite, n.d.; Tweetdeck, n.d.; TweetCaster, n.d.; Buffer, n.d.). These platforms offer various automation features, from scheduling tweets to managing multiple accounts and automating tasks like tweeting and retweeting.

This method is mathematically denoted as follows:

Define the set of recognized automating tweet sources as

$$A = \{\text{IFTTT};\ \text{Hootsuite};\ \text{TweetDeck};\ \text{TweetCaster};\ \text{Buffer}\},$$

For any user u_j, let the total occurrences of tweets from automating sources be denoted by $(ats)_j$, $where\ (ats)_j \in A,\ j\ \in\ \mathbb{N}$

$$A\ user\ u_j, is\ identified\ \text{as}\ a\ bot\ or\ cyborg\ if:$$

$$\frac{(ats)_j}{\sum_i (tw)_{ij}} \geq 0.7\ , and\ \sum_i (tw)_{ij} \geq 30$$

This approach ensures that the analysis is focused on users whose use of automation significantly exceeds what might be expected from typical, manual Twitter use, thereby highlighting potential bots and cyborgs within the dataset.

4.5.7.7 Data Analysis Methods and the Research Questions

In this section, we revisit and refine the main research question and its related sub-questions and detail the specific steps taken in the data analysis process.

4.5.7.7.1 Research Question 1

How can the COVID-19 infodemic on Twitter be computationally analysed within a South African context for Fake News?

This research question has resonance since COVID-19 introduced unparalleled levels of disruption locally and globally. In the age of social media, this naturally generated voluminous discussions centred on COVID-19, which, in turn, fuelled an associated infodemic. Navigating through misinformation, disinformation and malinformation on such platforms with the intention to monitor or conduct research

would prove futile if not for computational approaches such as data science. ML, a core component of data science, can potentially mitigate the challenges encountered when attempting to analyse millions of tweets since it can process large chunks of data in a short period pending available resources.

To analyse Research Question 1, annotated datasets (sections 4.5.2 and 4.5.3) had to be collected, cleaned and transformed in preparation for ML experimentation. In addition, a collection of South African-related COVID-19 tweets (SAcovid19dataset) had to be acquired computationally (section 4.5.3). Furthermore, ML experimentations are needed to yield a desired model to detect COVID-19-related Fake News (section 4.5.7.2). Thereafter, the chosen model was applied on the SAcovid19dataset (Figure 4.4) to predict tweet content as either *Fake* or *Not Fake*. Finally, a Fake News trend analysis was conducted together with descriptive analysis to identify the leading Fake News tweeters.

4.5.7.7.2 Research Question 2

Research Question 2 investigates changes in online sentiment from the SAcovid19dataset:

> **Where there significant changes in the average sentiment from the South African COVID-19-related Twitter data?**

Understanding shifts in sentiment can provide valuable insights into how public opinion and perceptions evolved during the pandemic. This research question, therefore, attempts to contribute to the broader field of social and behavioural sciences by providing empirical evidence on how a pandemic impacts local public sentiment. This can inform future research in psychology, sociology and public health.

To examine Research Question 2, the VADER sentiment analysis model was utilized (section 4.5.7.2) and assigned a sentiment polarity score to each tweet, categorizing them into one of three sentiment classes: positive, negative and neutral. The aggregate count of tweets in each sentiment category provided an overview of the overall sentiment during the study's longitudinal timeframe. However, given that sentiment can vary over time, a more detailed monthly analysis was necessary to capture the prevailing sentiment more precisely.

The sentiment scores of all tweets within a month involved calculating the average monthly sentiment polarity for each sentiment class separately and depicting these as three distinct trend lines. This analysis was conducted for months with more than one tweet, excluding November 2019 and December 2019, where only a single tweet was recorded in each month.

Additionally, the volume of tweets was overlaid on the graph, covering the 17-month period (543 days) of the study (section 4.5.7.5). The vertical values on the positive, negative and neutral curves represented the total percentage of tweets for each sentiment at each point, allowing for a comparative analysis of the volume of each sentiment over time on the x-axis. This separation of sentiments was crucial for observing the breadth and variation of differing opinions in any given month, significantly contributing to the depth of the sentiment analysis.

To ascertain the changes in sentiment over the duration of this study, the following hypothesis was tested:

$$H_0 = There\ were\ no\ changes\ in\ the\ overall\ average\ sentiment\ polarity$$

$$H_1 = There\ were\ change(s)\ in\ the\ overall\ average\ sentiment\ polarity$$

This involved conducting a time series analysis based on the average daily sentiment. Furthermore, this was followed by implementing a CPA using the CUSUM method. The technical specifics of this process are detailed in section 4.5.7.5 of the study. Furthermore, leveraging CPA as a means to explore links between online sentiment and real-world events was demonstrated (Khan, Thakur, Obiyemi and Adetiba, 2022a).

The segment of the study outlined in Figure 4.7 encompasses three critical phases: data pre-processing, data mining and data analysis. Initially, data pre-processing was conducted by segmenting the overall timeline into discrete days, followed by the computation of average sentiment for each day. This preparation was pivotal for transitioning into the data mining phase, where deeper insights and patterns within the dataset were investigated.

CPA was employed, with the findings subsequently explored in the data analysis phase. The sentiments for the SAcovid19dataset were calculated on a daily basis, with the dataset ordered chronologically over 543 consecutive days. Each day's data

reflected the average sentiment expressed in the tweets. CPA was applied to this time-ordered dataset to identify any significant shifts in sentiment across the period. The CPA methodology, leveraging CUSUM and bootstrap techniques, was applied as designed, without modifications or additional training, facilitating the identification of temporal sentiment changes within the dataset.

These methods have been thoroughly described and justified. CPA was chosen as it is statistically adept at detecting shifts in sentiment, a crucial aspect given the longitudinal scope of this study. Moreover, CPA is effective not only in identifying changes but also in providing exact dates of change as well as confidence level of these changes.

4.5.7.7.3 Research Question 3

Research Question 3 investigated the potential role of social bots:

> **Were social bots leveraged to produce content around South African-related COVID-19 Twitter content?**

In the era of sophisticated information warfare, identifying the use of social bots (bots) in disseminating content provides insights into the tactics used to spread disinformation. Given the unique socio-political and cultural context of South Africa, researching bot activity in this specific setting can provide valuable insights that are regionally relevant, contributing to a more global understanding of information manipulation tactics during pandemics.

The investigation into social bots, specifically focusing on Twitterbots and Cyborgs, was conducted through a rigorous application of techniques that leverage their general characteristics, as delineated in section 2.11. The methodologies employed to discern the presence of a Twitterbot or Cyborg are comprehensively described in Methods 1 to 3, which are detailed in section 4.5.7.6. In pursuit of addressing Research Question 3, the outcomes derived from these methods were meticulously analysed in relation to the Top 10 tweeters. This analysis aimed to ascertain whether social bots played a pivotal role as significant influencers in the dissemination of South African-related COVID-19 content. These methods have been demonstrated by Khan, Thakur, Obiyemi and Adetiba (2022b).

In summary, this section methodically addressed the initial pilot study undertaken to evaluate the pertinence of analysing COVID-19-related Twitter discussions within the South African context. Subsequently, it systematically delineated the technical aspects of the data analysis process. This encompassed the implementation of Fake News detection, sentiment analysis and descriptive analysis, along with an in-depth explanation of the CPA and the CUSUM method. The methodology for identifying social bots was also elucidated. The section culminated by articulating the practical application of these methodologies in answering the research questions established by this study.

4.5.8 Stage 8: Data Visualisation

Stage 8 focuses on data visualization (Erl, Khattak and Buhler, 2016). This stage leverages graphical tools to represent the analysis conducted, facilitating verification and allowing for alternative interpretations by different analysts. The visualizations encompass a wide range of formats, from graphical representations to tables, covering aspects such as Fake News, sentiment analysis and CPA trends as well as the frequency of tweets and descriptive tables. This stage, presented in Chapter 5 of the thesis, signifies the importance of visual tools in enhancing the comprehensibility and interpretability of the data analysis outcomes.

4.5.9 Stage 9: Utilisation of Analysis Results

Stage 9, the final phase of the Data Analytics Lifecycle, is dedicated to the utilization of analysis results, serving as the culmination of the analytical process described by Erl, Khattak and Buhler (2016). This stage transitions the outcomes of the data analysis into actionable insights, making them accessible to key stakeholders, including academics and practitioners. The objective is to empower these stakeholders with data-driven evidence to inform strategic decisions and foster further exploration and utilization of the findings. The insights generated through the analysis are systematically presented and discussed in Chapters 5 and 6 of the study, ensuring that the research not only contributes to academic discourse but also offers tangible value to decision-makers in the field.

4.6 Conclusion

In this chapter, the methodological foundation and strategic design choices underlying the study were elucidated, with the investigation being steered by a post-positivist philosophical orientation. The research methodology employed is fundamentally quantitative, supplemented by a deductive approach to research and framed within a longitudinal study design. The primary aim of this chapter was to articulate the methodologies employed for data collection and analysis, tailored to effectively address the research inquiries presented.

The data corpus, termed the SAcovid19dataset, comprises a compilation of tweets spanning from 8 November 2019 to 19 July 2021. Notably, the dataset was curated to exclude multimedia elements such as audio, images and videos. In preparation for conducting analyses on Fake News detection, sentiment analysis, CPA and the identification of social bots, a rigorous process of data curation was undertaken to eliminate corrupted or redundantly duplicated data entries.

This chapter then provided an exposition of the nine stages encompassed by the Data Analytics Lifecycle, highlighting their applicability and significance to the current study. It delved into the intricacies of the data analysis stages, showcasing their instrumental role in elucidating the research questions. Justifications for the methodological decisions, aimed at ensuring the analytical rigor and adherence to ethical standards of the study, were also presented.

The findings derived from this comprehensive methodological approach are disclosed in the subsequent chapter (Chapter 5), providing insights and conclusions drawn from the analytical endeavours.

Chapter 5: Findings and Results

5.1 Introduction

This chapter presents the descriptive analysis and findings from the SAcovid19dataset (section 5.2), followed by the detailed findings related to the three primary research questions (sections 5.3–5.5). The chapter concludes by summarizing the main findings and their implications (section 5.6).

5.2 Descriptive Analysis of the SAcovid19dataset

In this section, a descriptive analysis of the SAcovid19dataset, with 976 086 tweets, is conducted, focusing on various aspects such as linguistic analysis, tweet volume, tweet distribution, daily tweet distribution, tweets with multiple hashtags and the number of Twitter users. This analysis provides a foundational understanding of the data before addressing the research questions.

5.2.1 Linguistic Analysis

A total of 57 distinct languages were detected within the SAcovid19dataset metadata. English emerged as the dominant communication language, accounting for 863 542 tweets, which constitutes 88.47% of the overall dataset.

Table 5.1 showcases the distribution of tweets across the Top 20 languages, alongside their respective tweet counts. Notably, a substantial number of tweets, amounting to 63 869, were categorized as undefined. The undefined classification was due to the Twitter algorithm's limitation in recognizing tweets in languages that do not align with its coding parameters. Following English, Indonesian was recorded as the second most frequent defined language, with a total of 9 924 tweets.

Table 5.1: Tweets Distribution by Language (Top 20)

Language	Total		Language	Total
English	**863 542**		Arabic	1 129
Undefined	63 869		Romanian	1 038
Indonesian	9 924		Turkish	1 022
Dutch	8 632		Portuguese	928
Tagalog	8 368		Catalan	896
Haitian Creole	2 422		German	818

Language	Total		Language	Total
Spanish	2 146		Finnish	809
Estonian	1 599		Polish	705
Hindi	1 543		Welsh	651
French	1 133		Italian	591

5.2.2 Tweet Volume

The monthly tweet distribution for SAcovid19dataset, presented in Table 5.2, illustrates the frequency of tweets from 8 November 2019 to 19 July 2021. The initial months of November and December in 2019 registered a minimal presence, with only two and one tweets respectively. A significant surge in activity was observed in March 2020, with a total of 337 571 tweets, marking the peak of the dataset, closely followed by April 2020, which accounted for 205 618 tweets. These two months combined constituted 55.65% of the total 976 086 tweets analysed.

Table 5.2: Tweet Distribution

Year of Date	Month	Number of Tweets
2019	November	2
	December	1
2020	January	74
	February	592
	March	337 571
	April	205 618
	May	91 806
	June	45 055
	July	57 518
	August	31 083
	September	14 877
	October	11 221
	November	17 123
	December	35 607
2021	January	41 144
	February	23 018
	March	14 213
	April	8 701
	May	9 780
	June	15 250
	July	15 832
Total		**976 086**

5.2.3 Tweet Distribution

The tweet distribution for the period from January 2020[12] to December 2020 was subjected to a hypothesis test to determine the distribution pattern of the data:

$$\mathcal{H}_0 = The\ data\ follows\ a\ lognormal\ distribution$$

$$\mathcal{H}_1 = The\ data\ does\ not\ follow\ a\ lognormal\ distribution$$

A Kolmogorov-Smirnov test, a type of goodness-of-fit test, was conducted to test for log-normality within this 12-month period. The results of this test confirmed that the data adhered to a lognormal distribution.

In Table 5.3, descriptive statistics are presented alongside the results of the distribution fit for the total number of tweets per month.

Table 5.3: Descriptive and Distribution Test Results for Total Tweets per Month from January 2020 until December 2020 (α = 0.05)

Descriptive Statistics							
Count	*Mean*	*StDev*	*Median*	*Min*	*Max*	*Skew (Pearson)*	*Kurt (Pearson)*
12	70678,75	101322,41	33345	74	337571	1,61	1,34
Distribution Results							
μ	*Sigma*	*Log-Likelihood*	*D*	*p Value (Two Tailed)*	*BIC*	*AIC*	
9,86	2,28	-145,16	0,24	0,42	295,29	259,70	

Given the p-value of 0.42, the null hypothesis ($\mathcal{H}_0$) could not be rejected at the 5% significance level. This leads to the conclusion that the distribution of tweets from January 2020 until December 2020 is consistent with a lognormal distribution, indicating that the frequency of tweets is multiplicatively random and the logarithms of the tweet counts are normally distributed.

[12] The lognormal tweet distribution test was applied to data from the year 2020, based on the available range and volume (Table 5.2).

5.2.4 Daily Tweet Distribution

The analysis of the tweet distribution across the days of the week, as depicted in Figure 5.1, reveals a pattern of fluctuating tweet volumes. Notably, there is a pronounced surge in activity on Mondays, followed by a substantial decline in tweet volume by Saturday, indicating a weekly cyclical pattern in Twitter.

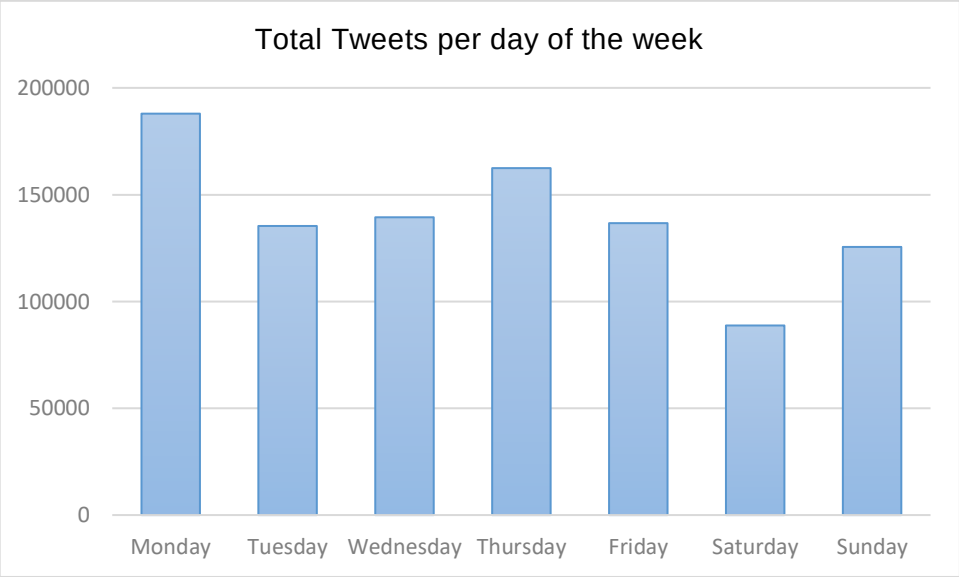

Figure 5.1: Tweet Distribution per Day of the Week

5.2.5 Tweets with Multiple Hashtags

Table 5.4 ranks the ten most prevalent hashtags, with #covid19sa leading with 77 301 mentions.

Table 5.4: Top 10 Hashtags in the SAcovid19dataset

Hashtags	Counts
#covid19sa	77 301
#lockdownsa	77 114
#covid19	47 154
#coronavirussa	44 735
#covid19insa	37 365
#covid19southafrica	34 710
#coronavirusinsa	33 984
#lockdownsouthafrica	30 322
#coronavirus	26 598
#covid_19	21 560

5.2.6 Number of Twitter Users

The total number of unique Twitter users in the SAcovid19dataset were 229 968. Table 5.5 ranks the Top 10 with the highest tweeter being *Bhekisisa_MG* with 7 980 tweets, while 4 of the Top 10 ranked users are associated to media outlets.

Table 5.5: Top 10 Highest Tweeting Users

Username	Number of Records
Bhekisisa_MG	7 980
Jobfundi_com	6 444
CapricornFMNews	4 120
CapeTownFreeway	3 566
POWER987News	3 284
SANDF_ZA	2 927
JacaNews	2 822
ewnreporter	2 708
weworx	2 548
Powerfm987	2 535

5.3 Findings of Research Question 1

Research Question 1 asked:

> **How can the COVID-19 infodemic on Twitter be computationally analysed within a South African context for Fake News?**

To address this question, the study was divided into several key parts. Initially, ML suitable datasets (C19MLdataset) were identified (section 4.5.2.1) to develop a model for detecting Fake News on Twitter. Next, an automated method using SNScrape python library was used to collect relevant South African COVID-19 Twitter content using specific keywords and hashtags (section 4.5.2.2), ensuring a comprehensive dataset. Various ML models, including LightGBM, Bi-LSTM and BERT (sections 3.4 – 3.6), were evaluated based on performance metrics such as accuracy, precision, recall, F1-score, speed and balanced accuracy (section 4.5.7.2). The best feasible performing model (section 5.3.1) was applied to the SAcovid19dataset to classify Fake News, followed by a Fake News trend analysis to examine the distribution of detected Fake News and other relevant patterns (sections 5.3.3–5.3.4).

5.3.1 ML Performance Metrics and Model Selection

Given the significant importance of COVID-19-related information dissemination and verification of its accuracy, the researcher decided to neutrally assess a variety of popular algorithms as informed by the literature review. This evaluation process was extensive, incorporating traditional, DL and transformer ML models with the aim of drawing comparative insights. In total, 27 shallow ML, five (5) DL and seven (7) transformer models were deployed.

The balanced accuracy scores (section 4.5.7.2) were used to rank the ML models for the shallow, DL and transformer models. This study utilises balanced accuracy scores and model execution times (training and evaluation) as the selection criterion; however, alternate performance metrics may be selected by other researchers.

The performance scores for the 27 shallow ML models are reflected in Table 5.6. The ExtraTreesClassifier had the highest balanced accuracy score of 87.55% but took about 81 seconds to complete, with RandomForestClassifier having the second highest balanced accuracy score of 86.52%, completing at about 40 seconds. The SVC model took the longest to run with a time of 1128.16 seconds. LightGBM (LGBMClassifier) had the third highest balanced accuracy score of 86.48% but at significantly better speed, about 7.13 seconds, making it approximately 11 times faster than the ExtraTreesClassifier and five times quicker than the RandomForestClassifier.

Table 5.6: Results of ML Models Evaluated

Model	Balanced Accuracy	Accuracy	Precision	Recall	Specificity	F1 Score	Time Taken (sec)
ExtraTreesClassifier	0.8755	0.8792	0.8849	0.9039	0.8470	0.8790	80.75
RandomForestClassifier	0.8652	0.8706	0.8702	0.9063	0.8241	0.8701	39.67
LGBMClassifier	0.8648	0.8699	0.8707	0.9042	0.8254	0.8695	7.13
XGBClassifier	0.8621	0.8681	0.8653	0.9079	0.8162	0.8675	11.54
BaggingClassifier	0.8508	0.8524	0.8741	0.8634	0.8382	0.8526	153.21
SVC	0.8405	0.8495	0.8381	0.9096	0.7714	0.8482	1128.16
NuSVC	0.8281	0.8394	0.8213	0.9152	0.7409	0.8373	959.95
LinearDiscriminantAnalysis	0.8269	0.8345	0.8327	0.8852	0.7687	0.8335	12.84
RidgeClassifierCV	0.8266	0.8343	0.8322	0.8854	0.7678	0.8333	10.94
RidgeClassifier	0.8266	0.8343	0.8322	0.8854	0.7678	0.8333	2.03
AdaBoostClassifier	0.8231	0.8303	0.8311	0.8784	0.7678	0.8294	25.14
LogisticRegression	0.8212	0.8273	0.8335	0.8679	0.7745	0.8266	1.81
DecisionTreeClassifier	0.8177	0.8214	0.8394	0.8461	0.7894	0.8213	25.00

Model	Balanced Accuracy	Accuracy	Precision	Recall	Specificity	F1 Score	Time Taken (sec)
CalibratedClassifierCV	0.8125	0.8210	0.8189	0.8775	0.7476	0.8198	365.90
BernoulliNB	0.8090	0.8202	0.8079	0.8948	0.7233	0.8181	2.03
SGDClassifier	0.7989	0.8039	0.8197	0.8374	0.7604	0.8036	16.82
NearestCentroid	0.7952	0.8041	0.8045	0.8632	0.7272	0.8026	1.24
LinearSVC	0.7933	0.7976	0.8176	0.8264	0.7601	0.7974	92.03
ExtraTreeClassifier	0.7670	0.7712	0.7966	0.7994	0.7345	0.7712	1.76
Perceptron	0.7647	0.7678	0.7985	0.7882	0.7412	0.7680	1.71
PassiveAggressiveClassifier	0.7525	0.7565	0.7853	0.7835	0.7214	0.7566	2.07
QuadraticDiscriminantAnalysis	0.7469	0.7524	0.7767	0.7889	0.7050	0.7521	12.14
GaussianNB	0.7264	0.7282	0.7701	0.7402	0.7126	0.7288	1.42
KNeighborsClassifier	0.6310	0.6688	0.6452	0.9203	0.3417	0.6344	16.00
LabelSpreading	0.5169	0.4537	1.0000	0.0337	1.0000	0.3038	57.38
LabelPropagation	0.5169	0.4537	1.0000	0.0337	1.0000	0.3038	60.74
DummyClassifier	0.5000	0.5654	0.5654	1.0000	0.0000	0.4084	0.96

The performance metrics for the five DL models are indicated by Accuracy, Precision, Recall and F1 scores, which are shown in Table 5.7. Ranked by highest balanced accuracy, the Bi-LSTM (88.18%) outperformed LSTM (87.67%) and GRU (87.62%) minutely, with a slower time of about 123 seconds compared to LSTM's 107 seconds and GRU's 70 seconds. The fastest model in this segment was the CNN at about 46 seconds. RNN was significantly slower than the rest, with a time of 3 209 seconds and the lowest balanced accuracy score of 84.98%.

Table 5.7: DL Model Performance Metrics

Model	Balanced Accuracy	Accuracy	Precision	Recall	Specificity	F1	Time (sec)
Bi-LSTM	0.8818	0.8773	0.9196	0.8449	0.8327	0.8980	123.47
LSTM	0.8767	0.8809	0.8840	0.9086	0.8449	0.8961	107.32
GRU	0.8762	0.8748	0.8787	0.8835	0.9046	0.8940	70.42
CNN	0.8617	0.8671	0.8674	0.9030	0.8205	0.8849	45.87
RNN	0.8498	0.8539	0.8635	0.8807	0.8190	0.8721	3209.34

The performance metrics for the seven transformer-related models are reflected in Table 5.8. RoBERTa achieved the highest balanced accuracy score of 89.87%, with all transformer models obtaining scores of at least 86.71%. The fastest trained transformer was Electra and had a time of about 150 seconds, whereas the longest was XLNet with a time of 854 seconds. The range between the highest and lowest performing model based on balanced accuracy was approximately 3%.

Table 5.8: Transformer Model Performance Metrics

Model	Balanced Accuracy	Accuracy	Precision	Recall	Specificity	F1 Score	Time (sec)
RoBERTa	0.8987	0.8991	0.9186	0.9014	0.8961	0.9099	535.77
BERT	0.8970	0.8991	0.9092	0.9126	0.8814	0.9109	537.99
AlBERT	0.8958	0.8984	0.9057	0.9157	0.8760	0.9106	597.44
Distilbert	0.8933	0.8959	0.9038	0.9131	0.8735	0.9084	272.49
XLM-RoBERTa	0.8829	0.8804	0.9201	0.8634	0.9025	0.8909	758.51
XLNet	0.8817	0.8871	0.8826	0.9231	0.8403	0.9024	854.48
Electra	0.8671	0.8683	0.8890	0.8765	0.8577	0.8827	149.76

A comparison of Tables 5.6, 5.7 and 5.8 revealed that the best performing models are the BERT type transformers with RoBERTA and BERT having the highest balanced accuracy scores of 89.87% and 89.70% respectively. However, the processing time for transformers and DL models were significantly slower than most conventional shallow ML models.

The LightGBM was opted as the preferred model to parse through the SAcovid19dataset because of its significant superiority speed over other better performing DL and transformer models. In addition, the LightGBM algorithm had a superior execution time advantage over the ExtraTreeClassifier algorithm (Table 5.6), while only having a marginally 1.07% lower balanced accuracy score but maintaining a significant F1 score of 86.95%. This was a decisive factor considering the computational requirements needed for transformer and DL models to process the SACovid19dataset which can be costly, time consuming and unfeasible as a self-funded researcher. Subsequently, the LightGBM model parameters were optimized manually using learning curves[13] and evaluated with the log loss function (section 3.4.4.1.4). This is reflected in Table 5.9 as follows:

[13] https://colab.research.google.com/drive/1TX3MviNbI-j9KUGKqJfWwBfZGPzNcHr4?usp=sharing

Table 5.9: LightGBM Model Parameters

Parameter	Values
Boosting type	'Gradient-boosted decision tree'
Class weight	None
Colsample bytree	1.0
Importance type	'split'
Learning rate	0.09
Max depth	-5
Min child samples	20
Min child weight	0.001
Min split gain	0.0
n estimators	440
n jobs	-1
num leaves	31
objective	None
random state	42
reg alpha	0.0
reg lambda	0.0
silent	True
subsample	1.0
subsample for bin	200 000
subsample freq	0

Table 5.10 presents the corresponding confusion matrix for the optimised LightGBM algorithm. The subsequent performance metrics are reflected in Table 5.11. The classification values from the Confusion Matrix are TP = 3899, TN = 2827, FP= 454 and FN = 369.

Table 5.10: Confusion Matrix of LightGBM Classifier

True label	***Fake***	3899	369
	Not fake	454	2827
		Fake	***Not Fake***
		Predicted label	

The optimised LightGBM model improved its original balanced accuracy of 86.48% (Table 5.6) to 88.76%. It has a significant F1 score of 90.45% and Precision score of 94.35%. The other metrics are all greater than 86.00%.

Table 5.11: LightGBM Evaluation Metrics

Measure	Value
Balanced Accuracy	88.76%
Accuracy	89.10%
Precision	89.57%
Sensitivity	91.35%
Specificity	86.16%
F1 Score	90.45%
Matthews Correlation Coefficient	0.7777

5.3.2 Fake News Trend Analysis

The SAcovid19dataset was parsed through the refined LightGBM model and detected 262 508 (26.89%) tweets as containing Fake News (Fake) and 713 578 (73.11%) tweets as not being detected as fake (Not Fake). Figure 5.2 depicts a longitudinal trend of 'Fake' and 'Not Fake' news content detected from the collected South African COVID-19 dataset.

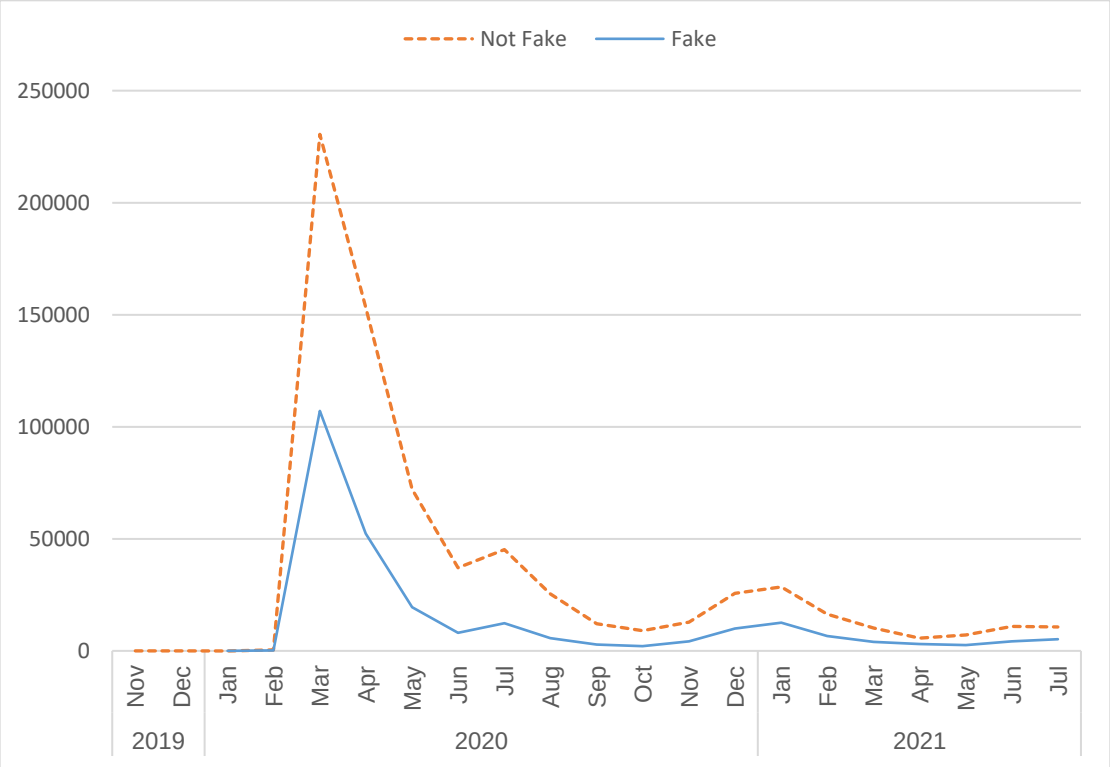

Figure 5.2: 'Fake' and 'Not Fake' News Frequency Trends of the SAcovid19dataset

5.3.3 The leading Fake News tweeters

Table 5.12 deals with the Top 10 users detected with Fake News content together with their average sentiment score, likeCount, number of hashtags, number of URL and retweetCount.

Table 5.12: The Top 10 Prolific Tweeters Detected with Fake News

Username	Avg. Sentiment Score	Avg. likeCount	Avg. No. of Hashtags (#)	Avg. No. of URLs	Avg. retweet Count	No. of detected Fake Tweets
CapricornFMNews	0.03	0.79	1.16	0.02	0.30	**1704**
JacaNews	0.04	1.77	1.34	0.06	1.30	**1213**
ewnreporter	0.01	25.17	1.19	0.04	18.29	**1174**
POWER987News	0.02	3.38	1.42	0.05	2.61	**970**
Bhekisisa_MG	0.05	2.57	2.01	1.06	1.90	**429**
CapeTownFreeway	0.66	0.36	3.39	0	0.04	**80**
Powerfm987	-0.02	8.14	1.87	0.73	3.92	**85**
SANDF_ZA	0.16	9.13	4.12	0.92	2.40	**52**
Jobfundi_com	0	0.01	1.07	1.99	0.01	**94**
weworx	0.32	0.04	3.57	1	0	**79**

The components of Table 5.12 are filtered for Fake News tweets only and are therefore not reflective of the overall tweets within the SAcovid19dataset. They are described as follows:

- **Avg. Sentiment Score**: Refers to the average sentiment value of Fake News tweets attributed to each user.
- **Avg. likeCount**: Refers to the average number of likes received by Fake News tweets for a user.
- **Avg. No. of Hashtags**: Shows the average number of hashtags (#) used by a user in their Fake News tweets.
- **Avg. No. of URLs**: Indicates the average number of URLs utilized by a user in their Fake News tweets.
- **Avg. retweetCount**: Shows the average number of times Fake News tweets from a user were retweeted.
- **Number of detected Fake Tweets**: The cumulative total of Fake News tweets published by a user.

The first three users, CapricornFMNews, JacaNews and ewnreporter, each have more than 1 000 Fake News tweets, with CapricornFMNews leading at 1 704 tweets. SANDF_ZA had the highest number of hashtags per tweet, averaging 4.12 hashtags. Ewnreporter had the highest average retweet count (18.29) and like count (25.17). Jobfundi_com had the highest average number of URLs per tweet at 1.99. The average sentiment score for the leading Fake News tweeters was generally neutral, with the exception of CapeTownFreeway and weworx, who exhibited positive average sentiment scores of 0.66 and 0.32, respectively.

Pertinent observations are that news agencies dominated the ranking for the most number of tweets. This dataset did not contain voluminous views of international anti-vaxxers or influencers.

5.4 Findings of Research Question 2

Research Question 2 asked:

> **Were there significant changes in the average sentiment from the South African COVID-19-related Twitter data?**

This was analysed firstly in section 5.4.1, which visually depicts the sentiment trends together with its related volume across the dataset timeline. Thereafter, in section 5.4.2, CPA was used to analyse sentiment trend changes.

5.4.1 The Tweet Frequency and Sentiment Trend

The sentiment analysis time series plot, Figure 5.3, illustrates the distribution of sentiments among tweets related to COVID-19 in South Africa from November 2019 to July 2021. The plot utilizes dual axes to juxtapose the volume and sentiment proportion of tweets on a monthly basis. The left vertical axis quantifies the total number of tweets for each month, while the right vertical axis indicates the sentiment proportions expressed as percentages.

For instance, in April 2020, the plot shows a high volume of tweets, which aligns with the number reported in Table 5.2 (205 618 tweets). The sentiment breakdown for this month reveals a stark contrast: 37.25% of tweets exhibit positive sentiment, 23.14% display negative sentiment and 39.61% are tweets categorized as neutral. This indicates that in April 2020, discussions around COVID-19 on Twitter were

predominantly neutral, a sentiment trend that could reflect public response to the pandemic's challenges at that time. The consistency in sentiment distribution ensures that at any point, the sum of positive, negative and neutral sentiments equals 100%, providing a complete emotional overview of the Twitter discourse for that month.

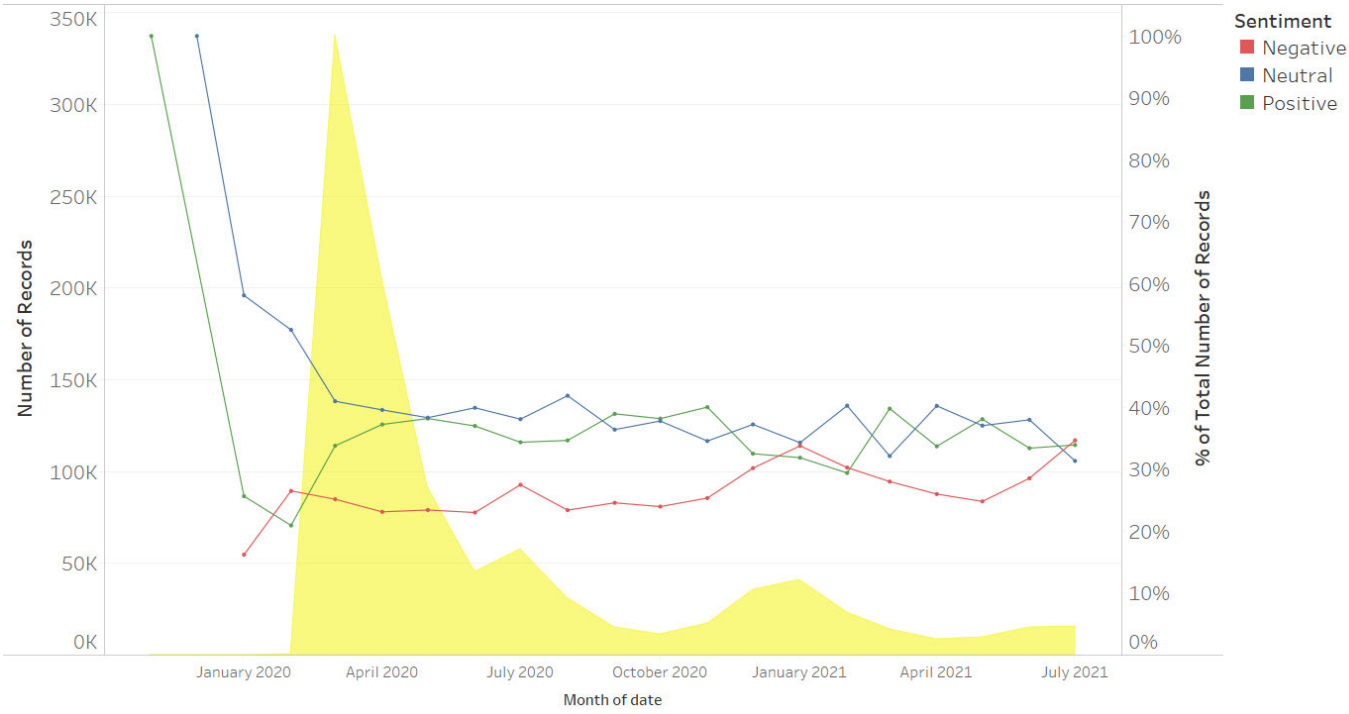

Figure 5.3: Sentiment Analysis Time Series Plot (by month) November 2019 to July 2021

Table 5.13 represents the sentiment ratio trend data which reveals significant variations in public opinion over time. In total, there were 248 601 negative, 383 072 neutral and 344 413 positive tweets. The onset of the COVID-19 pandemic saw a dramatic increase in tweet volume, peaking in March 2020 with 337,571 tweets. The sentiment distribution in March 2020 showed 25% negative, 41% neutral and 34% positive tweets. This initial spike likely corresponds to the rising awareness and concern about the pandemic. Following the initial peak, tweet volume remained high, with April 2020 recording 205 618 tweets. The sentiment remained relatively stable, with a slight increase in negative tweets (28% in July 2020). This period corresponds with the initial lockdowns and widespread public health measures.

Table 5.13: Sentiment Trend Percentages by Tweet Volume

	Sentiment			
Month-year	**Negative**	**Neutral**	**Positive**	**Total**
Nov-19			100%	2
Dec-19		100%		1
Jan-20	16%	58%	26%	74
Feb-20	27%	53%	21%	592
Mar-20	25%	41%	34%	337571
Apr-20	23%	40%	37%	205618
May-20	23%	38%	38%	91806
Jun-20	23%	40%	37%	45055
Jul-20	28%	38%	34%	57518
Aug-20	23%	42%	35%	31083
Sep-20	25%	36%	39%	14877
Oct-20	24%	38%	38%	11221
Nov-20	25%	35%	40%	17123
Dec-20	30%	37%	33%	35607
Jan-21	34%	34%	32%	41144
Feb-21	30%	40%	29%	23018
Mar-21	28%	32%	40%	14213
Apr-21	26%	40%	34%	8701
May-21	25%	37%	38%	9780
Jun-21	29%	38%	33%	15250
Jul-21	35%	31%	34%	15832

Tweet volume gradually decreased from 31 083 in August 2020 to 35 607 in December 2020. The sentiment distribution trend suggests a normalization in public discourse about COVID-19 as the initial shock waned. From February 2021 to July 2021, tweet volumes continued to decline, with notable fluctuations in sentiment. For instance, negative sentiment peaked at 35% in July 2021. Positive sentiment remained relatively stable around 34–38%, indicating ongoing engagement and concern.

The highest tweet volumes were recorded in March and April 2020, reflecting the immediate public reaction to the global spread of COVID-19. This period also marked the highest diversity in sentiment, with significant proportions of negative, neutral and positive tweets. After the initial peak, there was a gradual decline in tweet volumes, stabilizing between 10 000 and 50 000 tweets per month. The drop in volume

suggests a reduction in novelty and perhaps a fatigue in public discourse around the pandemic. In early 2021, there were renewed spikes in tweet volumes, particularly in January 2021 (41 144 tweets), coinciding with new COVID-19 waves and vaccine rollouts. This period saw a corresponding increase in negative sentiments, likely driven by concerns about vaccine efficacy and new virus variants.

A detailed monthly analysis shows that March 2020 marked the highest volume with 337 571 tweets, with sentiment fairly balanced but skewed towards neutral and positive. In April 2020, there was a slight drop to 205 618 tweets, maintaining a similar sentiment distribution. By July 2020, negative sentiment increased to 28%, with tweet volume at 57 518. January 2021 saw another peak in tweet volume (41 144), with 34% negative sentiment, reflecting renewed public concern. By July 2021, tweet volume decreased to 15 832, with the highest negative sentiment (35%) in the recorded period.

5.4.2 CPA Analysis of Change in Sentiment Trends

The hypotheses were as follows:

$$H_0 = There\ were\ no\ changes\ in\ the\ overall\ average\ sentiment$$

$$H_1 = There\ were\ change(s)\ in\ the\ overall\ average\ sentiment$$

Figure 5.4 depicts the average sentiment score trend from 24 January 2020 until 19 July 2021 from the SAcovid19dataset.

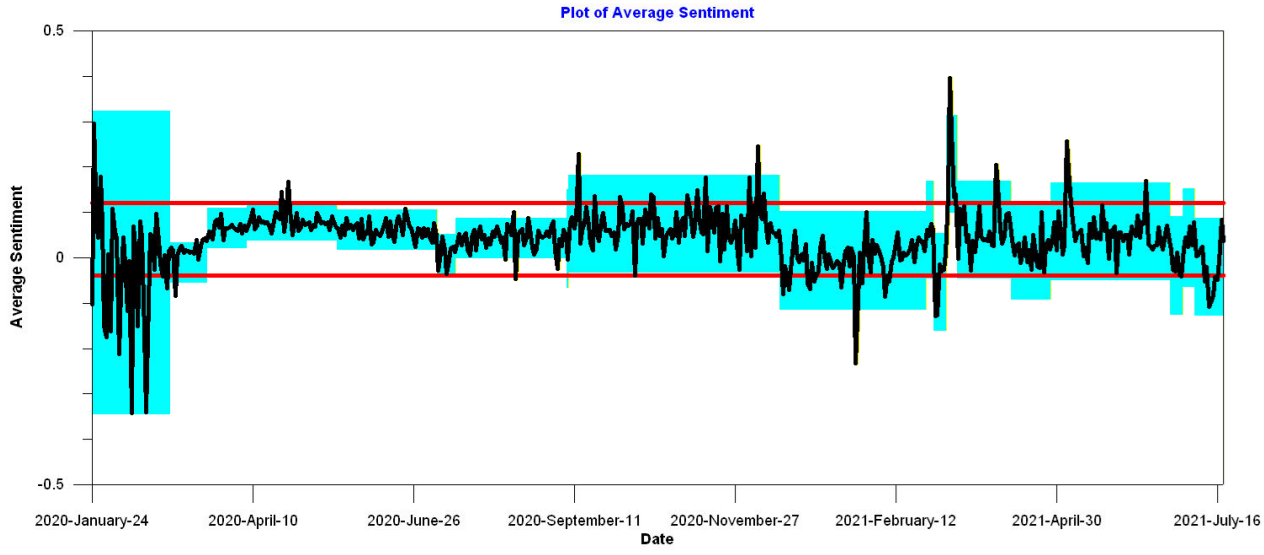

Figure 5.4: Average Sentiment Trend of SAcovid19dataset

Table 5.14 presents the significant changes during the 543-day period after applying CPA. A total of 16 significant changes were detected.

Table 5.14: Significant Changes in Average Sentiment

(Confidence Level for Candidate changes= 50%, Confidence Level = 95%, CI=95%, Bootstrap = 10 000, Without Replacement, MSE Estimates, Analyse Ranks)

No	Date	Confidence Interval	Conf.	From	To
1	15/03/20	(06/03/20, 16/03/20)	100%	-0.01088	0.064728
2	03/04/20	(22/03/20, 14/04/20)	96%	0.064728	0.080865
3	16/05/20	(11/05/20, 31/05/20)	100%	0.080865	0.06121
4	03/07/20	(01/07/20, 04/07/20)	100%	0.06121	0.007196
5	12/07/20	(10/07/20, 29/07/20)	99%	0.007196	0.041974
6	04/09/20	(25/08/20, 27/09/20)	100%	0.041974	0.073741
7	14/12/20	(10/12/20, 16/12/20)	100%	0.073741	-0.00666
8	22/02/21	(19/02/21, 22/02/21)	99%	-0.00666	0.061141
9	26/02/21	(26/02/21, 26/02/21)	95%	0.061141	-0.05339
10	04/03/21	(04/03/21, 04/03/21)	99%	-0.05339	0.20696
11	09/03/21	(09/03/21, 20/03/21)	99%	0.20696	0.061777
12	04/04/21	(22/03/21, 09/04/21)	99%	0.061777	0.015766
13	23/04/21	(18/04/21, 08/05/21)	100%	0.015766	0.057333
14	19/06/21	(11/06/21, 19/06/21)	96%	0.057333	-0.01773
15	25/06/21	(25/06/21, 29/06/21)	99%	-0.01773	0.043633
16	01/07/21	(27/06/21, 06/07/21)	100%	0.043633	-0.02061

The acceptance of the alternative hypothesis H_1 at a 95% confidence level indicates that significant changes in the average sentiment within tweets indeed occurred throughout the 543-day period under investigation. This conclusion is supported by the detection of seven distinct change points, each with a confidence level of 100%, underscoring the robustness of these findings.

Khan, Thakur, Obiyemi and Adetiba (2022a) supports the premise that shifts in online sentiment can be directly linked to real-world events. In this context, Change Point 1 (15 March 2020) aligns with a major event in South Africa—the announcement of a National State of Disaster by President Cyril Ramaphosa and the subsequent measures introduced to limit public gatherings to 100 individuals in an effort to control the spread of COVID-19 (South Africa, 2020c). This correlation between an impactful political announcement and the observed sentiment change on Twitter substantiates the argument that significant real-world developments have discernible reflections in online social sentiment.

The change points with their nearest perceived real-world events are presented below:

Change Point 1: Date: 15/03/20, Confidence Interval Range (06/03/20, 16/03/20)

President Ramaphosa declared a national state of disaster due to the COVID-19 pandemic, introducing the first set of restrictions to curb the spread of the virus (South Africa, 2020c). The announcement was made on 15 March 2020, which is exactly on the change point date and within the confidence interval range. Sentiment polarity scores increased from -0.01088 to 0.064728, indicating a shift from neutral to positive.

Change Point 2: Date: 03/04/20, Confidence Interval Range (22/03/20, 14/04/20)

President Ramaphosa announced the extension of the 21-day lockdown by an additional two weeks, pushing the end date to 30 April 2020 (South Africa, 2020d). The address was made on 9 April 2020, which is within the confidence interval range. Sentiment polarity scores increased positively going from 0.064728 to 0.080865.

Change Point 3: Date: 16/05/20, Confidence Interval Range (11/05/20, 31/05/20)

President Ramaphosa provided an update about preparations for easing restrictions with plans to move certain areas to Alert Level 3 (South Africa, 2020e). The address was made on 13 May 2020, which is three days prior to the change point date and within the confidence interval range. Sentiment polarity scores decreased from 0.080865 to 0.06121 but still remained positive.

Change Point 4: Date: 03/07/20, Confidence Interval Range (01/07/20, 04/07/20)

President Ramaphosa interacted virtually to communities in attempts to address some of the COVID-19 challenges (South Africa, 2020f). The address was delivered on 1 July 2020, which is slightly outside the change point date but within the confidence interval range. Sentiment polarity scores decreased from 0.06121 to 0.007196.

Change Point 5: Date: 12/07/20, Confidence Interval Range (10/07/20, 29/07/20)

President Ramaphosa reinstated a ban on alcohol sales and introduced a curfew as part of heightened restrictions (South Africa, 2020g). The announcement was made on 12 July 2020, the same the change point date. Sentiment polarity scores increased positively from 0.007196 to 0.041974.

Change Point 6: Date: 04/09/20, Confidence Interval Range (25/08/20, 27/09/20)

President Ramaphosa announced the easing of restrictions, moving South Africa to Alert Level 1. This included reopening borders and relaxing restrictions on gatherings (South Africa, 2020h). The announcement was made on 16 September 2020, within the confidence interval range and sentiment polarity scores increased from 0.041974 to 0.073741, indicating heightened positivity.

Change Point 7: Date: 14/12/20, Confidence Interval Range (10/12/20, 16/12/20)

President Ramaphosa announced stricter lockdown measures in response to the second wave of COVID-19, reintroducing an alcohol ban and curfews (South Africa, 2020i). The announcement was made on 14 December 2020, exactly on the change point date and within the confidence interval range. Sentiment polarity scores decreased from 0.073741 to -0.00666, showing neutrality.

Change Point 8: Date: 22/02/21, Confidence Interval Range (19/02/21, 22/02/21)

The closest announcement was President Ramaphosa on the rollout of Johnson & Johnson vaccine beginning on 17 February 2021 (South Africa, 2020j). This preceded the exact change point date and falls just outside the confidence interval range. Sentiment polarity scores moved from neutral (-0.00666) to positive (0.061141).

Change Point 9: Date: 26/02/21, Confidence Interval Range (26/02/21, 26/02/21)

The closest incident was Minister Tito Mboweni's state of the nation address on 24 February 2021, which slightly precedes the change point date (South Africa, 2020k). He discusses the increasing of fuel levies and other tax adjustment policies which may be negatively perceived. Sentiment polarity scores decreased from positive (0.061141) to negative (-0.05339).

Change Point 10: Date: 04/03/21, Confidence Interval Range (04/03/21, 04/03/21)

The notable passing away of journalist, Karima Brown attributed to COVID-19 occurred on the 4 March 2021 (South Africa, 2020l). The media release coincides with the change point date. Sentiment polarity scores increased from negative (-0.05339) to positive (0.20696).

Change Point 11: Date: 09/03/21, Confidence Interval Range (09/03/21, 09/03/21)

News of the 7% shrinkage of South African's economy were circulated on the 9 March 2021 (Businesstech, 2021). This is exactly on the change point date and within the confidence interval range. Sentiment polarity scores decreased slightly from 0.20696 to 0.06177, though remaining positive.

Change Point 12: Date: 04/04/21, Confidence Interval Range (22/03/21, 09/04/21)

President Ramaphosa introduced temporary restrictions over the Easter weekend to prevent the spread of COVID-19 (South Africa, 2021m). The announcement was made on 30 March 2021, which is within the confidence interval range. Sentiment polarity scores decreased from positive (0.06177) to neutral (0.015766).

Change Point 13: Date: 23/04/21, Confidence Interval Range (18/04/21, 08/05/21)

President Ramaphosa delivered an address on Freedom Day and mentioned the challenges of COVID-19 faced by the many concerning job losses, financial constraints and struggling businesses (South Africa, 2021n). This is the closest notable incident, which was made on 27 April 2021, four days post the change point date and within the confidence interval range. Sentiment polarity scores moved from 0.015766 to 0.057333, showing a positive shift.

Change Point 14: Date: 19/06/21, Confidence Interval Range (11/06/21, 26/06/21)

President Ramaphosa addressed the nation about the rise of COVID-19 cases due to the third wave and announced the return to Alert Level 3, enforcing stricter lockdown measures (South Africa, 2021o). The announcement was made on 15 June 2021, which is within the confidence interval range but prior to the change point date. Sentiment polarity scores decreased from 0.057333 to -0.017173, indicating a shift to neutrality.

Change Point 15: Date: 25/06/21, Confidence Interval Range (25/06/21, 29/06/21)

President Ramaphosa announced the shift to Alert Level 4 in response to the third wave of COVID-19 cases, introducing more stringent restrictions (South Africa, 2021p). The announcement was made on 27 June 2021, which is within the

confidence interval range but slightly after the change point date. Sentiment polarity scores increased slightly from -0.017173 to 0.043633.

Change Point 16: Date: 01/07/21, Confidence Interval Range (27/06/21, 06/07/21)

President Ramaphosa extended the Alert Level 4 restrictions as COVID-19 cases continued to rise during the third wave (South Africa, 2021q). The address was made on 11 July 2021, which is outside the confidence interval range, but close to the broader timeframe. Sentiment polarity scores shifted from 0.043633 to -0.02061, marking a significant decrease.

5.5 Findings of Research Question 3

Research Question 3 asked:

> **Were social bots leveraged to produce content around South African-COVID-19 Twitter content?**

Applying Method 1 from section 4.5.7.6 yielded 619 unique users that were detected as bots or cyborgs. Table 5.15 reveals the Top 5 corresponding users' volume of tweets breaching this detection method.

Table 5.15: Top 5 Social Bot Users Tweet Volume Breaching Method 1

Username	Tweet Volume
CapricornFMNews	998
GCISMedia	107
Bhekisisa_MG	103
ewnreporter	81
spiceofi	65

Method 2 returned 16 users while Method 3 yielded 115 potential users. This suggests that there were indeed elements of social bot activity surrounding the posting of South African COVID-19-related content.

5.6 Conclusion: Main Findings

In summary, the study has three main findings.

5.6.1 Fake News were Detected in South African COVID-19-related Tweets

The main finding from Chapter 5 revealed that data science, particularly through the leverage of ML, can be leveraged for COVID-19 Fake News detection and applied on South African-related Twitter data. This is significant to mitigate the spread of harmful information as well as monitor its dissemination across significantly large amount of data which otherwise cannot be practically and effectively be investigated by human moderators.

5.6.2 There were Fluctuations in Sentiment

Another main finding revealed changes in sentiment trends on Twitter during the 543-day study period coincided with at least one known real-world event. This highlights the influential link between social media conversations and real-world events, making sentiment analysis an effective research tool to gauge moods across large online conversation for consumer or civil insights.

5.6.3 The News Media did Amplify Elements of Fake News

The study revealed a notable impact of news media in the online discourse surrounding COVID-19, highlighting that user accounts associated with news outlets were among the most active in disseminating information throughout the period analysed. Media accounts were identified as prominent sources of tweets classified as both non-Fake and Fake News, underscoring their substantial role in shaping the information landscape during the pandemic.

In addition, the sentiment scores associated with these media outlets indicated that their average sentiment was approximately zero. This neutral average suggests a balanced distribution of positive, negative and neutral content in their tweets, implying an unbiased stance in the presentation of news. Such a finding is particularly salient for researchers and stakeholders concerned with media bias and agenda-setting. It suggests that, at least in terms of sentiment expression within their tweets, these media outlets did not exhibit a systematic bias toward any particular sentiment, potentially countering accusations of partisanship in their COVID-19 coverage. This aspect of the study provides a data-driven perspective on the media's role in information dissemination during critical events, which could be invaluable for understanding media practices and informing media literacy initiatives.

5.6.4 Bots were Deployed

A significant finding within the ambit of this research was the identification of social bots in the dataset designated as SAcovid19dataset. Notwithstanding their presence, the quantitative analysis revealed that social bots accounted for a relatively insubstantial proportion of the aggregate tweet volume. Nonetheless, these entities were prominent, with accounts typified as social bots ranking as the foremost contributors in terms of the number of tweets disseminated.

In summary, this chapter has meticulously chronicled the rigorous process of training, validating and appraising a suite of ML models. A model was judiciously selected, predicated upon its expedience and precision and subsequently applied to the SAcovid19dataset for the binary classification of content into *Fake* and *Not Fake* categories. The classified dataset was then subjected to a comprehensive array of analytic techniques—including sentiment analysis, descriptive statistics, CPA and bot detection methodologies. The results derived from these analytic modalities were systematically delineated, ensuring a methodical response to each posed research query while underscoring the noteworthy discoveries of the study. In furnishing the requisite empirical evidence, this chapter sets the stage for cogent inferential conclusions, proffers a scaffold for future investigative endeavours and articulates recommendations that bear relevance for scholars and stakeholders in related domains. The discourse on these findings will be expanded upon in Chapter 6.

Chapter 6: Discussion of the Findings

6.1 Introduction

Chapter 5 systematically addressed the research questions posed. Chapter 6 commences by examining descriptive analysis of the dataset. This is followed by an exploration of the results pertinent to Research Question 1 (section 6.3), which is then succeeded by the findings for Research Question 2 (section 6.4) and subsequently, the insights related to Research Question 3 (section 6.5), providing a nuanced interpretation of the data as well as broader ramifications within the research context.

6.2 Discussion on the Descriptive Analysis of the Data

Section 5.2 presented the descriptive analysis of the SAcovid19dataset. This section discusses the linguistic analysis (section 6.2.1), tweet distribution (section 6.2.2), tweet volume (section 6.2.3), daily tweet distribution (section 6.2.4), tweets with multiple hashtags (section 6.2.5) and number of twitter users (section 6.2.6).

6.2.1 Linguistic Analysis

This study's collection of 976 086 tweets (SAcovid19dataset) makes an important contribution to the non-medical discourse on COVID-19, filling a gap highlighted by Shuja *et al.* (2021), who surveyed various open datasets, primarily with medical relevance. This SAcovid19dataset is available as an open data set in its pre-processed (raw), cleaned and processed forms.

The analysis of the SAcovid19dataset (Table 5.1) that revealed the majority of tweets were in the English language. The category labelled 'Undefined' emerged as the second most prevalent because it included tweets employing non-standard language, such as South African vernaculars, colloquialisms, or a mix of characters and emojis that challenge the algorithm's classification capabilities, as suggested by Khan (2019). The presence of Indonesian and Dutch languages, while seemingly anomalous, is rationalized by South Africa's diverse linguistic landscape, with the prominence of Dutch reflecting a similar language, Afrikaans, which is a widely spoken language in the region.

The nuances of language detection underscore the necessity for continuous improvement of algorithms to enhance their precision in identifying and interpreting the myriad forms of language expression as they evolve. The improvement is noted and flagged for future work as it is beyond the scope of this study.

6.2.2 Tweet Distribution

Statistical investigation of the SAcovid19dataset discerned a lognormal distribution, (section 5.2.3). This finding resonates with the research by Bild *et al.* (2015), who posited that distributions of tweets, during Twitter campaigns, adhere to a lognormal pattern over specific sample intervals. The alignment of the SAcovid19dataset with this expected lognormal distribution intimates that the dataset's generation was not entirely the result of bot automation, which strengthen the dataset's validity. The congruence of the distribution in this study with those reported by Bild *et al.* (2015) corroborates the generalisation of this lognormal behaviour across different Twitter datasets and campaign contexts, reinforces the credibility of the data in this research.

6.2.3 Tweet Volume

In Table 5.2, the dataset presents negligible Twitter activity in November and December 2019, with only three tweets in total, suggesting COVID-19 had not yet become a topic of widespread local conversation, and its discourse was in its nascent stage. There is a significant increase in tweet volume beginning in March 2020, which coincides with the global spread of COVID-19 and the WHO's pandemic declaration on March 11, 2020. This seems the most appropriate likelihood given the National State of Disaster announced by the South African president on 15 March 2020 (South Africa, 2020c). The high volume in March and April 2020 therefore likely reflects the public's growing concern and the infodemic commencement.

Since April 2020, there is a noticeable decline in monthly tweet volumes, even though numbers remain significantly higher than pre-March levels. This could indicate a stabilization of public discourse as people adjusted to the pandemic or a shift from the platform's use to other communication channels. It could also indicate 'pandemic fatigue' or information flow updates comforting users. The tweet volumes fluctuate over the ensuing months of the study but with a general declining trend continuing into 2021. This too is typical infodemic behaviour (Bild *et al.*, 2015).

Lower volumes may also be attributed to infrequent delays between significant state changing government announcements. Furthermore, international COVID-19 conversations excluding South African context as per keywords identified in section 4.5.2.1 may have contributed to this observation.

6.2.4 Daily Tweet Distribution

The analysis of tweet frequencies per weekday of the SAcovid19dataset revealed a distinct pattern where Monday registered the highest number of tweets, while Saturday had the fewest. This trend could be linked to the timing of COVID-19 announcements in South Africa, typically scheduled for Sunday evenings. The aftermath of these announcements saw a surge in Twitter activity on Mondays as users engaged and reacted to the information disseminated by the state. It was customary for the State President to deliver COVID-19-related addresses on Sunday nights, a practice that became affectionately known as “Family meetings”. These addresses, characterized by President Ramaphosa’s earnest attempts to unify the nation and foster a spirit of solidarity amidst the crisis, may have significantly influenced public discourse on Twitter, particularly at the start of the week. Nkoala and Dlanga (2023) note the impact of these communications in their efforts to advance nation-building during a challenging period, suggesting a direct correlation between these national broadcasts and the observed Twitter activity.

6.2.5 Tweets with Multiple Hashtags

The leading hashtag in the SAcovid19dataset was #covidsa, followed by #lockdownsa, with the remaining top associated hashtags centred on COVID-19 (Table 5.4). There were no dominant pairing hashtags unrelated to COVID-19 attempting to ‘ambush’ or divert or highlight different topics. This suggests the data is largely COVID-19 specific and not, for example, vaccine specific. This also suggests that the severity of COVID-19 overwhelmed or dissuaded perceived ambush marketers.

The absence of vaccine-related hashtags suggests that the keywords selected for data extraction (section 5.5.2.2.1) did not primarily capture vaccine-centric discussions, pointing to a broader range of COVID-19-related conversations beyond vaccination themes. This is flagged for future work.

6.2.6 Number of Twitter Users

The tweet distribution was widespread across many users, with one standout among the Top 10 tweeters (Table 5.5) comprising 7 980 tweets. This represents less than 1% of the SAcovid19dataset tweets and given the large volume of 229 968 unique users, it suggests that no small number of users significantly dominated the content, alluding to the universality of the conversation.

6.3 Discussion of Research Question 1

This section discusses the collection of South African COVID-19-related Twitter data (section 6.3.1), ML Fake News datasets (6.3.2), performance of Fake News ML detection models (section 6.3.3), the Fake News trend (section 6.3.4) and the top Fake News tweeters (section 6.3.5).

6.3.1 Automated Approach to Collect South African COVID-19-related Twitter Data

Because of limitations imposed on Twitter API's, data retrieval is restrictive and potentially very costly if purchases are made directly to Twitter for the data (Stokel-Walker, 2023b). The costs deterred the researcher from making purchases given budget constraints and a desire to leverage open-source tools to allow research replication. This led to the utilization of SNScrape to automate the data collection, which took several hours to complete.

The resultant SACovid19dataset is a significant contribution to this study because it provides approximately 1 million tweets as an open-source resource to assist researchers with analysing South African COVID-19 content.

6.3.2 Dataset to use for Fake News Detection

In the selection of ML suitable datasets (section 4.5.2.1), deliberate emphasis was placed on sources with academic credibility and platform integrity, rendering these datasets particularly suitable for ML model development given their pre-categorized nature. This strategic compilation resulted in a significant aggregation of ML-compatible datasets, yielding 30 193 labelled texts. This collection serves as a foundational element for the enhancement of ML model performance, both in terms of accuracy and computational efficiency.

Strydom and Grobler (2023) utilized 10 000 labelled tweets to evaluate the efficacy of transformer-based neural network classifiers for COVID-19 misinformation detection on Twitter within a South African context. They emphasized the model's insufficient generalization to the South African socio-linguistic context and underscored the importance of extensive, labelled datasets in augmenting the learning depth and overall performance of these models.

In contrast, this study presents a significantly larger labelled dataset (C19MLdataset), three times the size of Strydom and Grobler's (2023), addressing some of the limitations. The C19MLdataset also surpasses the 21 379 dataset utilized by Tashtoush *et al.* (2022) to analyse COVID-19 Fake News. This expanded dataset provides a valuable resource for improving COVID-19 misinformation detection models by increasing the training depth, thus representing a significant contribution of this research. While this study advances the field, it aligns with Strydom and Grobler's (2023) recommendation that the creation of meticulously curated, context-specific misinformation datasets remains essential for refining the detection capabilities of ML models within the South African context.

6.3.3 ML Performance Metrics and Model Selection

6.3.3.1 Shallow ML Models

The Top 10 performing models from Table 5.6 were dominated by Ensemble learning, Gradient boosting, SVM and Linear Models. This suggests that binary classification models requiring training of textual data should use the ensemble learning, gradient boosting or linear models because of its superior training and performing times over SVM. This finding is supported by Alzamzami, Hoda and El Saddik (2020).

6.3.3.2 Deep Learning Models

Bi-LSTM outperformed LSTM, GRU, CNN and RNN in performance metrics, with all except CNN sharing the same DL architecture (Table 5.7). This ranking indicates an evolutionary improvement in text classification models, consistent with the findings of Jang *et al.* (2020). Further, CNN's performance highlights its versatility as a fast DL algorithm for text classification, despite originally being designed for image processing (section 3.5.2). The DL and transformer models were developed using

NVIDIA's Tesla T4 GPU on Google Colab[14], though they were not as fast as some high-performing shallow ML models (Table 5.6), which exhibited comparable performance metrics but did not require GPU support.

6.3.3.3 Transformer Models

It is unsurprising that the transformer models performed the best (Table 5.8) as they are considered the state-of-the-art on NLP tasks. However, their large size and complexity demand substantial computational resources and extended training times, which can hinder adoption[15]. Transformer models perform well with high-end GPUs and parallel processing with service providers offering these at a cost (The Chief I/O, n.d.). Regrettably, such GPUs were not physically available to the researcher and accessing them through third-party services exceeded the project budget. This limitation is noted for further work.

6.3.3.4 Model Selection

This study required the processing of approximately 1 million textual content, which meant that models with significant speed advantages and good performance metrics were the priority. Hence, the LightGBM was the preferred choice (Table 5.6). Following optimization, it produced an enhanced balanced accuracy score of 88.76%, ranking it among the best performing transformer models (5.8) and was subsequently leveraged as the COVID-19 Fake News detection model.

6.3.4 Fake News Trend

The selected LightGBM model applied on the SAcovid19dataset detected 26.89% of the 976 086 tweets to be 'Fake' implying that at least one in four tweets probably contained Fake News content related to COVID-19 within the dataset. It must be

[14] https://colab.research.google.com/

[15] Variability in performance metrics may occur in transformer and DL models, even with identical code, because of factors such as stochastic training processes, hardware variability, hyperparameter sensitivity, floating-point precision differences and inconsistencies in data processing.

noted that these detections should be considered as alerts to be investigated further by human moderators for clarity.

Figure 5.2 appears to depict a positive proportional relationship between the volume of tweets and Fake News content. This is a logical outcome, given the substantial communication about COVID-19, suggesting that an increase in tweet volume would correspond with a rise in Fake News. This again alludes to the data validity.

6.3.5 Leading Fake News Tweeter

At least six of the Top 10 tweeters that contained Fake News were news outlets, which is superficially counterintuitive. This, however, is not unusual because news outlets use sensational headlines for content to lure the user into redirecting them to their main news websites. This is called ‘Click Bait’.

Perhaps this technique was leveraged to generate engagement or utilized as means to mitigate Fake News, since users that read the original article would be informed of the context concerning the posted content. However well this is intended, it inadvertently amplifies Fake News content across social media and may not be beneficial to users who do not investigate the posts further. It can be further suggested that the average sentiment across these top Fake News tweeters were neutral and not intended to purposely influence a desired sentiment of content. It is noteworthy that not all Fake News is emotionally charged, as some misinformation can be presented in a neutral or factual-sounding manner, with bots potentially generating such content to enhance credibility.

Regardless of the narrative proposed, this form of click baiting should be discouraged as unethical journalistic practice (Chen, Conroy and Rubin, 2015). Future work is to determine measures to discourage influences and general users to resist from amplifying such messages.

6.4 Discussion of Research Question 2

6.4.1 Sentiment Trends

The overarching neutral sentiment trend (Figure 5.3) dominated the longitudinal period, which is understandable given the unprecedented nature of COVID-19 and its global impact. People primarily shared news, reflecting the uncertainty and

unfamiliarity surrounding the pandemic. Positive sentiment was minimal throughout 2020 and 2021, as many organizations endured economic and human resource challenges, while patients dealt with inadequate healthcare facilities, lack of accommodation and reports of widespread human losses. These factors contributed to a predominantly neutral sentiment during this period.

6.4.2 Changes in Overall Average Sentiment

The null hypothesis in section 4.5.8.2 was rejected because of 16 significant changes (Table 5.14) detected using CPA. For instance, Change point 1 was found to coincide with a significant real-world event, namely, the National State of Disaster announcement on 15 March 2020 (South Africa, 2020c). Change points 1, 2, 5, 6, 7, 12, 14, 15 and 16 are associated with adjustments to lockdown measures and restrictions, indicating that approximately 56% of the sentiment changes were related to lockdowns. The remaining change points were linked to public health announcements and the general discourse surrounding the COVID-19 pandemic.

These finding corroborates the research conducted by Khan, Thakur, Obiyemi and Adetiba (2022a), which elucidated a pronounced correlation between the sentiments expressed on Twitter and significant societal events, inclusive of declarations by governmental and academic authorities. Although this study did not explore the magnitude or mathematical correlation of these relationships, the findings suggest opportunities for future research on the reciprocal influence between social media sentiment and real-world events, including the potential for sentiment analysis to forecast outcomes or assess event impacts on public sentiment.

Not every real-world event or governmental announcement will necessarily trigger a change in sentiment trends. However, certain key events often align with significant changes in sentiment, either before or after the occurrence of notable events. These change points must be analysed as valuable information feedback. If sentiment shifts are related to Fake News, mitigation strategies should be employed. CPA should be deployed in any viral Twitter discourse to detect such shifts. This study signifies the subtle nature in analysing relationships between online sentiment and real-world events, highlighting CPA as important to South African infodemic research. As such, CPA is recommended for policy adoption by Health Departments.

6.5 Discussion of Research Question 3

A significant finding of the study was the involvement of social bots surrounding the posting of COVID-19 within a South African context. Indeed, the first and second highest tweeting accounts (Table 5.13) were detected as bots or cyborgs, using the three social bots detection methods (section 4.5.7.6). These methods were built to identify automated behaviour with respect to posting anomalies in terms frequency, volume, source and content. The identified social bots may have been deployed to amplify, manage or spread targeted content periodically and systematically. Determining whether they were indeed malicious or not is subjective and requires thorough investigation and different forms of analysis, which was beyond the scope of the study. This is recommended for future work.

It is noteworthy that Twitter's decision to restrict API access was a measure aimed at combating social bots and allowing only 'good bots' as part of a broader strategy to enhance the integrity and security of the Twitter platform (Stacey, 2023). By restricting API access, it becomes more challenging for these automated systems to operate at scale, potentially reducing the spread of false information and malicious activities on the platform. However, this has made research such as this one non-trivial in cost.

6.6 Conclusion

This chapter provided further insights into the findings through discussion and supplementary literature. The study effectively collected and analysed approximately 1 million South African COVID-19-related tweets, developing a robust Fake News detection model using a labelled subset of 30 193 texts. LightGBM was identified as the most feasible performing model, detecting 26.89% of tweets as potential Fake News. Media outlets were inadvertently among the top identified amplifiers of Fake News. Sentiment analysis revealed a predominantly neutral sentiment throughout 2020 and 2021, with 16 significant sentiment shifts detected, 56% of which were linked to lockdown measures using CPA. Social bot detection identified systematic amplification of content by automated accounts. These findings demonstrate the value of data science approaches such as ML, sentiment analysis and CPA in examining the COVID-19 infodemic and providing localised insights.

The next chapter concludes with a summary of the study and future work.

Chapter 7: Conclusion and Future Work

There is an urgent need for automated methods to process big data so as to address infodemic challenges such as Fake News on Twitter, particularly highlighted during the COVID-19 pandemic. While Twitter facilitates diverse COVID-19 discussions, it also inadvertently propagates misinformation. This study applied data science approaches to the South African context, focusing on Fake News detection, sentiment analysis and social bot identification. The analytic approaches and findings offer critical insights that can support and inform public health policies.

This chapter revisits the research objectives from Chapter 1, noting contributions and offers recommendations for future research.

7.1 Main Objective

> *To computationally analyse South African-related COVID-19 Twitter data for Fake News using data science.*

This study demonstrated that ML can be an effective approach when addressing infodemic challenges and may be extended to other issues such as service delivery protests and hate speech. Using the C19MLdataset, multiple shallow, DL and transformer models were evaluated, with LightGBM emerging as the most feasible because of its speed and accuracy in processing nearly 1 million tweets. The development of a COVID-19 Fake News detection model is a key contribution of this research.

Key Fake News Tweeters were further identified with their posting characteristics analysed to assess influence. This analysis is recommended for future use.

7.2 Secondary Objective 1

> *To examine the South African-related COVID-19 Twitter data for changes in average sentiment.*

The CPA detected 16 sentiment changes in the SAcovid19dataset, with 56% being associated with lockdown measures and restrictions, including the National State of Disaster address. Sentiment shifts can act as real-time red flags for events like violent protests or disasters, aiding in emergency response. This study's use of CPA to track

sentiment over 20 months is a novel contribution to COVID-19 infodemiology in South Africa, linking social media conversations to real-world events.

7.3 Secondary Objective 2

> *To determine the presence of social bots within the South African COVID-19-related Twitter data.*

Social bot detection, a key contribution of this study, revealed the presence of Twitter bots and cyborgs. The top two users exhibited automated behaviour typical of social bots. Since bots amplify their creators' causes, analysing their role is crucial to assessing the credibility of social media content. Any comprehensive Twitter analysis should include bot detection.

7.4 Future Work

This study, in its quest to answer the research questions, encountered further scope opportunities and research questions which had to be delimited. The following prospects are identified for scholarly inquiry:

7.4.1 Explore the ML Algorithms with Nuanced Computer Science

This study used balanced accuracy scores and execution times as the main performance metric selection criterion. Alternate performance metrics such as F1 score, Specificity and AUC may be prioritized by other researchers.

7.4.2 The Research should be Extended both in Longitudinal Time and Data Scope

To supplement the research, it is significant to expand the study period to encompass the entirety of COVID-19 lockdown periods for South Africa. Post-lockdown data should be harvested and analysed for COVID-19 fears, views and insights. Moreover, incorporating specific keywords could strengthen the ongoing and future aggregation of tweets related to COVID-19 in South Africa. Analysing data from varying time intervals exposes distinct temporal characteristics valuable for historical examination.

7.4.3 Analyse South African COVID-19-related Content on other Social Media Platforms

Research should not be limited to Twitter, as social media users use other platforms such as Facebook, Instagram, TikTok and Reddit. These platforms had significant roles in disseminating COVID-19-related content in South Africa. It is therefore important to investigate the impact and vulnerability of these platforms on the spread of COVID-19 misinformation. The research on whether all these platforms were equally susceptible to the dissemination of false information about COVID-19 is an area that warrants additional study.

7.4.4 Analyse the Media and Metadata that Twitter Provides

The scope of detecting Fake News on Twitter could be broadened to include analysis of COVID-19-related content embedded in videos, audio and images. For instance, visuals or videos crafted with misleading information may evade existing detection algorithms on social media platforms, consequently spreading harmful misinformation with possibly severe consequences. Examining different aspects of metadata could uncover specific or common patterns in Twitter usage and its communities, providing opportunities for investigation.

7.4.5 The SAcovid19dataset should be Reanalysed using Computational Demanding State-of-the-art ML Models

Advanced ML models, such as transformer models (BERT, RoBERTa) or recent innovations like Generative Pretrained Transformer (ChatGPT), should be utilized for analysing Fake News in the data and their results compared to this study.

While implementing these advanced models may prove expensive and demand substantial computational resources, funding research in this area is crucial for developing effective solutions to combat misinformation, particularly in scenarios like the COVID-19 infodemic.

7.4.6 Multi- and Inter-disciplinary Work is Needed

The researcher had the advantage of working with diverse datasets that necessitate varied analytical approaches, including contextual, sociological, political and psychological perspectives. It it essential to involve multidisciplinary experts for

further study. This need is particularly pressing given the recent changes by Twitter, now known as X, in its API access policies, which restrict free content extraction, posing financial challenges for open-source research related to Twitter.

7.4.7 Social Bots' Role should be Examined In-depth

Future research should prioritize the context and interpretation of social bots' intentions, as these robots are designed for diverse purposes, ranging from transparently stated objectives to subtle or sometimes concealed aims, with intentions that are alternatively benevolent or malevolent. An area of particular concern is the potential for social bots to leverage, amplify or generate Fake News to manipulate public sentiment, highlighting the need for thorough investigation into this aspect. Research should examine the intentions of the bot creators; intentions ranging from political through social to economic must be established.

7.4.8 Extend Study to 'Monitor' other Significant Online Infodemics

The research methodologies employed in this study are adaptable to other crises or major global events on social media and online platforms where widespread information dissemination occurs. This includes other healthcare issues and current or future national elections within South Africa and internationally. For South Africa, the listeriosis outbreak (Van der Vyver, 2018), the seasonal malaria, HIV and influenza epidemics (Davies *et al.*, 2019) have high applicability. The risk of social media being used by malevolent entities to influence elections (Millham and Thakur, 2016) or instigate regional conflicts (Khan, Thakur, Obiyemi and Adetiba, 2022b) is significant, particularly with the strategic deployment of social bot armies targeting specific groups or areas.

7.4.9 Extend Fake News Detection Languages and Update Existing Datasets

South Africa is a multilingual nation necessitating the creation of contextually labelled ML datasets that reflect this linguistic diversity. In particular, the ML datasets used in this study should undergo translation into other spoken South African languages such as Zulu and Xhosa while preserving their semantic integrity. Alternatively, original datasets should be curated in native languages. This would assist in the development of multilingual ML COVID-19 detection models that are more versatile and therefore warrants further investigation.

Further, existing datasets used in Fake News detection may become outdated as new information emerges, leading to misclassification and reduced model accuracy. Regular updates to classification labels are necessary to maintain relevance, especially in rapidly evolving contexts such as health crises or political events. This would require systematic monitoring and reclassification to ensure models remain effective. However, implementing dynamic labelling is resource-intensive and complex, making it an essential area for future research to explore automated strategies for updating datasets.

7.4.10 Real-time Analysis should be Explored

The exploration and development of a real-time, industry-standard system for detecting Fake News related to COVID-19 and other pertinent topics is crucial for enabling stakeholders such as governments, media organizations and the public to swiftly identify and mitigate misinformation, thereby preserving information integrity and public trust. These detection tools not only help curb the spread of false narratives and reduce potential social unrest but also provide researchers with valuable insights into the dynamics of misinformation. By analysing detection algorithms and patterns of Fake News distribution, researchers can enhance detection methods and contribute to advancements in AI and ML. Ultimately, real-time Fake News detection benefits both stakeholders and researchers by protecting public discourse, promoting societal well-being and enriching scientific understanding of misinformation.

7.5 Recommendations

The COVID-19 infodemic highlighted the challenges of managing misinformation in the digital age, affecting individuals, institutions and society at large. Based on the findings of this study, a series of targeted recommendations have been formulated to support policymakers, public health authorities, social media platforms, researchers and the general public in mitigating the adverse effects of misinformation during public health crises. The recommendations are as follows:

7.5.1 Expanding Infodemiology and Infoveillance

Research in the fields of infodemiology and infoveillance should be significantly expanded to enable impartial monitoring of public responses to health crises, such as pandemics and epidemics. The concept of neutrality is critical, as it underscores the

potential for information systems to adapt innovative solutions that observe public behaviour in a non-intrusive manner. By employing neutral and voluntary methods of observation, these studies will likely gain broader acceptance from society, enabling public health authorities to collect valuable insights without infringing on individual privacy.

7.5.2 Integration of Advanced Social Media Analytics in Educational and Government Institutions

It is recommended that educational institutions, in partnership with governmental bodies such as the Department of Higher Education, Universities of South Africa and the National Treasury, integrate advanced social media analytics into their existing frameworks. This integration should include both retrospective and real-time monitoring of misinformation (specifically Fake News) using the methodologies outlined in this research. This functionality will enhance the ability of institutions to track and mitigate the impact of misinformation across educational and governmental domains, fostering a more informed and resilient society.

7.5.3 Training and Capacity Building for Data Science and Social Media Literacy

To further strengthen the response to misinformation, data science researchers and ICT professionals should facilitate workshops aimed at equipping individuals with the necessary skills to analyse Fake News and other forms of misinformation. These workshops should cover data science techniques, such as computational modelling and ML, for identifying misleading content. Additionally, training should be offered to individuals across various sectors, including education, health, government and the private sector, to promote responsible social media usage and improve their capacity to engage effectively with digital platforms.

7.5.4 Policy Recommendations for Policymakers and Public Health Authorities

Policymakers and public health authorities should engage in deeper collaborations with social media platforms to co-develop robust content moderation strategies aimed at reducing the proliferation of COVID-19 misinformation. These partnerships should focus on creating balanced approaches that respect freedom of expression while ensuring that public health information remains accurate and reliable.

Governments and public health bodies should initiate extensive public awareness campaigns designed to educate citizens on how to identify, report and avoid Fake News related to COVID-19. These campaigns should emphasize the importance of using verified and trusted sources for pandemic-related information and encourage public engagement with factual content.

7.5.5 Recommendations for Social Media Platforms

Social media companies should prioritize the development and deployment of advanced artificial intelligence systems capable of detecting and flagging misleading or harmful content in real-time on their platforms. These AI-driven systems must be carefully calibrated to avoid overreach into legitimate discourse while still safeguarding public health. Further, these platforms should enhance transparency regarding their content moderation practices, providing users with clear information on how content is flagged, reviewed and removed. Moreover, platforms should develop and disseminate educational materials to help users navigate online information responsibly, particularly during times of public health crises.

7.5.6 Recommendations for Researchers

Researchers are encouraged to continue exploring the dynamics of the COVID-19 infodemic, focusing on the development of more accurate computational models for Fake News detection, sentiment analysis and the assessment of public perception during health crises. This research will contribute to a deeper understanding of how misinformation spreads and how it can be curtailed.

Researchers should foster interdisciplinary collaborations that bring together experts in data science, public health, psychology and communication studies. Such collaborations are crucial for addressing the multifaceted nature of the infodemic and for devising holistic strategies to mitigate its effects.

7.5.7 Recommendations for the General Public

The public should be encouraged to develop critical information literacy skills, which are essential for assessing the credibility of news sources and determining the veracity of content shared on social media platforms. Educational campaigns and initiatives focused on media literacy will play a key role in empowering individuals to make informed decisions.

It is essential that individuals prioritize engagement with credible and authoritative sources of information. By doing so, the public will be better equipped to navigate the complexities of the infodemic and make informed decisions based on accurate data and expert guidance.

7.6 Conclusion

This research underscores the importance of Fake News detection as a tool to tackle infodemics, particularly in the context of Twitter discussions about COVID-19 in South Africa. It highlights the advantages of using data science techniques over traditional manual labelling methods. Automated and cost-effective approaches like ML and sentiment analysis, as applied in this study, offer practical means to analyse the vast amounts of subjective and unstructured data typically found on social media. Additionally, the study reveals how longitudinal analysis of Fake News and sentiment on Twitter can yield valuable insights into historical trends and public opinion, which are beneficial for potential stakeholders.

The study also discovered that sentiment analysis, combined with change detection algorithms, could serve as a real-time tool for sensing shifts in public sentiment. A critical finding of this research is the identification of social bots in the South African COVID-19-related Twitter data, with two particularly influential ones identified based on their high volume of activity. The methods used in this study for developing a Fake News detection model, identifying social bots and analysing significant sentiment changes over time, provide a useful framework that can be adapted to other Twitter-related events.

In summary, this study has achieved its objectives, proven its significance and made a notable contribution to the academic field. It stands as a pioneering effort in South Africa, not only in creating a Fake News detection model and conducting sentiment analysis but also in identifying social bots within the dataset, thereby filling a research gap. The data science approach employed in analysing the COVID-19 infodemic on Twitter within the South African context demonstrates how quantitative methods can be effectively used to research content and user opinions on Twitter, offering insightful perspectives into the COVID-19 infodemic on this platform.

References

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M. and Shah, Z., 2020. Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study. *Journal of Medical Internet Research*, 22(4), p. e19016.

Abdelminaam, D.S., Ismail, F.H., Taha, M., Taha, A., Houssein, E.H. and Nabil, A., 2021. CoAID-DEEP: An optimized intelligent framework for automated detecting COVID-19 misleading information on Twitter. *IEEE Access*, 9, pp. 27840-27867.

Abdolrasol, M.G.M., Hussain, S.M.S., Ustun, T.S., Sarker, M.R., Hannan, M.A., Mohamed, R., Ali, J.A., Mekhilef, S. and Milad, A., 2021. Artificial neural networks based optimization techniques: A review. *Electronics*, 10(21), p. 2689.

Acar, A. and Muraki, Y., 2011. Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, 7(3), pp. 392-402.

Adwaith, D., Abishake, A.K., Raghul, S.V. and Sivasankar, E., 2022. Enhancing multimodal disaster tweet classification using state-of-the-art deep learning networks. *Multimedia Tools and Applications*, 81(13), pp. 18483-18501.

Africa, S., Sokupa, S. and Gumbi, M., 2021. Report of the expert panel into the July 2021 civil unrest. *The presidency Republic of South Africa* (online). Available: https://www.thepresidency.gov.za/sites/default/files/2022-05/Report%20of%20the%20Expert%20Panel%20into%20the%20July%202021%20Civil%20Unrest.pdf (Accessed 20 November 2023).

Ahamed, B.S. and Arya, S., 2021. LGBM classifier based technique for predicting type-2 diabetes. *European Journal of Molecular and Clinical Medicine*, 8(3), pp. 454-467.

Ahmed, S. and Rasul, M.E., 2022. Social media news use and Covid-19 misinformation engagement: Survey study. *Journal of Medical Internet Research*, 24(9), p. e38944. https://doi.org/10.2196/38944

Ahmed, W., 2018. Using Twitter data to provide qualitative insights into pandemics and epidemics. Doctoral dissertation. University of Sheffield.

Ahmed, W., Bath, P.A., Sbaffi, L. and Demartini, G., 2019. Novel insights into views towards H1N1 during the 2009 Pandemic: A thematic analysis of Twitter data. *Health Information and Libraries Journal*, 36(1), pp. 60-72.

Ahmed, W., Demartini, G. and Bath, P., 2017. Topics Discussed on Twitter at the Beginning of the 2014 Ebola Epidemic in United States. In: *iConference 2017 Proceedings*. Wuhan, China. pp. 774-777.

Aïmeur, E., Amri, S. and Brassard, G., 2023. Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, 13(1), p. 30.

Al-Ahmad, B., Al-Zoubi, A.M., Abu Khurma, R. and Aljarah, I., 2021. An evolutionary fake news detection method for Covid-19 pandemic information. *Symmetry*, 13(6), p. 1091.

Al-Asadi, M.A. and Tasdemir, S., 2022. Using artificial intelligence against the phenomenon of fake news: A systematic literature review. In: M. Lahby, A.S.K. Pathan, Y. Maleh and W.M.S. Yafooz (eds.), *Combating fake news with computational intelligence techniques*. Cham: Springer International, pp. 39-54. https://doi.org/10.1007/978-3-030-90087-8_2

Al-Garadi, M.A., Yang, Y.C., Cai, H., Ruan, Y., O'Connor, K., Graciela, G.H., Perrone, J. and Sarker, A., 2021. Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Medical Informatics and Decision Making*, 21(1), pp. 27.

Ali, A.A., Latif, S., Ghauri, S.A., Song, O.Y., Abbasi, A.A. and Malik, A.J., 2023. Linguistic features and bi-LSTM for identification of fake news. *Electronics*, 12(13), p. 2942.

Allcott, H. and Gentzkow, M., 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), pp. 211-236.

Alotaibi, A., Rahman, A.U., Alhaza, R., Alkhalifa, W., Alhajjaj, N., Alharthi, A., Abushoumi, D., Alqahtani, M. and Alkhulaifi, D., 2022. Spam and sentiment detection in Arabic tweets using MARBERT model. *Mathematical Modelling of Engineering Problems*, 9(6), pp. 1574-1582.

Alqurashi, S., Hamoui, B., Alashaikh, A., Alhindi, A. and Alanazi, E., 2021. Eating garlic prevents COVID-19 infection: Detecting misinformation on the Arabic content of Twitter. *arXiv preprint arXiv:2101.05626*.

Alsaedi, N., Burnap, P. and Rana, O., 2017. Can we predict a riot? Disruptive event detection using Twitter. *ACM Transactions on Internet Technology*, 17(2), pp. 1-26.

Alshahrani, R. and Babour, A., 2021. An infodemiology and infoveillance study on COVID-19: Analysis of Twitter and google trends. *Sustainability*, 13(15), p. 8528. https://doi.org/10.3390/su13158528

Alzamzami, F., Hoda, M. and El Saddik, A., 2020. Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation. *IEEE Access*, 8, pp. 101840-101858.

Amari, S.I., 1972. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, C-21(11), pp. 1197-1206.

Aminikhanghahi, S. and Cook, D.J., 2017. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2), pp. 339-367.

An, J., Li, W.J., Ji, L. and Wang, R., 2013. A survey on information credibility on Twitter. *Applied Mechanics and Materials*, 401-403, pp. 1788-1791. https://doi.org/10.4028/www.scientific.net/amm.401-403.1788

Annamalai, B., Chandrasekaran, S. and Pathak, A.A., 2023. Driving citizen engagement through Twitter: The case of COVID-19 vaccination drive in India. *Journal of Public Affairs*, 23(2), p. e2858. https://doi.org/10.1002/pa.2858

Aoun Barakat, K., Dabbous, A. and Tarhini, A., 2021. An empirical approach to understanding users' fake news identification on social media. *Online Information Review*, 45(6), pp. 1080-1096. https://doi.org/10.1108/OIR-08-2020-0333

Apolinardo-Arzube, O., García-Díaz, J.A., Medina-Moreira, J., Luna-Aveiga, H. and Valencia-García, R., 2019. Evaluating information-retrieval models and machine-learning classifiers for measuring the social perception towards infectious diseases. *Applied Sciences*, 9(14), p. 2858. https://doi.org/10.3390/app9142858

Arain, M., Campbell, M.J., Cooper, C.L. and Lancaster, G.A., 2010. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Medical Research Methodology*, 10(1), p. 67.

Artetxe, M. and Schwenk, H., 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, pp. 597-610.

Asare, A.O., Sarpong, E.O., Truong Holds, N., Osei-Bonsu, P., Ahado, S. and Mensah, W.G., 2022. COVID-19 pandemic and African innovation: Finding the good from the bad using Twitter data and text mining approach. *International Social Science Journal*, 73(250), pp. 959-978. https://doi.org/10.1111/issj.12386

Asur, S. and Huberman, B.A., 2010, August. Predicting the future with social media. In: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, vol. 1, pp. 492-499.

Aulia, N. and Budi, I., 2019**.** Hate speech detection on Indonesian long text documents using machine learning approach. In: *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence (ICCAI '19)*. New York, NY, USA: Association for Computing Machinery, pp. 164-169. https://doi.org/10.1145/3330482.3330491

Aveyard, H., 2023. *Doing a literature review in health and social care: A practical guide.* 5th ed. UK: McGraw-Hill Education.

Bacsu, J., Fraser, S., Chasteen, A.L., Cammer, A., Grewal, K.S., Bechard, L., Bethell, J., Green, S., McGilton, K.S., Morgan, D., O'Rourke, H.M., Poole, L., Spiteri, R.J. and O'Connell, M.E., 2022. Using Twitter to examine stigma against people with dementia during COVID-19: infodemiology study. *JMIR Aging*, 5(1), p. e35677. https://doi.org/10.2196/35677

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

Bane, K.C., 2017. Tweeting the agenda: How print and alternative web-only news organizations use Twitter as a source. *Journalism Practice*, 13(2), pp. 191-205. https://doi.org/10.1080/17512786.2017.1413587

Banik, S., 2020. COVID fake news dataset. *Zenodo* (online). https://doi.org/10.5281/zenodo.4282522

Bansal, S., 2019. The math and intuition behind gradient descent. *DataDrivenInvestor* (online). Available: https://medium.datadriveninvestor.com/the-math-and-intuition-behind-gradient-descent-13c45f367a11 (Accessed 9 September 2024).

Barbier, G. and Liu, H., 2011. Data mining in social media. In: C. Aggarwal (ed.), *Social network data analytics*. Boston, MA: Springer. https://doi.org/10.1007/978-1-4419-8462-3_12

BBC News, 2020. Facebook, Twitter and Google face questions from US senators. *BBC News* (online). Available: https://www.bbc.com/news/technology-54721023 (Accessed 19 April 2021).

Bej, S., Davtyan, N., Wolfien, M., Nassar, M. and Wolkenhauer, O., 2021. LoRAS: An oversampling approach for imbalanced datasets. *Machine Learning*, *110*, pp. 279-301.

Bekker, A., 2019. 4 types of data analytics to improve decision-making. *ScienceSoft*, 14 May 2019 (online). Available: https://www.scnsoft.com/blog/4-types-of-data-analytics (Accessed 26 April 2024).

Bendat, J.S. and Piersol, A.G., 2011. *Random data: analysis and measurement procedures*. John Wiley & Sons.

Benoit, S.L. and Mauldin, R.F., 2021. The “anti-vax” movement: A quantitative report on vaccine beliefs and knowledge across social media. *BMC Public Health*, 21(1), p. 2106 https://doi.org/10.1186/s12889-021-12114-8

Bessi, A. and Ferrara, E., 2016. Social bots distort the 2016 US presidential election online discussion. *First Monday*, 21(11). Available: https://ssrn.com/abstract=2982233 (Accessed 14 July 2024).

Bhandari, N., 2018. ExtraTreesClassifier. *Medium.* Available: https://medium.com/@namanbhandari/extratreesclassifier-8e7fc0502c7. (Accessed 22 October 2018).

Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M. and Wallach, D.S., 2015. Aggregate characterization of user behaviour in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1), p. 4.

Bin Naeem, S. and Kamel Boulos, M.N., 2021. Covid-19 misinformation online and health literacy: A brief overview. *International Journal of Environmental Research and Public Health*, 18(15), p. 8091. https://doi.org/10.3390/ijerph18158091

Bird, W. and Smith, T., 2020. Disinformation during Covid-19: Weekly trends from real411 in South Africa. *Daily Maverick* (online). Available: https://www.dailymaverick.co.za/article/2020-06-08-disinformation-during-covid-19-weekly-trends-from-real411-in-south-africa/ (Accessed 28 November 2023).

Blom, J.N. and Hansen, K.R., 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76, pp. 87-100.

Blumberg, R. and Atre, S., 2003**.** The problem with unstructured data. *DM Review*, 13(42-49), p. 62. Available: https://atrepower.com/pdf/resources/DMReview/DMReviewFeb2003.pdf (Accessed 26 April 2024).

Bojjireddy, S., Chun, S.A. and Geller, J., 2021. Machine Learning approach to detect fake news, misinformation in COVID-19 pandemic. In: *DG.O2021: The 22nd Annual International Conference on Digital Government Research*. Omaha, NE, USA: Association for Computing Machinery, pp. 575-578. https://doi.org/10.1145/3463677.3463762

Bondielli, A. and Marcelloni, F., 2019. A survey on fake news and rumour detection techniques. *Information Sciences*, 497, pp. 38-55.

Bosch, T., 2017. Twitter activism and youth in South Africa: The case of #RhodesMustFall. *Information, Communication and Society*, 20(2), pp. 221-232.

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

Bruns, A., 2020. Big Social Data approaches in Internet studies: The case of Twitter. In: J. Hunsinger, M.M. Allen and L. Klastrup (eds.), *Second international handbook of Internet research*. Dordrecht: Springer Netherlands, pp. 65-81. https://doi.org/10.1007/978-94-024-1555-1_3

Bruns, A. and Liang, Y.E., 2012. Tools and methods for capturing Twitter data during natural disasters. *First Monday*, 17(4). https://doi.org/10.5210/fm.v17i4.3937

Bruns, A. and Stieglitz, S., 2014. Metrics for understanding communication on Twitter. In: A. Bruns, M. Mahrt, K. Weller, J. Burgess and C. Puschmann (eds.), *Twitter and society [Digital Formations, Volume 89]*. United States of America: Peter Lang, pp. 69-82. Available: https://eprints.qut.edu.au/66326/ (Accessed 14 July 2024).

Buffer, n.d. *Buffer* (online). Available: https://www.buffer.com (Accessed 1 July 2023).

Businesstech, 2021. South Africa's economy contracted by 7% last year. *Businesstech* (online). Available: https://businesstech.co.za/news/finance/474294/south-africas-economy-contracted-by-7-last-year/ (Accessed 08 August 2024).

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., Li, J., 2020. Exploring the role of visual content in fake news detection. In: K. Shu, S. Wang, D. Lee and H. Liu (eds.), *Disinformation, misinformation and fake news in social media*. Lecture Notes in Social Networks. Cham: Springer. https://doi.org/10.1007/978-3-030-42699-6_8

Cao, L., 2017. Data science: A comprehensive overview. *ACM Computing Surveys*, 50(3), pp. 1-42.

Cavazos-Rehg, P.A., Krauss, M., Fisher, S.L., Salyer, P., Grucza, R.A. and Bierut, L.J., 2015. Twitter chatter about marijuana. *Journal of Adolescent Health*, 56(2), pp. 139-145.

CCDH, 2021. The disinformation dozen. *Center for Countering Digital Hate.* PDF. Available: https://counterhate.com/wp-content/uploads/2022/05/210324-The-Disinformation-Dozen.pdf (Accessed 08 April 2024).

Chadwick, A., Vaccari, C. and Kaiser, J., 2022. The amplification of exaggerated and false news on social media: The roles of platform use, motivations, affect, and ideology. *American Behavioral Scientist*. https://doi.org/10.1177/00027642221118264

Charbuty, B. and Abdulazeez, A., 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(1), pp. 20-28. https://doi.org/10.38094/jastt20165

Chauhan, N.S., 2022. Decision tree algorithm, explained. *KDnuggets* (online). Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html (Accessed 29 November 2023).

Chen, E., Deb, A. and Ferrara, E., 2022. #Election2020: The first public Twitter dataset on the 2020 US Presidential election. *Journal of Computational Social Science*, 5(1), pp. 1-18. https://doi.org/10.1007/s42001-021-00117-9

Chen, M., Mao, S. and Liu, Y., 2014. Big data: A survey. *Mobile Networks and Applications*, 19(2), pp. 171-209.

Chen, M.Y., Lai, Y.W. and Lian, J.W., 2023. Using deep learning models to detect fake news about COVID-19. *ACM Transactions on Internet Technology*, 23(2), 1-23. https://doi.org/10.1145/3533431

Chen, X., Lee, W. and Lin, F., 2022. Infodemic, institutional trust, and COVID-19 vaccine hesitancy: A cross-national survey. *International Journal of Environmental Research and Public Health*, 19(13), p. 8033. https://doi.org/10.3390/ijerph19138033

Chen, Y., Conroy, N.J. and Rubin, V.L., 2015. Misleading online content: recognizing clickbait as 'false news'. In: *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (WMDD '15)*. New York, NY, USA: Association for Computing Machinery, pp. 15-9. Available: https://doi.org/10.1145/2823465.2823467

Chen, T. and Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, California, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785

Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., Xiao, X., Nazarian, S. and Bogdan, P., 2021. A COVID-19 rumor dataset. *Frontiers in Psychology*, 12, p. 644801. https://doi.org/10.3389/fpsyg.2021.644801

Chew, C. and Eysenbach, G., 2010. Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One*, 5(11), p. e14118. https://doi.org/10.1371/journal.pone.0014118

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cho, S.E., Jung, K. and Park, H.W., 2013. Social media use during Japan's 2011 earthquake: How Twitter transforms the locus of crisis communication. *Media International Australia*, 149(1), pp. 28-40.

Choi, S., 2024. The coronavirus disease 2019 infodemic: A concept analysis. *Frontiers in Public Health*, 12. https://doi.org/10.3389/fpubh.2024.1362009

Chong, S.K., Ali, S.H., Đoàn, L.N., Yi, S.S., Trinh-Shevrin, C. and Kwon, S.C., 2022. Social media use and misinformation among Asian Americans during Covid-19. *Frontiers in Public Health*, 9. https://doi.org/10.3389/fpubh.2021.764681

Choudhary, M., Jha, S., Prashant, Saxena, D. and Singh, A.K., 2021. A review of fake news detection methods using machine learning. In: *2021 2nd International Conference for Emerging Technology (INCET)*, Belagavi, India, pp. 1-5. https://doi.org/10.1109/INCET51464.2021.9456299

Christensen, L.B., Johnson, B., Turner, L.A. and Christensen, L.B., 2011. *Research methods, design and analysis*. Boston, Massachusetts: Pearson.

Chu, Z., Gianvecchio, S., Wang, H. and Jajodia, S., 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6), pp. 811-824.

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

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F. and Scala, A., 2020. The Covid-19 social media infodemic. *Scientific Reports*, 10(1), pp. 16598.

Coleman, K., 2023. X's community-led approach: Tackling inaccurate & misleading information. *X*. Available: https://blog.x.com/en_us/topics/company/2023/xs-community-led-approach-tackling-inaccurate-and-misleading-information (Accessed 26 August 2024).

Comninos, A., 2011. Twitter revolutions and cyber crackdowns: User-generated content and social networking in the Arab spring and beyond. *Association for Progressive Communications*, pp. 1-18.

Compton, R., Lee, C., Xu, J., Artieda-Moncada, L., Lu, T.C., Silva, L.D. and Macy, M., 2014. Using publicly visible social media to build detailed forecasts of civil unrest. *Security Informatics*, 3, p. 4. https://doi.org/10.1186/s13388-014-0004-6

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Cornelius, I., 2002. Theorizing information for information science. *Annual Review of Information Science and Technology*, 36(1), pp. 392-425.

Cover, T.M. and Thomas, J.A., 2001. *Elements of information theory*. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0471200611

Creswell, J.W., 2013. *Qualitative inquiry & research design: Choosing among the five approaches*. Thousand Oaks, California: SAGE.

Creswell, J.W. and Creswell, J.D., 2017. *Research design: Qualitative, quantitative and mixed methods approaches*. 3rd ed. Los Angeles: SAGE.

Crowley, J.L., 2023. Generative networks and the autoencoder. In: M. Chetouani, V. Dignum, P. Lukowicz and C. Sierra (eds.), *Human-centred artificial intelligence. ACAI 2021*. Lecture Notes in Computer Science. Cham: Springer. https://doi.org/10.1007/978-3-031-24349-3_4

Cucinotta, D. and Vanelli, M., 2020. WHO declares COVID-19 a pandemic. *Acta Bio-Medica: Atenei Parmensis*, 91(1), pp. 157-160.

Cui, L. and Lee, D., 2020. CoAID: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Cui, L., Shu, K., Wang, S., Lee, D. and Liu, H., 2019. dEFEND: A system for explainable fake news detection. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Beijing, China: Association for Computing Machinery, pp. 2961-2964. https://doi.org/10.1145/3357384.3357862

Czakon, J. 2023. F1 score vs. ROC AUC vs. Accuracy vs. PR AUC: Which evaluation metric should you choose? *Neptune.ai* (online). Available: https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc (Accessed 15 July 2024).

D'Ulizia, A., Caschera, M.C., Ferri, F. and Grifoni, P., 2021. Repository of fake news detection datasets. Version 1. *4TU.ResearchData* (dataset). https://doi.org/10.4121/14151755.v1

Dadas, C., 2017. Hashtag activism: The promise and risk of "Attention". In: D.M. Walls and S. Vie (eds.), *Social writing/social media: Publics, presentations and pedagogies*. Fort Collins, CO: The WAC Clearinghouse, University Press of Colorado, pp. 17-36. https://doi.org/10.37514/PER-B.2017.0063.2.01

Daniel, F. and Millimaggi, A., 2020. On twitter bots behaving badly. *Journal of Web Engineering*. https://doi.org/10.13052/jwe1540-9589.1883

Davies, C., Graffy, R., Shandukani, M., Baloyi, E., Gast, L., Kok, G., Mbokazi, F., Zita, A., Zwane, M., Magagula, R., Mabuza, A., Ramkrishna, W., Morris, N., Porteous, J., Shirreff, G., Blumberg, L., Misiani, E. and Moonasar, D., 2019**.** Effectiveness of 24-h mobile reporting tool during a malaria outbreak in Mpumalanga Province, South Africa. *Malaria Journal*, 18(1), 45. Available: https://doi.org/10.1186/s12936-019-2683-4 (Accessed 29 April 2024)

Dawber, T.R., Kannel, W.B. and Lyell, L.P., 1963. An approach to longitudinal studies in a community: The Framingham study. *Annals of the New York Academy of Sciences*, 107(2), pp. 539-556.

DeVerna, M.R., Pierri, F., Truong, B.T., Bollenbacher, J., Axelrod, D., Loynes, N., Torres-Lugo, C., Yang, K.C., Menczer, F. and Bryden, J., 2021, May. CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines. In: *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1), pp. 992-999. https://doi.org/10.1609/icwsm.v15i1.18122

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Díaz-Uriarte, R. and Andrés, S. Á. d., 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1). https://doi.org/10.1186/1471-2105-7-3

Dietrich, D., 2013. The genesis of EMC's data analytics lifecycle. *Dell Technologies Services*, 1 November 2013 (online). Available: https://infocus.dellemc.com/david_dietrich/the-genesis-of-emcs-data-analytics-lifecycle/ (Accessed 25 June 2019).

Dongo, I., Cardinale, Y., Aguilera, A., Martínez, F., Quintero, Y. and Barrios, S., 2021. Web scraping versus Twitter API: A comparison for a credibility analysis. In: *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*. Chiang Mai, Thailand: Association for Computing Machinery, pp. 263-273. https://doi.org/10.1145/3428757.3429104

Duggal, H., 2021. Mapping internet shutdowns around the world. *Aljazeera* (online). Available: https://www.aljazeera.com/news/2021/3/3/mapping-internet-shutdowns-around-the-world (Accessed 28 November 2023).

Elhadad, M.K., Fun Li, K. and Gebali, F., 2019. Fake news detection on social media: A systematic survey. In: *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. New York: IEEE, pp. 1-8. https://doi.org/10.1109/PACRIM47961.2019.8985062

Ellerman, D., 2017. Logical information theory: New logical foundations for information theory. *Logic Journal of the IGPL*, 25(5), pp. 806-835.

Emadi, M. and Tanha, J., 2020. Margin-based semi-supervised learning using Apollonius Circle. In: S. Barbosa and M. Ali Abam (eds.), *Topics in theoretical computer science. TTCS 2020*. Lecture Notes in Computer Science. Cham: Springer. https://doi.org/10.1007/978-3-030-57852-7_4

Erl, T., Khattak, W. and Buhler, P., 2016. *Big data fundamentals. Concepts, drivers and techniques*. Prentice Hall.

Essa, E., Omar, K. and Alqahtani, A., 2023. Fake news detection based on a hybrid BERT and LightGBM models. *Complex Intelligent Systems*, 9, pp. 6581-6592. https://doi.org/10.1007/s40747-023-01098-0

Euronews, 2021. Big Tech CEOs face questions in U.S. Congress over misinformation. *Euronews* (online). Available: https://www.euronews.com/2021/03/25/facebook-twitter-and-google-face-questions-in-us-congress-over-misinformation (Accessed 19 April 2021).

Eysenbach, G., 2009. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyse search, communication and publication behaviour on the Internet. *Journal of Medical Internet Research*, 11(1), p. e11.

Eysenbach, G., 2011. Infodemiology and infoveillance: Tracking online health information and cyberbehavior for public health. *American Journal of Preventive Medicine*, 40(5) Supplement 2, pp. S154-S158.

Eysenbach, G., 2020. How to fight an infodemic: The four pillars of infodemic management. *Journal of Medical Internet Research*, 22(6), p. e21820. https://doi.org/10.2196/21820

Facebook, 2020. Foundational integrity research: Misinformation and polarization request for proposals. *Facebook Research* (online). Available: https://research.fb.com/programs/research-awards/proposals/foundational-integrity-research-misinformation-and-polarization-request-for-proposals/ (Accessed 19 April 2021).

Ferrada, P., Suliburk, J.W., Bryczkowski, S.B., Selby, L.V., Lee, E.E., Torres, M., Kothari, A.N. and Kulaylat, A.N., 2016. The surgeon and social media: Twitter as a tool for practicing surgeons. *Bulletin of the American College of Surgeons*, 101(6), pp. 19-24.

Ferrara, E., Varol, O., Davis, C., Menczer, F. and Flammini, A., 2016. The rise of social bots. *Communications of the ACM*, 59(7), pp. 96-104.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), pp. 179-188.

Gandhi, R., 2018a. Naive Bayes classifier. *Medium* (online). Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (Accessed 27 November 2023).

Gandhi, R., 2018b. Support Vector Machine – Introduction to machine learning algorithms. *Medium* (online). Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (Accessed 21 November 2023).

Gandomi, A. and Haider, M., 2015. Beyond the hype: Big data concepts, methods and analytics. *International Journal of Information Management*, 35(2), pp. 137-144.

Gaonkar, S., Itagi, S., Chalippatt, R., Gaonkar, A., Aswale, S. and Shetgaonkar, P., 2019. Detection of online fake news: A survey. In: *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, Vellore, India, 2019, pp. 1-6. https://doi.org/10.1109/ViTECoN.2019.8899556

GeeksforGeeks, 2021. LightGBM (light gradient boosting machine). *GeeksforGeeks.* Available: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/. (Accessed 21 December 2021).

Gehring, J., Auli, M., Grangier, D., Yarats, D. and Dauphin, Y.N., 2017**.** Convolutional sequence to sequence learning. In: D. Precup and Y.W. Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*. Proceedings of Machine Learning Research, pp. 1243-1252. Available: https://proceedings.mlr.press/v70/gehring17a.html (Accessed 29 April 2024).

Gelfert, A., 2018. Fake news: A definition. *Informal Logic*, 38(1), pp. 84-117.

Geurts, P., Ernst, D. and Wehenkel, L., 2006. Extremely randomized trees. *Machine Learning*, 36, pp. 3-42. https://doi.org/10.1007/s10994-006-6226-1

Ghanem, B., Ponzetto, S.P., Rosso, P., Rangel, F., Merlo, P., Tiedemann, J. and Tsarfaty, R., 2021. FakeFlow: Fake news detection by modelling the flow of affective information. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, pp. 679-689. https://doi.org/10.18653/v1/2021.eacl-main.56

Ghebreyesus, T.A., 2020. General's opening remarks at the media briefing on COVID-19 −11 March 2020. *World Health Organization* (online). Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (Accessed 21 April 2021).

Ghorpade, S.J., Chaudhari, R.S. and Patil, S.S., 2022. Enhancement of imbalance data classification with boosting methods: an experiment. *ECS Transactions*, 107(1), pp. 15923-15934. https://doi.org/10.1149/10701.15923ecst

Giglietto, F., Rossi, L. and Bennato, D., 2012. The open laboratory: Limits and possibilities of using Facebook, Twitter and YouTube as a research data source. *Journal of Technology in Human Services*, 30(3-4), pp. 145-159.

Gilda, S., 2017. Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection. In: *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, IEEE, pp. 110-115. https://doi.org/10.1109/SCORED.2017.8305411

Gillis, A.S., n.d. What is data engineer? *TechTarget* (online). Available: https://searchdatamanagement.techtarget.com/definition/data-engineer (Accessed 3 February 2021).

Goldberg, Y., 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57, pp. 345-420.

Golos, A.M., Guntuku, S.C., Piltch-Loeb, R., Leininger, L., Simanek, A.M., Kumar, A., Albrecht, S.S., Dowd, J.B., Jones, M. and Buttenheim, A.M., 2023. Dear pandemic: A topic modelling analysis of COVID-19 information needs among readers of an online science communication campaign. *PLoS One*, 18(3), p. e0281773. https://doi.org/10.1371/journal.pone.0281773

González-Bailón, S., Borge-Holthoefer, J., Rivero, A. and Moreno, Y., 2011. The dynamics of protest recruitment through an online network. *Scientific Reports*, 1(1), p. 197. https://doi.org/10.1038/srep00197

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning*. MIT Press.

Google, n.d. *Fact Check Tools* (online). Available: https://toolbox.google.com/factcheck/explorer (Accessed 04 April 2022).

Granjon, P., 2013. The CUSUM algorithm – A small review. *HAL* (online). Available: https://hal.archives-ouvertes.fr/hal-00914697 (Accessed 10 July 2019).

Graves, A. and Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6), pp. 602-610.

Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., Liu, L., Shan, H., Lei, C.L., Hui, D.S.C., Du, B., Li, L.J., Zeng, G., Yuen, K.Y., Chen, R.C., Tang, C.L., Wang, T., Chen, P.Y., Xiang, J., Li, S.Y., Wang, J.L., Liang, Z.J., Peng, Y.X., Wei, L., Liu, Y., Hu, Y.H., Peng, P., Wang, J.M., Liu, J.Y., Chen, Z., Li, G., Zheng, Z.J., Qiu, S.Q., Luo, J., Ye, C.J., Zhu, S.Y., Zhong, N.S. and China Medical Treatment Expert Group for Covid-19, 2020. Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*, 382(18), pp. 1708-1720.

Gundecha, P. and Liu, H., 2012. Mining social media: A brief introduction. In: *New directions in informatics, optimization, logistics and production*, pp. 1-17. https://doi.org/10.1287/educ.1120.0105

Gupta, B., Rawat, A., Jain, A., Arora, A. and Dhami, N., 2017. Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, 163(8), pp. 15-19.

Hacking, I., 1983. *Representing and intervening, introductory topics in the philosophy of natural science*. Cambridge, UK: Cambridge University Press.

Hadlington, L., Harkin, L.J., Kuss, D., Newman, K. and Ryding, F.C., 2023. Perceptions of fake news, misinformation and disinformation amid the Covid-19 pandemic: A qualitative exploration. *Psychology of Popular Media*, 12(1), pp. 40-49. https://doi.org/10.1037/ppm0000387

Hakak, S., Alazab, M., Khan, S., Gadekallu, T.R., Maddikunta, P.K.R. and Khan, W.Z., 2021. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 117, pp. 47-58.

Hansrajh, A., Adeliyi, T.T. and Wing, J., 2021. Detection of online Fake News using blending ensemble learning. *Scientific Programming*, 2021(3434458), pp. 1-10. https://doi.org/10.1155/2021/3434458

Hasanzad, M., Namazi, H. and Larijani, B., 2023. COVID-19 anti-vaccine attitude and hesitancy. *Journal of Diabetes and Metabolic Disorders*, 22(1), pp. 1-4.

Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H., 2009. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. New York: Springer, pp. 1-758.

Hauschild, J. and Eskridge, K., 2024. Word embedding and classification methods and their effects on fake news detection. *Machine Learning with Applications*, 17, p. 100566. https://doi.org/10.1016/j.mlwa.2024.100566

Haustein, S., Bowman, T.D., Holmberg, K., Tsou, A., Sugimoto, C.R. and Larivière, V., 2016. Tweets as impact indicators: Examining the implications of automated 'bot' accounts on Twitter. *Journal of the Association for Information Science and Technology*, 67(1), pp. 232-238. https://doi.org/10.1002/asi.23456

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H.A., Epstein, D.H., Leggio, L. and Curtis, B., 2021. Bots and misinformation spread on social media: Implications for Covid-19. *Journal of Medical Internet Research*, 23(5), p. e26933. https://doi.org/10.2196/26933

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

Hoffmann, J.P., 2021. *Linear regression models: Applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003162230

Holford, D.L., FASCE, Fasce, A., Costello, T.H. and Lewandowsky, S., 2023. Psychological profiles of anti-vaccination argument endorsement. *Scientific Reports*, 13(1), p. 11219.

Holzinger, A., 2015. Data mining with decision trees: Theory and applications. *Online Information Review*, 39(3), pp. 437-438. https://doi.org/10.1108/OIR-04-2015-0121

Hootsuite, n.d. *Hootsuite* (online). Available: https://www.hootsuite.com/platform/publishing (Accessed 1 July 2023).

Horton, R., 2020. Offline: COVID-19 is not a pandemic. *Lancet*, 396(10255), p. 874.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression* 398. John Wiley & Sons.

Howard, P.N. and Hussain, M.M., 2013. *Democracy's fourth wave?: Digital media and the Arab Spring*. Oxford University Press.

Hu, Y. and Hong, Y., 2017. Modeling Twitter engagement in real-life events. In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. Hilton Waikoloa Village, Hawaii, USA, 4-7 January 2017. HICSS, pp. 960-969.

Hui, P.M., Shao, C., Flammini, A., Menczer, F. and Ciampaglia, G.L., 2018, June. The Hoaxy misinformation and fact-checking diffusion network. In: *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Hutto, C.J. and 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* Eighth international AAAI conference on weblogs and social media, 8(1), pp. 216-225. Ann Arbor, Michigan, USA, 1-4 June 2014.

IFTTT, n.d. *IFTTT* (online). Available: http://www.Ifttt.com (Accessed 1 July 2023).

Islam, M.R., 2018. Sample size and its role in Central Limit Theorem (CLT). *International Journal of Physics and Mathematics*, 1(1), pp. 37-47. https://doi.org/10.31295/ijpm.v1n1.42

Jang, B., Kim, M., Harerimana, G., Kang, S.U. and Kim, J.W., 2020. Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Applied Sciences*, 10(17), p. 5841.

Jang, H., Rempel, E., Roth, D., Carenini, G. and Janjua, N.Z., 2021. Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modelling and aspect-based sentiment analysis. *Journal of Medical Internet Research*, 23(2), p. e25431.

Janiesch, C., Zschech, P. and Heinrich, K., 2021. Machine learning and deep learning. *Electronic Markets*, 31(3), pp. 685-695.

Jaworsky, B.N. and Qiaoan, R., 2021. The politics of blaming: The narrative battle between China and the US over COVID-19. *Journal of Chinese Political Science*, 26(2), pp. 295-315.

Jaybhaye, S.M., Badade, V., Dodke, A., Holkar, A. and Lokhande, P., 2023. Fake News detection using LSTM based deep learning approach. In: *ITM Web of Conferences*. EDP Sciences, 56.

Jin, L., Wang, Z., Gu, R., Yuan, C. and Huang, Y., 2014, May. Training large scale deep neural networks on the Intel Xeon Phi many-core coprocessor. In: *IEEE International Parallel & Distributed Processing Symposium workshops* 2014. IEEE, pp. 1622-1630.

Joseph, A.M., Fernández, V., Kritzman, S., Eaddy, I., Cook, O.M., Lambros, S., Jara Silva, C.E., Arguelles, D., Abraham, C., Dorgham, N., Gilbert, Z.A., Chacko, L., Hirpara, R.J., Mayi, B.S. and Jacobs, R.J., 2022. Covid-19 misinformation on social media: A scoping review. *Cureus*, 14(4), e24601. https://doi.org/10.7759/cureus.24601.

Jurafsky, D. and Martin, J.H., 2024. *Speech and language processing: An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 3rd ed. Available: https://web.stanford.edu/~jurafsky/slp3/ed3bookaug20_2024.pdf (Accessed 9 September 2024).

Kadiyala, A. and Kumar, A., 2018. Applications of python to evaluate the performance of decision tree-based boosting algorithms. *Environmental Progress & Sustainable Energy*, 37(2), pp. 618-623. https://doi.org/10.1002/ep.12888

Kaliyar, R.K., Goswami, A., Narang, P. and Sinha, S., 2020. FNDNet – A deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61, pp. 32-44. https://doi.org/10.1016/j.cogsys.2019.12.005.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146-3154.

Kemp, S., 2021. Digital 2021: The latest insights into the “state of digital”. *We are social*. Available: https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/ (Accessed 02 May 2024).

Kemp, S., 2022. Digital 2022: South Africa. *DATAREPORTAL* (online). Available: https://datareportal.com/reports/digital-2022-south-africa (Accessed 16 November 2023).

Kemp, S., 2023. Digital 2023: South Africa. *DATAREPORTAL* (online). Available: https://datareportal.com/reports/digital-2023-south-africa (Accessed 16 November 2023).

Khan, M.A. and Khan, M.A., 2021. Comprehensive sentimental analysis of tweets towards COVID-19 in Pakistan: A study on governmental preventive measures. *Journal of Ambient Intelligence and Humanized Computing*, 12(11), pp. 11815-11828.

Khan, Y., 2019. A longitudinal sentiment analysis of the #FeesMustFall campaign on Twitter data. Masters dissertation. Durban University of Technology.

Khan, Y. and Thakur, S., 2022. Fake News detection of South African COVID-19 related tweets using Machine Learning. In: *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Durban, South Africa, 2022, pp. 1-5. https://doi.org/10.1109/icABCD54961.2022.9856272

Khan, Y., Thakur, S., Obiyemi, O. and Adetiba, E., 2022a. Exploring links between online activism and real-world events: A case study of the #FeesMustFall. *Scientific Programming*, vol. 2022, 1562592. https://doi.org/10.1155/2022/1562592

Khan, Y., Thakur, S., Obiyemi, O. and Adetiba, E., 2022b. Identification of Bots and Cyborgs in the #FeesMustFall Campaign. *Informatics*, 9(1), p. 21. Available: https://www.mdpi.com/2227-9709/9/1/21 (Accessed 02 May 2024).

Kim, Y., 2014. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1746-1751. https://doi.org/10.3115/v1/D14-1181

Kim, Y., Nordgren, R. and Emery, S., 2020. The story of Goldilocks and three Twitter's APIs: A pilot study on Twitter data sources and disclosure. *International Journal of Environmental Research and Public Health*, 17(3), p. 864.

Klimiuk, K.B. and Balwicki, Ł.W., 2024. What is infodemiology? An overview and its role in public health. *Przeglad Epidemiologiczny*, 78(1), pp. 81-89. https://doi.org/10.32394/pe/188119

Kolluri, N.L. and Murthy, D., 2021. CoVerifi: A COVID-19 news verification system. *Online Social Networks and Media*, 22, p. 100123.

Kotsiantis, S.B., 2013. Decision trees: A recent overview. *Artificial Intelligence Review*, 39(4), pp. 261-283. https://doi.org/10.1007/s10462-011-9272-4

Kouloumpis, E., Wilson, T. and Moore, J., 2021. Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), pp. 538-541. https://doi.org/10.1609/icwsm.v5i1.14185

Kumar, N., Walter, N., Nyhan, K., Khoshnood, K., Tucker, J.D., Bauch, C.T., Ding, Q., Jones-Jang, S.M., De Choudhury, M., Schwartz, J.L., Papakyriakopoulos, O. and Forastiere, L., 2022. Interventions to mitigate Covid-19 misinformation: Protocol for a scoping review. *Systematic Reviews*, 11(1), 107. https://doi.org/10.1186/s13643-022-01917-4

Kwak, H., Lee, C., Park, H. and Moon, S., 2010. What is Twitter, a social network or a news media? In: *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, North Carolina, USA: Association for Computing Machinery, pp. 591-600. https://doi.org/10.1145/1772690.1772751.

Kwak, S.G. and Kim, J.H., 2017. Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology*, 70(2), pp. 144-156.

Lambrou, G.I., Zaravinos, A., Ioannidou, P. and Koutsouris, D., 2021. Information, thermodynamics and life: A narrative review. *Applied Sciences*, 11(9), p. 3897.

Layton, D., 2020. Data engineering: What is it? *Medium* (online). Available: https://towardsdatascience.com/data-engineering-what-is-it-ebd8e32df589 (Accessed 3 February 2021).

Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J. and Zittrain, J.L., 2018. The science of fake news. *Science*, 359(6380), pp. 1094-1096.

Leung, X.Y., Sun, J. and Bai, B., 2019. Thematic framework of social media research: State of the art. *Tourism Review*, 74(3), 517-531.

Li, Q. and Zhou, W., 2020. Connecting the dots between fact verification and fake news detection. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, pp. 1820-1825. https://doi.org/10.18653/v1/2020.coling-main.165

Li, Z., Zhang, Y., Sui, B. and Xing, Z., 2022. FPGA implementation for the sigmoid with piecewise linear fitting method based on curvature analysis. *Electronics*, 11(1365). https://doi.org/10.3390/electronics11091365

Lieneck, C., Heinemann, K., Patel, J., Huynh, H., Leafblad, A., Moreno, E. and Wingfield, C., 2022. Facilitators and barriers of Covid-19 vaccine promotion on social media in the United States: A systematic review. *Healthcare*, 10(2), p. 321. https://doi.org/10.3390/healthcare10020321

Lin, C. and He, Y., 2009. Joint sentiment/topic model for sentiment analysis. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. Hong Kong, China: Association for Computing Machinery, pp. 375-384. https://doi.org/10.1145/1645953.1646003

Lipschultz, J.H., 2020. *Social Media Communication: Concepts, Practices, Data, Law and Ethics*. 3rd ed. Routledge. https://doi.org/10.4324/9780429202834

Liu, B., 2012. *Sentiment analysis and opinion mining*. Cham: Springer. Series: Synthesis Lectures on Human Language Technologies. https://doi.org/10.1007/978-3-031-02145-9

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. http://arxiv.org/abs/1907.11692

Liu, W.Z. and White, A.P., 1994. The importance of attribute selection measures in decision tree induction. *Machine Learning*, 15(1), pp. 25-41.

Lohiniva, A., Nurzhynska, A., Al-Hassan, H., Anim, B. and Aboagye, D.C., 2022. Infodemic management using digital information and knowledge co-creation to address COVID-19 vaccine hesitancy: Case study from Ghana. *JMIR Infodemiology*, 2(2), p. e37134. https://doi.org/10.2196/37134

Loomba, S., De Figueiredo, A., Piatek, S.J., De Graaf, K. and Larson, H.J., 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3), pp. 337-348. https://doi.org/10.1038/s41562-021-01056-1

MacKay, D.J., 2003. *Information theory, inference and learning algorithms*. Cambridge University Press.

Madhoushi, Z., Hamdan, A.R. and Zainudin, S., 2015. Sentiment analysis techniques in recent works. In: *2015 Science and Information Conference (SAI)* (online), pp. 288-291. https://doi.org/10.1109/SAI.2015.7237157

Madiba, T., 2020. DA launches campaign against spread of fake Covid-19 news. *Polity* (online). Available: https://www.polity.org.za/article/da-launches-campaign-against-spread-of-covid-19-fake-news-2020-04-06 (Accessed 18 November 2020).

Maimon, O.Z. and Rokach, L., 2014. *Data mining with decision trees: Theory and applications* 81. World Scientific.

Malhotra, R., Mahur, A. and Achint, 2022. COVID-19 Fake News detection system. In: *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 428-433. https://doi.org/10.1109/Confluence52989.2022.9734144

Manhas, J. and Kotwal, S., 2021. Implementation of intrusion detection system for Internet of Things using machine learning techniques. In: K.J. Giri, S.A. Parah, R. Bashir and K. Muhammad (eds.), *Multimedia security. Algorithms for intelligent systems*. Singapore: Springer. https://doi.org/10.1007/978-981-15-8711-5_11

Martinez, I., Viles, E. and Olaizola, I.G., 2021. Data science methodologies: Current challenges and future approaches. *Big Data Research*, 24, p. 100183.

Marwitz, K.K., 2021. The pharmacist's active role in combating Covid-19 medication misinformation. *Journal of the American Pharmacists Association: JAPhA*, 61(2), pp. e71-e74. https://doi.org/10.1016/j.japh.2020.10.022

Matamoros-Fernández, A., Bartolo, L. and Alpert, B., 2024. Acting like a bot as a defiance of platform power: Examining youtubers' patterns of 'inauthentic' behaviour on twitter during Covid-19. *New Media and Society*, 26(3), pp. 1290-1314. https://doi.org/10.1177/14614448231201648

Mavragani, A., 2020. Infodemiology and infoveillance: Scoping review. *Journal of Medical Internet Research*, 22(4), p. e16206. https://doi.org/10.2196/16206

Mayer-Schönberger, V. and Cukier, K., 2013. *Big Data: A* revolution *that will transform how we live, work and think*. Boston, MA: Houghton Mifflin Harcourt.

McKenna, B., Myers, M.D. and Newman, M., 2017. Social media in qualitative research: Challenges and recommendations. *Information and Organization*, 27(2), pp. 87-99.

Media Update, 2021. *Seven trending hashtags about COVID-19 on social media*. Available: https://www.mediaupdate.co.za/social/148423/seven-trending-hashtags-about-covid-19-on-social-media (Accessed 16 November 2022).

Mehta, D., Dwivedi, A., Patra, A. and Anand Kumar, M., 2021. A transformer-based architecture for fake news classification. *Social Network Analysis and Mining*, 11(1), pp. 1-12. https://doi.org/10.1007/s13278-021-00738-y

Mehta, I., 2023. Twitter has officially changed its logo to "X". *Tech Crunch* (online). Available: https://techcrunch.com/2023/07/24/twitter-has-officially-changed-its-logo-to-x/ (Accessed 20 November 2023).

Meijer, T., 2017. *Violent protest in #FeesMustFall*. Thesis. Leiden University.

Mengji, S., Ambarte, S., Arumilli, S.V.T., Mhamane, S. and Rane, R., 2021. Fake News detection using RNN-LSTM. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 9(10), pp. 1731-1737. https://doi.org/10.22214/ijraset.2021.35687

Micallef, N., Avram, M., Menczer, F. and Patil, S., 2021. Fakey: A game intervention to improve news literacy on social media. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 6, pp. 1-27. https://doi.org/10.1145/3449080

Mienye, I.D., Sun, Y. and Wang, Z., 2019. Prediction performance of improved decision tree-based algorithms: A review. *Procedia Manufacturing*, 35, pp. 698-703.

Millham, R. and Thakur, S., 2016. Social media and big data. In: G.S. Tomar, N.S. Chaudhari, R.S. Bhadoria and G.C. Deka (eds.), *The human element of big data: Issues, analytics and performance*. New York: Chapman and Hall/CRC, pp. 179-194.

Mishra, A., 2018. Metrics to evaluate your Machine Learning algorithm. *Medium* (online). Available: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 (Accessed 26 March 2021).

Mohamed, A., Muhammad Zain, Z., Shaiba, H., Alturki, N., Aldehim, G., Sakri, S., Farik Mat Yatin, S. and Mohamad Zain, J., 2023. Lexdeep: Hybrid lexicon and deep learning sentiment analysis using twitter for unemployment-related discussions during Covid-19. *Computers, Materials and Continua*, 75(1), pp. 1577-1601. https://doi.org/10.32604/cmc.2023.034746

Mokoka, M., 2021. *Meet the instigators: The Twitter accounts of the RET forces network that incited violence and demanded Zuma's release*. Centre for Analytics and Behavioural Change.

Montesi, M., 2021. Understanding fake news during the Covid-19 health crisis from the perspective of information behaviour: The case of Spain. *Journal of Librarianship and Information Science*, 53(3), pp. 454-465.

Mouratidis, D., Nikiforos, M.N. and Kermanidis, K.L., 2021. Deep learning for fake news detection in a pairwise textual input schema. *Computation*, 9(2), p. 20.

Mumenin, K.M., Reza, K.J., Shathi, S.S., Akter, H., Raihan, M., Hassan, M.M., Rahman, S. and Awal, M.A., 2021. COVID-19 Fake News detection on social media. In: *2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, pp. 1-4. https://doi.org/10.1109/IC4ME253898.2021.9768523

Nambiar, A. and Mundra, D., 2022. An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing*, 6(4), p. 132.

Nasir, J.A., Khan, O.S. and Varlamis, I., 2021. Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, 1(1), p. 100007.

Ng, J.Y., Liu, S., Maini, I., Pereira, W., Cramer, H. and Moher, D., 2023. Complementary, alternative and integrative medicine-specific Covid-19 misinformation on social media: A scoping review. *Integrative Medicine Research*, 12(3), 100975. https://doi.org/10.1016/j.imr.2023.100975

Ng, L.H.X., Robertson, D.C. and Carley, K.M., 2024. Cyborgs for strategic communication on social media. *Big Data and Society*, 11(1). https://doi.org/10.1177/20539517241231275

Ng, Q.X., Lim, S.R., Yau, C.E. and Liew, T.M., 2022. Examining the prevailing negative sentiments related to COVID-19 vaccination: Unsupervised deep learning of Twitter posts over a 16 month period. *Vaccines*, 10(9), p. 1457.

Ngai, C.S.B., Singh, R.G. and Yao, L., 2022. Impact of COVID-19 vaccine misinformation on social media virality: Content analysis of message themes and writing strategies. *Journal of Medical Internet Research*, 24(7), e37806. https://doi.org/10.2196/37806

Nkoala, S. and Dlanga, S., 2023. "My Fellow South Africans" Cyril Ramaphosa emulates Mandela in his COVID-19 presidential speeches. *African Journal of Rhetoric*, 15(1), pp. 167-192.

O'Leary, D.E., 2015. Twitter mining for discovery, prediction and causality: Applications and methodologies. *Intelligent Systems in Accounting, Finance and Management*, 22(3), pp. 227-247.

Obar, J.A. and Wildman, S.S., 2015. Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy*, 39(9), pp. 745-750.

Olayiwola, A., Oyedeji, A., Omoyeni, O., Ayemimowa, O. and Olaoluwa, M., 2023. Comparative analysis of machine learning models for detection of Fake News: A case study of COVID-19. *Journal of Information Technology and Computer Engineering (JITCE)*, 7(01), pp. 29-33. https://doi.org/10.25077/jitce.7.01.29-33.2023

Oracle, 2020. What is data science? *Oracle* (online). Available: https://www.oracle.com/data-science/what-is-data-science.html (Accessed 18 November 2020).

Osanga, E., 2020. Addressing the skills shortage in data science and analytics. *Standard Bank* (online). Available: https://www.standardbank.com/sbg/standard-bank-group/whats-happening/newsroom/Addressing-the-skills-shortage-in-data-science-and-analytics (Accessed 2 February 2021).

Osborne, M.T., Malloy, S.S., Nisbet, E.C., Bond, R.M. and Tien, J.H., 2022. Sentinel node approach to monitoring online Covid-19 misinformation. *Scientific Reports*, 12(1), 9832. https://doi.org/10.1038/s41598-022-12450-8

Oxford English Dictionary, 2023. *Infodemic.* Oxford University Press. https://doi.org/10.1093/OED/6664836666

Pan, R.J., 2020. The modern plague of antivaccine extremists. *Annals of Allergy, Asthma and Immunology*, 125(1), pp. 6-7.

Pant, A., 2019. Introduction to logistic regression. *Medium* (online). Available: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148 (Accessed 27 November 2023).

Park, J., 2020. Framework for sentiment-driven evaluation of customer satisfaction with cosmetics brands. *IEEE Access*, 8, pp. 98526-98538. https://doi.org/10.1109/ACCESS.2020.2997522

Park, H., Reber, B.H. and Chon, M.G., 2015. Tweeting as health communication: Health organizations' use of Twitter for health promotion and public engagement. *Journal of Health Communication*, 21(2), pp. 188-198. https://doi.org/10.1080/10810730.2015.1058435

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M.S., Das, A. and Chakraborty, T., 2021. Fighting an infodemic: COVID-19 fake news dataset. In: *Combating Online Hostile Posts in Regional Languages during Emergency Situations (First International Workshop), CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*. Springer International, pp. 21-29. https://doi.org/10.1007/978-3-030-73696-5_3

Paul, K., 2020. Zuckerberg: Facebook will review policies after backlash over Trump posts. *The Guardian* (online). Available: https://www.theguardian.com/technology/2020/jun/05/mark-zuckerberg-facebook-trump-policies-review (Accessed 19 April 2021).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85), pp. 2825-2830.

Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1), pp. 3-14. https://doi.org/10.1080/00220670209598786

Pennycook, G. and Rand, D.G., 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), pp. 388-402. https://doi.org/10.1016/j.tics.2021.02.007

Perez, S., 2018. Twitter's doubling of character count from 140 to 280 had little impact on length of tweets. *TechCrunch*, 30 October 2018 (online). Available: https://techcrunch.com/2018/10/30/twitters-doubling-of-character-count-from-140-to-280-had-little-impact-on-length-of-tweets/ (Accessed 26 March 2021).

Petit, J., Li, C., Millet, B., Ali, K. and Sun, R., 2021. Can we stop the spread of false information on vaccination? How online comments on vaccination news affect readers' credibility assessments and sharing behaviours. *Science Communication*, 43(4), pp. 407-434.

Phan, H.T., Tran, V.C., Nguyen, N.T. and Hwang, D., 2020. Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model. *IEEE Access*, 8, pp. 14630-14641. https://doi.org/10.1109/ACCESS.2019.2963702

Piltch-Loeb, R., Su, M., Hughes, B., Testa, M., Goldberg, B., Braddock, K., Miller-Idriss, C., Maturo, V. and Savoia, E., 2022. Testing the efficacy of attitudinal inoculation videos to enhance Covid-19 vaccine acceptance: Quasi-experimental intervention trial. *JMIR Public Health and Surveillance*, 8(6), p. e34615. https://doi.org/10.2196/34615

PolitiFact, n.d. *PolitiFact | Coronavirus* (online). Available: https://www.politifact.com/coronavirus/ (Accessed 04 April 2022).

Pomeranz, J.L. and Schwid, A.R., 2021. Governmental actions to address Covid-19 misinformation. *Journal of Public Health Policy*, 42(2), pp. 201-210. https://doi.org/10.1057/s41271-020-00270-x

Prajapati, V., 2013. Understanding the data analytics project life cycle. *Pingax.* Available: http://pingax.com/understanding-data-analytics-project-life-cycle/ (Accessed 28 November 2023).

Provost, F. and Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), pp. 51-59.

Rajendran, A., Sahithi, V.S., Gupta, C., Yadav, M., Ahirrao, S., Kotecha, K., Gaikwad, M., Abraham, A., Ahmed, N. and Alhammad, S.M., 2022. Detecting extremism on Twitter during U.S. capitol riot using deep learning techniques. *IEEE Access*, 10, pp. 133052-133077. https://doi.org/10.1109/ACCESS.2022.3227962

Rao, A., Morstatter, F. and Lerman, K., 2022. Partisan asymmetries in exposure to misinformation. *Scientific Reports*, 12(1), 15671. https://doi.org/10.1038/s41598-022-19837-7

Rao, C.R., 1948. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society Series B*, 10(2), pp. 159-193.

Raza, S., 2021. Automatic fake news detection in political platforms - A transformer-based approach. In: H. Hürriyetoğlu (ed.), *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics, pp. 68-78. Available: https://aclanthology.org/2021.case-1.10 (Accessed 18 November 2020).

Reddy, H., Raj, N., Gala, M. and Basava, A., 2020. Text-mining-based fake news detection using ensemble methods. *International Journal of Automation and Computing*, 17(2), pp. 210-221. https://doi.org/10.1007/s11633-019-1216-5

Reis, J. and Housley, M., 2022. *Fundamentals of data engineering*. O'Reilly Media.

Rhys, H., 2020. *Machine learning with R, the tidyverse and mlr*. New York: Simon and Schuster. Available: https://www.simonandschuster.com/books/Machine-Learning-with-R-the-tidyverse-and-mlr/Hefin-Rhys/9781638350170 (Accessed 10 September 2024).

Richter, E., Carpenter, J.P., Meyer, A. and Richter, D., 2024. Digital social support among educators in social media: An international comparative study of tweets and replies in #teachertwitter and #twlz. *Computers & Education*, 221, p. 105137. https://doi.org/10.1016/j.compedu.2024.105137

Ridhwan, K.M. and Hargreaves, C.A., 2021. Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore. *International Journal of Information Management Data Insights*, 1(2), p. 100021.

Righetti, N., 2021. Four years of fake news: A quantitative analysis of the scientific literature. *First Monday*, 26(7). https://doi.org/10.5210/fm.v26i7.11645

Rish, I., 2001. An empirical study of the naive Bayes classifier. In: *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*. Seattle, 4 August 2001. pp. 41-46.

Rizkallah, J., 2017. The big (unstructured) data problem. *Forbes*, 5 June 2017 (online). Available: https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#1fb91093493a (Accessed 28 November 2023).

Robillard, J.M., Johnson, T.W., Hennessey, C., Beattie, B.L. and Illes, J., 2013. Aging 2.0: Health information about dementia on Twitter. *PLoS One*, 8(7), p. e69861.

Rohera, D., Shethna, H., Patel, K., Thakker, U., Tanwar, S., Gupta, R., Hong, W. and Sharma, R., 2022. A taxonomy of fake news classification techniques: survey and implementation aspects. *IEEE Access*, 10, pp. 30367-30394.https://doi.org/10.1109/access.2022.3159651

Roscoe, J.T., 1975. *Fundamental research statistics for the Behavioural Science*. 2nd ed. New York City: Holt, Rinehart & Winston.

Rosen, A., 2017. Tweeting made easier. *Twitter*, 7 November 2017 (online). Available: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html (Accessed 26 March 2021).

Roth, Y. and Achuthan, A., 2020. Building rules in public: Our approach to synthetic & manipulated media. *Twitter* (online). Available: https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media (Accessed 5 August 2021).

Roth, Y. and Harvey, D., 2018. How Twitter is fighting spam and malicious automation. *Twitter* (online). Available: https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html (Accessed 25 November 2023).

Roth, Y. and Pickles, N., 2020. Updating our approach to misleading information. *Twitter* (online). Available: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html (Accessed 4 June 2021).

Rotman, D., Vieweg, S., Yardi, S., Chi, E., Preece, J., Shneiderman, B., Pirolli, P. and Glaisyer, T., 2011. From slacktivism to activism: participatory culture in the age of social media. In: *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. Vancouver, Canada, 7-12 May 2011. New York: ACM, pp. 819-822. https://doi.org/10.1145/1979742.1979543

Saenz, J., Kalathur Gopal, S. and Shukla, D., 2021. Covid-19 fake news infodemic research dataset (Covid19-FNIR dataset). *IEEE DataPort* (online). https://doi.org/10.21227/b5bt-5244

Saif, H., 2015. *Semantic sentiment analysis in social streams*. Berlin: Akademische Verlagsgesellschaft. IOS Press.

Sakaki, T., Okazaki, M. and Matsuo, Y., 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: Association for Computing Machinery, pp. 851-860. https://doi.org/10.1145/1772690.1772777

Salathé, M. and Khandelwal, S., 2011. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology*, 7(10), p. e1002199. https://doi.org/10.1371/journal.pcbi.1002199

Sanh, V., Debut, L., Chaumond, J. and Wolf, T., 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sarin, G. and Kumar, P., 2020. ConvGRUText: A deep learning method for fake text detection on online social media. In: *PACIS 2020 Proceedings*, 60. Available: https://aisel.aisnet.org/pacis2020/60 (Accessed 14 July 2024).

Sayce, D., 2021. The number of tweets per day in 2020. *David Sayce* (online). Available: https://www.dsayce.com/social-media/tweets-day/ (Accessed 26 March 2021).

Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V. and Napolitano, A., 2010. Rusboost: a hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans*, 40(1), pp. 185-197. https://doi.org/10.1109/tsmca.2009.2029559

Shacklett, M., 2023. Structured vs. unstructured data. *Datamation,* 3 November 2023 (online). Available: https://www.datamation.com/big-data/structured-vs-unstructured-data.html (Accessed 25 November 2023).

Shah, S., 2021. Facebook blocks "#vaccines kill" hashtag. *Engadget* (online). Available: https://www.engadget.com/facebook-blocks-vaccines-kill-hashtag-094031580.html (Accessed 08 April 2024).

Shahzad, K., Khan, S.A., Ahmad, S. and Iqbal, A., 2022. A scoping review of the relationship of big data analytics with context-based fake news detection on digital media in the data age. *Sustainability*, 14(21), p. 14365. https://doi.org/10.3390/su142114365

Shalizi, C., 2013. *Advanced data analysis from an elementary point of view*. Available: https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf (Accessed 9 September 2024).

Shannon, C.E., 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), pp. 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M. and Liu, Y., 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, 10(3), pp. 1-42.

Sharma, K., Zhang, Y. and Liu, Y., 2022, May. COVID-19 vaccine misinformation campaigns and social media narratives. In. *Proceedings of the International AAAI Conference on Web and Social Media*, 16.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H., 2020. FakeNewsNet: A data repository with news content, social context and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3), pp. 171-188.

Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations Newsletter*, 19(1), pp. 22-36. https://doi.org/10.1145/3137597.3137600

Shuja, J., Alanazi, E., Alasmary, W. and Alashaikh, A. 2021. COVID-19 open source data sets: a comprehensive survey. *Applied Intelligence*, 51(3), pp. 1296–1325. https://doi.org/10.1007/s10489-020-01862-6

Singer, M. and Clair, S., 2003. Syndemics and public health: Reconceptualizing disease in bio-social context. *Medical Anthropology Quarterly*, 17(4), pp. 423-441.

Singh, J.P., Dwivedi, Y.K., Rana, N.P., Kumar, A. and Kapoor, K.K., 2019. Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, 283(1), pp. 737-757. https://doi.org/10.1007/s10479-017-2522-3

Singh, M.K., Ahmed, J., Alam, M.A., Raghuvanshi, K.K. and Kumar, S., 2024. A comprehensive review on automatic detection of fake news on social media. *Multimedia Tools and Applications*, 83(16), pp. 47319-47352. https://doi.org/10.1007/s11042-023-17377-4

Singh, R. and Agarwal, B.B., 2022. Automatic image classification and abnormality identification using machine learning. In: M.S. Kaiser, A. Bandyopadhyay, K. Ray, R. Singh and V. Nagar (eds.), *Proceedings of trends in electronics and health informatics*. Lecture Notes in Networks and Systems. Singapore: Springer. https://doi.org/10.1007/978-981-16-8826-3_2

Sinnenberg, L., DiSilvestro, C.L., Mancheno, C., Dailey, K., Tufts, C., Buttenheim, A.M., Barg, F., Ungar, L., Schwartz, H., Brown, D., Asch, D.A. and Merchant, R.M., 2016. Twitter as a potential data source for cardiovascular disease research. *JAMA Cardiology*, 1(9), pp. 1032-1036.

Skafle, I., Nordahl-Hansen, A., Quintana, D.S., Wynn, R. and Gabarrón, E., 2022. Misinformation about Covid-19 vaccines on social media: Rapid review. *Journal of Medical Internet Research*, 24(8), p. e37367. https://doi.org/10.2196/37367

Sloan, L., Jessop, C., Al Baghal, T. and Williams, M., 2020. Linking survey and Twitter data: Informed consent, disclosure, security and archiving. *Journal of Empirical Research on Human Research Ethics*, 15(1-2), pp. 63-76. https://doi.org/10.1177/1556264619853447

Small, T.A., 2011. What the hashtag?: A content analysis of Canadian politics on Twitter. *Information, Communication and Society*, 14(6), pp. 872-895.

Soetekouw, L. and Angelopoulos, S., 2024. Digital resilience through training protocols: Learning to identify fake news on social media. *Information Systems Frontiers*, 26(2), pp. 459-475. https://doi.org/10.1007/s10796-021-10240-7

Soni, D., 2018. Introduction to k-Nearest Neighbors. *KDnuggets* (online). Available: https://www.kdnuggets.com/2018/03/introduction-k-nearest-neighbors.html (Accessed 9 September 2024).

South Africa, 2013. *Protection of personal information act 4 of 2013*. Cape Town: Department of Labour (online). Available: https://www.gov.za/sites/default/files/gcis_document/201409/3706726-11act4of2013protectionofpersonalinforcorrect.pdf (Accessed 18 November 2020).

South Africa, 2020a. *Fake news – Coronavirus COVID-19. South African Government* (online). Available: https://www.gov.za/covid-19/resources/fake-news-coronavirus-covid-19 (Accessed 21 April 2021).

South Africa, 2020b. *Government monitors and responds to misinformation and fake news during Coronavirus Covid-19 lockdown* (online). Available: https://www.gov.za/news/media-statements/government-monitors-and-responds-misinformation-and-fake-news-during (Accessed 28 November 2023).

South Africa, 2020c. *President Cyril Ramaphosa: Measures to combat Coronavirus COVID-19 epidemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-measures-combat-coronavirus-covid-19-epidemic-15-mar-2020 (Accessed 08 August 2024).

South Africa, 2020d. *President Cyril Ramaphosa: Extension of Coronavirus COVID-19 lockdown to the end of April* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-extension-coronavirus-covid-19-lockdown-end-april-09-apr (Accessed 08 August 2024).

South Africa, 2020e. *President Cyril Ramaphosa: South Africa's response to Coronavirus COVID-19 pandemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-south-africas-response-coronavirus-covid-19-pandemic-13-may (Accessed 08 August 2024).

South Africa, 2020f. *President Cyril Ramaphosa interacts virtually with communities on Coronavirus COVID-19, 1 Jul* (online). Available: https://www.gov.za/news/media-advisories/government-activities/president-cyril-ramaphosa-interacts-virtually (Accessed 08 August 2024).

South Africa, 2020g. *President Cyril Ramaphosa: Progress in national effort to contain the Coronavirus COVID-19 pandemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-progress-national-effort-contain-coronavirus-covid-19 (Accessed 08 August 2024).

South Africa, 2020h. *President Cyril Ramaphosa: Progress in South Africa's effort to contain the Coronavirus Covid-19 pandemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-progress-south-africas-effort-contain-coronavirus-covid-19 (Accessed 08 August 2024).

South Africa, 2020i. *President Cyril Ramaphosa: Progress in national effort to contain Coronavirus Covid-19 pandemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-progress-national-effort-contain-coronavirus-covid-19-2 (Accessed 08 August 2024).

South Africa, 2020j. *President Cyril Ramaphosa on arrival of Johnson & Johnson COVID-19 Coronavirus Vaccine* (online). Available: https://www.gov.za/news/media-statements/president-cyril-ramaphosa-arrival-johnson-johnson-covid-19-coronavirus (Accessed 08 August 2024).

South Africa, 2020k. *Minister Tito Mboweni: 2021 Budget Speech* (online). Available: https://www.gov.za/news/speeches/minister-tito-mboweni-2021-budget-speech-24-feb-2021 (Accessed 08 August 2024).

South Africa, 2020l. *Government on the passing of journalist Karima Brown* (online). Available: https://www.gov.za/news/media-statements/government-passing-journalist-karima-brown-04-mar-2021 (Accessed 08 August 2024).

South Africa, 2021m. *President Cyril Ramaphosa: Developments in the country's response to the Coronavirus COVID-19 pandemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-developments-country%E2%80%99s-response-coronavirus-covid-19 (Accessed 08 August 2024).

South Africa, 2021n. *President Cyril Ramaphosa: Freedom Day 2021* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-freedom-day-2021-27-apr-2021 (Accessed 08 August 2024).

South Africa, 2021o. *President Cyril Ramaphosa: South Africa's response to Coronavirus COVID-19 pandemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-south-africas-response-coronavirus-covid-19-pandemic-15-jun (Accessed 08 August 2024).

South Africa, 2021p. *President Cyril Ramaphosa: South Africa's response to Coronavirus COVID-19 pandemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-south-africas-response-coronavirus-covid-19-pandemic-27-jun (Accessed 08 August 2024).

South Africa, 2021q. *President Cyril Ramaphosa: Progress in national effort to contain the Coronavirus COVID-19 pandemic* (online). Available: https://www.gov.za/news/speeches/president-cyril-ramaphosa-progress-national-effort-contain-coronavirus-covid-19-3 (Accessed 08 August 2024).

South Africa: DTI, 1978, amended 2008. *Copyright Act No. 98 of 1978*. Pretoria: Department of Labour. Available: https://www.gov.za/sites/default/files/gcis_document/201504/act-98-1978.pdf (Accessed 12 April 2024).

Sriram, A., Li, Y. and Hadaegh, A., 2021. Mining social media to understand user opinions on IoT security and privacy. In: *Proceedings of the 2021 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 252-257. https://doi.org/10.1109/SMARTCOMP52413.2021.00056

Stacey, S., 2023. Elon Musk says bots making "good content" will be exempt from Twitter's plan to charge for API access. *Business Insider* (online). Available: https://www.businessinsider.com/elon-musk-twitter-api-charges-wont-apply-to-good-bots-2023-2 (Accessed 21 November 2023).

Stokel-Walker, C., 2023a. Twitter changed science – what happens now it's in turmoil? *Nature*, 613(7942), pp. 19-21.

Stokel-Walker, C., 2023b. Twitter's $42,000-per-month API prices out nearly everyone. *Wired UK* (online). Available: https://www.wired.co.uk/article/twitter-data-api-prices-out-nearly-everyone (Accessed 29 November 2023).

Strydom, I.F. and Grobler, J., 2023. Transformers for COVID-19 misinformation detection on Twitter: A South African case study. In: *Machine Learning, Optimization and Data Science: 8th International Conference, LOD 2022, Certosa Di Pontignano, Italy, September 18–22, 2022, Revised Selected Papers, Part I*. Berlin, Heidelberg: Springer-Verlag, pp. 197-210. https://doi.org/10.1007/978-3-031-25599-1_15

Su, P. and Vijay-Shanker, K., 2022. Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction. *BMC Bioinformatics*, 23(1), p. 120.

Sule, S., DaCosta, M.C., DeCou, E., Gilson, C., Wallace, K. and Goff, S.L., 2023. Communication of Covid-19 misinformation on social media by physicians in the US. *JAMA Network Open*, 6(8), p. e2328928. https://doi.org/10.1001/jamanetworkopen.2023.28928

Susha, I., Janssen, M. and Verhulst, S., 2017. Data collaboratives as a new frontier of cross-sector partnerships in the age of open data: Taxonomy development. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 2691-2700. http://hdl.handle.net/10125/41481

Sykes, L.R., 1971. Aftershock zones of great earthquakes, seismicity gaps and earthquake prediction for Alaska and the Aleutians. *Journal of Geophysical Research*, 76(32), pp. 8021-8041.

Sykora, M., Elayan, S. and Jackson, T.W., 2020. A qualitative analysis of sarcasm, irony and related #hashtags on Twitter. *Big Data & Society*, 7(2). https://doi.org/10.1177/2053951720972735

Taghavi, D., 2017. *Exploring Fallism: Student Protests and the Decolonization of Education in South Africa*. Masters Dissertation, University of Cologne, Cologne.

Tandoc Jr, E.C., Lim, Z.W. and Ling, R., 2018. Defining 'fake news': A typology of scholarly definitions. *Digital Journalism*, 6(2), pp. 137-153.

Tashtoush, Y., Alrababah, B., Darwish, O., Maabreh, M. and Alsaedi, N., 2022. A deep learning framework for detection of COVID-19 Fake News on social media platforms. *Data*, 7(5), p. 65. https://www.mdpi.com/2306-5729/7/5/65

Taylor, W.A., 2000. Change-point analysis: A powerful new tool for detecting changes. *Taylor Enterprises* (online). Available: https://variation.com/wp-content/uploads/change-point-analyzer/change-point-analysis-a-powerful-new-tool-for-detecting-changes.pdf (Accessed 29 November 2023).

The Chief I/O, n.d. Comparison of Cloud GPU providers. *The Chief I/O* (online). Available: https://thechief.io/c/editorial/comparison-cloud-gpu-providers/ (Accessed 4 December 2023).

Tian, L., Zhang, X., Wang, Y. and Liu, H., 2020. Early detection of rumours on Twitter via stance transfer learning. In: J.M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M.J. Silva and F. Martins (eds.), *Advances in information retrieval*. Cham: Springer International, pp. 575-588.

Tong, X., Li, Y., Li, J., Bei, R. and Zhang, L., 2022. What are people talking about in #BackLivesMatter and #StopAsianHate? Exploring and categorizing Twitter topics emerged in online social movements through the Latent Dirichlet Allocation Model. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics and Society*. Oxford, United Kingdom: Association for Computing Machinery, pp. 723-738. https://doi.org/10.1145/3514094.3534202

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A. and Krause, A., 2018. Fake news detection in social networks via crowd signals. In: *Companion Proceedings of the Web Conference 2018*, Lyon, France, Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, pp. 517-524. https://doi.org/10.1145/3184558.3188722

TweetCaster, n.d. *TweetCaster* (online). Available: http://www.tweetcaster.com (Accessed 1 July 2023).

TweetDeck, n.d. *TweetDeck* (online). Available: https://tweetdeck.twitter.com (Accessed 1 July 2023).

Twitter, n.d.a. Twitter Terms of Service. *Twitter* (online). Available: https://twitter.com/en/tos/previous/version_13 (Accessed 20 November 2023).

Twitter, n.d.b. Twitter API. *Twitter* (online). Available: https://developer.twitter.com/en/products/twitter-api (Accessed 20 November 2023).

Ugarte, D.A. and Young, S., 2023. Effects of an online community peer-support intervention on Covid-19 vaccine misinformation among essential workers: Mixed-methods analysis. *Western Journal of Emergency Medicine*, 24(2), pp. 264-268. https://doi.org/10.5811/westjem.2023.1.57253

Valdez, D., Ten Thij, M., Bathina, K., Rutter, L.A. and Bollen, J., 2020. Social media insights into US mental health during the COVID-19 pandemic: Longitudinal analysis of Twitter data. *Journal of Medical Internet Research*, 22(12), p. e21418.

Van der Vyver, A.G., 2018, September. The listeriosis outbreak in South Africa: A Twitter analysis of public reaction. In: *International Conference on Management and Information Systems*, September 21.

Van Dyk, J. and Malan, M., 2018. Case closed: The #listeria outbreak was caused by Enterprise polony. *Bhekisisa*, 4 March 2018 (online). Available: https://bhekisisa.org/article/2018-03-04-00-case-closed-listeria-outbreak-was-caused-by-enterprise-polony-says-motsoaledi (Accessed 29 November 2023).

Van Rossum, G., 2007. Python programming language. In: *USENIX Annual Technical Conference,* 41(1), pp. 1-36.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 30. Curran Associates. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (Accessed 14 July 2024).

Visentin, M., Pizzi, G. and Pichierri, M., 2019. Fake news, real problems for brands: The impact of content truthfulness and source credibility on consumers' behavioural intentions toward the advertised brands. *Journal of Interactive Marketing*, 45, pp. 99-112.

Vivion, M., Anassour Laouan Sidi, E., Betsch, C., Dionne, M., Dubé, E., Driedger, S.M., Gagnon, D., Graham, J., Greyson, D., Hamel, D., Lewandowsky, S., MacDonald, N., Malo, B., Meyer, S.B., Schmid, P., Steenbeek, A., van der Linden, S., Verger, P., Witteman, H.O. and Yesilada, M., 2022. Prebunking messaging to inoculate against Covid-19 vaccine misinformation: An effective strategy for public health. *Journal of Communication in Healthcare*, 15(3), pp. 232-242. https://doi.org/10.1080/17538068.2022.2044606

Vo, T.H., Phan, T.L.T. and Ninh, K.C., 2022. Development of a fake news detection tool for Vietnamese based on deep learning techniques. *Eastern-European Journal of Enterprise Technologies*, 5(2(119)), pp. 14-20. https://doi.org/10.15587/1729-4061.2022.265317

Vosoughi, S., Roy, D. and Aral, S., 2018. The spread of true and false news online. *Science*, 359(6380), pp. 1146-1151. https://doi.org/10.1126/science.aap9559

Wade, A. and Di MarzoSerugendo, G., 2017. A model of extracting patterns in social network data using topic modelling, sentiment analysis and graph databases. In: D.C. Wyld (ed.), *Proceedings of the 2017 International Conference on Computer Science, Engineering and Information Technology (CS & IT)*. Geneva: University of Geneva, pp. 75-84. https://doi.org/10.5121/csit.2017.70608

Wakefield, J., 2021. Google, Facebook Twitter grilled in US on fake news. *BBC News* (online). Available: https://www.bbc.com/news/technology-56523378 (Accessed 19 April 2021).

Wang, L., Wang, Y., Chen, Y., Liu, C. and Fan, X., 2017. Prediction of lymphocytosis using machine learning algorithm based on checkup data. In: *2017 4th International Conference on Systems and Informatics (ICSAI)*, pp. 649-654. https://doi.org/10.1109/ICSAI.2017.8248369

Wang, Y., Zheng, J., Li, Q., Wang, C., Zhang, H. and Gong, J., 2021. XLNet-Caps: Personality classification from textual posts. *Electronics*, 10(11), p. 1360.

Wani, A., Joshi, I., Khandve, S., Wagh, V. and Joshi, R., 2021. Evaluating deep learning approaches for COVID-19 fake news detection. In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, Springer International, pp. 153-163.

Wardle, C. and Derakhshan, H., 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe* (online). Available: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c (Accessed 5 November 2020).

Waseem, Z. and Hovy, D., 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: J. Andreas, E. Choi and A. Lazaridou (eds.), *Proceedings of the NAACL Student Research Workshop*. San Diego, California, June 2016. Association for Computational Linguistics, pp. 88-93. Available: https://aclanthology.org/N16-2013 (Accessed 14 July 2024).

Watanabe, H., Bouazizi, M. and Ohtsuki, T., 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6, pp. 13825-13835.

Weatherbed, J., 2023. Twitter Blue's new 10,000 character limit turns tweets into essays. *The Verge* (online). Available: https://www.theverge.com/2023/4/14/23683082/twitter-blue-10000-character-limit-bold-italic-features-substack-newsletter (Accessed 20 November 2023).

Weil, K., 2010. Measuring tweets. *Twitter*, 22 February 2010 (online). Available: https://blog.twitter.com/en_us/a/2010/measuring-tweets.html (Accessed 26 March 2021).

Weinzierl, M.A. and Harabagiu, S.M., 2021. Automatic detection of COVID-19 vaccine misinformation with graph link prediction. *Journal of Biomedical Informatics*, 124, p. 103955.

White, B.K., Phuong, L., Roach, J., Teggelove, N. and Wallace, H., 2022. Pandemics, infodemics and health promotion. *Health Promotion Journal of Australia*, 34(1), pp. 169-172. https://doi.org/10.1002/hpja.644

Wiciaputra, Y.K., Young, J.C. and Rusli, A., 2021. Bilingual text classification in English and Indonesian via transfer learning using XLM-RoBERTa. *International Journal of Advances in Soft Computing and its Applications*, 13(3), pp. 73-87.

Wilson, S.L. and Wiysonge, C., 2020. Social media and vaccine hesitancy. *BMJ Global Health*, 5(10), p. e004206.

Wirth, R. and Hipp, J., 2000. CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. London, UK: Springer-Verlag, pp. 29-39.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A., 2020. Transformers: State-of-the-art natural language processing. In: Q. Liu and D. Schlangen (eds.), In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online, October 2020. Association for Computational Linguistics, pp. 38-45. Available: https://aclanthology.org/2020.emnlp-demos.6 (Accessed 14 July 2024).

Woode-Smith, N., 2016. #RhodesMustFall. In: *Fallism: A Year of Rational Commentary*. Cape Town: Rational Standard, pp. 20-23.

Woods, A., 2017. Twitter users flood Barcelona hashtags with cute animals to thwart terrorists. *New York Post*, 18 August 2017 (online). Available: https://nypost.com/2017/08/18/twitter-users-flood-barcelona-hashtags-with-cute-animals-to-thwart-terrorists/ (Accessed 25 November 2023).

Wu, Z. and McGoogan, J.M., 2020. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *JAMA*, 323(13), pp. 1239-1242.

Xanthopoulos, P., Pardalos, P.M. and Trafalis, T.B., 2013. Linear discriminant analysis. In: *Robust Data Mining*. New York, NY: Springer, pp. 27-33. https://doi.org/10.1007/978-1-4419-9878-1_4

Xing, J., Wang, S., Zhang, X. and Ding, Y., 2021. HMBI: A new hybrid deep model based on behaviour information for fake news detection. *Wireless Communications and Mobile Computing*, 2021, pp. 1-7. https://doi.org/10.1155/2021/9076211

Yang, P. and Chen, Y., 2017. A survey on sentiment analysis by using machine learning methods. In: *Proceedings of the 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pp. 117-121. https://doi.org/10.1109/ITNEC.2017.8284920

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F. and Liu, H., 2019. Unsupervised fake news detection on social media: A generative approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), pp. 5644-5651. https://doi.org/10.1609/aaai.v33i01.33015644

Yashpal, S., Raghunath, A., Gencerliler, N. and Burns, L.E., 2022. Exploring public perceptions of dental care affordability in the united states: Mixed method analysis via twitter. *JMIR Formative Research*, 6(7), p. e36315. https://doi.org/10.2196/36315

Young, S.D., Rivers, C. and Lewis, B., 2014. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine*, 63, pp. 112-115.

Zafarani, R., Liu, H., Phoha, V.V. and Azimi, J., 2021. Introduction on recent trends and perspectives in fake news research. *Digital Threats*, 2(2), pp. 1-3. https://doi.org/10.1145/3448634

Zeraatkar, K. and Ahmadi, M., 2018. Trends of infodemiology studies: a scoping review. *Health Information & Libraries Journal*, 35(2), pp. 91-120. https://doi.org/10.1111/hir.12216

Zhai, C. and Massung, S., 2016. *Text data management and analysis: A practical introduction to information retrieval and text mining*. Morgan & Claypool.

Zhang, J., Dong, B. and Yu, P.S., 2020. FakeDetector: Effective fake news detection with deep diffusive neural network. In: *Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1826-1829. https://doi.org/10.1109/ICDE48307.2020.00180

Zhang, J., Yin, Z., Chen, P. and Nichele, S., 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59, pp. 103-126.

Zhao, C., Musonye, H.A., Wang, P. and Pan, H., 2024. Infodemiology as a powerful tool in public health: bridging the digital divide for better well-being. *Innovation Discovery*, 1(2), p. 19. https://doi.org/10.53964/id.2024019

Zhao, Y., Zhu, S., Wan, Q., Li, T., Zou, C., Wang, H. and Deng, S., 2022. Understanding how and by whom Covid-19 misinformation is spread on social media: Coding and network analyses. *Journal of Medical Internet Research*, 24(6), p. e37623. https://doi.org/10.2196/37623

Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., Li, D., Wu, J. and Havlin, S., 2020. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9(1). https://doi.org/10.1140/epjds/s13688-020-00224-z

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W. and China Novel Coronavirus Investigating and Research Team, 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, 382(8), pp. 727-733.

Zhukov, A.V., Sidorov, D.N. and Foley, A.M., 2017. Random forest based approach for concept drift handling. In: D.I. Ignatov, M.Y. Khachay, V.G. Labunets, N. Loukachevitch, S.I. Nikolenko, A. Panchenko, A.V. Savchenko and K. Vorontsov (eds.), *Analysis of images, social networks and texts. AIST 2016*. Communications in Computer and Information Science, 661. Cham: Springer. https://doi.org/10.1007/978-3-319-52920-2_7

Zu, Z.Y., Jiang, M.D., Xu, P.P., Chen, W., Ni, Q.Q., Lu, G.M. and Zhang, L.J., 2020. Coronavirus disease 2019 (COVID-19): A perspective from China. *Radiology*, 296(2), pp. E15-E25.