Leveraging Large Language Models for Multilingual Disinformation Detection: The 2022 Kenyan General Elections Case Study

By

Rose Kwamboka Orwaru

054952

Submitted in Partial fulfillment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: **Rose Kwamboka Orwaru**

Sign:  Date: 27 / 05 / 2025

Approval

The dissertation of **Rose Kwamboka Orwaru** was reviewed and approved for examination by the following:

Dr. John Olukuru,
Senior Lecturer, Data Science and Analytics,
Strathmore University

Dr. Godfrey Madigu,
Dean, Institute of Mathematical Sciences,
Strathmore University

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

The rise of disinformation on social media, particularly during elections, poses challenges to democratic processes. This study examined the effectiveness of Large Language Models (LLMs) in detecting disinformation during the 2022 Kenyan General Elections. It analyzed the structure of disinformation networks, key actors' influence mechanisms, and the synergies between these methodologies. Using a dataset from X (formerly known as Twitter), the study applied Hugging Face's mBERT mBERT model for its multilingual capabilities. The data used to train the model had 20,000 rows and 37 features. Textual data was tokenised and classified as 'Hate Speech' or 'Not Hate Speech,' with key performance metrics, including accuracy, precision, recall, and F1-score indicating strong performance but potential overfitting due to dataset similarities. Further analysis revealed that disinformation networks relied on repetitive messaging and high-retweet amplification, with key influencers shaping discourse. SNA identified coordinated clusters, while SA detected strategic use of neutral and coded language strategically used to reframe discussions. LLMs enhanced detection by recognising subtle framing techniques, including word substitutions and sentiment shifts. Findings showed that disinformation was amplified through coordinated behaviour involving high-profile figures and media accounts. The integration of SA, SNA, and LLMs demonstrated the potential of AI-driven tools in misinformation mitigation. The study highlights the need for a multi-pronged detection approach, combining network analysis, sentiment evaluation, and language modeling, with implications for policymakers, journalists, and researchers.

Keywords: Large Language Models (LLMs), Disinformation, Hate Speech

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI - Artificial Intelligence

API - Application Programming Interface

BERT - Bidirectional Encoder Representations from Transformers

CC - CommonCrawl

CRISP-DM - Cross Industry Standard Process for Data Mining

DT - Decision Tree

EDA - Exploratory Data Analysis

KNN - K-Nearest Neighbours

LLM - Large Language Model

mBERT - Multilingual Bidirectional Encoder Representations from Transformers

MCA - Member of the County Assembly

MDM - Misinformation, Disinformation and Malinformation

ML - Machine Learning

MLQA - MultiLingual Question Answering

MP - Member of Parliament

NB - Naive Bayes

NER - Named Entity Recognition

NLP - Natural Language Processing

QT - Quoted Tweet. Reposting another user's tweet and adding your own thoughts

REST API - Representational State Transfer Application Programming Interface

RF - Random Forest

RT - Retweet. Sharing another user's tweet with your own followers

SA - Sentiment Analysis

SMOTE - Synthetic Minority Over-sampling Technique

SNA - Social Network Analysis

SNAPP - Social Networks Adapting Pedagogical Practice

SVM - Support Vector Machine

UI - User Interface

XLM-R - Cross-lingual Language Model – RoBERTa

XNLI - Cross-lingual Natural Language Inference

Acknowledgements

First and foremost, to the one true God, the source of all wisdom and strength. This work would not have been possible without divine guidance, grace, and perseverance granted throughout this journey. “Wisdom is the principal thing; therefore get wisdom: and with all thy getting get understanding” (Proverbs 4:7).

Secondly, my deepest gratitude to my supervisor, Dr John Olukuru, for his invaluable guidance, insightful feedback, and unwavering support. Your expertise and encouragement were instrumental in shaping both the direction and depth of this study.

A heartfelt thank you goes to Joy Kipkemboi and Ivan Musebe, whose technical insight on LLMs and help with understanding the political landscape significantly enhanced this study. Your brilliance and patience did not go unnoticed.

I am also sincerely thankful to the faculty and staff of The Institute of Mathematical Sciences, whose commitment to academic excellence created a nurturing environment for my research to thrive.

Special thanks to the institutions and platforms that provided access to the datasets and tools critical to this research. I acknowledge the generous contributions of open-source communities whose resources made the implementation and evaluation of the models possible.

To my family and friends, thank you for your endless patience, understanding, and words of encouragement during the many highs and lows of this journey.

Lastly, I dedicate this work to all researchers and practitioners working tirelessly to combat disinformation and uphold digital integrity in democratic societies. May this work add a humble voice to that noble effort.

Dedication

To my husband Bill ‘Casso’ Okutoi,

We made it.

Chapter 1: Introduction

In this chapter, background, problem statement, objectives, significance, scope, justification and limitations of the study are presented.

1.1 Background to the Study

In the digital age, social media platforms have become indispensable tools for communication, information dissemination, and public discourse (Lawton, 2011). However, alongside their benefits, these platforms have also become breeding grounds for the spread of disinformation, particularly during significant events such as elections. The 2022 Kenyan General Elections was no exception, witnessing a surge in disinformation campaigns across various social media platforms, notably X (formerly known as Twitter).

Disinformation, defined as the deliberate spread of false or misleading information with the intent to deceive, manipulate, or influence public opinion, poses a significant threat to the integrity of democratic processes and societal stability (Monteiro, 2023). Disinformation campaigns often target key events such as elections, exploiting vulnerabilities in online ecosystems to amplify false narratives, sow discord, and undermine trust in institutions and electoral processes.

Traditional methods of combating disinformation, such as content analysis and fact-checking, have proven insufficient in addressing the scale and complexity of disinformation campaigns on social media (Park et al., 2022). Moreover, the presence of automated accounts (bots), coordinated networks of users, and the rapid dissemination of information pose formidable challenges to traditional mitigation strategies.

In response to these challenges, researchers and practitioners have increasingly turned to advanced analytical techniques and automated tools such as Large Language Models (LLMs), sentiment analysis (SA), and social network analysis (SNA), to identify, analyse, and counter disinformation campaigns effectively. LLMs provide multilingual capabilities to identify patterns in online discourse, especially across diverse languages, while SA categorises content sentiment to highlight contentious topics. SNA, a methodology rooted in graph theory and network science, offers a holistic approach to understanding the structure, dynamics, and influence of social networks.

By mapping the connections between users, identifying key actors, and analysing information flow within networks, SNA provides valuable insights into the mechanisms underlying disinformation dissemination Park et al., 2022; Subrahmanian, 2016). Additionally, SA, a branch of natural language processing (NLP), enables researchers to assess the polarity of textual content on social media platforms.

By automatically categorising text as positive, negative, or neutral based on SA (Liu, 2017), it can reveal patterns in user sentiment, identify contentious topics, and detect the spread of misinformation and hate speech. The combination of SA, SNA and LLMs holds immense promise for unravelling the intricate web of disinformation during the 2022 Kenyan General Elections.

This study focuses on analysing the structural characteristics of disinformation networks during the 2022 Kenyan General Elections using LLMs. It investigates how key actors influence the flow and virality of content and explores the potential synergy between SA, SNA, and LLMs to enhance disinformation detection. By leveraging these methods, the research aims to contribute to safeguarding democratic processes through multilingual monitoring and more effective interventions.

1.2 Statement of the Problem

The proliferation of disinformation on social media platforms has become a critical issue in recent years, particularly during significant events such as elections. Disinformation actors exploit online ecosystems to spread false narratives, manipulate public opinion, and undermine trust in democratic processes.

While efforts have been made to identify and combat disinformation, traditional methods such as content analysis - which can identify the content of disinformation, but may not be able to identify the the actors behind the disinformation campaigns - have proven to be inadequate, given the scale and complexity of social media platforms. Furthermore, the rapid evolution of disinformation across multiple languages, including Swahili and Kenyan vernaculars, adds complexity to detection efforts.

Thus, there is a need for more sophisticated tools and techniques, such as SA, SNA, and LLMs to identify, analyse, and counter disinformation actors and their networks effectively.

In particular, SA will analyse the emotional polarity of content, SNA will map user connections to identify influential actors, and LLMs will enhance multilingual detection efforts, including in under-resourced languages.

Therefore, the aim of this study is to investigate the efficacy of SA, SNA and LLMs in unveiling disinformation campaigns during the 2022 Kenyan General Elections. By leveraging advanced analytical techniques, including SA to assess the polarity of textual content and SNA to map the connections between users and identify key actors, this study seeks to gain a deeper understanding of the mechanisms driving disinformation dissemination on social media platforms. This study will contribute to the development of predictive models that assess the effectiveness of LLMs in identifying emerging disinformation trends. The intended outcome is not only to provide valuable insights into the workings of disinformation during elections but also to build a dashboard that offers real-time monitoring and customised training for LLMs tailored to the Kenyan context.

1.3 Study Objectives

1.3.1 Overall Objective

This study aims to enhance LLMs for more effective disinformation detection, and to evaluate its performance in uncovering coordinated disinformation campaigns.

1.3.2 Specific Objectives

(i) To assess the limitations of existing disinformation detection models in capturing linguistic and contextual nuances.

(ii) To develop and test a multilingual disinformation detection pipeline using Large Language Models, tailored to Kenyan languages and socio-political context.

(iii) To analyse the effectiveness of the proposed LLM-based approach in identifying coordinated disinformation campaigns and the key actors involved during the 2022 Kenyan General Elections.

1.3.3 Research Questions

(i) To what extent do current Large Language Models (LLMs) detect and classify disinformation in Kenya's multilingual media landscape?

(ii) What are the key limitations of existing disinformation detection models when applied to content from the 2022 Kenyan General Elections?

(iii) How can multilingual and locally grounded adaptations of LLMs improve the accuracy and relevance of disinformation detection in electoral contexts?

1.4 Justification

In recent years, the surge of disinformation on social media has posed a growing threat to the integrity of democratic processes worldwide. The 2022 Kenyan General Election provides a pertinent case study to examine how disinformation campaigns impact electoral outcomes and influence public discourse.

The rise of multilingual disinformation — often in local languages like Swahili, Sheng and vernacular — further complicates efforts to detect and mitigate false information. This challenge necessitates advanced tools capable of handling diverse linguistic contexts.

By integrating LLMs, this study leverages new possibilities in Sentiment Analysis (SA) and Social Network Analysis (SNA) for disinformation detection. Traditional SA models struggle with language nuances, dialects, and code-switching common in Kenya, limiting their effectiveness. LLMs, trained on multilingual corpora, offer improved capacity to detect sentiment shifts and disinformation patterns across both English and non-English content. Similarly, SNA can map influential nodes and information flows within disinformation networks, helping identify key actors driving false narratives.

This study aims to address the pressing need for innovative solutions to combat online disinformation detection through the combined use of SA, SNA, and LLMs. Analysing how LLMs enhance sentiment classification and social network insights offers an opportunity to develop real-time, multilingual monitoring systems. Such integrated approaches are essential in understanding the mechanisms driving disinformation dissemination, identifying key

actors within disinformation networks, thus for safeguarding democratic integrity and fostering informed civic participation during elections and beyond.

Further, the findings will provide actionable insights for policymakers, researchers, and practitioners by identifying the sources of disinformation, mapping how false narratives spread, and developing targeted interventions. This research will contribute to strengthening digital literacy and public resilience against disinformation, ensuring better monitoring of multilingual content, and supporting robust electoral processes.

1.5 Limitation of the Study

This study examines tweets collected from X (formerly known as Twitter) between January 2021 and December 2022, limiting the analysis to variables observable within this specific platform and timeframe. Disinformation spreading through other platforms, such as Facebook, and dark socials such as WhatsApp and Telegram, is beyond the scope of this research, which could lead to incomplete insights into cross-platform narratives.

A lesser limitation is X (formerly known as Twitter)'s implementation of restrictions to its API, has affected access to historical tweets. This may potentially lead to missing critical nodes or interactions within the disinformation networks.

Chapter 2: Literature Review

2.1 Introduction

The emergence and rise of social media platforms has ushered in a new era of communication, providing opportunities for individuals worldwide to connect, share information, and engage in public discourse.

However, alongside these benefits, the prevalence of social media has brought forth its challenges; particularly the rapid spread of false information, commonly known as disinformation, hate speech, and other forms of harmful content. This phenomenon has become especially pronounced during periods of heightened political activity, such as elections, where social media platforms are often inundated with divisive rhetoric and false narratives. Scholars and practitioners have increasingly focused their attention on understanding the dynamics of disinformation propagation and its ramifications on public discourse, particularly in the context of elections and democratic processes. The 2022 Kenyan General Election were no exception, witnessing a surge in the spread of disinformation and hate speech across various online platforms.

In this digital age, the rapid and widespread circulation of false information poses significant challenges to the integrity of public discourse, political processes, and societal cohesion. Understanding the dynamics of disinformation campaigns, particularly within the context of social networks, has become imperative for mitigating the harmful effects of disinformation and safeguarding the integrity of democratic processes.

Against this backdrop, the need to understand and counteract the virality of disinformation and hate speech during electoral periods has become increasingly urgent. The 2022 Kenyan General Election serves as a poignant case study, highlighting the complexities and challenges associated with detecting and mitigating online misinformation in real-time. This study seeks to elucidate the underlying dynamics of disinformation and hate speech during the 2022 Kenyan General Election, with a specific focus on early detection and intervention strategies. It seeks to uncover the tactics employed by malicious actors to manipulate public discourse and influence electoral outcomes. By training a multilingual LLM to analyse these dynamics, the research aspires to enhance the detection and mitigation of disinformation in real-time.

This chapter presents a comprehensive review of peer-reviewed scholarly literature that forms the foundation for this study. The review is organised into three main sections to provide a structured analysis of the existing literature relevant to the research objectives.

The first section of the review lays the groundwork by clarifying the three types of information disorder namely; misinformation, malinformation, and disinformation. This distinction is critical for understanding the specific challenges posed by each type within the context of social media and electoral processes.

The second section explores fundamental concepts related to the detection and analysis of disinformation dissemination on social media platforms. It emphasises the importance of methodologies such as Sentiment Analysis (SA) and Social Network Analysis (SNA), which form the methodological backbone of this study. This section is crucial as it examines existing frameworks and tools for identifying and analysing disinformation campaigns, thus helps us understand the mechanisms driving disinformation and the effectiveness of the proposed analytical approaches

The final section focuses on reviewing scholarly literature that covers conceptual methods for assessing effectiveness of interventions against disinformation. Understanding the impact is vital for developing solutions. Understanding the impact of these interventions is essential for developing solutions that can mitigate the influence of disinformation on public opinion and democratic integrity, aligning closely with the study's goal of exploring the potential synergies between SA, SNA, and LLMs in enhancing disinformation detection.

2.2 Theoretical Review

2.2.1 Introduction

This section provides a theoretical review of key concepts informing the study of disinformation, particularly in elections. It draws from foundational works on information disorder and social media networks, focusing on three major dimensions: (1) types of information disorder (misinformation, disinformation, malinformation), (2) the use of technology for disinformation detection (Sentiment Analysis and Social Network Analysis), and (3) the role of LLMs in the detection of misinformation and hate speech.

2.2.2 Information Disorder

Defining the three major categories of information disorder is crucial to understanding the different tactics used by actors online to manipulate both true and false narratives to sway public opinion and amplify polarisation. These distinctions are particularly relevant in the 2022 Kenyan General Elections, where false narratives aimed at stoking ethnic divisions were prevalent, and malinformation targeted individuals by exposing private details to the public.

Abbreviated as MDM by Rosalie Li (2021), misinformation, disinformation and malinformation, though used interchangeably, have distinctly different meanings in the context of information disorder, each with different implications for public perception and behaviour. Understanding these nuances is essential for developing effective strategies to counter their negative effects.

As it represents the worst form of information disorder, the word "disinformation" is frequently used to refer to any false information (Rosalie Li, 2021). However, we can distinguish it from other types of information disorder based on its underlying deliberate intent to cause harm and its position on the true-false spectrum.

Karlova and Fisher (2013) define misinformation as false or misleading information that is spread unintentionally (Qazvinian et al., 2011), often due to misunderstanding, ignorance, or a lack of verification. According to Wardle and Derakhshan (2017), it is disseminated without malicious intent, and is based on genuine mistakes or erroneous beliefs (Rosalie Li, 2021). This can create an environment where public trust is eroded, and individuals make decisions based on incorrect facts, ultimately impacting behaviours, such as voting patterns or health-related choices (Majerczak & Strzelecki, 2022). Misinformation can still have harmful effects, even if it is not deliberately created to deceive.

Disinformation, on the other hand, is deliberately false or deceptive information (Karlova & Fisher, 2013), often referred to as “fake news”, rumors, and “hyperpartisan” news (Tucker et al., 2018). The primary motivation behind disinformation is to harm a person, social group, organisation or country, manipulate public opinion, sow discord, or achieve specific political or financial objectives (Wardle, 2017), which can lead to widespread mistrust and polarisation within society, according to Humprecht (2020).

Unlike misinformation, which may be spread unintentionally, disinformation often involves coordinated efforts. This form of information disorder poses a serious threat to democratic processes, as it can shape narratives and influence elections, as observed during the 2016 U.S. elections, where false stories generated significant engagement and profit for their creators (Hughes & Waismel-Manor, 2020).

Malinformation refers to true information that is shared with the intent to cause harm or damage to individuals, organisations, or society (Baines & Elliott, 2020; Cherilyn & Julie, 2024). Unlike misinformation and disinformation, malinformation often involves the repurposing of true information for manipulating ends (Baines & Elliott, 2020), such as doxxing or inciting violence, with the intention of creating a negative impact like inciting violence, by exposing private information, or spreading confidential data (Wardle & Derakhshan, n.d.). This highlights the ethical and legal complexities surrounding information disorder, as it blurs the line between public interest and privacy rights.

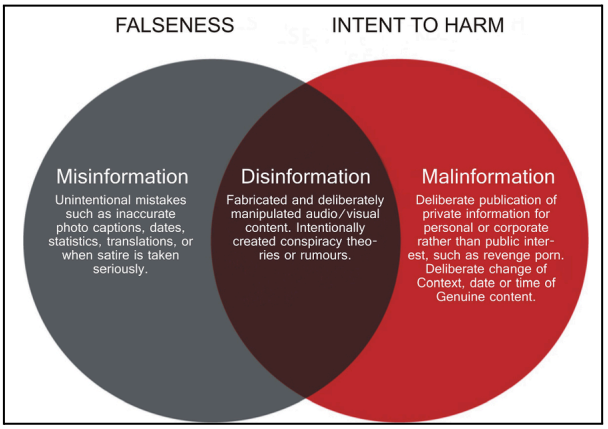


Figure 2.1: Information Disorder Venn Diagram (Cherilyn & Julie, 2024)

2.2.3 The Elements and Phases of Information Disorder

In order to understand the various components involved in the creation, dissemination, and reception of MDM, we need to identify and analyse the elements of information disorder - including the agent, messages, and interpreters (Wardle & Derakhshan, 2017).

Researchers identify the actors responsible for generating misleading content (agents) and their motivations, the content itself (messages) including its format and characteristics, and the individuals or groups influenced by or spreading such content (interpreters) including when they received the message, how they interpreted it and what action they took, if any

(Gentzkow & Allcott, 2017), to help in devising strategies to combat information disorder effectively.

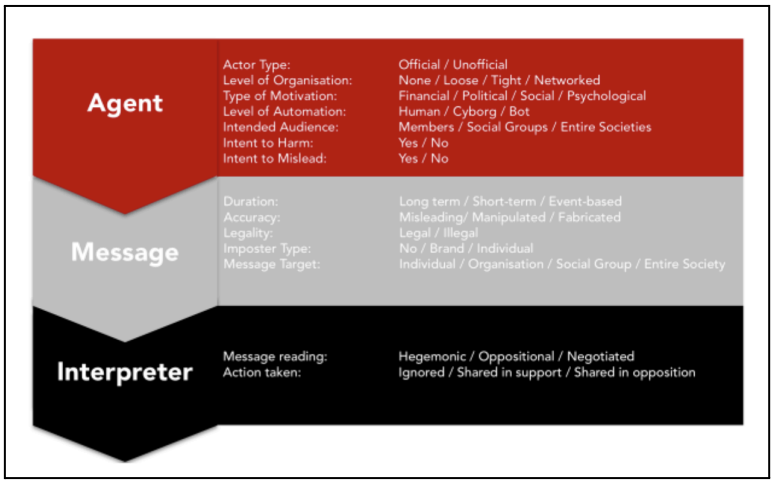


Figure 2.2: Information Disorder Questions (Wardle & Derakhshan, 2017)

Considering the life of information disorder, which occurs in three different phases (creation, production/reproduction, distribution) (Wardle, 2017), helps in identifying vulnerabilities and points of intervention to address information disorder effectively, thus holding accountable those who engage in malicious activities and addressing the root causes of information disorder.

This is because each phase involves different actors with distinct motivations. Agents are known to be driven by two major motivations; financial gain and promoting candidates or causes they support (Gentzkow & Allcott, 2017). For instance, during the 2016 US elections, over 100 sites posting fake news run by teenagers in the small town of Veles, Macedonia, generated fake stories that favoured both Trump and Clinton, resulting in earnings of tens of thousands of dollars for them (Hughes & Waismel-Manor, 2020).

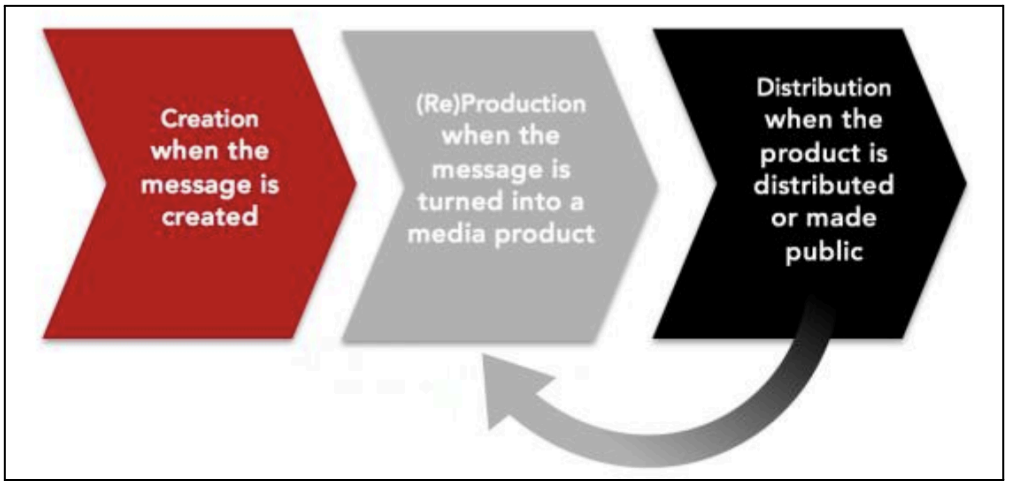


Figure 2.3: Three Phases Of Information Disorder (Wardle & Derakhshan, 2017)

2.2.4 Global Context

The rise of social media as a source for information and news dissemination and consumption has not only generated a wealth of data, providing a treasure trove for research on human behaviour (Morstatter & Liu, 2017), but has also become fertile ground for the spread of false information, commonly known as disinformation, influencing public opinion and potentially inciting violence.

The open, unmoderated nature of social media platforms (Elhag et al., 2018; Arkaitz Zubiaga et al., 2018) provide opportunities to study how users discuss and propagate disinformation, its ramifications on public discours, and to explore how NLP and data mining techniques may be used to find ways of determining their accuracy.

Recent studies indicate that false information spreads faster and broadly than the truth, (Soroush Vosoughi et al., 2018; The Spread of True and False News Online, 2018; Dwi Surjatmodjo et al., 2024), posing significant challenges to societal cohesion and democratic integrity. Researchers are increasingly employing NLP and data mining techniques to analyse disinformation propagation, seeking methods to verify the accuracy of information in real time (Jain R, 2024).

2.2.4.1 Sentiment Analysis (SA)

SA has emerged as a powerful tool for gauging public opinion and emotional responses to various issues, including political discourse and social movements. Studies show that SA can

effectively capture the nuances of public sentiment, allowing researchers to understand the impact of disinformation on public perception (Arias et al., 2022; Das et al., 2024)**.** However, a gap exists in the application of SA in multilingual contexts, where the subtleties of language can skew sentiment detection, limiting the accuracy of findings (Nur & Aniza, 2021).

For instance, a study by Elhag et al. (2018) to evaluate the performance of SA algorithms on Arabic and English language datasets found that machine learning (ML) algorithms for Arabic datasets were weak and need more work. The accuracy results obtained by two ML algorithms, Decision Tree (DT) and Naive Bayes (NB), were 50.76% and 54.43% for DT and NB respectively for Arabic datasets, compared to 97.16% and 89.52% accuracy for DT and NB respectively for English datasets.

2.2.4.2 Social Network Analysis (SNA)

SNA further complements these efforts by examining the structures of relationships within social media networks**.** It enables researchers to identify key actors and their roles in disseminating information (Serrat, 2017), helping to map how disinformation spreads through social channels (Duzen et al., 2023).

Lawton (2011) laid the groundwork for utilising social media data analysis as a mechanism for identifying disinformation actors and discerning emerging patterns. Building on Lawton's efforts, Monteiro et al. (2023) explored the use of social network analysis (SNA) to identify the spread of false information within social networks, distinguishing it from legitimate information sources.

Further, Corman et al. (2017) investigated the dynamics of disinformation on social media. Their research not only identified disinformation actors and patterns, but also shed light on the false information on public perceptions and attitudes.

Additionally, the role of social bots in disseminating disinformation has garnered significant attention in recent years. Studies by Kollanyi et al. (2016) and Shao et al. (2018) have highlighted the spread of low-credibility content by social bots, underscoring the need to understand their influence on information diffusion. Moreover, Vosoughi et al. (2018) and Guess et al. (2019) have examined the prevalence and predictors of fake news dissemination

on social media platforms, providing valuable insights into the mechanisms driving the spread of misinformation.

Shu, Sliva, Wang, Tang, and Liu (2017) identified several significant challenges in their study, which examined the dissemination of false information about X (formerly known as Twitter) and its influence on public opinion using a variety of data mining techniques. These include bias in data collection that affects the detection models' accuracy and a dearth of validated datasets for training and testing algorithms for detecting fake news. Additionally, the social interactions of users with fake news generate “data that is big, incomplete, unstructured, and noisy," ensuring that detection models generalise well across different domains and topics is a persistent challenge.

SNA has also been used to identify the accounts responsible for initiating and disseminating health misinformation during the COVID-19 pandemic on X (formerly known as Twitter), revealing how certain influential nodes can amplify false narratives (Kothari et al., 2022; Asma Ul Hussna et al., 2024). Despite advancements in SNA, many SNA studies lack real-time application capabilities that are critical for timely interventions in disinformation campaigns, much like how tools such as the Social Networks Adapting Pedagogical Practice (SNAPP) help educators monitor student interactions in online environments (Dawson et al., 2019). The absence of such real-time monitoring tools in disinformation research hinders the ability to identify and address misinformation promptly, underscoring a significant gap in current methodologies.

One significant issue in SA and SNA is the use of non-English languages to propagate disinformation. This complicates the analysis because many SA and SNA algorithms are trained on predominantly English datasets (Mohammad et al., 2016; Mohammad, 2017). This leads to data bias and significantly reduced accuracy when analysing content in local languages like Swahili and vernacular languages. This can result in misinterpretation or underestimation of the sentiment and network structures present in non-English posts.

Further, analysing content in multiple languages requires sophisticated NLP techniques capable of handling linguistic nuances, idiomatic expressions, and cultural references. Existing SA models may not accurately capture sentiment in languages other than English, while SNA might fail to identify key actors and relationships due to language barriers. Additionally, disinformation actors may use otherwise neutral words that could be construed

negatively in other languages, or vice versa, further complicating sentiment analysis (Hogenboom et al., 2015; Mozetič et al., 2016; Stieglitz et al., 2018; Farzindar & Inkpen, 2015).

2.2.4.3 Large Language Models (LLMs)

In the context of LLMs, recent advancements have shown their potential in automating the detection of disinformation. LLMs can analyse vast datasets in multiple languages, and understand intricate linguistic patterns, making them suitable for multilingual contexts where traditional methods may struggle (A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers, 2017; Hadi et al., 2023). Their ability to understand context and semantics allows for the nuanced identification of misinformation and its motivations, enhancing the overall effectiveness of detection strategies (Large Language Model Agent for Fake News Detection, 2017).

However, a significant gap exists in the integration of LLMs with SA and SNA methodologies; data timeliness. Given the dynamic nature of language, LLMs may lack access to the most up-to-date data, resulting in models trained on outdated or missing information (Pallagani et al., 2024).

This limitation becomes particularly problematic in the context of rapidly evolving events like elections, where misinformation and public narratives shift quickly. There also remains a lack of empirical studies that specifically assess the efficacy of combined approaches, particularly in non-English contexts like Kenya, where language and cultural nuances significantly affect the dissemination and perception of information (Towards Measuring and Modeling “Culture” in LLMs: A Survey, 2024; Ingers J, 2024).

2.2.5 Local Context

In Kenya, the growing adoption of social media platforms such as Facebook, TikTok and X (formerly known as Twitter), and dark social platforms such as WhatsApp and Telegram, has transformed how people access news and interact with information. While these platforms enhance civic engagement, they have also become channels for the spread of mis- and dis- and mal-information, especially during politically charged periods like elections. A study by Odanga Madung (2022), highlighted the proliferation of disinformation on TikTok during

Kenya's 2022 elections, with coordinated networks spreading false narratives in both English and Swahili. Similarly, research by the Mozilla Foundation showed how influencers and bots were employed to manipulate public opinion, often targeting local ethnic, cultural, and political divisions.

Local languages play a critical role in shaping these narratives. With content often shared in Swahili or vernacular languages, traditional sentiment and network analysis models — primarily trained on American English datasets — struggle to capture key nuances in British English (Ukkonen & Tiedemann, 2023), let alone other languages.

This introduces a challenge, as disinformation actors exploit linguistic and cultural subtleties to influence public perception while evading detection by automated systems. For example, politically motivated actors have weaponised otherwise neutral cultural symbols and sayings, creating subtle disinformation campaigns that evade detection by algorithms trained on English languages (Atefeh Farzindar & Inkpen, 2015; Bhutani et al., 2019).

The 2022 Kenyan General Election held on August 9, 2022, to elect the president, Members of Parliament (MPs), county governors, senators, members of the county assembly (MCAs), and women's representatives (National Council for Law Reporting with the Authority of the Attorney-General, 2010) continued this trend, with social media being a battleground for disinformation campaigns. Studies like those by Ogola (2022) highlight issues such as ethnic divisions, political polarisation, and the lack of robust fact-checking mechanisms have made Kenya particularly vulnerable to disinformation.

Information disorder here is not merely limited to spreading falsehoods but includes malinformation—true but harmful content shared with malicious intent, such as exposing private conversations or selective sharing of politically sensitive events. During election periods, such narratives escalate, increasing the potential for violence and social unrest (Lynch, Cheeseman, & Willis, 2019). Dark social platforms such as Telegram and WhatsApp, are especially difficult to monitor, due to their encrypted nature, thus allowing disinformation to spread rapidly and covertly (Kuru, Campbell, Bayer, Baruh, & Ling, 2022).

Despite advances in SA and SNA, real-time monitoring tools adapted for Kenya's multilingual and multicultural context remain inadequate. (A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers, 2017; Hadi et al., 2023).

Shikali and Mokhosi (2020) showed that many SA and SNA systems are designed for high-resource languages, resulting in low accuracy for Swahili or vernacular data.

There is also limited local research on how LLMs can enhance the accuracy of multilingual disinformation detection, representing a critical gap. This study aims to address these challenges by proposing the integration of LLMs with SA and SNA models, enhancing the capacity to detect and mitigate disinformation in real-time.

2.3 Empirical Review

The application of LLMs in combating disinformation has gained traction, especially in multilingual environments.

2.3.1 Models Applied in Disinformation Detection

2.3.1.1 Multilingual Bidirectional Encoder Representations From Transformers

Among the largest multilingual LLMs, Google's Multilingual Bidirectional Encoder Representations from Transformers (mBERT) was pre-trained on Wikipedia data in 104 languages — of the 7,000 known languages currently spoken in the world (Toloka, 2024) — , enabling it to perform cross-lingual tasks like document classification, language identification, word alignment, translation, and named entity recognition (NER). For instance, Pires et al. (2019) found that an mBERT model performs well when transferring between languages with distinct scripts, such as Urdu (written in Arabic script) and Hindi (written in Devanagari script) with 91% accuracy, by leveraging shared linguistic structures across the languages.

Despite its ability to generalise across typologically similar high-resource languages, that is, the more common languages, it faces challenges with low-resource languages, that is the underrepresented languages, that have dissimilar structures or sparse training data (Wu & Dredze, 2020). Furthermore, mBERT does not explicitly align multilingual representations, limiting its understanding of complex multilingual disinformation networks. This limits its ability to align multilingual disinformation patterns seamlessly across unrelated languages. (The Achilles Heel of Large Language Models: Lower-Resource Languages Raise More Safety Concerns, 2024)

A study by Conneau et al. (2020) produced the graph below that shows a distribution of data across the internet divided into language groups. While the high-resource languages have hundreds of GB of data available for training models, the low-resource languages, that is those towards the tail of the graph only have data available in the range of hundreds of megabytes, if at all.

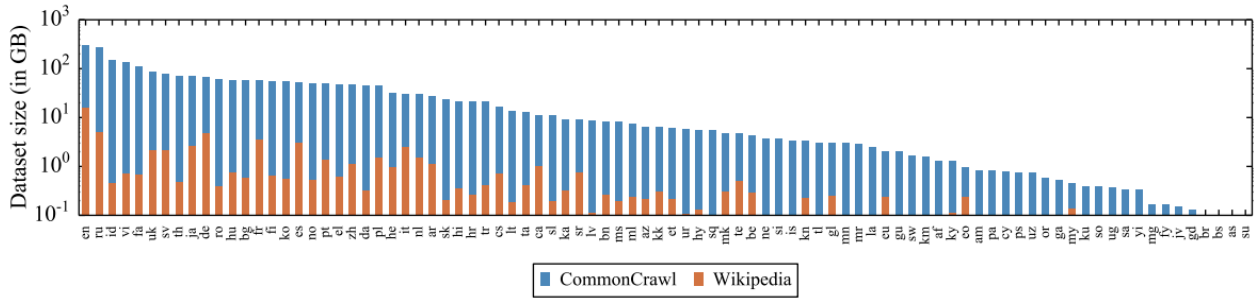


Figure 2.4: Data In Gb (Log-Scale) For 88 Languages That Appear In The Wiki-100 Corpus Used For Mbert (Conneau Et Al., 2020)

2.3.1.2 Cross-Lingual Language Model with RoBERTa architecture (XLM-R)

Meta's XLM-R, an improvement over mBERT, was trained on 100 languages with over two terabytes refined/filtered CommonCrawl (CC), as opposed to Wikipedia data. Common Crawl corpus (CC-100) addresses many of mBERT's limitations by offering stronger performance in multilingual environments. (Noo, 2021).

XLM-R shows particular strength in low-resource languages, outperforming mBERT on a variety of cross-lingual benchmarks. For instance, on Cross-lingual Natural Language Inference (XNLI) — an evaluation corpus for language transfer and cross-lingual sentence classification in 15 languages —, XLM-R outperformed mBERT by 14.6% average accuracy, 13% average F1 score on MLQA, and 2.4% F1 score on NER. (Conneau et al., 2019; facebookresearch, 2023).

2.3.1.3 Capsule Networks

Capsule Networks (CapsNets) have been employed with word embeddings like BERT to detect fake news on social platforms such as X (formerly known as Twitter). CapsNets maintain spatial relationships within text and are designed to identify subtle contextual cues, improving rumor detection accuracy by 6-7% over traditional deep learning methods. This

approach ensures that even nuanced manipulations within disinformation campaigns are more easily captured.

2.3.2 Formula and Performance Metrics

Performance of these models in multilingual settings is often evaluated using Accuracy and F1-score, calculated as:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2.4 Gaps in the Research

After a concise review of past studies in subsection 2.4 and 2.5 several gaps were identified.

Bastos and Mercea (2019), and Yang et al. (2023) faced a data accessibility challenge on X (formerly known as Twitter). X (formerly known as Twitter)'s limitations on access to individual account data with researchers makes studying bots on that platform very hard as one must fetch data from social media platforms before they can perform the evaluation. This situation was made worse with the removal of open-source API access on X (formerly known as Twitter), leading to the cancellation or suspension of many ongoing research projects relying on X (formerly known as Twitter) data. Further, given the rapidly evolving nature of political discourse, especially during elections, existing models are often limited by outdated data and struggle to detect real-time shifts in disinformation patterns (Hadi et al., 2023). This research aims to fill the gap by developing real-time monitoring capabilities.

Many models and datasets for disinformation detection are trained on English, with minimal work done on African languages or multilingual political environments like Kenya's. This indicates the need for customised detection frameworks that capture linguistic nuances in languages such as Swahili and vernacular Kenyan dialects**.**

Prior studies have either focused on individual methods — Sentiment Analysis (SA), Social Network Analysis (SNA), or LLMs — but none have explored how these models can be integrated effectively, which this study aims to explore.

2.5 Operationalisation of Research Objectives

To operationalise the objectives of this study based on the gaps identified the researcher utilised data from the social media platform X (formerly known as Twitter), despite its current limitations on data accessibility. The choice of X (formerly known as Twitter) data is critical as it is a significant platform for political discourse, providing insights into the dynamics of information dissemination and engagement

Models like mBERT and XLM-R were utilised to analyse language-specific disinformation strategies. The aim was to enhance detection capabilities by leveraging the models' ability to process and analyse multilingual data, addressing the current underrepresentation of African languages in existing models. Thereafter, frameworks that allow for real-time monitoring of disinformation patterns were implemented, enabling swift responses to evolving narratives.

Operationalising the research objectives offered valuable insights for policymakers and electoral bodies on effective monitoring and intervention strategies to protect the integrity of democratic processes.

Table 2.1: Operationalisation Of Research Objectives

Research Objective (RO)	**Aspect**	**Measure**	**Test**
RO1. Analyse the structural characteristics of disinformation networks on social media platforms during the 2022 Kenyan General Elections using LLMs.	Structural characteristics of disinformation networks	Quantify the relationships and influence of various actors in the disinformation network.	Node centrality. Degree distribution. Clustering coefficients. Modularity. Pearson/Spearman correlation.

RO2. Examine the mechanisms through which key actors within disinformation networks influence the flow and virality of content using LLMs.	Influence of key actors	Assess the volume and engagement metrics (likes, shares, retweets) of content disseminated by identified key actors.	Regression analysis.
RO3. Explore the potential synergies between SA, SNA and LLMs in enhancing disinformation detection and protecting the integrity of democratic processes across languages.	Algorithms	Algorithm	Best Model from Evaluation of: Logistic Regression. Naïve Bayes model. KNN. Support Vector Machine (SVM). Random Forest (RF). Transformer models.
RO3 Continued (evaluation)	Language Understanding	Number of correctly classified	Accuracy, Precision, Recall,

	Across Contexts (LLMs)	multilingual disinformation messages across various languages.	and F1 Score of the LLM models in identifying disinformation patterns.

Source: (Researcher, 2024)

Chapter 3: Methodology

3.1 Introduction

The aim of this study was to develop a machine-learning model for detecting hate speech in text-based content, specifically tweets. This chapter outlines the methodological approach taken to achieve this objective, detailing the research design, target population, data collection methods, data analysis techniques, ethical considerations, and study limitations.

The section provides an overview of the data science pipeline, from dataset acquisition and preprocessing to model training, evaluation, and deployment. Emphasis is placed on the parameters guiding text analysis, acknowledging the exclusion of other media forms, such as images and videos, which are also commonly used in disinformation campaigns. By systematically presenting each step, this chapter seeks to bridge the gaps identified in the previous chapter and ensure transparency in the model development process.

3.2 Study Design

This study utilised the Cross Industry Standard Process for Data Mining (CRISP-DM) model, a well-established methodology used for performing data mining projects. The CRISP-DM model provides a structured approach to handling various data mining tasks and allows for a deeper understanding of the underlying data. This study follows the six key phases of CRISP-DM model, highlighted in Figure 3.1 below, ensuring a methodical progression from data collection to insight generation.

The first phase referred to as Business Understanding, involves defining the objectives and requirements of the project. The second phase, Data Understanding, focuses on data acquisition and exploration. This involves gathering the relevant datasets, examining their structure and content, and identifying any potential issues or inconsistencies. This phase provides valuable insights that inform the subsequent data preparation and modelling stages. The third phase, Data Preparation, involves pre-processing and transforming the data to ensure its quality and suitability for modelling. This involves tasks such as data cleansing, feature engineering, and data integration, handling any missing, inconsistent, or duplicate data to create a clean and structured dataset that is conducive to accurate and efficient modelling. This stage is crucial, as the quality of the prepared data directly impacts the

performance of the data mining models. Modelling is the fourth phase, involves selecting and applying appropriate data mining techniques to the prepared data. It involves experimenting with various algorithms and methods to find the most suitable models for the specific problem and data at hand, fine-tuning the model parameters to optimise performance. To assess the models' effectiveness, a range of evaluation metrics and cross-validation techniques are used to ensure that the chosen models generalise well to unseen data.

Once the models have been constructed and refined, there are evaluated to assess the model performance against the research objectives. This involves testing the models on a separate dataset and measuring their performance using relevant evaluation metrics. This phase is crucial for ensuring that the models meet the project's goals and provide valuable insights for decision-making. The final phase, Deployment, involves integrating the data mining models into the business processes and systems. By carefully planning and executing the models' deployment, the researcher ensures that the project's objectives are realised, and the insights gained through data mining are effectively utilised to inform business decisions and strategies.

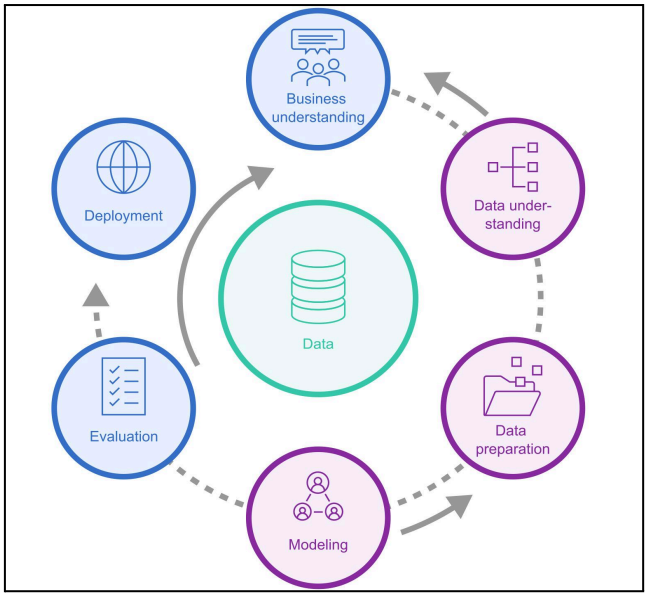


Figure 3.1: CRISPM-DM Model by Hotz (2018)

3.3 Data Collection and Sources

This study utilised secondary data sourced by Code for Africa (CfA) for research purposes. In the context of social media, specifically focusing on X (formerly known as Twitter), (González-Bailón et al., 2014) highlighted two types of biases introduced by API-based data

collection: the first, known as first-order bias, pertains to the representativeness of tweets returned by the API, as the selection process may not accurately reflect the entire X (formerly known as Twitter) ecosystem; while the second, known as second-order bias, arises when sampled data inadequately captures the intensity and structure of communication between agents and interpreters (Wardle & Derakhshan, 2017), leading to distortions in the analysis of interactions and influence within the platform. These biases are particularly relevant in studies like this that analyse online discourse, where missing data can impact the reliability of network-based inferences.

To mitigate both types of biases, our search query was guided by a HateLex — a 23-term Kenyan political hate speech lexicon —, developed by The National Cohesion and Integration Commission (NCIC) in collaboration with the Mapema Coalition (Abbreviations and Acronyms FGD Focus Group Discussions NCIC National Cohesion and Integration Commission NCI Act National Cohesion and Integration Act, n.d.). By employing this predefined HateLex, the study ensured that relevant data was systematically captured while minimising the risk of overlooking critical content due to API filtering limitations. Furthermore, the HateLex-based approach helped standardise the data retrieval process, reducing inconsistencies that might arise from arbitrary keyword selection.

The CfA database used for this study contains approximately 101,000 publicly available and relevant tweets for the period between January 1, 2022, and December 31, 2022. This dataset was collected using Meltwater, a media intelligence platform that aggregates publicly available content from social media platforms, online news sources, blogs, and forums using proprietary web crawling and API integrations. Meltwater applies natural NLP and artificial intelligence (AI) to filter, categorise, and analyse large volumes of text-based data. Through its access to historical and real-time data streams, Meltwater facilitated the systematic collection of tweets relevant to the study's focus, ensuring comprehensive data coverage.

To ensure balanced representation across different time periods, stratified sampling was applied by selecting a subset of tweets from each month. Stratified sampling involves dividing the population into homogeneous, mutually exclusive groups (strata), and randomly selecting a sample from each group (stratum) to maintain proportionality (Bhattacherjee, 2012). This sampling approach was employed to capture variations in discourse before, during, and after the 2022 Kenyan General Elections. The final analytical dataset consisted of

73,321 tweets, distributed across the three electoral phases, ensuring comprehensive temporal coverage.

By applying this stratified sampling approach and leveraging Meltwater as a data collection tool, the study accounted for fluctuations in online engagement, political discourse intensity, and the emergence of key disinformation narratives over time. This methodological framework not only reduced sampling bias but also improved the generalisability of findings, providing a more accurate representation of the digital information landscape surrounding the elections. As a result, the dataset offers valuable insights into the structural characteristics of disinformation networks, the role of political hate speech in online discussions, and the broader implications for digital disinformation.

3.4 Data Pre-Processing

Since we make inferences from the data, data preparation is a crucial step in ensuring the dataset is clean, consistent, and ready for analysis. This step involved handling missing values, duplicates, and standardising the dataset for optimal model performance (Marczyk, DeMatteo, & Festinger, 2010).

For this study, putting various treatment options into consideration, we handled duplicates as well as missing values that might affect the overall representation of the population.

3.4.1 Data Inspection and Exploration

This process involves inspecting the dataset to gain an understanding of its structure, attributes, and underlying patterns. This step entails examining summary statistics (e.g.,mean, median, standard deviation) to identify potential anomalies, data entry errors and potential sources of noise

The dataset used consisted of structured metadata fields that provide essential context for each tweet. The relevant variables are outlined in Table 3.1 below.

Table 3.1: Relevant Tweet Metadata Variables And Description

Variable	**Description**
Date	Date when the tweet was posted
Headline	Main headline or summary of the tweet
URL	Link to the actual tweet
Opening Text	First few words or introductory text of the tweet
Hit Sentence	Main text (the tweet/news content)
Source	Platform or publication source of the tweet
Influencer	X (formerly known as Twitter) username that shared the tweet
Country	Country where the tweet originated
Subregion	Specific subregion or state related to the tweet
Language	Language in which the tweet was written
Sentiment	Classification of tone of the message (positive, neutral, or negative)
Reach	Estimated number of people who viewed the tweet
Desktop Reach	Estimated number of viewers on desktop devices
Mobile Reach	Estimated number of viewers on mobile devices
Twitter Social Echo	Engagement level of the tweet on X (formerly known as Twitter)
Facebook Social Echo	Engagement level of the tweet on Facebook
Reddit Social Echo	Engagement level of the tweet on Reddit
National Viewership	Number of views the tweet received nationally

Engagement	Total interactions (likes, shares, comments) with the tweet.
AVE	Advertising Value Equivalency, estimating media value of the tweet
Key Phrases	Extracted key phrases summarising tweet content
Input Name	Name of the dataset input or source
Keywords	Keywords associated with the tweet content
Twitter Authority	Influence score of the X (formerly known as Twitter) account
Tweet ID	Unique identifier for the tweet
Twitter ID	Unique identifier for the X (formerly known as Twitter) user
Twitter Client	Application used to post the tweet (e.g., X (formerly known as Twitter), Web App, Android)
Twitter Screen Name	Public display name of the X (formerly known as Twitter) user
User Profile URL	Link to the X (formerly known as Twitter) user's profile
Twitter Bio	Biography or description provided by the X (formerly known as Twitter) user
Twitter Followers	Number of followers the X (formerly known as Twitter) user has
Twitter Following	Number of accounts the user follows
Alternate Date Format	Additional format of the tweet's posting date
Time	Time when the tweet was posted
State	State or province associated with the tweet
City	City associated with the tweet
Document Tags	Tags or labels assigned to the tweet for categorisation

3.4.2 Treatment for Missing Values

During data preparation, missing values were systematically handled to maintain the dataset's integrity and ensure accurate analysis. The *pandas* module in Python was employed to manage missing data efficiently. In the “Sentiment” column, only a few values were missing (less than 5%). Given their minimal impact, these missing entries were simply dropped using the ‘dropna()’ function rather than being imputed. This decision ensured that only complete and reliable sentiment data were used in subsequent analyses, preventing any potential biases from incorrect estimations.

3.4.3 Handling Duplicates

Data manipulation in this study involved handling duplicates, particularly the “Hit Sentence" column, which contained tweet content. The approach to handling duplicates varied depending on the analytical method used.

When building the LLM, it was crucial to ensure that duplicate messages did not distort the training data. The first step in duplicate handling involved filtering out retweets. Since retweets replicated the original tweet without modification, any entry where the "Hit Sentence" began with ‘RT’ was removed. This prevented multiple counts of the same message from skewing sentiment classification.

However, in the context of Social Network Analysis (SNA), duplicates were retained. Retweets and repeated messages played a critical role in understanding how information spread across the network. These duplicates were preserved in the dataset used to construct the network graph in Gephi, as they provided valuable insights into the flow of disinformation, influence patterns, and user interactions. By keeping these repeated messages, the study was able to analyse how specific tweets were amplified, which users played key roles in dissemination, and the overall structure of the disinformation network.

3.4.4 Data Type Conversion

The next step was to convert data from one type to another to ensure consistency and compatibility analytical methods. This is achieved by identifying incompatible data types then selecting appropriate conversion methods. Date and time were standardised to a panda’s

format for easier manipulation during analysis. Other conversions included converting integer to floats.

3.4.5 Data Labeling

The labeling of hate speech in the dataset was conducted using a lexicon-based approach. A predefined HateLex - hate speech lexicon - containing keywords and phrases associated with hate speech was used to categorise tweets.

Each tweet was analysed for the presence of these keywords, and if a match was found, it was assigned a corresponding hate speech label. This method allowed for automated and consistent annotation of the data, ensuring that tweets containing hate-related terms were flagged for further analysis.

However, while lexicon-based labeling provided a systematic and scalable way to classify content, it had limitations, such as context misinterpretation (e.g., sarcasm or reclaimed slurs). To mitigate this, additional manual verification and contextual analysis were conducted to refine the labels and improve accuracy.

3.4.6 Data Balancing

After labeling the data, we observed a significant class imbalance, as illustrated in Figure 3.2 below. To mitigate this, we experimented with various balancing techniques, including 50-50 oversampling using SMOTE (Synthetic Minority Over-sampling Technique). However, this approach led to poor training performance and heightened overfitting, as the synthetic sentiment samples did not generalise well to unseen data.

```
# Check label distribution
print(df["label"].value_counts())

<ipython-input-17-dd7e23d5bc20>:28: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/st
  df["label"] = df["Hit Sentence"].apply(auto_label)
label
1    16889
0     3734
Name: count, dtype: int64
```

Figure 3.2: Distribution Of Tweet Labels (Hate Speech = 1, Not Hate Speech =0)

Since undersampling was not a viable option due to potential data loss, we opted for a minor oversampling strategy. Specifically, we increased the number of "Not Hate Speech" samples to 10,000 to achieve a more balanced distribution while minimising the risk of overfitting.

3.5 Feature Transformation

Once the dataset is cleaned, feature transformation is applied to convert raw data into a format suitable for analysis. This includes encoding categorical variables, scaling numerical features, and selecting the most relevant attributes for modeling.

3.5.1 Categorical Variable Encoding

Machine learning models often require categorical data to be represented numerically. In this dataset, the categorical variable, "Sentiment" (Positive, Negative, Neutral), was encoded using Label Encoding, where each sentiment category was assigned a unique numerical value (Negative = 0, Positive = 1, Neutral = 2).

3.5.2 Feature Scaling

The study employed feature scaling to ensure numerical features were on a similar scale, typically within predefined range such as [0,1] or [-1,1], preventing models from being biased toward variables with larger magnitudes. Min-Max Scaling was applied to transform values between 0 and 1, while Standardisation (subtracting the mean and dividing by the standard deviation) was used to normalise distributions. Both transformations were implemented using

the sklearn.preprocessing module in Python ensuring consistency and reproducibility in the scaling process.

3.5.3 Feature Selection

Selecting the most informative features helps improve model performance and reduces computational complexity. In this dataset, the most relevant attributes for analysis included Date, Hit Sentence, Influencer, Country, Language, Source, and Sentiment. Features that do not contribute significantly to the analysis were dropped to streamline processing. Additionally, feature engineering techniques such as extracting key phrases or sentiment intensity were applied to enhance model performance.

To ensure consistency, text normalisation techniques such as lowercasing, stopword removal, and lemmatisation were applied to standardise the text data. Additionally, tweets were tokenised for further processing, and labeled according to predefined hate speech classification categories.

3.6 Exploratory Data Analysis

Exploratory Data Analysis (EDA) involves coming up with summaries and visualisations of sample distributions that allow us to better understand and make inferences about the entire population (Rahmany et al., 2020).

Generally, EDA is conducted to obtain an overall representation of the population from the analysis and examining the central measures of tendency of sample data. The study focuses on both univariate and multivariate analysis of the data to check for skewed distributions as well as outliers within the data. This is where descriptive and correlational statistics were obtained to aid in the selection of significant variables within the assessments, behavioural and demographic aspects for the predictive analysis.

Descriptive statistics assist researchers in data description and establish relationships between variables (Marczyk, DeMatteo, & Festinger, 2010). We assessed the relationship between the various predictor/independent variables (engagement metrics, such as the number of posts, likes, shares, comments, and retweets), with the dependent variable (virality) using Pearson/Spearman correlation.

The insights gained from EDA guided the development and fine-tuning of the LLM, sentiment analysis models, and the final dashboard.

3.6.1 Machine Learning Algorithms

There are several machine learning techniques that were employed to perform predictive-based tasks. This study sequentially looks into the following algorithms; Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), Random Forests and Transformer Models. These models were used to perform classification tasks.

3.6.1.1 Logistic Regression

Logistic Regression is a supervised learning algorithm used for classification. In this study we'll use it to identify whether content is likely disinformation or not (binary classification).

In Logistic Regression, predictions are made by calculating the chance of an event occurring from the linear combination of one or more explanatory variables (Sperandei, 2014). Typically a logistic regression applied for this study is represented as;

$$log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

where,

p represents the probability,

X_j represents the predictor variables/input features, and

β_i represents the coefficients associated to the impact brought about by each explanatory variable.

3.6.1.2 Naïve Bayes (NB)

NB is a supervised classification technique based upon the Baye's Theorem (Messinis & Hatziargyriou, 2018). It assumes the independence of variables.

Mathematically, Bayes theorem can be written like this:

$P(Y/X) = \frac{P(X/Y) \bullet P(Y)}{P(X)}$ where;

$P(Y/X)$ is the posterior probability i.e probability of target given the predictor.

$P(X/Y)$ is the likelihood i.e the probability of the predictor given the target variable.

$P(Y)$ is the prior probability of the predictor.

$P(X)$ is the prior probability of the target.

3.6.1.3 Support Vector Machines (SVM)

The main objective of SVM is creating a hyperplane with the biggest separation between support vectors in the data used in this study. We aim at maximising the distance between the different classes (e.g., disinformation or not) so as to minimise the classification error. (Berwick, n.d.; Sumbatilinda, 2024).

The optimal hyperplane is obtained by;

$y = w \cdot x + b$ where;

w represents a normal vector and

b is some offset

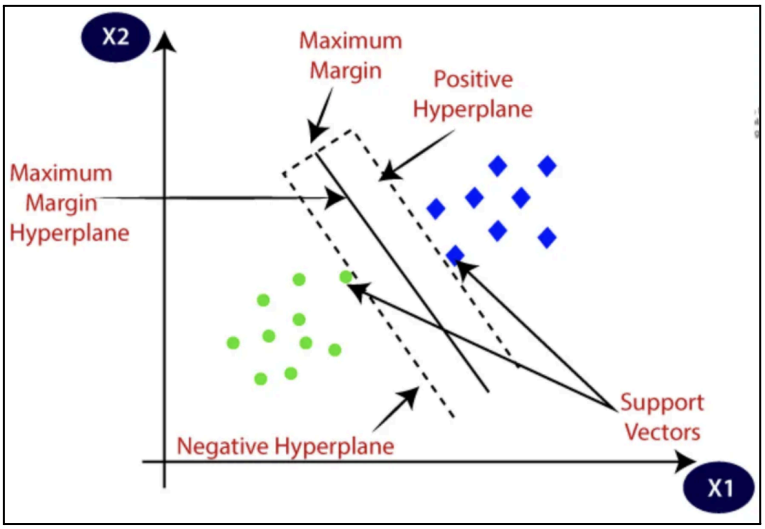


Figure 3.3: Illustration Of The Concept Of SVM (Sumbatilinda, 2024)

3.6.1.4 Random Forest (RF)

RF is an ensemble learning algorithm for classification (Belgiu & Lucian Drăguţ, 2016) suitable for predicting content virality and distinguishing influential actors within networks.

A RF will predict the likelihood of content becoming viral based on engagement metrics (likes, retweets, shares) and identify the most influential actors or accounts in a network.

RF relies on aggregation of multiple decision trees:

$\widehat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(X)$ where;

$T_m(X)$ is the prediction of the *m-th* tree, and

M is the number of trees in the forest.

3.6.1.5 Transformer Models

Transformer Models are a type of deep learning architecture, such as BERT (Devlin et al., 2024), XLM-R (Conneau et al., 2019), and mBERT that tracks relationships in sequential input, to capture long-range dependencies and contextual relationships within the data (Merritt, 2022; Ferrer, 2024).

These models trained on NLP to identify and classify disinformation will be fine-tuned to detect disinformation across multilingual data and assess real-time shifts in sentiment.

The core of transformer models operation is mathematically expressed in this way:

$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V$ where;

Q is the query matrix,

K is the key matrix,

V is the value matrix, and

$\sqrt{d_k}$ is the scaling where d_k is the dimension of the keys is employed to standardise the vectors to unit variance for ensuring stable softmax gradients during training (Pallagani et al., 2024).

3.6.2 Performance Evaluation

The models applied for this study will be subjected to evaluation using the following evaluation metrics; Accuracy, Precision, Recall, and F1-Score.

3.6.2.1 Accuracy

Accuracy measures the ratio of correct predictions over the total number of instances evaluated (Hossin & Sulaiman, 2015). It indicates how well an algorithm correctly identifies both positive and negative cases out of all cases.

Accuracy is expressed mathematically as;

$$Accuracy\ (Acc) = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{True\ Positive\ (TP) + True\ Negative\ (TN) + False\ Positive\ (FP) + False\ Negative\ (FN)}$$

3.6.2.2 Precision

Precision is among the most applied metrics for evaluating the performance of classification algorithms. It measures the fraction of the positive cases that are correctly predicted from the total positive instances (Hossin & Sulaiman, 2015; Geetha et al., 2024). Specifically, Precision focuses on minimising false positives — cases where content is wrongly classified as disinformation.

For this study, Precision will show how accurately the model identifies disinformation within the collected data, relative to all the flagged cases.

Mathematically it is represented as;

$$Precision\ (P) = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)}$$

3.6.2.3 Recall

Recall, is a measure of the proportion of positive cases that were correctly classified (Hossin & Sulaiman, 2015; Geetha et al., 2024). Recall focuses on minimising false negatives — instances where real disinformation is not detected.

For this study, it will measure the ability of the predictive models to correctly identify all relevant disinformation cases from the dataset, even if subtle or hidden within multilingual contexts.

The formula for is ;

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

3.6.2.4 F1-Score

The F1-Score is the weighted harmonic mean of Precision and Recall, giving a balanced measure of the model's accuracy (Hossin & Sulaiman, 2015; Geetha et al., 2024).

In our study, F1-Score will help us address situations where both false positives and false negatives are critical.

It is mathematically represented as;

$$F_1 - Score = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

ACTUAL VALUES	POSITIVE	NEGATIVE
POSITIVE	TP	FN
NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

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

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Figure 3.4: A Representation Of The Confusion Matrix (Research Gate, 2023)

3.7 Model Deployment

3.7.1 Introduction

The ultimate goal of the study is to develop an effective system for detecting disinformation in real-time on social media platforms. This system will leverage advanced machine learning models, including LLMs and classical algorithms like Naive Bayes, to identify and classify content.

3.7.2 Model Implementation

Once the machine learning models are developed, they will be deployed for real-time detection of disinformation on social media platforms like X (formerly known as Twitter). To ensure efficient deployment, the machine learning models will be dockerised to streamline the production environment. Dockerisation involves encapsulating the machine learning code, libraries, and environment into a container using Docker.

This will ensure that the deployment is consistent and can be easily replicated across different systems.

The deployment will be achieved through a REST API, which will allow efficient communication between the model and external platforms. We will create an API service that will receive input data (such as posts or tweets) from X (formerly known as Twitter) and send the output (predictions regarding disinformation) back to the user in real-time.

3.7.3 Data and Results Dissemination

The data used for this study was stored in a data warehouse for public reuse. Additionally, the data and machine learning code were uploaded to my GitHub repository. Public access to the repository will be read-only, ensuring data integrity. Collaborators can request additional access by contacting the repository administrator.

For result dissemination, a dashboard was developed where the results of the model's predictions are displayed in an interactive dashboard. This dashboard will be accessible to authorised users, enabling them to monitor disinformation trends and gain insights. The

dashboard allows for real-time tracking of posts and provide summaries on disinformation patterns. In order to process incoming data efficiently, the Dockerised model run continuously, monitoring social media activity. The model assesses various engagement metrics like the number of posts, likes, shares, and comments, and apply sentiment analysis and other machine learning techniques to detect disinformation.

Through effective deployment, the model provides ongoing, real-time detection and classification of disinformation campaigns, contributing valuable insights for managing and mitigating the spread of harmful content on social media.

3.7.4 Data Utilisation

Effective model deployment is critical to ensuring that the insights gained from this research are utilised in practical applications to protect the integrity of democratic processes.

In particular, the key findings were utilised to inform policy recommendations, enhance media literacy, and improve platform moderation.. Collaborative efforts with policymakers and fact-checking organisations ensures the broad application of findings across multilingual environments.

3.8 Ethical Considerations

Confidentiality is one of the key ethical considerations that researchers are advised to observe (Marczyk et al., 2010). When collecting data from social media, researchers must ensure that personal information is anonymised or pseudonymised to protect the identities of users (Bos, 2020). This study maintained confidentiality by utilising only publicly available data. This is especially important in the context of disinformation, where individuals may be targeted or affected by harmful content.

Chapter 4: System Implementation and Testing

4.1 Introduction

This chapter discusses the implementation and testing of the hate speech detection system using Hugging Face's transformer models, specifically BERT and mBERT. The focus is on the practical deployment of these models in detecting election related disinformation on social media.

4.2 Hugging Face Implementation Using mBERT

4.2.1 Model Selection

The Hugging Face library was used to deploy the mBERT model, leveraging its pre-trained models and powerful NLP tools for detecting disinformation and hate speech.

Despite the high performance of both BERT and mBERT models, mBERT was specifically chosen for its multilingual capabilities, a critical feature given the multilingual nature of the Kenyan context during the election period. BERT, on the other hand, is highly effective for single-language applications, especially English. However they both struggled with low resource languages like Kenya's local dialect.

4.2.2 Model Integration

Once the model was loaded, it was integrated into the analysis pipeline, where it processed the input i.e. tweets to generate predictions. The text was tokenised into the appropriate format before being passed through the mBERT model, which outputs a classification result 'Hate Speech' or 'Not Hate Speech'.

4.2.3 Performance Metrics

The mBERT model was evaluated using precision, recall, and F1-score, with results showing near-perfect classification accuracy. With this high performance, the model demonstrated exceptional reliability in detecting hate speech and disinformation.

The model's effectiveness was evaluated using key performance metrics like Precision and Recall, which were well-balanced, indicating low false positives and false negatives.

However, the exceptionally high scores suggest potential overfitting, likely due to similarities between the training and evaluation datasets. To address this, further testing on an unseen dataset from a different distribution is necessary to assess the model's generalisability.

Table 4.1: Comparison of mBERT and BERT's Performance

	Accuracy	**Precision**	**Recall**	**F1-Score**
mBERT	99.94%	99.94%	99.94%	99.94%
BERT	99.94%	99.94%	99.94%	99.94%

4.2.4 User Interface Implementation

The user interface (UI) of the system was developed to ensure ease of access and interaction for users, particularly content moderators, journalists, and researchers. It serves as the main point of interaction between the user and the hate speech detection system.

The interface contains a text input fields for users to paste content for classification into the system for real-time classification. The display of classification outcome, includes whether the content is classified as "Hate Speech" or "Not Hate Speech". Alongside the classification is the associated confidence score, allowing users to easily understand the results.

Given the use of mBERT, the interface can handle input in multiple languages, providing classification results for content in English, Swahili, and other languages relevant to the Kenyan context.

4.2.5 System Functionality Testing

The functionality testing of the system involved conducting unit tests to ensure its proper operation under various conditions. These tests focused on verifying the key components of the system, with particular emphasis on the model prediction functionality.

One test case involved validating that the model consistently output predictions within the expected categories, such as "Hate Speech," and "Not Hate Speech." Controlled test data was provided to the model, and the outputs were compared to manually labeled data to ensure they aligned with the expected results.

Another test case ensured that the model could handle a variety of input types, including short phrases, long sentences, and multi-paragraph texts, while still providing accurate predictions. This was crucial for ensuring the model's flexibility and ability to process different types of content effectively.

Finally, the prediction time of the model was tested to confirm that it operated within acceptable limits. This was particularly important for real-time applications, such as social media monitoring, where timely analysis of data is essential. The test results confirmed that the model was able to provide predictions promptly, making it suitable for use in real-time systems.

To address limitations in existing corpora and improve contextual accuracy in Kenya's multilingual setting, the system was enhanced with a HateLex corpus. This lexicon enriched the dataset with locally relevant hate speech terms and coded language often overlooked in standard datasets. Despite the dominance of English and other international languages in the dataset (as shown in Figure 4.1), HateLex helped increase the presence of local linguistic features such as politically charged phrases (e.g., “madoadoa”, “hatupangwingwi”) and ethnic slurs, thereby increasing the model’s sensitivity to nuanced and context-dependent hate speech. This integration proved critical during testing, as it enabled the model to better detect disinformation patterns unique to Kenya’s sociopolitical and linguistic environment.

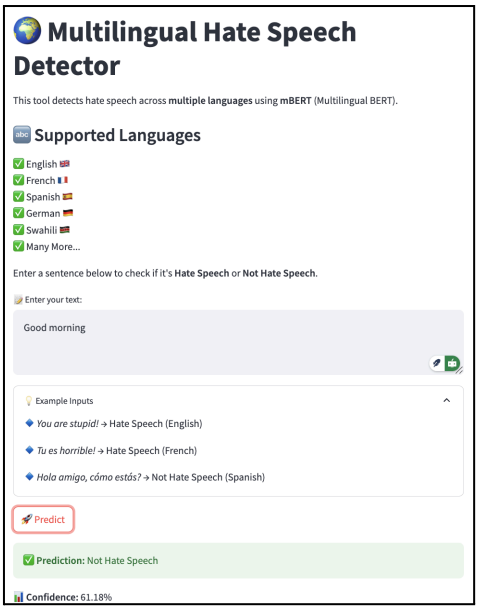


Figure 4.1: Hate Speech Detector Snippet

Chapter 5: Results and Discussion

5.1 Introduction

This chapter gives a detailed discussion of the results obtained from the findings of the study, based on the research questions investigated.

The structural characteristics of disinformation networks, the role of key actors in shaping information flow, and the potential of integrating SA, SNA, and LLMs for detecting and mitigating disinformation. The analysis leverages Gephi visualisations, frequency counts, and network metrics to uncover trends in online discourse surrounding Kenya's 2022 General Election.

5.2 Exploratory Data Analysis (EDA)

EDA was conducted to identify patterns, trends, and key features of the dataset used for training and evaluating the hate speech detection model.

The data comprised election-related tweets from the 2022 Kenyan General Elections, categorised as 'Negative' and 'Positive'. As illustrated in Figure 5.1 below, there is a marked class imbalance in sentiment distribution, with the 'Negative' (representing hate speech) class significantly outnumbering the 'Positive' (representing non-hate speech) class. This skewed distribution highlights the need for a robust disinformation detection model capable of effectively handling class imbalances, particularly to ensure accurate and fair classification.

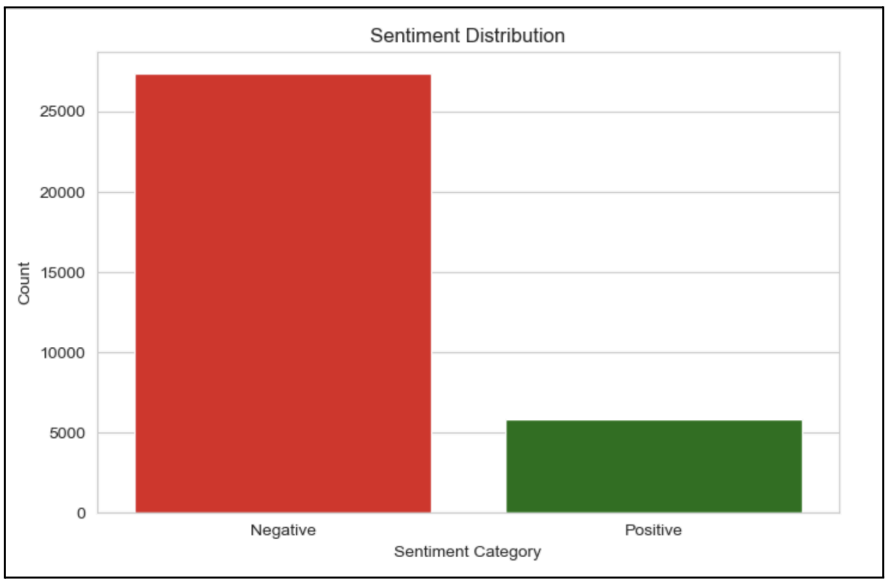


Figure 5.1: Count Of Target Variable

In addition, the tweets appeared in a total of 48 languages including English and Swahili as the top languages. This highlighted the necessity of a multilingual model like mBERT.

5.3 Model Performance Evaluation

5.3.1 Structural Characteristics of Disinformation Networks

A review of the dataset revealed that disinformation during the elections exhibited patterns of repetitive messaging, and high-retweet amplification. Specific influencers played a central role in pushing narratives that either fueled misinformation or framed discussions in a way that downplayed or justified controversial topics.

Figure 5.2 illustrates that prominent accounts such as those by political figure @migunamiguna (with over 2.1 million followers), legal commentator @ahmednasirlaw (with over 1.4 million followers), and media accounts @citizentvkenya (with over 610,000 followers) ranked among the most retweeted sources, showing that authority figures significantly influenced online discourse.

Their tweets received significant engagement, top tweets getting over 1,000 retweets, indicating a strong amplification effect. The data showed that specific influencers repeatedly pushed narratives targeting political figures or communities, often employing inflammatory or coded language. This coordinated activity suggests a deliberate attempt to manipulate public discourse through sustained and wide-reaching disinformation campaigns.

	target	Twitter Screen Name	Twitter Followers	tweet	Country	Retweets
3475	@migunamiguna	Dr. Miguna Miguna	2135053.0	If all Kenyans want is to introduce substances or things that can earn us money to pay foreign debt, here is the magic bullet: Repossess TRILLIONS of Shillings stashed abroad by the Kenyatta, Moi, Njonjo, Odinga and other plutocratic families. This will eliminate all debt.	Unknown	4367
20343	@ahmednasirlaw	Ahmednasir Abdullahi SC	1455413.0	RT @JKNjenga: Today, I've learnt that language is an ass. 3 politicians used the word madoadoa in the last two months. 1. Nyandarua Governor Amos Kimemia. 2. ODM Leader Raila Odinga 3. Meru Senator Mithika Linturi For Linturi, it is hate speech, but for Kememia and Raila, it is context.	Kenya	3161
1073	@itsmutai	Lord Abraham Mutai	610114.0	RT @DonaldBKipkorir: In attempts to revise own personal records, uda leadership repeatedly tries to portray AZIMIO as "watu wa kurusha mawe" and "watu wa kung'oa reli", euphemism for Luo: dog-whistling to drive a tribal wedge. But record shows those with real violent record of murder, rape & arson.	Kenya	1878
2569	@citizentvkenya	Citizen TV Kenya	4258492.0	"Don't kill the goose that lays the golden eggs," Raila Odinga on the closure of Keroche Breweries. https://t.co/90rhEGHqg1	Kenya	1830
12042	@williamsruto	William Samoei Ruto, PhD	4640398.0	Though we are competitors,to subject HE Kalonzo to some humiliating 'interview' is impunity. We must unite to eliminate the culture of political deceit,the hallmark of some politicians. Whatever the circumstances every leader deserves some dignity and respect. Heshima si utumwa. https://t.co/oJ3q4UJhks	Kenya	1736
41136	@susankihika	Sen. Susan Kihika	492670.0	For the avoidance of doubt, hapa ndio tuko! Nahatusongi! Ignore FAKE propaganda by DESPERATE bloggers! Ruto the 5th! Hatupangwingwi! https://t.co/jHepA7Cdfx	Kenya	1299
14420	@mutahingunyi	Mutahi Ngunyi	1814503.0	Livondo is beginning to SOUND like Jacob Juma: DEAD man WALKING. But is he WRONG? I did a HYPOTHESIS on the same in 2020 after a PSYCHO called David Ndii advised Ruto to Kill his Boss. Read on! #LivondoLeaks https://t.co/1gvVl9JE4y	Kenya	1230
8266	@robertalai	Robert ALAI	1783144.0	William Ruto will kill his bodyguards. This is wrong! These are human beings. https://t.co/xh7MGxwqhV	Kenya	1026
60	@thriving_luos	DHOLUO DICTIONARY	3638.0	RT @MigunaMiguna: Dr Robert Ouko knew that Nicholas Biwott and Dictator Daniel arap Moi were plotting to Kill him. Pio Gama Pinto, Tom Joseph Mobya and JM Kariuki knew that Kenyatta and Njonjo were plotting to assassinate them. But they did not expose those plots. I'll expose all plots to kill me.	Unknown	853
1722	@oleitumbi	Dennis Itumbi, HSC	1675181.0	Hatutishwi. Hatupangwingwi. Nairobi today stood against the Azimio, Raila Odinga violence! They have decided the Next Governor is @SakajaJohnson The Next President is @WilliamsRuto! Ruto declared Kenya does not belong to Three Families! Freedom is Coming... https://t.co/IPsnf0NTjr	Kenya	788

Figure 5.2: Repetitive Messaging And High-Retweet Amplification

5.3.2 Influence of Key Actors and Manipulation Techniques

The most retweeted individuals in the dataset also had a big following, averaging 1.9 million followers as shown in Figure 5.3 below. Therefore, their influence in shaping political discourse cannot be understated, as their messages reach large audiences who either adopt or challenge the narratives being presented.

```
# Select the top 10 retweeted accounts
top_retweeted = _influencers.nlargest(10, 'Retweets')

# Calculate the average number of followers for these accounts
average_followers = top_retweeted['Twitter Followers'].mean()

# Print the result
print(f"Average Twitter Followers for Top Retweeted Accounts: {average_followers:.2f}")
```

```
Average Twitter Followers for Top Retweeted Accounts: 1886860.60
```

Figure 5.3: Average X Following

The data further reveals instances where neutral sentiments were used to reframe discussions, rather than overtly spreading false information. This suggests that some influencers engage in narrative shaping — a subtle form of disinformation where language is used strategically to sway public perception without making direct false claims.

Additionally, the sentiment analysis of viral tweets highlights instances where disinformation is gradually reframed over time. Some messages begin with strong negative sentiment but later shift toward neutral or even positive tones, a technique commonly used to sustain engagement while avoiding outright misinformation flags by platforms.

The data also underscores the importance of coded language in disinformation, where specific words or phrases are used as dog whistles to communicate political or ethnic biases without direct hate speech violations. The observation that some messages justify terms such as ‘madoadoa’ and ‘hatupangwingwi’ while others condemn it suggests subtle manipulation tactics in framing, as referenced in Figure 5.4. Applying LLMs trained on Swahili, English, and Sheng enhanced the detection of such subtle framing techniques such as word substitutions, coded language, or shifts in sentiment particularly in multilingual contexts.

	target	Twitter Screen Name	tweet	Twitter Followers	Sentiment	Retweets
3475	@migunamiguna	Dr. Miguna Miguna	If all Kenyans want is to introduce substances or things that can earn us money to pay foreign debt, here is the magic bullet: Repossess TRILLIONS of Shillings stashed abroad by the Kenyatta, Moi, Njonjo, Odinga and other plutocratic families. This will eliminate all debt.	2135053.0	Negative	4367
20343	@ahmednasirlaw	Ahmednasir Abdullahi SC	RT @JKNjenga: Today, I've learnt that language is an ass. 3 politicians used the word madoadoa in the last two months. 1. Nyandarua Governor Amos Kimemia. 2. ODM Leader Raila Odinga 3. Meru Senator Mithika Linturi For Linturi, it is hate speech, but for Kememia and Raila, it is context.	1455413.0	Neutral	3161
1073	@itsmutai	Lord Abraham Mutai	RT @DonaldBKipkorir: In attempts to revise own personal records, uda leadership repeatedly tries to portray AZIMIO as “watu wa kurusha mawe” and “watu wa kung’oa reli”, euphemism for Luo: dog-whistling to drive a tribal wedge. But record shows those with real violent record of murder, rape & arson.	610114.0	Negative	1878
2569	@citizentvkenya	Citizen TV Kenya	“Don’t kill the goose that lays the golden eggs,” Raila Odinga on the closure of Keroche Breweries. https://t.co/90rhEGHqg1	4258492.0	Neutral	1830
12042	@williamsruto	William Samoei Ruto, PhD	Though we are competitors,to subject HE Kalonzo to some humiliating 'interview' is impunity. We must unite to eliminate the culture of political deceit,the hallmark of some politicians. Whatever the circumstances every leader deserves some dignity and respect. Heshima si utumwa. https://t.co/oJ3q4UJhks	4640398.0	Neutral	1736
41136	@susankihika	Sen. Susan Kihika	For the avoidance of doubt, hapa ndio tuko! Nahatusongi! Ignore FAKE propaganda by DESPERATE bloggers! Ruto the 5th! Hatupangwingwi! https://t.co/jHepA7Cdfx	492670.0	Neutral	1299
14420	@mutahingunyi	Mutahi Ngunyi	Livondo is beginning to SOUND like Jacob Juma: DEAD man WALKING. But is he WRONG? I did a HYPOTHESIS on the same in 2020 after a PSYCHO called David Ndii advised Ruto to Kill his Boss. Read on! #LivondoLeaks https://t.co/1gvVI9JE4y	1814503.0	Negative	1230
8266	@robertalai	Robert ALAI	William Ruto will kill his bodyguards. This is wrong! These are human beings. https://t.co/xh7MGxwqhV	1783144.0	Negative	1026
60	@thriving_luos	DHOLUO DICTIONARY	RT @MigunaMiguna: Dr Robert Ouko knew that Nicholas Biwott and Dictator Daniel arap Moi were plotting to Kill him. Pio Gama Pinto, Tom Joseph Mobya and JM Kariuki knew that Kenyatta and Njonjo were plotting to assassinate them. But they did not expose those plots. I'll expose all plots to kill me.	3638.0	Negative	853
1722	@oleitumbi	Dennis Itumbi, HSC	Hatutishwi. Hatupangwingwi. Nairobi today stood against the Azimio, Raila Odinga violence! They have decided the Next Governor is @SakajaJohnson The Next President is @WilliamsRuto! Ruto declared Kenya does not belong to Three Families! Freedom is Coming... https://t.co/lPsnf0NTjr	1675181.0	Neutral	788
1139	@udapartyke	U D A - Official Fans Page.	...and giving it 36s of coverage 2. EXTREMELY COMPROMISED criminal justice department.. - The kind that sends 100 police officers to pick Linturi for asking the people of Eldoret to vote in UDA candidates but cheers Raila for telling his supporters to reject madoadoa. 2/7	119075.0	Negative	785

Figure 5.4: Neutral Sentiments Used To Reframe Discussions

5.3.3 Integrating Sentiment Analysis, Social Network Analysis, and LLMs

Figure 5.5 shows a SNA mapped retweet clusters using the hashtag #RutoThe5thPresident. The hashtag garnered over 16,000 mentions within 12 hours with a primary network of 39 accounts contributing to approximately 30% of all mention. This pushing of the same message within short timeframes indicates a high likelihood this hashtag was amplified through coordinated behaviour on social media by different accounts. However, there were incidents where X (formerly known as Twitter) proactively removed accounts that participated in coordinated posting, even before the analysis of these networks. For example,

as reported on various mainstream media sources, around 500 accounts on X (formerly known as Twitter) that posted using the hashtag were suspended by X (formerly known as Twitter).

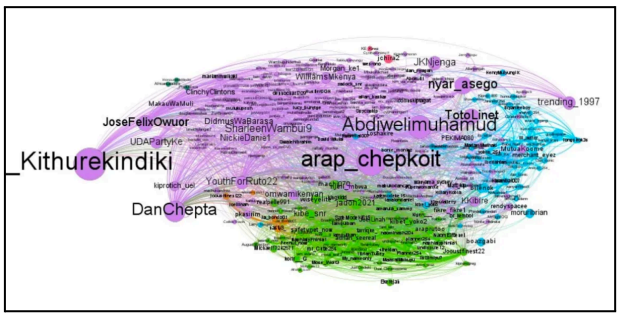


Figure 5.5: A Gephi Social Network Map Of X Accounts Using #RutoThe5thPresident

Chapter 6: Conclusion, Recommendation and Future Work

6.1 Introduction

This chapter provides a summary of the research findings, addresses the implications for various stakeholders, and proposes areas for further study. The study analysed the structural characteristics of disinformation networks, explored how LLMs detect language-specific patterns in these networks, and examined the synergies between Sentiment Analysis (SA), Social Network Analysis (SNA), and LLMs to disinformation detection across multiple languages.

6.2 Implications of the Findings

The findings of this study have significant implications for various stakeholders, including policy makers, journalists, researchers, and data solution providers, in the fight against disinformation. By addressing these implications, stakeholders can develop more robust, ethical, and data-driven approaches to countering disinformation in the evolving digital landscape.

For policy makers, the study underscores the need for evidence-based regulations to address digital misinformation while safeguarding freedom of speech. The high accuracy of mBERT suggests that AI-driven tools can play a crucial role in content moderation and policy enforcement, but ensuring transparency and fairness in their implementation is essential.

For journalists and fact-checkers, the findings highlight the importance of integrating AI-powered tools into fact-checking processes to enhance the speed and accuracy of misinformation detection. Since disinformation spreads rapidly, leveraging machine learning models for early detection can help newsrooms proactively address false narratives before they gain widespread traction. Additionally, AI-driven systems can provide automated early warnings on trending misinformation, enabling fact-checkers to prioritise verification efforts and strengthen the credibility of news reporting.

For researchers, the study provides valuable insights into the structural characteristics of disinformation networks and the role of LLMs in detecting language-specific patterns. Future

research can build upon these findings by exploring multimodal analysis, weakly supervised learning, and ethical AI applications to improve disinformation detection across diverse contexts.

For data solution providers, the study emphasises the need for diverse and continuously updated datasets to improve the performance of AI-driven disinformation detection systems. The findings suggest that incorporating real-time monitoring, multilingual capabilities, and network-based insights can enhance the effectiveness and adaptability of such solutions.

6.3 Challenges Encountered

One of the primary challenges encountered in this study was the issue of overfitting, as both bert and mBERT models demonstrated exceptionally high accuracy in detecting hate speech and disinformation. While these results suggest strong model performance, they also suggest overfitting, as the model struggled when applied to unseen data. Addressing this limitation requires further evaluation using more diverse datasets and incorporating recent linguistic trends to improve adaptability.

Another significant challenge was accessing high-quality labeled data, particularly for underrepresented languages. The dynamic nature of online discourse, characterised by the rapid emergence of slang, coded language, and evolving disinformation tactics, made it difficult to keep the dataset updated. This challenge has been further compounded by the recent changes to X (formerly known as Twitter)'s API policies, which have restricted access to real-time social media data, limiting the ability to monitor and capture new disinformation patterns effectively. Additionally, fine-tuning large models like mBERT requires substantial computational resources, which can be a barrier to scalability.

6.4 Recommendations

To ensure that mBERT and similar models remain relevant, adaptable, and effective in addressing the evolving landscape of digital disinformation, continuous updates to the dataset are necessary to reflect emerging disinformation trends, evolving linguistic patterns, and new terminology such as "Kasongo" and other contemporary slang or coded language used in misinformation campaigns. This can be achieved through active monitoring of social media

conversations, fact-checking reports, and real-time data collection from diverse sources to ensure that the model remains relevant in detecting new disinformation strategies.

Given the scarcity of labeled data for low-resource/underrepresented languages, semi-supervised and weakly supervised learning approaches should be explored to generate additional training data with minimal manual annotation. Techniques such as self-training, distant supervision, and weak labeling from existing fact-checking databases could help improve LLM adaptation to low-resource languages.

To mitigate overfitting, future model iterations could incorporate regularisation techniques, adversarial training, and exposure to more diverse datasets to improve robustness. Evaluating the model's performance on unseen data from various sources will help assess its generalisability and effectiveness beyond the training set.

Furthermore, research should consider the ethical implications of AI-driven content moderation, including how automated disinformation detection impacts free speech, bias, and fairness across different demographic groups. Ensuring that AI tools remain transparent, accountable, and fair will be crucial in balancing effective disinformation mitigation with the protection of fundamental rights.

6.5 Future Research

Given Kenya's linguistic diversity, disinformation often shifts between English, Swahili, and Sheng, making it difficult to track narratives using traditional keyword-based approaches. Future research should focus on improving the adaptability of LLMs to low-resource languages and regional contexts by expanding training datasets and testing the models on data from diverse sources. This would help address overfitting concerns and improve the model's ability to detect disinformation across different linguistic and cultural landscapes.

Another promising avenue is multimodal disinformation analysis, which involves integrating text-based models like mBERT with image and video analysis techniques to detect disinformation more comprehensively, since disinformation often spreads through multiple formats like memes and deep fakes.

Additionally, exploring semi-supervised and weakly supervised learning approaches could help overcome the scarcity of labeled data, particularly for low-resource languages.

Techniques such as self-training, distant supervision, and weak labeling could enable models to learn from limited annotated examples while improving performance in detecting emerging disinformation trends.

References

A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers. (2017). Arxiv.org. https://arxiv.org/html/2405.10936v1

Abbreviations and Acronyms FGD Focus Group Discussions NCIC National Cohesion and Integration Commission NCI Act National Cohesion and Integration Act. (n.d.). Retrieved May 8, 2024, from https://cohesion.go.ke/images/docs/downloads/Hatelex_A_Lexicon_of_Hate_Speech_Terms_In_Kenya.pdf

Anol Bhattacherjee. (2025). *Social Science Research: Principles, Methods, and Practices*. Digital Commons @ University of South Florida. https://digitalcommons.usf.edu/oa_textbooks/3/#:~:text=This%20book%20is%20designed%20to,3.

APA PsycNet. (2024). Apa.org. https://psycnet.apa.org/record/1997-06270-001

Arias, F., Nunez, M. Z., Guerra-Adames, A., Nathalia Tejedor-Flores, & Vargas-Lombardo, M. (2022). Sentiment Analysis of Public Social Media as a Tool for Health-Related Topics. *IEEE Access*, *10*, 74850–74872. https://doi.org/10.1109/access.2022.3187406

Arkaitz Zubiaga, Aker, A., Kalina Bontcheva, Liakata, M., & Procter, R. (2018). Detection and Resolution of Rumours in Social Media. *ACM Computing Surveys*, *51*(2), 1–36. https://doi.org/10.1145/3161603

Asma Ul Hussna, Alam, R., Islam, R., Bader Fahad Alkhamees, Hassan, M. M., & Uddin, M. Z. (2024). Dissecting the Infodemic: An In-Depth Analysis of COVID-19 Misinformation Detection on X (Formerly Twitter) Utilizing Machine Learning and Deep Learning Techniques. *Heliyon*, *10*(18), e37760–e37760. https://doi.org/10.1016/j.heliyon.2024.e37760

Atefeh Farzindar, & Inkpen, D. (2015). Natural Language Processing for Social Media. *Synthesis Lectures on Human Language Technologies*, *8*(2), 1–166. https://doi.org/10.2200/s00659ed1v01y201508hlt030

Baines, D., & Elliott, R. J. R. (2020, April 21). *Defining misinformation, disinformation and malinformation: An urgent need for clarity during the COVID-19 infodemic* .

Research Gate. https://www.researchgate.net/profile/Darrin-Baines/publication/341130695_Defining_misinformation_disinformation_and_malinformation_An_urgent_need_for_clarity_during_the_COVID-19_infodemic/links/5eb01d1b299bf18b9594b28f/Defining-misinformation-disinformation-and-malinformation-An-urgent-need-for-clarity-during-the-COVID-19-infodemic.pdf

Bastos, M. T., & Mercea, D. (2019). *The Brexit Botnet and User-Generated Hyperpartisan News - Marco T. Bastos, Dan Mercea, 2019*. Social Science Computer Review. https://journals.sagepub.com/doi/full/10.1177/0894439317734157

Belgiu, M., & Lucian Drăguţ. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Berwick, R. (n.d.). *An Idiot's guide to Support vector machines (SVMs) R. Berwick, Village Idiot SVMs: A New Generation of Learning Algorithms*. https://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf

Bhutani, B., Rastogi, N., Sehgal, P., & Purwar, A. (2019). Fake News Detection Using Sentiment Analysis. *2019 Twelfth International Conference on Contemporary Computing (IC3)*. https://doi.org/10.1109/ic3.2019.8844880

Bos, J. (2020). Confidentiality. *Springer EBooks*, 149–173. https://doi.org/10.1007/978-3-030-48415-6_7

Cherilyn, I., & Julie, P. (Eds.). (2024). *Journalism, fake news & disinformation*. Unesco.org. https://unesdoc.unesco.org/ark:/48223/pf0000265552

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Unsupervised Cross-lingual Representation Learning at Scale*. ArXiv.org. https://arxiv.org/abs/1911.02116

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., & Ai, F. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. https://arxiv.org/pdf/1911.02116

Das, N., Bikash Sadhukhan, Chatterjee, R., & Chakrabarti, S. (2024). Integrating sentiment analysis with graph neural networks for enhanced stock prediction: A comprehensive survey. *Decision Analytics Journal*, *10*, 100417–100417. https://doi.org/10.1016/j.dajour.2024.100417

Dawson, S., Aneesha Bakharia, & Heathcote, L. (2019). *SNAPP : Realising the affordances of real-time SNA within networked learning environments*. Australian Institute of Business. https://research.aib.edu.au/en/publications/snapp-realising-the-affordances-of-real-time-sna-within-networked

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2024). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. ArXiv.org. https://arxiv.org/abs/1810.04805

Duzen, Z., Riveni, M., & Aktas, M. S. (2023). Analyzing the Spread of Misinformation on Social Networks: A Process and Software Architecture for Detection and Analysis. *Computers*, *12*(11), 232. https://doi.org/10.3390/computers12110232

Dwi Surjatmodjo, Andi Alimuddin Unde, Hafied Cangara, & Alem Febri Sonni. (2024). Information Pandemic: A Critical Review of Disinformation Spread on Social Media and Its Implications for State Resilience. *Social Sciences*, *13*(8), 418–418. https://doi.org/10.3390/socsci13080418

Elhag, M., Kharman, A., Balakrishnan, V., & Abdelaziz, A. (2018). Sentiment analysis algorithms: evaluation performance of the Arabic and English language. *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*. https://doi.org/10.1109/iccceee.2018.8515844

Eliza, a chatbot therapist. (2018). Njit.edu. https://web.njit.edu/~ronkowit/eliza.html#:~:text=By%20then%2C%20ELIZA%20was%20a,communications%20was%20at%20that%20time.

facebookresearch. (2023). *GitHub - facebookresearch/XNLI: Evaluating Cross-lingual Sentence Representations*. GitHub. https://github.com/facebookresearch/XNLI

Feingold, S., & World Economic Forum. (2022, July 14). *What are spam bots? All you need to know*. World Economic Forum. https://www.weforum.org/agenda/2022/07/spam-bots-what-are-they/

Feng, S., Shi, W., Wang, Y., Ding, W., Balachandran, V., & Tsvetkov, Y. (2024). *Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration*. ArXiv.org. https://arxiv.org/abs/2402.00367

Ferrer, J. (2024, January 9). *How Transformers Work: A Detailed Exploration of Transformer Architecture*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/how-transformers-work

Geetha, A. V., Mala, T., Priyanka, D., & Uma, E. (2024). Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. *Information Fusion*, *105*, 102218. https://doi.org/10.1016/j.inffus.2023.102218

Gentzkow, M., & Allcott, H. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://ideas.repec.org/a/aea/jecper/v31y2017i2p211-36.html

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, *38*, 16–27. https://doi.org/10.1016/j.socnet.2014.01.004

Hadi, M. U., qasem al tashi, Qureshi, R., Shah, A., amgad muneer, Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Seyedali Mirjalili. (2023). *A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage*. https://doi.org/10.36227/techrxiv.23589741.v1

Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01-11. https://doi.org/10.5121/ijdkp.2015.5201

Hughes, H. C., & Waismel-Manor, I. (2020). The Macedonian Fake News Industry and the 2016 US Election. *PS, Political Science & Politics*, *54*(1), 19–23. https://doi.org/10.1017/s1049096520000992

Humprecht, E. (2020). *Resilience to Online Disinformation: A Framework for Cross-National Comparative Research - Edda Humprecht, Frank Esser, Peter Van Aelst, 2020*. The International Journal of Press/Politics. https://journals.sagepub.com/doi/10.1177/1940161219900126

Ingers, J. (2024). *From Text to Truth: Integrating Large Language Models for Accurate Fake News Detection A in-depth study of different Deep Learning techniques*. https://kth.diva-portal.org/smash/get/diva2:1887927/FULLTEXT01.pdf

Jain R. (2024). *Exploring the Efficacy of Natural Language Processing and Supervised Learning in the Classification of Fake News Articles*. *2*(1), 1–6. https://doi.org/10.23880/art-16000108

Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research an International Electronic Journal*, *18*(1). https://www.researchgate.net/publication/285954290_A_social_diffusion_model_of_

misinformation_and_disinformation_for_understanding_human_information_behaviour

Kothari, A., Walker, K., & Burns, K. (2022). #CoronaVirus and public health: the role of social media in sharing health information. *Online Information Review*, *46*(7), 1293–1312. https://doi.org/10.1108/oir-03-2021-0143

Kuru, O., Campbell, S. W., Bayer, J. B., Baruh, L., & Ling, R. (2022). *Encountering and Correcting Misinformation on WhatsApp*. 88–107. https://doi.org/10.1002/9781119714491.ch7

Large Language Model Agent for Fake News Detection. (2017). Arxiv.org. https://arxiv.org/html/2405.01593v1

Lawton, G. (2011). In the News. *IEEE Intelligent Systems*, *26*(3), 5–9. https://doi.org/10.1109/mis.2011.53

Liu, B. (2017). Sentiment Analysis and Opinion Mining. *SpringerLink*. https://doi.org/10.1007-978-3-031-02145-9

Lynch, G. (2023). *Hybrid rallies and a rally-centric campaign: the case of Kenya's 2022 elections*. *61*(3), 334–356. https://doi.org/10.1080/14662043.2023.2232160

Lynch, G., Cheeseman, N., & Willis, J. (2019). From peace campaigns to peaceocracy: Elections, order and authority in Africa. *African Affairs*. https://doi.org/10.1093/afraf/adz019

Madung, O. (2022, June 8). *New Research: Disinformation on TikTok Gaslights Political Tensions Ahead of Kenya's 2022 Elections*. Mozilla Foundation. https://foundation.mozilla.org/en/blog/new-research-disinformation-on-tiktok-gaslights-political-tensions-ahead-of-kenyas-2022-elections/

Majerczak, P., & Strzelecki, A. (2022). Trust, Media Credibility, Social Ties, and the Intention to Share towards Information Verification in an Age of Fake News. *Behavioral Sciences*, *12*(2), 51–51. https://doi.org/10.3390/bs12020051

Marczyk, G. R., DeMatteo, D., & Festinger, D. (2010). *Essentials of Research Design and Methodology*. John Wiley & Sons. https://www.wiley.com/en-br/Essentials+of+Research+Design+and+Methodology-p-9780471470533

Marczyk, G., DeMatteo, D., & Festinger, D. (2015). *Essentials of Research Design and Methodology TEAM LinG -Live, Informative, Non-cost and Genuine !* (A. S. Kaufman & N. L. Kaufman, Eds.).

https://rlmc.edu.pk/themes/images/gallery/library/books/Behavioral%20Science/Essentials%20of%20Research%20design%20&%20Methdology.pdf

Merritt, R. (2022, March 25). *What Is a Transformer Model?* NVIDIA Blog. https://blogs.nvidia.com/blog/what-is-a-transformer-model/

Messinis, G. M., & Hatziargyriou, N. D. (2018). Review of non-technical loss detection methods. *Electric Power Systems Research*, *158*, 250–266. https://doi.org/10.1016/j.epsr.2018.01.005

Mohammad, S. M. (2017). Challenges in Sentiment Analysis. *Socio-Affective Computing*, 61–83. https://doi.org/10.1007/978-3-319-55394-8_4

Mohammad, S. M., Salameh, M., & Kiritchenko, S. (2016). How Translation Alters Sentiment. *Journal of Artificial Intelligence Research˜the œJournal of Artificial Intelligence Research*, *55*, 95–130. https://doi.org/10.1613/jair.4787

Morstatter, F., & Liu, H. (2017). Discovering, assessing, and mitigating data bias in social media. *Online Social Networks and Media*, *1*, 1–13. https://doi.org/10.1016/j.osnem.2017.01.001

National Council for Law Reporting with the Authority of the Attorney-General. (2010). *THE CONSTITUTION OF KENYA, 2010*. http://www.parliament.go.ke/sites/default/files/2023-03/The_Constitution_of_Kenya_2010.pdf

Noo, S. (2021). Sentiment analysis of Korean movie reviews using XLM-R. *International Journal of Advanced Culture Technology*, *9*(2), 86–90. https://doi.org/10.17703/IJACT.2021.9.2.86

Nur, & Aniza, I. (2021). Multilingual Sentiment Analysis: A Systematic Literature Review. *Pertanika Journal of Science & Technology*, *29*(1). https://doi.org/10.47836/pjst.29.1.25

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., B. Reeves, B., & Robinson, T. N. (2023). Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking. *Communication Methods and Measures*, *18(2)*, 124–141. https://doi.org/10.1080//19312458.2023.2181319

Onur Varol, Ferrara, E., Davis, C. A., Filippo Menczer, & Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. *Proceedings of the ... International AAAI Conference on Weblogs and Social Media/Proceedings of*

the International AAAI Conference on Weblogs and Social Media, *11*(1), 280–289. https://doi.org/10.1609/icwsm.v11i1.14871

Pallagani, V., Roy, K., Muppasani, B., Fabiano, F., Loreggia, A., Murugesan, K., Srivastava, B., Rossi, F., Horesh, L., & Sheth, A. (2024). *On the Prospects of Incorporating Large Language Models (LLMs) in Automated Planning and Scheduling (APS)*. Arxiv.org. https://arxiv.org/html/2401.02500v2/#Sx3

Park, C. Y., Mendelsohn, J., Field, A., & Tsvetkov, Y. (2022). *Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media* (pp. 5238–5264). https://aclanthology.org/2022.findings-emnlp.382.pdf

Pires, T., Schlinger, E., & Garrette, D. (2019). *How multilingual is Multilingual BERT?* (pp. 4996–5001). https://aclanthology.org/P19-1493.pdf

Qazvinian, V., Rosengren, E., Radev, D., & Mei, Q. (2011). *Rumor has it: Identifying Misinformation in Microblogs* (pp. 1589–1599). Association for Computational Linguistics. https://aclanthology.org/D11-1147.pdf

Rahmany, M., Zin, A. M., & Sundararajan, E. A. (2020). COMPARING TOOLS PROVIDED BY PYTHON AND R FOR EXPLORATORY DATA ANALYSIS. *IJISCS (International Journal of Information System and Computer Science)*, *4*(3), 131–131. https://doi.org/10.56327/ijiscs.v4i3.933

Rastogi, S., & Bansal, D. (2022). A review on fake news detection 3T's: typology, time of detection, taxonomies. *International Journal of Information Security*, *22*(1), 177–212. https://doi.org/10.1007/s10207-022-00625-3

Research Gate. (2023). *Confusion matrix: Precision, Recall, Accuracy, and F1 score.* ResearchGate. https://www.researchgate.net/figure/Confusion-matrix-Precision-Recall-Accuracy-and-F1-score_fig4_367393140

Rosalie Li, E. (2021, July 20). *Information Basics: Mis-, Dis-, and Malinformation*. Hoaxlines Lab; Hoaxlines Lab. https://infoepi.org/posts/2021/07-16-basics

Sanctis, V. D., Soliman, A. T., Daar, S., Ploutarchos Tzoulis, Fiscina, B., Christos Kattamis, & Thalassemia, I. (2022). Retrospective observational studies: Lights and shadows for medical writers. *PubMed*, *93*(5), e2022319–e2022319. https://doi.org/10.23750/abm.v93i5.13179

Serrat, O. (2017). Social Network Analysis. *Springer EBooks*, 39–43. https://doi.org/10.1007/978-981-10-0983-9_9

Shikali, C. S., & Mokhosi, R. (2020). Enhancing African low-resource languages: Swahili data for language modelling. *Data in Brief*, *31*, 105951. https://doi.org/10.1016/j.dib.2020.105951

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media. *SIGKDD Explorations*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Sivasankari, S., & Vadivu, G. (2021). Tracing the fake news propagation path using social network analysis. *Soft Computing*, *26*(23), 12883–12891. https://doi.org/10.1007/s00500-021-06043-2

Soroush Vosoughi, Roy, D., & Sinan Aral. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 12–18. https://doi.org/10.11613/bm.2014.003

Sumbatilinda. (2024, May). *Support Vector Machine (SVM) Algorithm. - Sumbatilinda - Medium*. Medium. https://medium.com/@sumbatilinda/support-vector-machine-svm-algorithm-064566b5d411

The Achilles Heel of Large Language Models: Lower-Resource Languages Raise More Safety Concerns. (2024). Arxiv.org. https://arxiv.org/html/2401.13136v1#:~:text=Surprisingly%2C%20while%20training%20with%20high,rooted%20in%20the%20pretraining%20stage.

The spread of true and false news online. (2018). Science. https://www.science.org/doi/10.1126/science.aap9559

Toloka. (2024). *Why Integrating Low Resource Languages Into LLMs Is Essential for Responsible AI*. Toloka.ai. https://toloka.ai/blog/low-resource-languages/

Towards Measuring and Modeling "Culture" in LLMs: A Survey. (2024). Arxiv.org. https://arxiv.org/html/2403.15412v1

Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sergey Sanovich, Stukal, D., & Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *Social Science Research Network*. https://doi.org/10.2139/ssrn.3144139

Twitter Terms of Service. (2018). Twitter.com. https://twitter.com/en/tos/previous/version_13#:~:text=You%20may%20not%20do%20any,network%20or%20breach%20or%20circumvent

twitter.com. (2023). *X Privacy Policy*. Twitter.com; twitter-com. https://twitter.com/en/privacy

Ukkonen, A., & Tiedemann, J. (2023). *HELSINKI UNIVERSITY FACULTY OF ARTS DEPARTMENT OF DIGITAL HUMANITIES Linguistic Feature Analysis of Real and Fake News: Human-written vs. Grover-written Ottilia Nikula 014841705 Master's Programme in Linguistic Diversity and Digital Humanities*. https://helda.helsinki.fi/server/api/core/bitstreams/f33f83e8-33e3-4993-899e-5f978f963ff0/content

V.S. Subrahmanian, Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Filippo Menczer, Stevens, A., Dekhtyar, A., Gao, S., Hogg, T., Farshad Kooti, Liu, Y., Onur Varol, Prashant Shiralkar, Vinod Vydiswaran, & Mei, Q. (2016). The DARPA Twitter Bot Challenge. *Computer*, *49*(6), 38–46. https://doi.org/10.1109/mc.2016.183

Vetter, T. R., & Schober, P. (2018). Regression: The Apple Does Not Fall Far From the Tree. *Anesthesia and Analgesia/Anesthesia & Analgesia*, *127*(1), 277–283. https://doi.org/10.1213/ane.0000000000003424

Vincenzo De Sanctis, Soliman, A. T., Daar, S., Ploutarchos Tzoulis, Fiscina, B., Christos Kattamis, & Thalassemia, I. (2022). Retrospective observational studies: Lights and shadows for medical writers. *PubMed*, *93*(5), e2022319–e2022319. https://doi.org/10.23750/abm.v93i5.13179

Wardle, C. (2017, February 16). *Fake News. It's Complicated, First Draft* . 30 Zuckerman, E. https://firstdraftnews.com/fake-newscomplicated/

Wardle, C., & Derakhshan, H. (2017). *Information Disorder : Toward an interdisciplinary framework for research and policy making*. https://rm.coe.int/information-disorder-report-2017/1680766412

Wu, S., & Dredze, M. (2020). Are All Languages Created Equal in Multilingual BERT? *ArXiv (Cornell University)*. https://doi.org/10.18653/v1/2020.repl4nlp-1.16

Yang, K.-C., Ferrara, E., & Menczer, F. (2022). Botometer 101: social bot practicum for computational social scientists. *Journal of Computational Social Science*, *5*(2), 1511–1528. https://doi.org/10.1007/s42001-022-00177-5

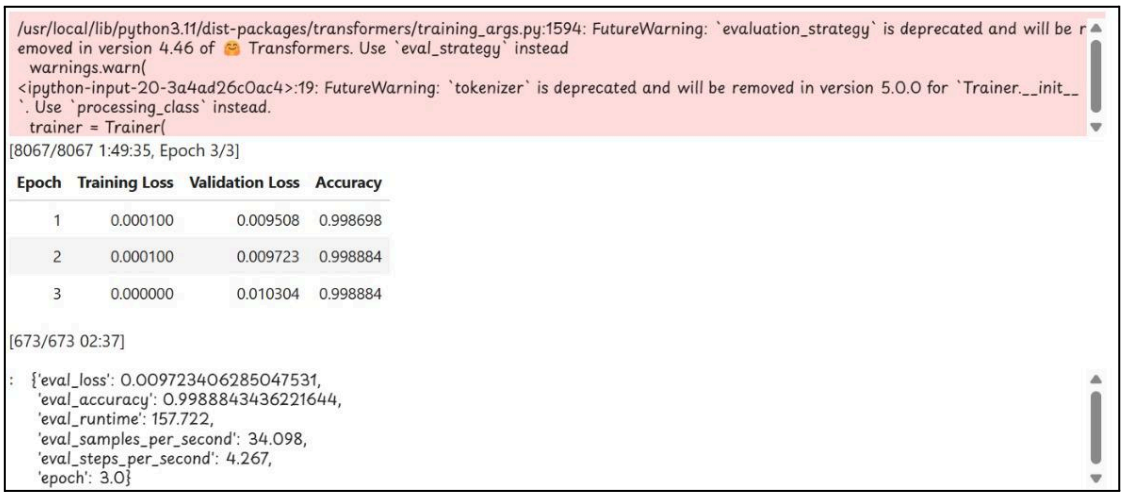
Yang, K.-C., Onur Varol, Hui, P.-M., & Filippo Menczer. (2020). Scalable and Generalizable Social Bot Detection through Data Selection. *Proceedings of the ... AAAI Conference on Artificial Intelligence*, *34*(01), 1096–1103. https://doi.org/10.1609/aaai.v34i01.5460

Yang, K.-C., Varol, O., Nwala, A. C., Sayyadiharikandeh, M., Ferrara, E., Flammini, A., & Menczer, F. (2023). *Social Bots: Detection and Challenges*. ArXiv.org. https://arxiv.org/abs/2312.17423

Appendices

Appendix A

A.1 BERT Implementation



```
/usr/local/lib/python3.11/dist-packages/transformers/training_args.py:1594: FutureWarning: `evaluation_strategy` is deprecated and will be r
emoved in version 4.46 of 🤗 Transformers. Use `eval_strategy` instead
  warnings.warn(
<ipython-input-20-3a4ad26c0ac4>:19: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Trainer.__init__
`. Use `processing_class` instead.
  trainer = Trainer(
[8067/8067 1:49:35, Epoch 3/3]
```

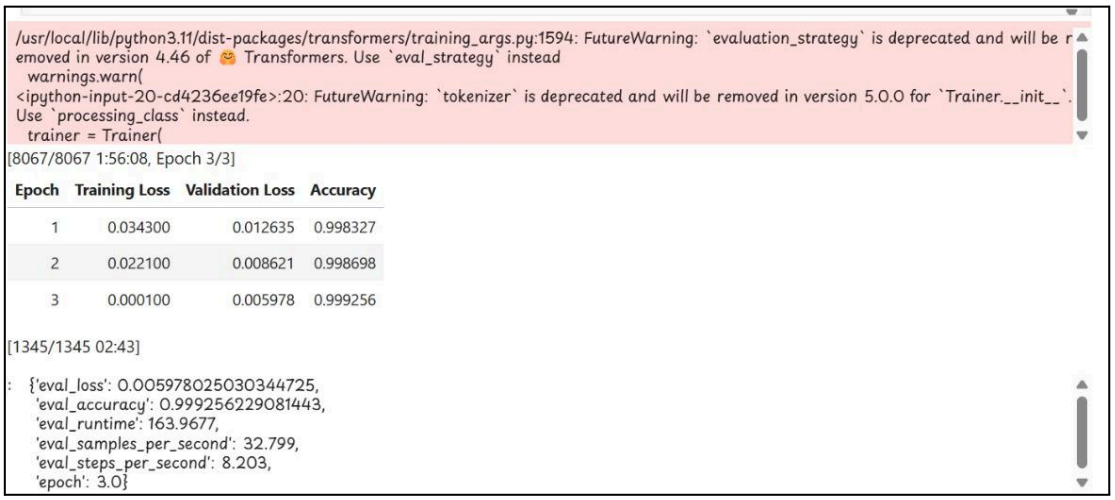
Epoch	Training Loss	Validation Loss	Accuracy
1	0.000100	0.009508	0.998698
2	0.000100	0.009723	0.998884
3	0.000000	0.010304	0.998884

```
[673/673 02:37]

:  {'eval_loss': 0.009723406285047531,
    'eval_accuracy': 0.9988843436221644,
    'eval_runtime': 157.722,
    'eval_samples_per_second': 34.098,
    'eval_steps_per_second': 4.267,
    'epoch': 3.0}
```

Figure A.1 Bert-Based Model Fine-Tuned Using The Hugging Face Transformers Library

A.2 mBERT Implementation



```
/usr/local/lib/python3.11/dist-packages/transformers/training_args.py:1594: FutureWarning: `evaluation_strategy` is deprecated and will be r
emoved in version 4.46 of 🤗 Transformers. Use `eval_strategy` instead
  warnings.warn(
<ipython-input-20-cd4236ee19fe>:20: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Trainer.__init__`.
Use `processing_class` instead.
  trainer = Trainer(
[8067/8067 1:56:08, Epoch 3/3]
```

Epoch	Training Loss	Validation Loss	Accuracy
1	0.034300	0.012635	0.998327
2	0.022100	0.008621	0.998698
3	0.000100	0.005978	0.999256

```
[1345/1345 02:43]

:  {'eval_loss': 0.005978025030344725,
    'eval_accuracy': 0.999256229081443,
    'eval_runtime': 163.9677,
    'eval_samples_per_second': 32.799,
    'eval_steps_per_second': 8.203,
    'epoch': 3.0}
```

Figure A.2 mBert-Based Model Fine-Tuned Using Hugging Face Transformers Library

Appendix B

B.1 Ethical Clearance Confirmation

Strathmore UNIVERSITY

5th December 2024

Ms Orwaru Rose,
rose.kwamboka@strathmore.edu

Dear Ms Orwaru,

RE: An Examination of the Efficacy of Sentiment Analysis, Social Network Analysis and Large language Models in Unveiling Disinformation Campaigns during the 2022 Kenyan General Elections

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2434/24.** The approval period is from **5th December 2024 to 4th December 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

B.2 Similarity Report