A Natural Language Processing Model to Detect Outbreaks of Infectious Diseases in Kenya

By

Kinara Calista Nyakerario

169237

Submitted In Partial Fulfilment of the Requirements for the Degree of Master of Science in Information Technology at Strathmore University

School of Computing and Engineering Sciences,

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

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 thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: Kinara Calista Nyakerario

Signature…  ……Date *28/05/2025*

Approval

The thesis of Kinara Calista Nyakerario was reviewed and approved for examination by the following:

Dr. Dickson Owuor,
Senior Lecturer, School of Computing & Engineering Sciences,
Strathmore University

Dr. Julius Butime,
Dean, School of Computing &Engineering Sciences,
Strathmore University

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

Abstract

Infectious diseases are more prevalent in urban settings than in rural areas. This is because the population in the metropolitan areas is higher, and several resources like public transport and social activities are shared despite the large population. The government is usually slow to react whenever there is an outbreak due to many reasons. Outbreaks of contagious diseases exert considerable pressure on the public and hospital facilities. Kenya is usually affected by epidemics of different kinds and at various times. The government gets information about outbreaks mainly through the health sector when several cases are reported in a specific region. Social media platforms enable users to share information at a global level. Data collected from these platforms can be utilised to monitor and predict disease outbreaks. X, one of the largest social media platforms in the world, hosts a vast amount of user-generated content, most of which is publicly accessible. Tweets generated by Kenyan citizens will be analysed for this study, and this study explores the application of Natural Language Processing (NLP) models to detect and predict disease outbreaks in Kenya by analysing publicly available data sources such as social media. This research seeks to help the government and other stakeholders know when there is an outbreak of infectious diseases and to put up stringent measures to mitigate further spread. This study delves into the development of a natural language processing model that will extract data from social media, e.g. X. The model focuses on information related to a particular infectious disease, and it utilises machine learning techniques to process unstructured textual data, identifying patterns and signals indicative of emerging public health threats. By leveraging real-time data, the NLP model aims to provide early warnings of outbreaks, enabling quicker responses from public health authorities. The results demonstrate the model's potential in enhancing disease surveillance.

Key words: *Infectious diseases, infodemic, public health surveillance, social media, syndromic surveillance*

Table of Contents

Declaration and Approval ii
Abstract iii
Table of Contents iv
List of Figures viii
List of Tables x
List of Abbreviations/Acronyms xi
Definition of Terms xii
Acknowledgements xiv
Chapter 1: Introduction 1
1.1 Background 1
1.2 Problem Statement 2
1.3 General Objective 3
1.4 Specific objectives 3
1.5 Research Questions 3
1.6 Justification 4
1.7 Scope and Limitations 4
Chapter 2: Literature Review 5
2.1 Introduction 5
2.2 Empirical Literature 5
2.2.1 Disease Outbreak Detection 5
2.2.2 Challenges in Using NLP for Outbreak Detection 12
2.3 Theoretical Literature 13
2.3.1 NLP and the Extraction of Health-Related Information from Social Media 13
2.3.2 Sentiment Analysis for Disease Outbreak Detection 13
2.3.3 Geolocation Tagging for Localised Outbreak Detection 14
2.3.5 Recent Advances in NLP for Outbreak Detection 14
2.4 Frameworks 15
2.4.1 Detecting Topic and Sentiment Dynamics Framework 15
2.4.2 Vaccine Sentiment Framework 15
2.4.3 Sentiment Analysis and Topic Modelling Framework 15
2.4.4 Word2PLTS Framework 17
2.4.5 NLP Framework 17
2.5 Models 18
2.5.1 CNN Model 18
2.5.2 LSTM Model 19
2.5.3 BERT Model 20

2.6 Algorithms 22
2.6.1 Support Vector Machines (SVMs) 23
2.6.2 Recurrent Neural Networks (RNNs) 24
2.6.3 LSTMs 24
2.6.4 Latent Dirichlet Allocation (LDA) 25
2.6.5 Dynamic Topic Modelling (DTM) 25
2.7 Architectures and Designs 27
2.7.1 Sentiment Analysis Design 27
2.7.2 LSTM Architecture 28
2.8 Research Gap 30
2.9 Conceptual Model 30
Chapter 3: Research Methodology/Design 33
3.1 Introduction 33
3.2 Research Design / Philosophy 33
3.3 Population and Sampling 33
3.3.1 Population 33
3.3.2 Sampling 34
3.4 Data Collection/Analysis 34
3.4.1 Data Collection 34
3.4.2 Data Analysis 34
3.5 Research Quality and Reliability 36
3.6 Systems Development Methodology 37
3.7 Utilisation of Research Results 38
3.8 Dissemination of Research Results 38
3.9 Ethical Considerations 38
Chapter 4: System Analysis and Design 39
4.1 Introduction 39
4.2 Requirement Specifications 39
4.2.1 Functional Requirements 39
4.2.2 Non-Functional Requirements 40
4.3 System Architecture 40
4.4 System Design 41
4.4.1 Use Case Diagram 41
4.4.1.1 Detailed Use Case Descriptions 42
4.4.2 Class Diagram 43
4.4.3 Sequence Diagram 44
4.4.4 Database Schema 44
4.5 Wireframes 45

4.5.1 Home Wireframe 45
4.5.2 Login Wireframe....46
4.5.3 Register Wireframe 47
4.5.4 Monitor Outbreaks Wireframe....48
4.5.5 Results Wireframe....49
Chapter 5: System Implementation and Testing 51
5.1 Introduction 51
5.2 Model Components....51
5.2.1 LSTM Layers....51
5.2.2 Dropout Layers 52
5.2.2 Dense Output Layer 53
5.3 Disease Outbreak Detection Tool....53
5.3.1 Home Interface 53
5.3.2 Monitor Outbreak Interface 54
5.3.3 Detection Results Interface....55
5.4 System Implementation....55
5.4.1 Development Environment....55
5.4.2 Data Collection 56
5.4.3 Data Pre-processing 57
5.4.3.1 Data Cleaning....57
5.4.3.2 Sentiment Analysis....58
5.4.3.3 Temporal Aggregation....59
5.4.3.4 Feature Engineering 59
5.4.4.4 Data Scaling 59
5.4.4.5 Splitting Data 60
5.4.4 Training Model 60
5.4.5 API 61
5.5 Testing....61
5.5.1 Model Accuracy....61
Chapter 6: Discussions 62
6.1 Introduction 62
6.2 Challenges Faced in Outbreak Detection of Infectious Diseases 62
6.4 Existing Models and Frameworks in Detecting Outbreaks of Infectious Diseases....63
6.5 Disease Outbreak Detection Models....63
6.6 Model Validation 64
Chapter 7: Conclusion and Recommendation....65
7.1 Conclusion....65
7.2 Recommendations....65

7.3 Future work 67
7.4 Limitations 67
7.5 Research Contribution 67
References 69
Appendices 75
Appendix A: Plagiarism Report 75
Appendix B: Ethical Approval 76
Appendix C: Excerpts from the Kenya Health Policy 2014–2030 touching on the objectives of the Health Policy 77
Appendix D: Mini-Budget for Thesis Project 78

List of Figures

Figure 2.1: Sentiment Analysis and Topic Modelling Framework. 16
Figure 2.2: Word2PLTS 17
Figure 2.3: NLP Framework. 17
Figure 2.6: The BERT Model Architecture. 21
Figure 2.7: SVM Algorithm 23
Figure 2.8: Architecture of Recurrent Neural Network 24
Figure 2.9: LSTM Algorithm 25
Figure 2.10: Sentiment Analysis Design 28
Figure 2.11: LSTM Architecture 29
Figure 2.12: Conceptual Model 32
Figure 3.4: Data collection and data pre-processing 36
Figure 3.6: Agile Methodology 37
Figure 4.1: System Architecture 41
Figure 4.2: Use Case Diagram. 42
Figure 4.3: Class Diagram. 43
Figure 4.4: Sequence Diagram 44
Figure 4.5: Database Schema 45
Figure 4.6: Home Page Wireframe 46
Figure 4.7: Login Wireframe 47
Figure 4.8: Register Wireframe 48
Figure 4.9: Monitor Outbreaks Wireframe 49
Figure 4.10: Results Wireframe 50
Figure 5.1: LSTM Layers 52
Figure 5.2: Dropout Layers 53
Figure 5.3: Dense Output Layer 53
Figure 5.4: Home Page 54
Figure 5.5: Monitor Outbreaks Interface 54
Figure 5.6: Detection Results Interface 55
Figure 5.7: Data Collection 57
Figure 5.8: Data Cleaning 58
Figure 5.9: Sentiment Analysis 58
Figure 5.10: Temporal Aggregation 59
Figure 5.11: Feature Engineering 59

Figure 5.12: Data Scaling 60
Figure 5.13: Splitting Data.....60
Figure 5.14: Model Training.....60
Figure 5.15: Model Accuracy 61

List of Tables

Table 2.1: Literature Review Summary.....11
Table 2.2: Summary of Frameworks.....18
Table 2.4: Summary of Algorithms.....26
Table 2.5: Summary of Architectures and Designs.....30
Table 4.1: Description of use cases.....42

List of Abbreviations/Acronyms

ANN	- Artificial Neural Networks
API	- Application Programming Interface
BERT	- Bidirectional Encoder Representations from Transformers
CNN	- Convolutional Neural Network
HER	- Electronic Health Records
GNIP	- Global Network of Isotopes in Precipitation
GIS	-Geographic Information System
LDA	- Latent Dirichlet allocation
LSTM	- Long Short-Term Memory
NER	- Named Entity Recognition
NLP	- Natural Language Processing
WFH	- Work From Home
XLM-R	- Cross-lingual Language Model - RoBERTa

Definition of Terms

Infectious diseases:	According to the World Health Organization (WHO), infectious diseases are "diseases caused by pathogenic microorganisms, such as bacteria, viruses, parasites, or fungi; the diseases can be spread, directly or indirectly, from one person to another."
Infodemic:	The term "infodemic" refers to the widespread spread of excessive information during a pandemic (Al-Garadi et al., 2022).
Lemmatising:	Lemmatisation is the process of grouping and analysing words with similar roots, allowing for qualitative analysis (SV et al., 2022).
Public health surveillance:	It entails the ongoing and systematic process of collecting, managing, and monitoring disease data to identify patterns and trends (Pilipiec et al., 2023).
Stemming:	Stemming reduces words to their base form by removing suffixes, for example, converting 'goals' to 'goal' and 'pens' to 'pen' (SV et al., 2022).
Social media:	Social media functions as a digital tool that enables users to develop content, distribute it, and exchange messages, perspectives, multimedia content, including text, emojis, images, audio, and videos, between their connected networks (Wilson et al., 2021).

Syndromic surveillance	Syndromic surveillance is a technique that uses health-related data, such as symptom clusters, to anticipate the probability of an upcoming outbreak (Pilipiec et al., 2023).
Topic Modelling	Topic modelling is a generative statistical approach used to uncover the underlying themes present within a text (SV et al., 2022).

Acknowledgements

I want to express my sincere appreciation to everyone who played a role in the successful completion of this research project. I am especially thankful to my supervisor, Dr. Dickson Owuor, for his exceptional guidance, continuous support, and expertise throughout the process. His insightful feedback and valuable suggestions have significantly contributed to the development of this work.

I am also profoundly grateful to Strathmore University for granting access to the necessary resources and facilities that made this research possible. Additionally, I extend my heartfelt thanks to my colleagues and friends for their consistent encouragement, motivation, and assistance.

Lastly, I acknowledge my classmates' moral support and valuable input, whose suggestions enriched this research study. This work would not have been possible without the collective efforts and collaboration of everyone mentioned above. Thank you all.

Chapter 1: Introduction

1.1 Background

In Kenya, outbreaks of infectious diseases are frequent. Social media-based infoveillance, which involves analysing online data to identify disease outbreaks more quickly than traditional surveillance methods, has shown considerable promise in health-related applications (Al-Garadi et al., 2022). As an interactive digital platform, social media allows individuals to create, share, and exchange information in different forms, including text, images, audio, and videos, within online communities and networks. Seasonal influenza, for instance, is responsible for approximately half a million deaths worldwide each year (Pilipiec et al., 2023). Most infectious diseases are associated with high mortality and fatality rates because of the unavailability of vaccination for some diseases. This means that monitoring for early detection of outbreaks is very crucial. WHO reports that informal sources detect approximately 60 percent of all outbreaks (Abbood et al., 2020).

Social media and publicly available news media serve as valuable sources for the early detection of epidemics. Social media data, in particular, plays a critical role in developing syndromic surveillance systems capable of detecting infectious disease outbreaks by analysing users' spatiotemporal patterns of self-reported symptoms (Al-Garadi et al., 2022). Recent studies have highlighted that analysing social media data is among the most effective methods for predicting, controlling, and preventing health crises or pandemics (SV et al., 2022). Additionally, social media offers insights into public perceptions and experiences during pandemics and identifies factors that either facilitate or hinder efforts to control the global spread of diseases. As noted by Oyebode et al. (2022), with 3.8 billion active users worldwide, social media has become a rich and diverse data source for research across various domains, including health.

Social media offers a rich, real-time, and often publicly available data stream, providing an opportunity to complement traditional surveillance methods. Users frequently share information about their symptoms, concerns about health, and awareness of ongoing diseases, which can serve as valuable indicators of emerging outbreaks. However, the data is unstructured, informal, and noisy, requiring advanced methods to extract useful information effectively (Wickramaarachchi et al., 2020). NLP provides the tools to automatically process and analyse this unstructured text data, converting it into structured information to detect patterns in disease spread. During the COVID-19 pandemic,

researchers used social media platforms to track real-time conversations about symptoms, treatments, and the virus's spread. The research demonstrated how social media is important for detecting public health events in advance (Li et al., 2020).

X data (formerly Twitter) enables health informatics researchers to analyse social media content for infection rate estimation, disease spread monitoring, and epidemic forecasting (Alsudias & Rayson, 2021). According to multiple research findings, analysing health-oriented social media content significantly improves current public health surveillance programs. Public health surveillance exists to identify emerging diseases early so that proper prevention steps can be applied while distributing adequate health resources for disease containment (Pilipiec et al., 2023). The conventional public health surveillance system obtains information from diagnosed cases, which healthcare professionals, laboratories, emergency departments, and hospitals report. Due to its passive nature, this method gives limited timely information regarding disease development patterns.

The study developed a Natural Language Processing (NLP) model that automatically detects early signs of infectious disease outbreaks from social media posts. The model implemented multiple layers to overcome potential false-positive and false-negative issues and increase its accuracy and reliability. Multiple NLP techniques linked with Named Entity Recognition (NER) and sentiment analysis were used with relevant disease filters to analyse specific keywords, symptoms, and later geolocations. Using domain-specific lexicons and cross-referencing extracted information with reliable public health datasets to eliminate irrelevant or misleading content, such as generic health discussions or conspiracy theories, minimises false positives. Similarly, false negatives were reduced by expanding the model's capacity to detect nuanced or indirectly expressed symptoms and disease-related content, utilising pre-trained language models fine-tuned on outbreak-specific data. The model aimed to provide a robust framework for real-time outbreak detection through these measures, complementing traditional public health monitoring systems while maintaining high precision and recall in its predictions.

1.2 Problem Statement

Infectious diseases are the most common community-acquired diseases globally. Discovering outbreak-related information from the world's available data is a discouraging task. Abbood et al. (2020) note that public health surveillance systems relying on official sources to detect significant disease outbreaks are often hampered by delays. In contrast,

leveraging unofficial sources, such as websites, to identify early signs or rumours of outbreaks can offer a critical time advantage. The emergence of new infectious diseases, such as the severe acute respiratory syndrome (SARS) outbreak in Asia and Toronto, the resurgence of older diseases like tuberculosis, and the intentional release of pathogens through bioterrorism, as demonstrated by the 2001 anthrax attacks, underscore the pressing need for effective infectious disease surveillance (Wilson et al., 2021).

Rapid disease outbreak identification remains mandatory for early disease control and containment efforts. Successful traditional surveillance methods face hurdles because of late reporting information and limited geographical monitoring capabilities (Amin et al., 2023). Social media platforms X and Threads, along with Facebook, present users with real-time content that delivers important health information about symptoms, disease mentions, and public health concerns (Wickramaarachchi et al., 2020). Social media data detection of outbreak signals becomes complicated because it contains unorganized and noisy information.

1.3 General Objective

To develop and validate a model for detecting outbreaks of infectious diseases by identifying key challenges and reviewing existing predictive methods.

1.4 Specific objectives

i). To identify the challenges faced in outbreak detection of infectious diseases.

ii). To review the current models and frameworks used to detect infectious disease outbreaks.

iii). To develop an NLP model that will help detect infectious disease outbreaks.

iv). To test and validate the proposed model.

1.5 Research Questions

i). What are the challenges faced in detecting infectious disease outbreaks?

ii). What are the current models and frameworks used to detect infectious disease outbreaks?

iii). How can a model to detect outbreaks of infectious diseases be developed?

iv). How can the developed model be validated?

1.6 Justification

This study is significant because it presents a novel approach to monitoring and managing outbreaks of infectious diseases in Kenya, leveraging the real-time nature of social media. This study provides a tool that can aid the government and relevant stakeholders in responding to outbreaks more efficiently by developing a Natural Language Processing (NLP) model to analyse public sentiment and detect disease-related topics from X posts. The real-time detection of disease outbreaks can enable quicker interventions, better resource distribution, and more informed public health decisions.

The model will serve as an asset for healthcare programs and NGOs by offering a new way to monitor outbreaks and justify funding to the health sector. It will also allow organisations to make data-driven decisions and adjust their responses based on emerging trends identified through social media conversations. For scholars, this study provides a foundation for further research in public health surveillance, serving as a basis for the expansion of future studies. It also opens the door to exploring similar applications in other sectors or geographic regions, thus contributing to the broader disease surveillance and public health field.

1.7 Scope and Limitations

This study focuses specifically on Kenya, particularly on the northeastern regions, including Garissa, Wajir, and Mandera. These areas are periodically affected by outbreaks of infectious diseases, making them ideal for applying a model designed to detect such outbreaks through real-time social media analysis. The study will analyse public tweets in response to disease outbreaks, providing a tool to help government agencies, healthcare organisations, and NGOs quickly identify emerging outbreaks. This tool will support rapid response efforts and the efficient allocation of resources to mitigate the spread of diseases. However, the study is limited to using data from X, excluding other platforms such as Facebook, Instagram, or WhatsApp. Additionally, the analysis will focus solely on text-based content, excluding multimedia elements such as memes, audio, and videos, which are increasingly common on social media. The study will also be geographically limited to Kenya, relying on the data available within this context, specifically targeting regions with a history of recurring outbreaks. While the model is designed for real-time outbreak detection, the findings will be contextualised to Kenyan healthcare practices and social media behaviours, limiting its immediate applicability to other regions or countries without further modification.

Chapter 2: Literature Review

2.1 Introduction

In recent years, using Natural Language Processing (NLP) models to detect infectious disease outbreaks has garnered significant attention, driven by the growing availability of digital data and the need for rapid disease surveillance. NLP models can analyse unstructured textual data from social media, news articles, and health reports to detect early signs of disease outbreaks. Research has shown that these models can enhance traditional epidemiological methods by offering real-time insights and early warnings.

In Kenya, where infectious diseases such as malaria, cholera, and typhoid remain prevalent, using NLP for outbreak detection could improve public health responses. However, the diverse linguistic landscape, regional disparities in healthcare infrastructure, and limited access to digital health data present unique challenges for implementing such models effectively. Existing literature explores the use of NLP in public health surveillance globally, but limited research focuses specifically on low-resource settings like Kenya. This review will examine key studies on NLP models for outbreak detection, the challenges in adapting these models to Kenya's context, and the opportunities for improving early disease detection using NLP technologies.

2.2 Empirical Literature

2.2.1 Disease Outbreak Detection

Social media surveillance for disease outbreaks has gained significant attention from researchers. The research by Liang et al. (2019) studied the spread of Ebola-related information on X while examining major users responsible for message dissemination. The research team obtained Ebola-related tweets from GNIP published worldwide between March 23, 2014, and May 31, 2015. They restored retweet paths by analysing the relationship between content and follower-followee connections to perform social network analysis to identify retweet patterns. The study divided users into four groups: influential users, hidden influential users, disseminators, and everyday users according to their following and retweeting activities. The analysis revealed that 91% of retweets originated directly from the original tweet, and 47.5% of retweeting paths had a depth of 1, meaning messages were retweeted directly by followers of the original user. This indicated that broadcasting, rather than viral spreading, was the dominant method of information diffusion. Influential and hidden influential users generated significantly more retweets than disseminators and

ordinary users, with the latter relying on a viral model to extend information beyond their immediate networks. The study concluded that public health communicators could enhance their outreach by collaborating with influential and hidden users who can amplify messages beyond official health accounts. While the study was insightful in showing how information circulates, it lacked an emphasis on sentiment analysis, which could provide deeper insights into the emotional responses surrounding disease-related discussions. Influential users are key to amplifying messages, but credibility remains a significant challenge when using them as reliable sources for outbreak communication. Moreover, the study did not explore geospatial data, which could have been valuable for pinpointing specific regions affected by Ebola or other diseases.

Fu et al. (2016) studied X platform activity after the World Health Organisation declared the Zika virus a global emergency. The study produced five central themes, including social effects of the outbreak, governmental and public actions, pregnancy-related health risks, transmission pathways, and reported case numbers. While the thematic analysis was comprehensive, the study did not incorporate sentiment analysis, which would have offered a deeper understanding of the public's emotional responses to the outbreak. Furthermore, the study focused on broad themes without directly linking them to predictive outbreak detection or real-time monitoring, thus missing an opportunity to demonstrate how these themes correlate with the onset of actual outbreaks.

Krauer & Schmid (2022) researchers explored natural language processing (NLP) for developing a dataset that examined plague outbreaks. A 1908 German plague treatise by Sticker received manual annotation from the researchers to build their definitive toponym list. A performance evaluation was conducted for location data extraction between five pre-trained NLP libraries, including Google, Stanford CoreNLP, spaCy, germaNER and Geoparser. The SpaCy system achieved the best sensitivity rating of 0.92 and reached an F1 score of 0.83, while Stanford CoreNLP followed closely with a sensitivity of 0.81 and an F1 score of 0.87. The F1 score of Google NLP reached 0.72 while its sensitivity was measured at 0.78, but both Geoparser and germaNER exhibited lower performance with sensitivities of 0.41 and 0.61, respectively. The research tested Google Geocoding, Geonames, and Geoparser automated geocoding services for outbreak location accuracy. They showed poor performance in historical regions, generating correct GIS retrieval rates at 60.4%, 52.7%, and 33.8%, respectively. The study compared their plague dataset digitalisation and Biraben's re-

digitalized work to provide current information about second pandemic plague outbreak distribution patterns. However, the study primarily focused on historical data and geospatial accuracy, which limits its relevance to real-time outbreak detection. The study also demonstrated that geospatial tagging systems performed poorly when applied to historical locations, thus hindering their potential use in detecting outbreaks in modern contexts. Despite the thorough evaluation of geospatial data accuracy, the study did not discuss the integration of multilingual data, which could be a significant challenge in regions where multiple languages are spoken.

Müller et al. (2023) developed COVID-X-BERT (CT-BERT) which uses transformer technology to train on X (formerly Twitter) COVID-19 posts from a large database. CT-BERT gained its unique design for social media COVID-19 analysis and underwent testing for multiple NLP functions which included classification and question-answering with chatbot development. The researchers evaluated the model through five classification datasets where one was from the target domain while comparing its performance to BERT-LARGE. CT-BERT delivered better results than BERT-LARGE by 10–30% throughout every dataset measurement while achieving its highest performance gains within the target domain. The study authors emphasized the significance of their findings while analysing their effects on social media pandemic content analysis. Despite its improvements, the model remains limited in scope to COVID-19 and does not account for multilingualism in social media posts. Given the global nature of pandemics, the ability to process multilingual data would significantly improve the model's applicability in non-English-speaking countries or regions with diverse linguistic populations, such as Kenya. Furthermore, while the study emphasizes the importance of sentiment classification, it does not address the potential of emotion analysis, which could provide more nuanced insights into public reactions to outbreaks.

Wang et al. (2023) explored X posts from mid-sized U.S. cities with high African American populations to understand attitudes toward the COVID-19 vaccine and vaccine hesitancy. Their analysis combined NLP techniques, identifying macro topics and emotional trends, with human-driven textual analysis to create market "profiles" for these cities. The study suggested three key strategies for public health agencies: (1) empower individuals with knowledge, (2) communicate how COVID-19 differs from the flu, and (3) emphasise the long-term health risks of the virus. The analysis of Kansas City's X data highlighted a lack

of understanding regarding the severity and transmission of COVID-19, calling for targeted public health messaging in the area. However, while the study provided actionable insights for public health communication, it did not directly connect these findings to disease outbreak detection or explore how social media sentiment could be used to predict outbreaks before they occur.

Using social media data, Martín-Corral et al. (2024) merged biological and informational diffusion processes to detect early signals of influenza-like illness (ILI) outbreaks. Their study demonstrated that high-out-degree users on X, who have many followers, could serve as effective sensors for detecting ILI outbreaks. These sensors were more likely to discuss local news, politics, and government than other users. The approach, which identified a smaller, more efficient set of social media sensors, proved more operationally efficient and privacy-respecting than previous methods that required monitoring large populations. However, the study does not explain whether this model would be applicable to other diseases or in regions where social media engagement is lower or more dispersed. The approach also did not incorporate geospatial data or emotion analysis, which would have improved the accuracy of detecting localized outbreaks and understanding public sentiment during early outbreaks.

Shen et al. (2020) analysed COVID-19-related posts on Weibo, a Chinese social media platform, to predict daily case counts. By examining approximately 15 million posts from 250 million users, they developed a machine learning classifier to identify "sick posts" referencing COVID-19 symptoms and diagnoses. The study revealed that these posts were reliable predictors of daily case counts, providing forecasts up to 14 days ahead of official statistics. This predictive accuracy remained consistent across mainland China, including Hubei province, despite regional disparities in healthcare access. However, the study's focus on Weibo limits its generalizability to other regions or platforms, such as X, where cultural differences and language barriers may impact the model's effectiveness. Furthermore, the study did not incorporate multilingual capabilities or account for the geospatial distribution of posts, which would be crucial in localized outbreak predictions.

Mansoor et al. (2020) executed a worldwide investigation of Twitter emotional responses towards COVID-19 by tracing both the emotional development patterns and the virus's consequences for daily routines. The researchers employed LSTM along with ANN to

perform sentiment classification and topic identification on remote work and online learning as well as evaluate their models' accuracy. The analysis of X posts about monkeypox by Olusegun et al. (2023) used deep learning models which integrated both Convolutional Neural Networks (CNN) and LSTM techniques. The research examined how people viewed monkeypox by investigating their safety worries and their distrust of public institutions together with their discriminatory attitudes toward minority communities. The analysis of 352,182 tweets through sentiment detection revealed important outbreak-related problems which increased public awareness about the disease. However, the study focused on broad patterns of sentiment without diving into specific regional variations in sentiment or incorporating multilingual data, which would be important for global applicability. The study also did not address the operational use of sentiment detection in real-time outbreak management.

Field et al. (2022) performed natural language processing (NLP) analysis on tweets about the 2020 Black Lives Matter demonstrations which occurred after George Floyd's death. Using a few-shot domain adaptation technique they found strong evidence of anger and disgust emotions in their research demonstrating how NLP provides useful insight into public emotions during social movements. Imran et al. (2020) studied emotional reactions alongside sentiments which diverse cultural groups expressed about COVID-19 and worldwide pandemic control strategies. Advances in sentiment analysis connected to deep learning models including LSTM enabled their study to achieve exceptional detection results for social media sentiment polarity and emotional expressions.

Mirugwe et al. (2024) applied five supervised learning techniques—random forest (RF), XGBoost classifier, support vector classifier (SVC), extra trees classifier (ETC), and decision tree (DT)—in combination with deep LSTM models to analyse sentiments in COVID-19-related tweets. Their results showed that the extra trees classifier outperformed the other models, achieving an accuracy rate of 93%. This model demonstrated significant potential for enhancing pandemic management through sentiment analysis. However, the study did not focus on multilingual data or geospatial analysis, which are crucial for global applications and localized outbreak detection. Integrating these factors could further enhance the model’s robustness.

Bengesi et al. (2023) investigated public sentiment surrounding the monkeypox outbreak to

reduce misinformation. Utilising a multilingual X dataset, they employed machine learning and NLP techniques, such as Vader, TextBlob, stemming, and lemmatisation. Among the 56 models developed, the highest-performing model achieved an accuracy rate of 93%. However, the study did not include emotion classification, which remains a key area in NLP research. Emotion classification categorises the thoughts, opinions, and feelings expressed in text, particularly regarding events. This capability is crucial for activism, fostering participation and sustaining engagement in social movements. While the study achieved an accuracy of 93%, it did not incorporate emotion classification, which is an essential aspect of understanding the public's emotional response to health crises. Adding emotion analysis could have provided more accurate predictions regarding public panic or fear, which often precede the official reporting of outbreaks.

While previous studies have made significant strides in leveraging social media for disease outbreak detection, they fail to address the complexities of multilingual data, geospatial integration, and emotion classification, all of which are critical for improving the real-time monitoring and prediction of disease outbreaks in diverse settings. This research aimed to fill these gaps by developing a model that incorporated geospatial analysis and provided emotion classification capabilities. These improvements enhanced localized outbreak detection and public health responses in regions with diverse linguistic and demographic characteristics, such as Kenya, and provide a more robust and scalable solution for global public health surveillance.

Table 2.1: Literature Review Summary

Table 2.1a) Literature Review Summary

Author	**Social Platform**	**Techniques**	**Limitations**
Liang et al. (2019)	X	Social network analysis, retweet path reconstruction	The credibility of hidden influential users remains a challenge.
Fu et al. (2016)	X	Thematic analysis identifying societal impact, government response, etc.	Limited focus on specific outbreak mitigation strategies.
Krauer & Schmid (2022)	N/A (Historical data)	Manual annotation, NLP libraries (spaCy, Stanford CoreNLP), geocoding	Poor performance of geocoding services with historical regions.
Müller et al. (2023)	X	COVID-X-BERT, Transformer models	Requires extensive fine-tuning for optimal performance.
Wang et al. (2023)	X	NLP analysis, sentiment analysis, human-driven textual analysis	Limited geographic focus, specific to U.S. mid-sized cities.
Martín-Corral et al. (2024)	X	Biological and informational diffusion modelling, social network analysis	Dependency on high-out-degree users as sensors.
Shen et al. (2020)	Weibo	Machine learning classifier, Granger causality, geotagged data analysis	Relies on a subset of geotagged posts, limiting generalizability.
Mansoor et al. (2020)	X	Sentiment analysis, LSTM, Artificial Neural Networks (ANN)	Limited to sentiment evolution without addressing outbreak detection directly.

Table 2.1b: Literature Review Summary

Author	Social Platform	Techniques	Limitations
Olusegun et al. (2023)	X	Hybrid CNN-LSTM, exploratory sentiment analysis	Focused on sentiment; lacks detailed outbreak prediction.
Field et al. (2022)	X	NLP, few-shot domain adaptation	Specific to social movements; no focus on health or disease.
Imran et al. (2020)	X	Sentiment analysis, deep LSTM models	Limited cross-cultural context due to dataset constraints.
Mirugwe et al. (2024)	X	Supervised learning (RF, SVC, XGBoost, etc.), deep LSTM	Limited to accuracy comparisons; lacks broader outbreak application.
Bengesi et al. (2023)	X	Multilingual sentiment analysis (Vader, TextBlob, stemming, lemmatisation)	Does not address emotion classification or nuanced outbreak dynamics.

2.2.2 Challenges in Using NLP for Outbreak Detection

Despite the potential of NLP models for outbreak detection, several theoretical challenges must be addressed to ensure accurate and timely predictions. One of the primary challenges is the noisy nature of social media data. Posts often contain irrelevant or misleading information, making extracting reliable signals of outbreaks difficult. For instance, many users may post about experiencing symptoms that are not related to infectious diseases (e.g., "I have a headache because I didn't sleep"), leading to false positives in outbreak detection models (Ordun et al., 2020).

Additionally, data imbalance can affect the accuracy of NLP models. Disease-related posts may constitute only a small fraction of social media data, making it harder for models to detect significant outbreaks promptly. To address this, advanced NLP models incorporate machine learning techniques such as topic modelling and clustering to identify posts likely related to infectious disease outbreaks (Chakraborty et al., 2020).

Another challenge is the multilingual nature of social media. Social media users post in

various languages, and NLP models must be able to process multiple languages and dialects to ensure global coverage of outbreak detection. Multilingual models based on transformers like XLM-R have been developed to handle this challenge, but they still require significant computational resources and training on diverse datasets (Li et al., 2020).

2.3 Theoretical Literature

Recent studies have focused on examining social media data since it provides an avenue for public health surveillance systems to detect infectious diseases in real-time. X and Facebook along with Threads and similar newer platforms create massive user content databases which reveal important public health patterns. Social media data extraction necessitates the application of sophisticated Natural Language Processing (NLP) techniques to generate usable data for monitoring and forecasting disease outbreaks because this data comes in unorganized and highly noisy formats. This review offers a theoretical overview of how NLP can be utilized to detect infectious disease outbreaks through social media data, emphasizing key NLP methods, associated challenges, and recent developments in the field.

2.3.1 NLP and the Extraction of Health-Related Information from Social Media

NLP is a branch of artificial intelligence that focuses on the interaction between computers and human language. NLP techniques are crucial for processing large volumes of unstructured social media text, transforming it into structured, analysable data. Disease surveillance systems use tokenisation as a basic NLP technique to analyse text by dividing it into separate words or phrases according to Wickramaarachchi et al. (2020). The model performs further analysis through NER to detect disease and symptom, and location mentions which enables real-time health information detection. NER plays a central role in health surveillance models by extracting key health entities such as diseases (e.g., "COVID-19," "flu") and symptoms (e.g., "fever," "cough") from social media posts (Ordun et al., 2020). For example, during the COVID-19 pandemic, NER models were used to track symptom mentions and disease names to monitor the spread and intensity of outbreaks in specific geographic regions (Li et al., 2020). However, the informal language used on social media, including slang, abbreviations, and emojis, complicates the accurate extraction of these entities.

2.3.2 Sentiment Analysis for Disease Outbreak Detection

In addition to identifying disease names and symptoms, sentiment analysis is often employed to assess the emotional tone of social media posts. Sentiment analysis helps differentiate between casual mentions of diseases or symptoms and posts that indicate genuine concern or

distress. For example, negative sentiment posts that express worry about symptoms (e.g., "I feel terrible with a high fever") may be early indicators of an outbreak (Charles-Smith et al., 2015). Theoretically, sentiment analysis filters out noise from social media data, improving the precision of outbreak detection models by focusing on posts that reflect real health concerns. Recent developments in sentiment analysis utilise deep learning models like transformers (e.g., BERT) that can understand context and nuances in text, making them more effective in analysing social media posts. This has enabled more accurate detection of posts related to public health crises, allowing sentiment analysis models to contribute to faster identification of outbreaks (Li et al., 2020).

2.3.3 Geolocation Tagging for Localised Outbreak Detection

Geolocation tagging is another critical aspect of NLP models for detecting disease outbreaks. The ability to associate social media posts with geographic locations enables public health officials to pinpoint outbreak hotspots. Geolocation can be obtained from metadata (if users allow location sharing) or text, where mentions of places or landmarks are recognised and extracted (Mavragani, 2020). Theoretical advancements in geolocation tagging focus on improving the accuracy of location extraction from text. For instance, combining NER with geospatial analysis helps link specific symptoms or disease mentions to regions, providing valuable spatial information for outbreak detection. However, a major challenge is the sparsity of location data in social media posts, as not all users share their location or mention it explicitly in their posts (Boehm et al., 2020).

2.3.5 Recent Advances in NLP for Outbreak Detection

Current advancements in NLP for disease outbreak detection emphasize the use of deep learning solutions for enhancing model performance together with scalability. BERT alongside GPT demonstrates superior capability than traditional NLP approaches when it comes to disease outbreak detection by understanding sophisticated language structures while handling word context (Wickramaarachchi et al., 2020). These models are particularly effective in processing informal social media text, which often lacks grammatical structure. Additionally, the integration of time-series analysis with NLP models allows for tracking disease mentions over time, providing dynamic outbreak predictions. By monitoring the frequency and sentiment of posts mentioning specific symptoms or diseases, time-series models can identify spikes in activity that may indicate the onset of an outbreak (Chakraborty et al., 2020).

Hybrid models that combine NLP techniques with epidemiological models have also been

proposed to enhance outbreak detection. These models use NLP to extract health-related data from social media and then apply epidemiological models to predict the potential spread of diseases based on the extracted data (Mavragani, 2020). This approach bridges the gap between digital surveillance and traditional public health methods, providing more comprehensive outbreak monitoring systems.

2.4 Frameworks

Several authors have presented frameworks for bio surveillance, with most attention placed on infectious studies. This section reviews the number of frameworks tested and developed and their limitations.

2.4.1 Detecting Topic and Sentiment Dynamics Framework

Yin et al. (2020) developed a framework to identify topics and analyse sentiment dynamics related to COVID-19 by examining a dataset of 13 million tweets over a two-week period. Their findings revealed that positive sentiment consistently surpassed negative sentiment throughout the study. They also observed that different aspects of COVID-19 were associated with distinct sentiment polarities. For example, topics like "stay safe home" were mainly linked to positive sentiment, while themes such as "people's death" consistently reflected negative sentiment. Overall, the framework provided valuable insights into the sentiment dynamics at the topic level.

2.4.2 Vaccine Sentiment Framework

Martinez et al. (2022) created a framework to extract coronavirus-related tweets from around the world. They applied unsupervised learning techniques, including K-means and hierarchical clustering, to analyse the situation across different countries. The study uncovered a wide range of emotions among individuals, with a significant prevalence of pessimism.

2.4.3 Sentiment Analysis and Topic Modelling Framework

Qin and Ronchieri (2022) proposed a framework for analysing public sentiment towards various pandemics, including cholera, Ebola, HIV/AIDS, influenza, malaria, Spanish influenza, swine flu, tuberculosis, typhus, yellow fever, and Zika. They collected 369,472 tweets from 3 March to 15 September 2022, applying natural language processing (NLP) techniques such as topic modelling and sentiment analysis. The analysis revealed that the

public discussions predominantly cantered around malaria, influenza, and tuberculosis, with the disease itself being the primary focus. The emotional responses were mostly fear, trust, and disgust, with trust being particularly associated with HIV/AIDS-related tweets.

However, several limitations can be identified in the framework. First, the reliance on Twitter as the sole data source may not capture the sentiments of individuals without social media access, leading to potential biases in representation. Additionally, the framework's focus on just three emotions—fear, trust, and disgust—may oversimplify the emotional landscape, overlooking other important emotional responses such as hope or confusion. Moreover, the method used for topic modelling might miss the subtleties of public discourse, especially in complex topics related to disease control measures. Enhancing the emotional scope and improving the contextualization of topics could make the framework more comprehensive.

Figure 2.1 illustrates the sentiment analysis and topic modelling framework used in this study.

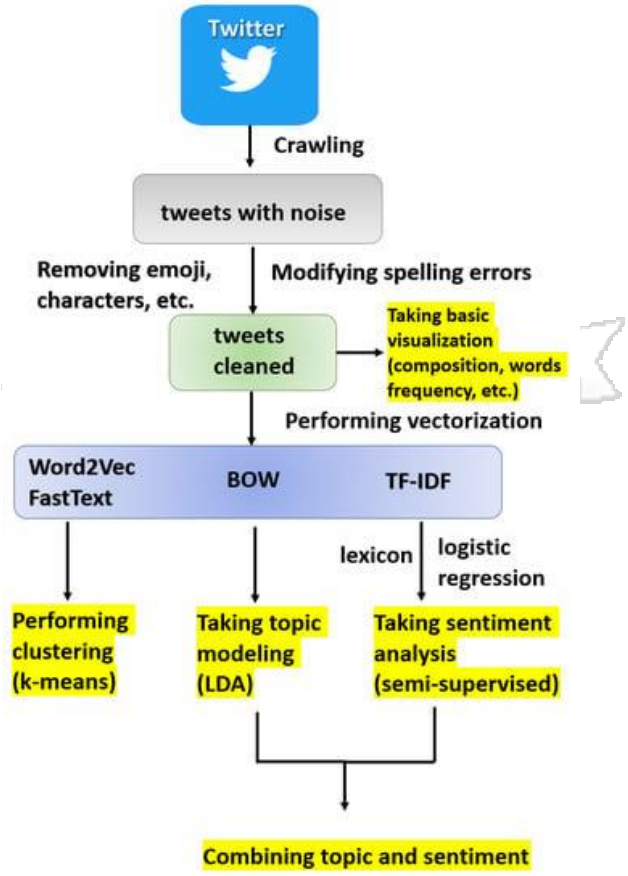


Figure 2.1: Sentiment Analysis and Topic Modelling Framework (Qin and Ronchieri, 2022).

2.4.4 Word2PLTS Framework

Song et al. (2020) introduced a novel text representation model for short-text sentiment analysis, combining supervised and unsupervised learning. Their framework utilised probabilistic linguistic terms and relevant theoretical concepts to improve sentiment classification accuracy as shown in Figure 2.2.

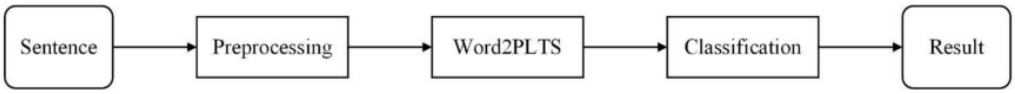


Figure 2.2: Word2PLTS (Song et al., 2020)

2.4.5 NLP Framework

Raza & Schwartz (2023) developed a framework which includes database building and NLP modules for clinical and social determinants extraction and a thorough evaluation process. Their method which utilized COVID-19 case reports for data construction and pandemic surveillance produced better results than benchmark methods by reaching an F1-score enhancement of about 1–3%. The framework design appears in Figure 2.3.

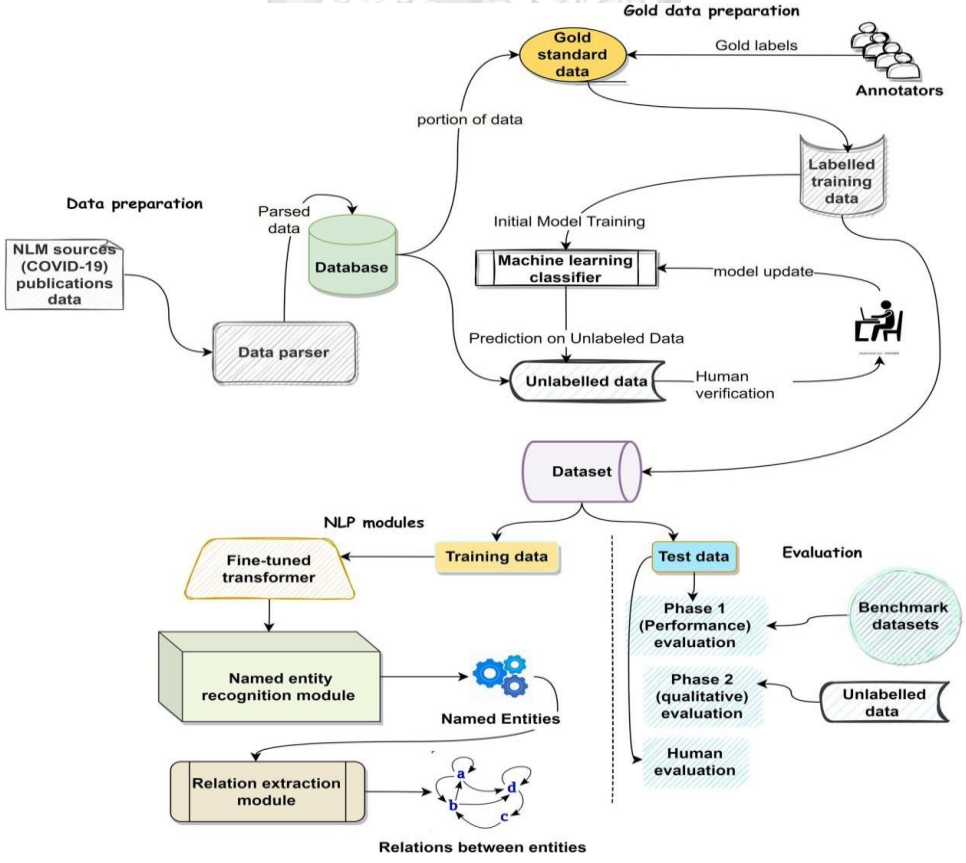


Figure 2.3: NLP Framework (Raza & Schwartz, 2023).

Table 2.2: Summary of Frameworks

Author	Technique	Limitation
Yin et al. (2020)	Framework for detecting topic and sentiment dynamics using a collection of 13 million tweets.	Limited to short-term analysis (two weeks); may not capture long-term sentiment trends.
Martinez et al. (2022)	Vaccine sentiment framework using unsupervised learning techniques (K-means, hierarchical clustering).	Highlights general emotions but lacks granular details on sentiment evolution over time.
Qin & Ronchieri (2022)	Sentiment analysis and topic modelling framework for HIV/AIDS using X data.	Focused on a specific event (World AIDS Day 2013), limiting its broader applicability.
Song et al. (2020)	The Word2PLTS framework combines supervised and unsupervised learning for short-text sentiment analysis.	Primarily focuses on sentiment classification accuracy; does not address topic detection.
Raza & Schwartz (2023)	NLP framework for extracting clinical and non-clinical information using COVID-19 case reports.	Relies heavily on case reports, limiting real-time application to social media surveillance.

2.5 Models

2.5.1 CNN Model

Convolutional Neural Networks (CNNs) serve as artificial neural networks that specialists use in computer vision applications and demonstrate high popularity within radiology (Yamashita et al., 2018). Through backpropagation CNNs maintain an automatic capability to learn adaptive spatial features that operate using convolutional pooling elements and fully

connected networks. Mirugwe et al. (2024) created a six-layer CNN model for their research. The model incorporated 32 filter layers in the first two stages while the successive two stages included 64 filters, and the final two stages contained 128 and 256 filters. The model received an enhancement through the addition of dropout layers set to 0.5 for regularization purposes. The architectural design of CNN models appears in Figure 2.4.

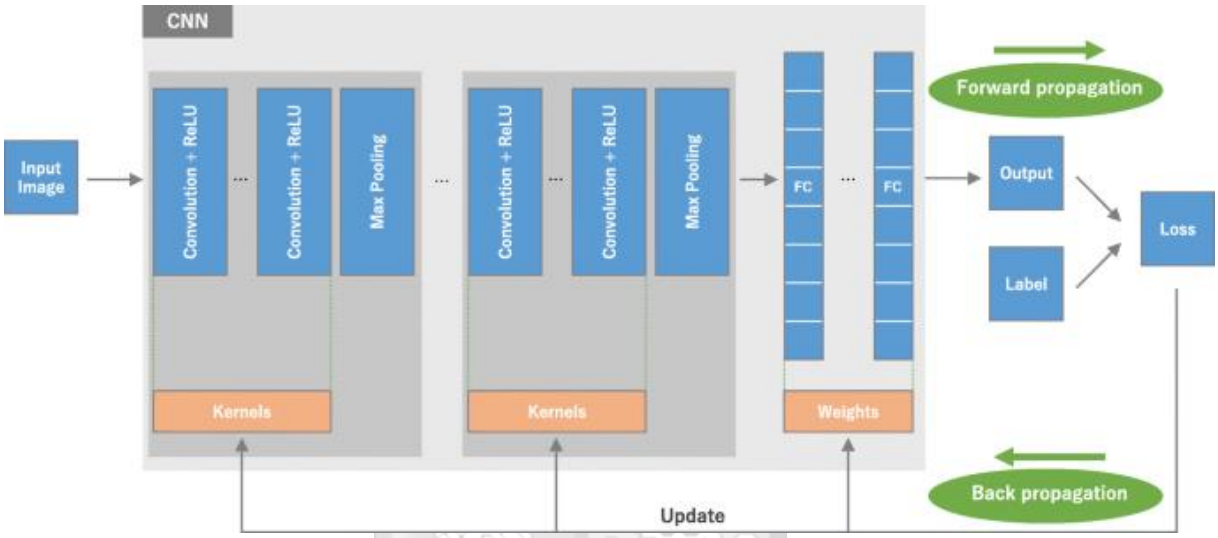


Figure 2.4: CNN Model Architecture (Yamashita et al., 2018)

2.5.2 LSTM Model

LSTM networks serve as specialized RNNs which specifically perform sentiment analysis according to Van Houdt et al. (2020). LSTMs maintain their ability to handle sequential data through their mechanism of retaining information across extended time spans. LSTM networks process information across extended periods during which they can recognize full sentence or paragraph content thus ensuring precise sentiment detection. The models receive training using large datasets of labelled text which assign each entry into positive or negative or neutral categories (Yu et al., 2019). By conducting training operations through backpropagation, the model develops ability to recognize sentiment-specific patterns and features. Figure 2.5 presents the configuration information about LSTM hyperparameters and the complete system architecture.

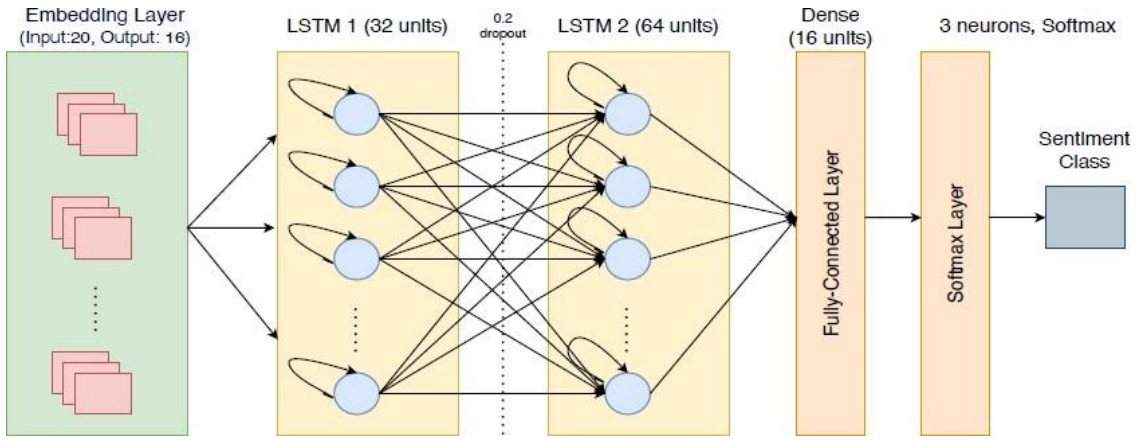


Figure 2.5: Sequential (LSTM) Architecture for Sentiment Analysis (Mirugwe et al., 2024).

The sentiment analysis model implemented by Mirugwe et al. (2024) uses the sequential (LSTM) hyperparameters and architecture which are displayed in Figure 2.5. Model performance evaluation took place during an iterative process that established the hyperparameter settings. The researchers conducted multiple experimental trials to determine the best set of parameters which they finally selected. The training phase processed the entire dataset 10 times before completion according to Mirugwe et al. (2024).

2.5.3 BERT Model

The BERT-based sentiment analysis model contains 12 transformer layers while using 110 million parameters. The framework implements an embedding layer that originates from the BERT pre-trained model to detect semantic relations in text. The Hugging Face Transformers library provides Python access to a text preprocessing tool which performs tokenizer operations on textual data. The model architecture contains two bidirectional LSTM layers with 32 units that use "return_sequences=True" configuration. Sequence information maintains its integrity throughout the embeddings because this configuration helps the model interpret text data context more effectively (Mirugwe et al., 2024). The GlobalMaxPooling1D layer transforms the variable-length LSTMs output into a fixed-length vector for classification purposes. The layer selects the most important features from the LSTM output sequence before passing them to the following layers. Each LSTM layer in the model contains a dropout layer with a 0.5 rate to prevent overfitting by disabling random neuron groups during training.

Using ReLU activation functions the model contains two dense layers with 128 units followed by another dense layer with 64 units to perform non-linear transformations on complex data patterns. The training proceeded through 30 epochs while utilizing the Adam optimizer together with sparse categorical cross-entropy loss for performing multi-class classification purposes. The BERT layers received training protection by freezing them to maintain pre-trained weights while minimizing trainable parameters for improved performance. The optimization strategies utilized a batch size of 64 together with early stopping which relied on validation performance to keep the training time short while maintaining an accurate model output. The complete system design appears in Figure 2.6.

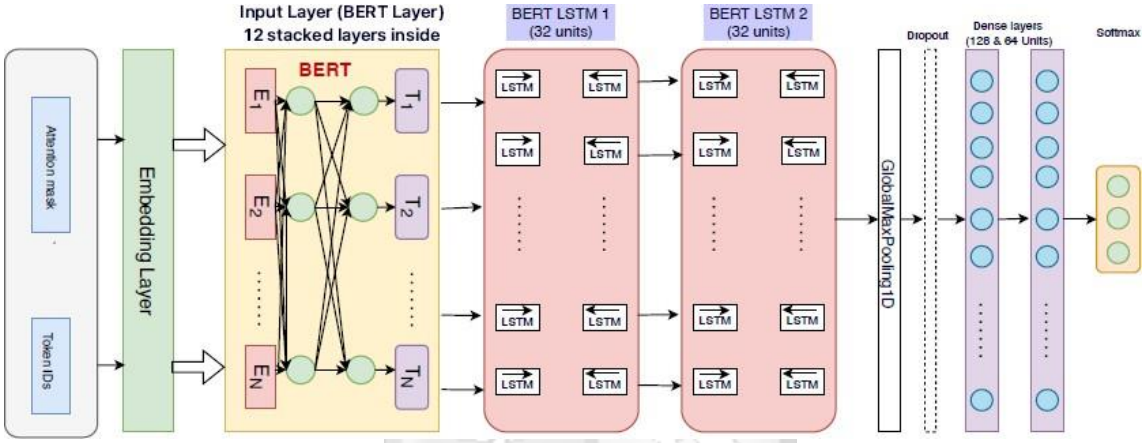


Figure 2.6: The BERT Model Architecture (Mirugwe et al., 2024).

Table 2.3: Summary of Models

Model	Description	Limitation of the Model
CNN Model	Convolutional Neural Network with six convolutional layers (filters: 32, 64, 128, 256), utilising dropout (0.5 rate) for regularisation. Designed to learn spatial hierarchies of features automatically.	Limited ability to capture temporal dependencies or sequential data relationships; suited primarily for image and spatial data.
LSTM Model	The long short-term memory model is designed for sequential data processing and capturing long-term dependencies in text. Trained over 10 epochs, with iterative hyperparameter tuning for optimal performance.	Computationally intensive, prone to vanishing gradients for very long sequences, and requires extensive training for accuracy.
BERT Model	Pre-trained BERT-based sentiment analysis model with 12 transformer layers, fine-tuned using bidirectional LSTMs, dropout layers (0.5), dense layers (ReLU activation), and GlobalMaxPooling1D for feature extraction.	Large parameters (~110 million) increase computational requirements; pre-trained weights limit adaptability to new tasks.

2.6 Algorithms

Multiple machine learning techniques along with algorithms serve as the foundation for building NLP models that detect infectious disease outbreaks through social media data. The algorithms perform multiple NLP operations starting from text preprocessing to entity extraction and sentiment analysis and topic modelling to convert unstructured social media

text into structured analysis data. This field utilizes multiple algorithms and techniques which are explained through research conducted within the last five years.

2.6.1 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) constitute supervised learning algorithms which serve both for classification needs and the purposes of regression tasks. SVM classifiers build maximum-margin hyperplanes in transformed input spaces to distinguish data points between different classes according to Maimon & Rokach (2010). The hyperplane achieves maximum separation distance by positioning itself between nearest data points from different classes. The parameters that define the optimal hyperplane are determined through a quadratic programming optimization process, which seeks to achieve the maximum margin by minimizing a loss function while adhering to constraints that ensure correct classification of data points. SVMs have traditionally been used for sentiment classification, particularly with text data where classes (e.g., positive, negative) are linearly separable. However, SVMs face challenges when dealing with complex, context-dependent sentiment (Wickramaarachchi et al., 2020). Figure 2.7 illustrates the architecture of the SVM algorithm.

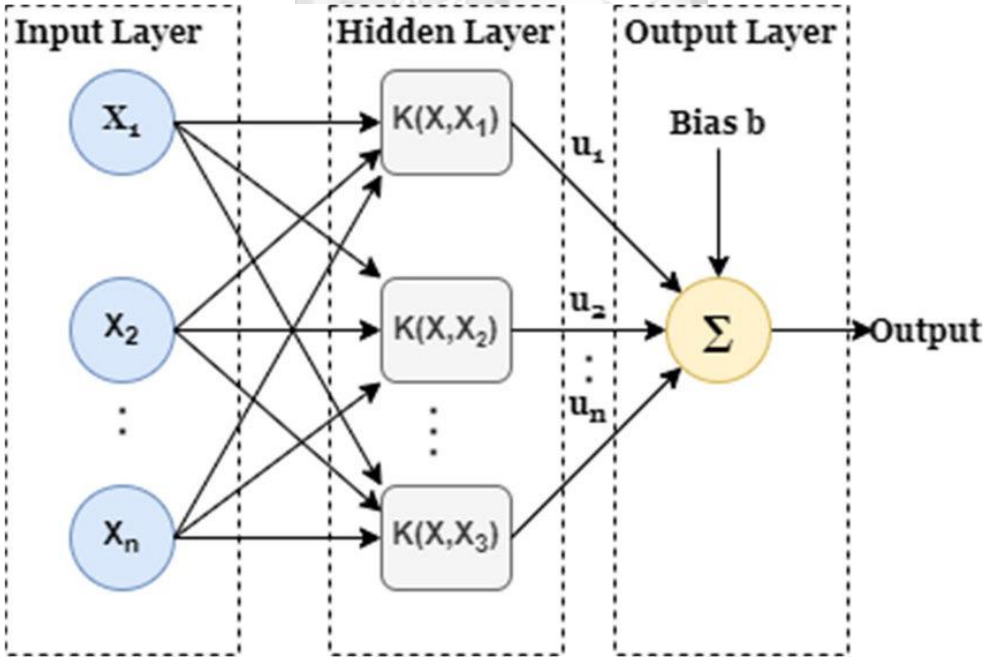


Figure 2.7: SVM Algorithm (Kadam et al., 2021)

2.6.2 Recurrent Neural Networks (RNNs)

A Recurrent Neural Network (RNN) is an artificial neural network characterised by internal loops that create recursive dynamics (Marhon et al., 2013). These loops enable the network to maintain dependencies over time by introducing delayed activations across its processing elements (PEs). This structure allows RNNs to model sequential and temporal relationships effectively. These algorithms are used for sentiment analysis because they can process data sequences. They capture the temporal dynamics of language, allowing them to perform well on sentiment classification tasks where the order of words matters. Figure 2.8 illustrates the architecture of a recurrent neural network (RNN), where the hidden layer outputs are fed back into the input layer. After each epoch, the network copies the outputs from the hidden units and incorporates them into the input layer. This feedback mechanism allows the use of a straightforward backpropagation training algorithm.

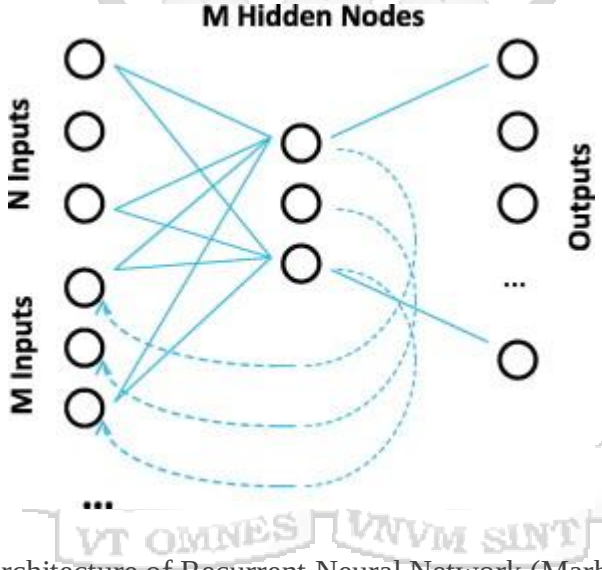


Figure 2.8: Architecture of Recurrent Neural Network (Marhon et al., 2013)

2.6.3 LSTMs

The LSTM network introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 was created to solve traditional RNN limitations which included gradient vanishing and exploding problems during backpropagation through time (BPTT) (Malashin et al., 2024). These challenges hinder standard RNNs from learning and retaining long-term dependencies, a critical requirement for sequential data tasks. LSTM networks overcome these issues with their unique architecture, which includes memory cells capable of preserving information over extended periods and a gating mechanism that regulates the flow of information. This design enables LSTMs to handle long-term dependencies effectively, making them well-suited for tasks like time-series prediction and modelling dynamic systems. The memory cells and gating

mechanisms ensure that gradients propagate efficiently over long sequences, mitigating the problems associated with traditional RNNs (Landi et al., 2021). Figure 2.9 illustrates the architecture of an LSTM network.

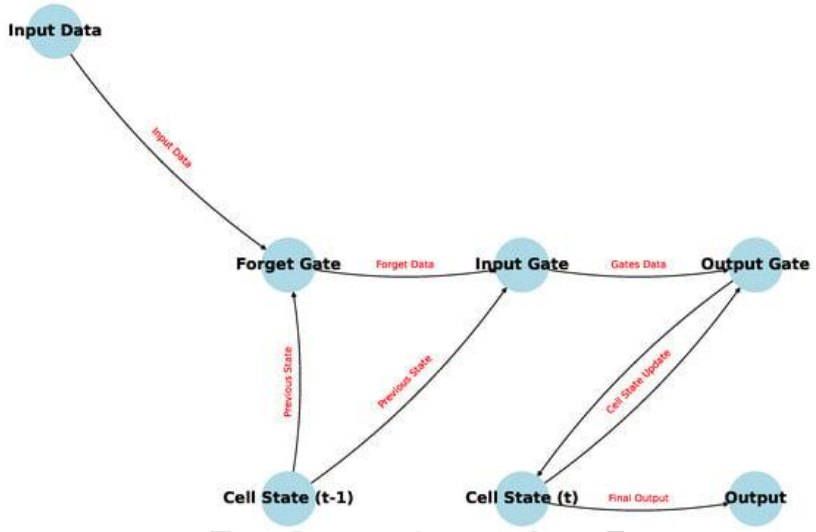


Figure 2.9: LSTM Algorithm (Malashin et al., 2024)

2.6.4 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) serves as a popular probabilistic topic modelling system which discovers thematic content in groups of documents that include social media content. LDA operates under the premise that every post combines multiple topics while each topic consists of various words. Public health research extensively uses LDA to extract topics about disease symptoms and treatments and public understanding of diseases (Wang et al., 2021).

2.6.5 Dynamic Topic Modelling (DTM)

Dynamic Topic Modelling (DTM) builds upon Latent Dirichlet Allocation (LDA) by capturing how topics change over time. This is especially valuable for tracking the progression of disease outbreaks as social media discussions transition from focusing on symptoms to treatments and recovery (Ordun et al., 2020).

Table 2.4: Summary of Algorithms

Algorithm	**Short Description**	**Limitation of the Algorithm**
Support Vector Machines (SVMs)	Supervised learning method for classification and regression that constructs a maximum-margin hyperplane to separate data classes. Primarily used for sentiment classification.	Struggles with complex, context-dependent sentiment and non-linearly separable data.
Recurrent Neural Networks (RNNs)	Neural networks with internal loops that model sequential and temporal relationships effectively. Suitable for tasks like sentiment analysis and language processing.	Prone to vanishing and exploding gradient problems, making long-term dependency learning challenging.
LSTMs	Advanced RNN networks apply Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) which create memory cells together with gating operations for preserving long-term dependencies within data. Advanced RNNs deliver favourable results for time-series prediction together with dynamic system modelling since they retain information across extended durations although basic RNNs lack such performance.	Computationally intensive; requires significant resources for training and fine-tuning.
Latent Dirichlet Allocation (LDA)	Probabilistic topic modelling algorithm that identifies topics as a mixture of words and applies them to documents (e.g., social media posts).	Assumes topics are static, which limits its ability to track changes over time.
Dynamic Topic Modelling (DTM)	An extension of LDA that tracks topic evolution over time, useful for analysing the progression of discussions, such as during disease outbreaks.	Computationally more intensive than LDA and sensitive to the data's temporal segmentation quality.

2.7 Architectures and Designs

2.7.1 Sentiment Analysis Design

Leelawat et al. (2022) conducted a study analysing English-language tweets about Thai tourism, categorizing them into sentiment and intention groups through manual labelling. The study compared the effectiveness of three machine learning algorithms—CART, random forest, and SVM—in predicting the sentiments and intentions expressed in the tweets. The researchers also examined the top 10 most frequently occurring words in each sentiment and intention category to gain insights and offer recommendations. Among the algorithms tested, SVM achieved the highest accuracy in identifying sentiments and intentions in tweets related to popular Thai destinations, including Bangkok, Phuket, and Chiang Mai. The study found that oversampling generally produced better results than under-sampling. Moreover, unigrams (single terms) often achieved higher accuracy in sentiment analysis than bigrams (two-word combinations). The analysis of the top 10 words revealed that X users expressed positive sentiments toward Thailand's food, tourist destinations, and hospitality. However, negative sentiments were associated with the political situation in Thailand and the impact of the COVID-19 pandemic. These negative perceptions likely had a significant adverse effect on the Thai tourism industry. The architectural design of this study is depicted in Figure 2.10.

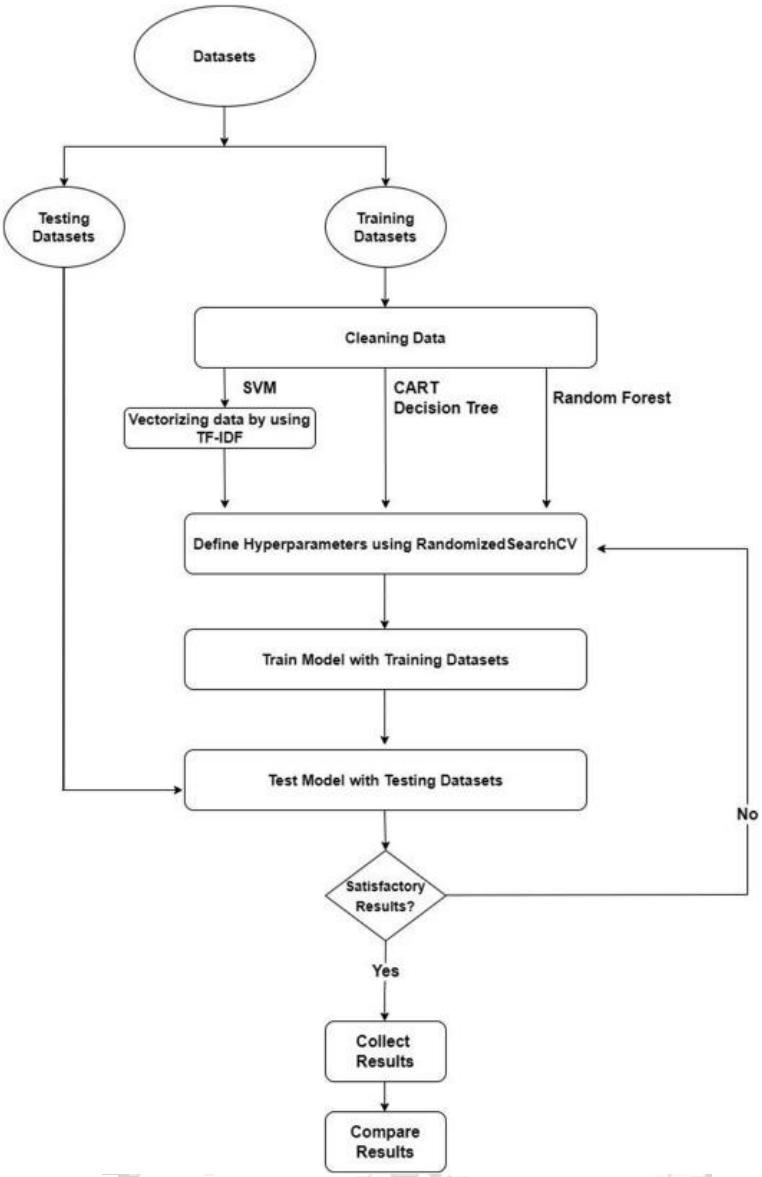


Figure 2.10: Sentiment Analysis Design (Leelawat et al., 2022)

2.7.2 LSTM Architecture

The LSTM architecture consists of memory blocks that link together as a sequence to store information through time using non-linear gating units for data flow control. The standard LSTM block (also known as vanilla LSTM) exhibits its essential components in Figure 2.11 including gates alongside the input signal x(t) and output y(t) and activation functions and peephole connections (Gers & Schmidhuber, 2000). The output from the block gets connected repeatedly to its input data and all gates to enable the network to maintain temporal dependencies.

An LSTM model operates through a network containing N processing blocks and M inputs. The recurrent neural network executes its forward pass through the following sequence:

- **Block Input**: The block input is updated by combining the current input x(t)x(t)x(t) with the output y(t−1) from the previous iteration of the LSTM unit. This update operation is expressed as follows:

$$z(t) = W_x x(t) + W_y y(t-1) + b$$

Where:

- z(t) is the combined input for the current time step,
- W_x and W_y are weight matrices for the input x(t) and previous output y(t−1), respectively,
- b is the bias term.

This combination allows the LSTM to incorporate information from both the current input and its memory of previous states, ensuring it can learn temporal dependencies effectively. The updated block input is then passed through the gating mechanisms to regulate how much information is retained or discarded in subsequent steps as shown in Figure 2.11.

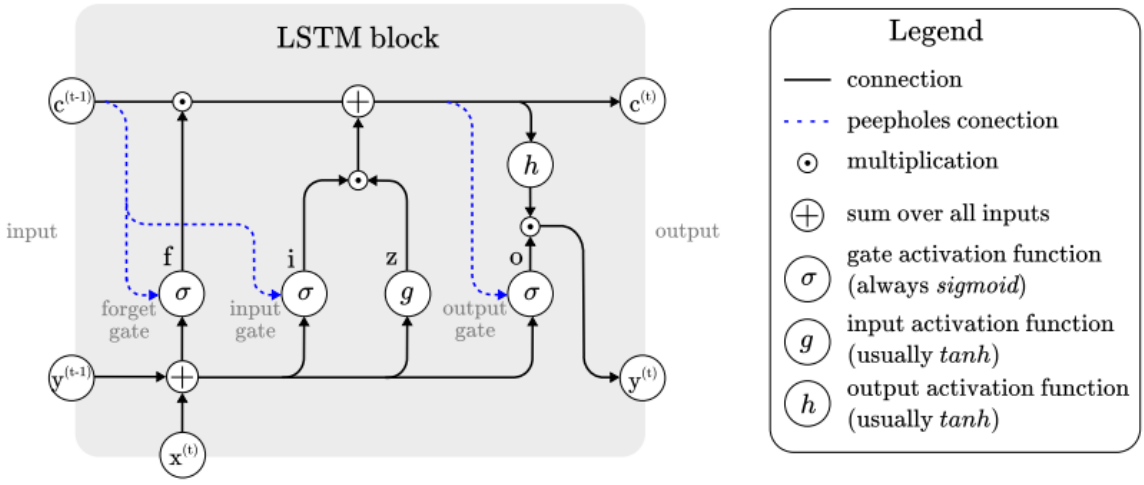


Figure 2.11: LSTM Architecture (Van Houdt et al., 2020)

Table 2.5: Summary of Architectures and Designs

Architecture	Author	Strengths	Limitations
Sentiment Analysis Design	Leelawat et al. (2022)	-High accuracy with SVM for sentiment and intention classification.	- I struggle with complex sentiment analysis, especially context-dependent emotions. -Negative perceptions, such as political unrest and COVID-19, limit actionable recommendations.
LSTM Architecture	Gers and Schmidhuber (2000)	-Effectively retains long-term dependencies in sequential data. -Gating mechanisms ensure efficient information flow.	- Computationally intensive, requiring significant training resources. -Sensitive to parameter tuning for optimal performance.

2.8 Research Gap

While previous studies have demonstrated the feasibility of using social media for outbreak detection, many existing approaches face high rates of false positives due to the informal language and lack of context in user posts (Ordun et al., 2020). Additionally, limited efforts have been made to incorporate geographical data effectively, which is crucial for localised outbreak detection (Chakraborty et al., 2020).

2.9 Conceptual Model

The proposed conceptual model will integrate multiple components to transform unstructured data into meaningful insights for public health surveillance. The process will begin with the data collection layer, which will gather large amounts of social media data from platforms like X, Facebook, and Threads. Posts will be retrieved using public APIs, focusing on specific keywords, hashtags, and metadata such as geotags and time frames to narrow the content to public health-relevant discussions (Pilipiec et al., 2023). The next step

will involve the data preprocessing layer, which will address social media data's noisy and unstructured nature. Irrelevant elements, such as URLs, emojis, and stop words, will be removed, and the text will be broken into individual tokens. Words will be reduced to their root forms through stemming or lemmatisation to ensure consistent treatment of variations like "fever" and "feverish." Additional filters will eliminate posts unrelated to health topics to minimise false positives.

The cleaned text will be transformed into a structured format that NLP models can process in the feature extraction layer. Named Entity Recognition (NER) will identify key entities such as symptoms, disease names, and locations, while sentiment analysis will evaluate the emotional tone of posts to highlight genuine health concerns. Geolocation tagging will extract or infer location data, enabling the mapping of disease spread. The model development layer will serve as the core of the framework, where NLP models will be trained and implemented. Transformer-based models, such as BERT and RoBERTa, will be fine-tuned for tasks like entity recognition and sentiment analysis. Topic modelling techniques, including Latent Dirichlet Allocation (LDA), will uncover themes related to symptoms and diseases. At the same time, time-series analysis models like LSTM or ARIMA will monitor the frequency of disease-related mentions over time to detect outbreaks (Al-Garadi et al., 2022).

To enhance prediction accuracy, the NLP model's output will be integrated with traditional epidemiological models, forming a hybrid approach. These models will incorporate real-world factors such as infection rates and recovery times to predict disease spread more effectively (Olusegun et al., 2023). A monitoring and detection layer will provide real-time visualisation tools and automated alerts for public health officials, highlighting trends and potential outbreaks in specific regions. Lastly, a feedback loop will ensure continuous improvement. As new data becomes available, the model will be retrained and fine-tuned. Public health experts will review predictions and provide feedback, enabling the model to evolve with emerging discussions and diseases, maintaining its accuracy and relevance over time. The conceptual model is shown in Figure 2.12 below.

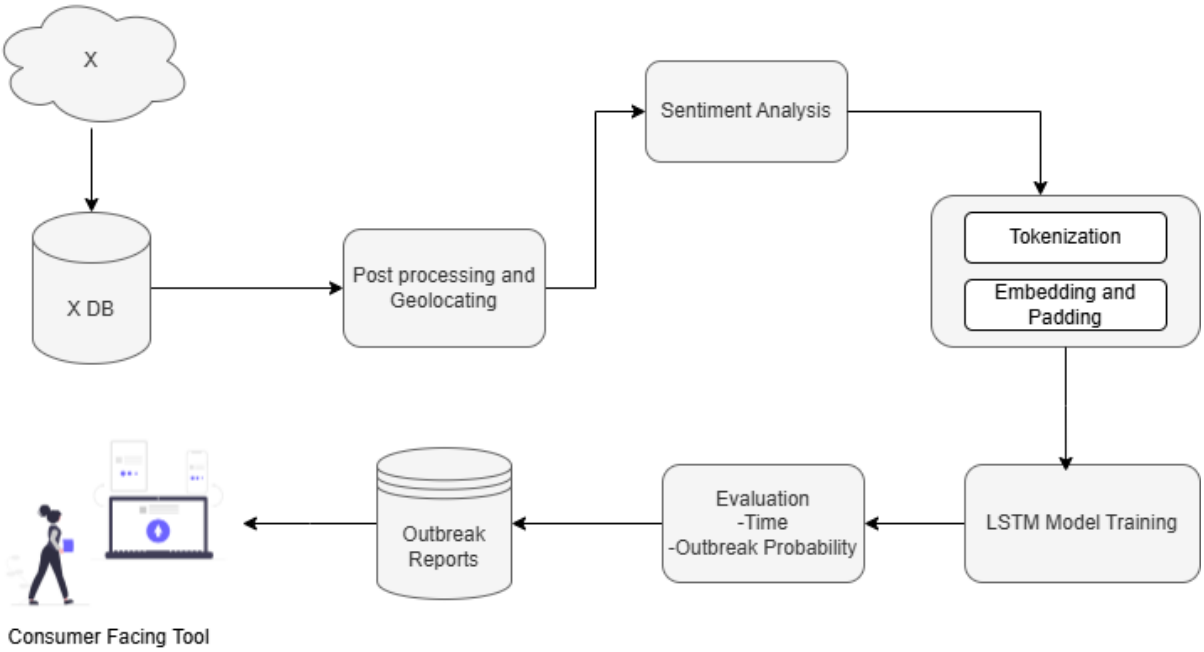


Figure 2.12: Conceptual Model

Chapter 3: Research Methodology/Design

3.1 Introduction

In this research, an NLP model will be developed to determine common infectious disease outbreaks in Kenya. Data will be acquired from X, formally known as Twitter. The data was analysed using various methods like sentiment analysis and topic modelling. The goal of the model is to help the government know about outbreaks of common infectious diseases, thus reducing the spread and aiding with resource distribution. This chapter outlines the research methodology that was used to carry out the study.

3.2 Research Design / Philosophy

This study adopted an experimental research design, combining NLP techniques to analyse publicly available X data on infectious diseases in Kenya. The research focused on detecting emerging topics and analysing sentiment through sentiment analysis and topic modelling, using tweets as the data source. The research philosophy underpinning this study aligns with interpretivism, which emphasises understanding social phenomena from the perspective of the participants. In this case, the goal is to interpret the meanings and sentiments expressed by users on the X platform concerning infectious diseases in Kenya. The interpretivist approach is selected because it allows for an in-depth exploration of how individuals and communities perceive and react to health issues, reflecting real-world concerns and behaviours that are crucial for the early detection of disease outbreaks. By analysing these subjective expressions, the study seeks to uncover insights into public awareness and reactions that are essential for effective health responses. Inductive reasoning was employed to develop new theories and insights based on the data collected, making it an ideal approach for this study. By analysing X posts and identifying patterns related to disease outbreaks, the research will generate findings that can inform the government and health authorities on early warning signals and resource allocation. This design choice ensures the flexibility to explore and identify unexpected trends in public health conversations that may not be captured through traditional research methods.

3.3 Population and Sampling

3.3.1 Population

The target population for this study consisted of all publicly available tweets relating to Kenya that mention keywords related to infectious diseases, such as "cholera," "measles," "malaria," and "polio." The decision to focus on X as the primary data source is based on several key factors. X is a widely used social media platform in Kenya, especially for real-time discussions

about public health issues and current events. It allows for the rapid sharing of information and public discourse, making it an ideal platform for tracking real-time reactions to disease outbreaks. Additionally, X's format, with its 280-character limit per tweet, promotes concise and direct expressions of sentiment and opinion, which can be particularly useful for sentiment analysis and topic modelling.

Unlike Facebook, which tends to feature more personal, family-oriented content with privacy settings that limit data accessibility, X is a public platform by default, providing open access to a wealth of data. This openness is crucial for monitoring widespread public concerns and discussions. Furthermore, X's use of hashtags and geo-tagging enables easier identification of tweets related to specific topics and locations. This facilitates a more targeted approach to identifying outbreaks and public sentiment in real time. Therefore, X's combination of public accessibility, real-time updates, and widespread use in Kenya makes it the most suitable platform for this study.

3.3.2 Sampling

Given the large volume of data available on X (formerly Twitter), a time-based sampling strategy was employed to collect tweets specifically related to disease outbreaks and health concerns within a defined period spanning from 2020 to 2024. A total of 179,207 tweets were collected, which was determined to be an appropriate sample size for training the machine learning model. This dataset was considered sufficient to capture relevant patterns and sentiment associated with disease outbreaks in Kenya, enabling the model to learn from a variety of real-world discussions. The time-based approach ensured that the dataset reflected the most recent and relevant trends in public health concerns.

3.4 Data Collection/Analysis

3.4.1 Data Collection

Tweets with the words "cholera", "measles", "malaria", and "polio" relating to Kenyan was be scrapped using the official twitter API tweepy. The first step involved removing the duplicate tweets and any other tweets in languages other than English because of the nature of the tweets.

3.4.2 Data Analysis

The collected tweets underwent several preprocessing steps to ensure high-quality data for analysis. First, non-English tweets were excluded to maintain consistency in the dataset. Then, duplicate tweets were identified and removed using standard deduplication methods to ensure each tweet was unique and contributed valuable information. Irrelevant content, such as tweets

not related to health or disease outbreaks, was filtered out to ensure the dataset focused only on disease-related discussions. Additionally, tweets that were geographically irrelevant to Kenya were excluded to ensure the study maintained its geographical focus. The remaining data was tokenized and lemmatized, ensuring uniformity in text before applying machine learning techniques like sentiment analysis and topic modelling.

i). Data Cleaning

Data cleaning is a vital step in text analytics to ensure accurate and reliable results. This process involves eliminating irrelevant components from the text, such as punctuation, stop words (e.g., "the," "an," "a"), and URLs. Stemming simplifies words by reducing them to their root form, for example, converting "goals" to "goal" and "pens" to "pen." In contrast, lemmatisation organises words with similar meanings into a common base form, allowing for more qualitative and meaningful analysis (SV et al., 2022).

ii). Sentiment Analysis

Sentiment analysis is the technique of automatically extracting and evaluating subjective opinions regarding various features of an item or entity, offering insights into the text's context and the emotions conveyed by the author. In this study, sentiment analysis was employed to gather information about an infectious disease outbreak. The TextBlob Python library analysed each word in the corpus documents to determine overall sentiment between positive and negative and neutral. SV et al. (2022) explains that TextBlob functions through the bag-of-words model to classify words using a predefined dictionary that identifies positive or negative attributes. Each word receives a score from the library after which the system produces an overall sentiment value to evaluate sentiment trends throughout the data.

iii). Thematic Analysis

Thematic analysis is a generative statistical approach used to uncover the core meaning of a text (SV et al., 2022). One widely adopted method for analysing the structure of a text is Latent Dirichlet Allocation (LDA) topic modelling, which identifies the underlying themes within a corpus. LDA assumes that each document in the corpus is a mixture of topics, and each topic is represented as a probability distribution over words. Operating under the bag-of-words assumption, the model is guided by two key principles, including the Dirichlet process, a probability distribution that defines a range of possible topic distributions. Figure 3.4 below illustrates the steps involved in data collection and preprocessing.

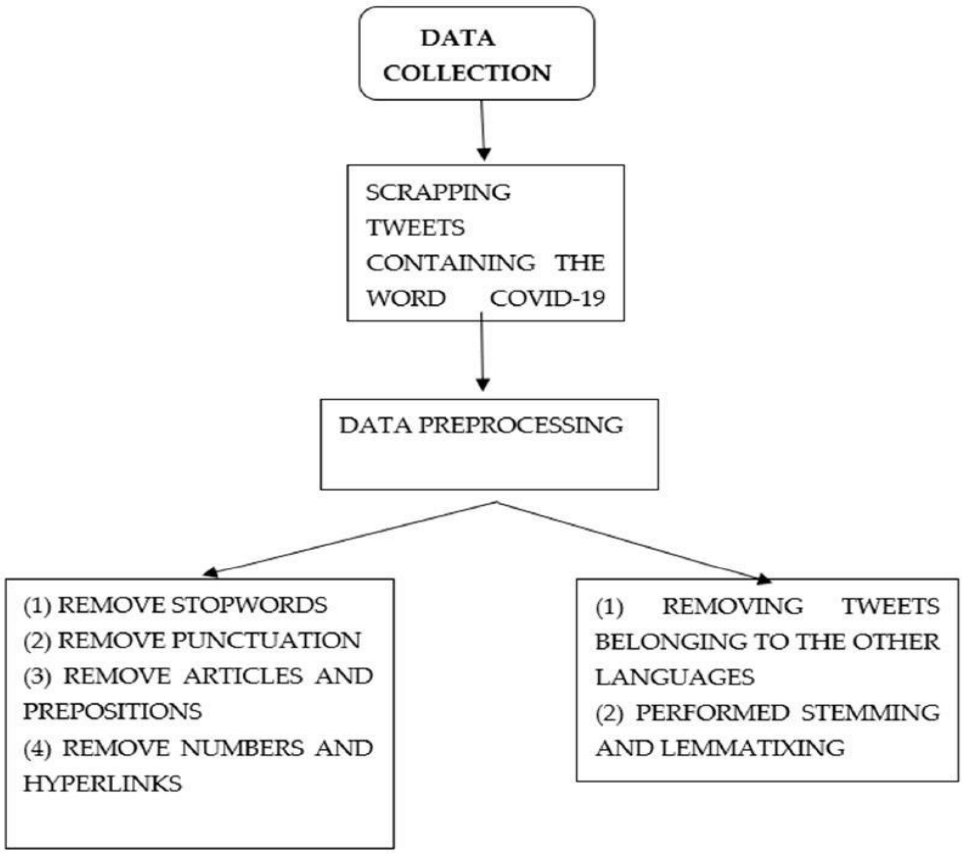


Figure 3.4: Data collection and data pre-processing (SV et al., 2022)

3.5 Research Quality and Reliability

The data cleaning process involved filtering out irrelevant or misleading content, such as tweets originating from outside Kenya or posts unrelated to the outbreak of infectious diseases. This step addressed the challenge of processing content that may include conspiracy theories, false positives, or outdated discussions about previous health crises, such as COVID-19. By narrowing the dataset to contextually and geographically relevant tweets, the analysis remain focused on identifying patterns specific to the Kenyan context. The reliability of the outbreak detection model was assessed using root mean squared error. This metric ensured the model's ability to classify outbreaks accurately and consistently across diverse tweet content. For thematic analysis, the application of Latent Dirichlet Allocation (LDA) was validated by cross-referencing identified topics with manually labelled tweet samples to ascertain their relevance and alignment with outbreak-related themes. To further enhance the robustness of the study, multiple reviewers independently interpreted the themes and sentiments derived from the data. This collaborative review process reduced potential bias and ensured consistency in the interpretation of findings. Collectively, these measures ensured that the developed Natural Language Processing model effectively predicts outbreaks of infectious diseases in Kenya while maintaining high standards of validity and reliability.

3.6 Systems Development Methodology

The Agile methodology was used to develop the NLP model for detecting disease outbreaks. This iterative approach is ideal for adapting to changing requirements and continuous feedback. The model will go through several stages, as shown in Figure 3.3:

i). **Requirement Gathering and Analysis**: This phase involves defining the objectives of the model, such as identifying key diseases and understanding the data sources. The model's scope was established, focusing on the extraction and analysis of disease-related tweets.

ii). **Design**: During this phase, the model's architecture will be designed, specifying how data will be collected, pre-processed, and analysed. The design also included the selection of NLP techniques, such as sentiment analysis and topic modelling, that best suit the project's needs.

iii). **Implementation**: The model was developed in this phase, with coding and integration of various components like data scraping, sentiment analysis, and LDA topic modelling.

iv). **Testing**: The model underwent validation to ensure it accurately detects disease outbreaks and correctly analyses public sentiment. This phase involved validating the model using real-world data.

v). **Deployment and Monitoring**: After successful testing, the model was deployed to monitor real-time social media data for emerging health concerns. Ongoing performance monitoring ensured the model remains effective in detecting outbreaks.

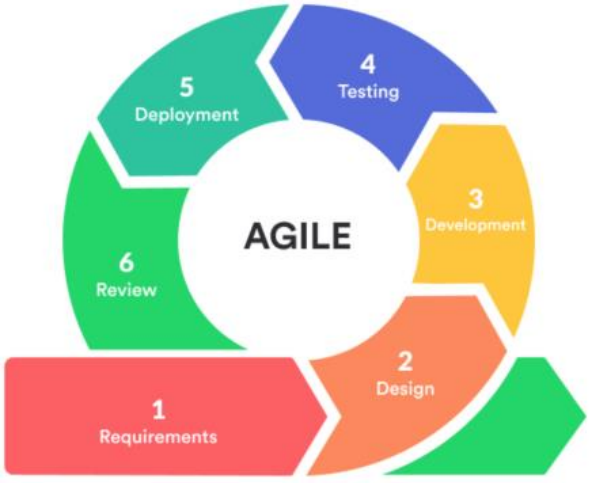


Figure 3.6: Agile Methodology (Ali & Sameh, 2024)

3.7 Utilisation of Research Results

The results of this research will be utilised by government health agencies, non-governmental organisations (NGOs), and public health organisations to enhance early disease detection of infectious diseases. By identifying emerging outbreaks through social media discussions, the government can respond faster by deploying resources where they are most needed and adjusting public health policies accordingly. The research also contributes to advancing the application of NLP in public health surveillance.

3.8 Dissemination of Research Results

The research findings will be disseminated through Strathmore University's digital repository and academic publications and conference presentations and health authorities and organizations will receive reports. Additionally, the results will be made publicly available through open-access repositories to encourage the wider application of this research in similar contexts. The goal is to increase awareness of the potential of social media data in health surveillance.

3.9 Ethical Considerations

Ethical considerations were strictly adhered to throughout the research process. Since the data will be scraped from publicly available tweets, no personal identifying information was collected or used. The study ensured compliance with privacy regulations and ethical guidelines regarding the use of social media data. Informed consent will be obtained where necessary, and sensitive information was handled with the utmost care to avoid harm. The researcher obtained clearance from the Strathmore Ethical Review Committee before commencing the research.

Chapter 4: System Analysis and Design

4.1 Introduction

The process of system analysis and design is fundamental in developing a structured and efficient software system. System analysis entails examining the problem domain, recognizing user requirements, and outlining the system's functional and non-functional specifications. This process ensures that the system meets its intended goals while adhering to technical and operational limitations. In contrast, system design converts these specifications into a detailed architecture, outlining how different components will interact to achieve the desired functionalities.

4.2 Requirement Specifications

Requirement specifications define the functional and non-functional expectations that guide the system's development. These requirements were identified through secondary research methods, which involved reviewing existing models and frameworks used in disease outbreak detection and machine learning-based predictive systems. By analysing past studies and relevant technologies, the system's requirements were established to ensure its effectiveness in tracking disease mentions and predicting potential outbreaks.

4.2.1 Functional Requirements

Functional requirements define the core capabilities of the system, ensuring that it performs its intended operations effectively. The Disease Outbreak Detection Tool must support the following functionalities:

i). The system shall allow a user to select a disease outbreak they want to monitor

ii). The system shall retrieve disease-related tweets in real time using the Twitter API.

iii). The system shall process tweets by removing irrelevant elements such as URLs, hashtags, and user mentions.

iv). The system shall perform sentiment analysis to categorize tweets as positive, negative, or neutral, providing insights into public perceptions of the mentioned diseases.

v). The system shall predict potential outbreaks using an LSTM machine learning model, which analyses historical tweet data to forecast outbreaks. The model should produce binary outputs indicating whether an outbreak is likely or not.

vi). The system shall provide visual representations of disease on a map of areas with potential outbreaks.

4.2.2 Non-Functional Requirements

Non-functional requirements define the performance, security, and operational constraints of the system. These include:

i). Performance – The system shall process large volumes of real-time data efficiently, ensuring that data collection, preprocessing, and prediction occur with minimal latency.

ii). Scalability – As the system expands to monitor additional diseases or geographic regions, it shall accommodate increasing data volumes without compromising performance.

iii). Reliability – The system shall maintain a high degree of reliability, ensuring that outbreak detection results are accurate and consistently delivered.

iv). Security – The system shall implement data security protocols, including encryption and authentication, to protect confidential user data and API logins.

v). Usability – The interface shall be intuitive, allowing health officials to efficiently access data and predictions without requiring extensive technical expertise.

4.3 System Architecture

The architecture in Figure 4.1 defines the high-level structure of the system, specifying how its components interact to achieve the desired functionalities. The Disease Outbreak Detection Tool follows a modular architectural design, ensuring that different components operate independently while integrating seamlessly.

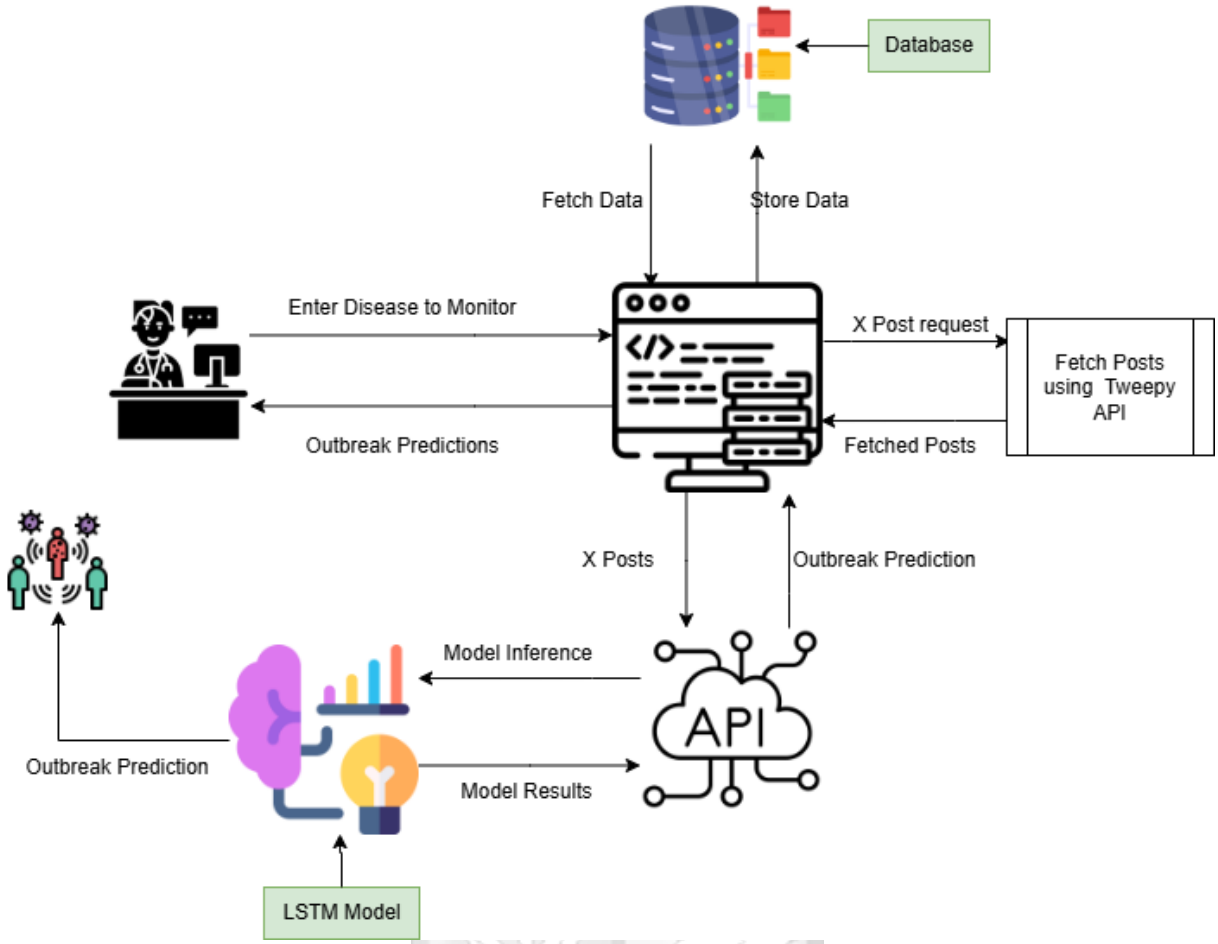


Figure 4.1: System Architecture.

4.4 System Design

System design focuses on translating the system requirements into structured models that define how different components interact. The Disease Outbreak Detection system follows an Object-Oriented Analysis and Design (OOAD) approach, where each module is treated as an object with specific attributes and behaviours. This approach ensures modularity, ease of maintenance, and future extensibility. Since the system is intended for health officials, the design prioritizes an intuitive user interface that simplifies data interpretation. The system's components are structured to allow seamless data flow, from data collection through to analysis and visualization.

4.4.1 Use Case Diagram

A use case diagram illustrates the system's core functionalities and how the health official interacts with the system. The key interactions for the use case diagram are shown in Figure 4.2 below.

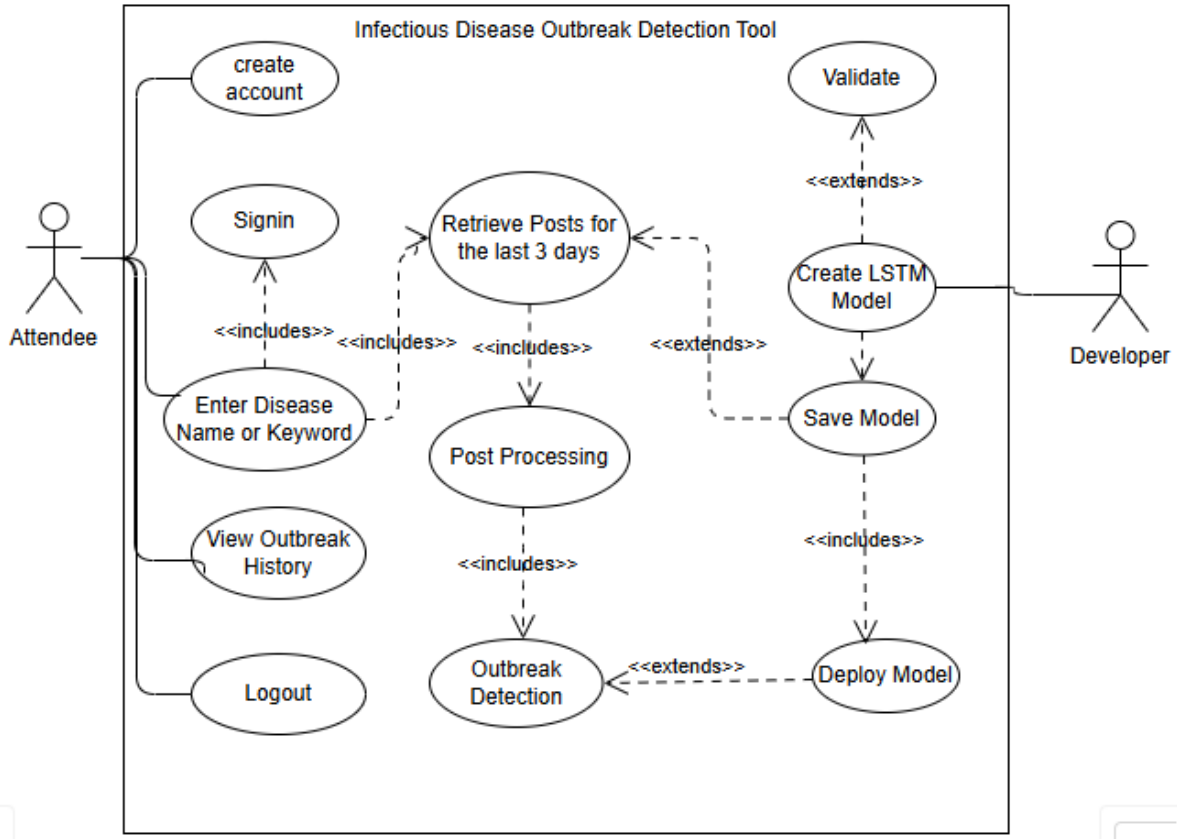


Figure 4.2: Use Case Diagram.

4.4.1.1 Detailed Use Case Descriptions

Table 4.1 shows the detailed description of use cases in Figure 4.2

Table 4.1: Description of use cases

Use Case	Pre-Conditions	Main Success Scenario	Post Conditions
Collect Tweets	Twitter API authentication is successful.	Tweets related to disease keywords are collected and stored.	Data is available for processing.
Perform Sentiment Analysis	Collected tweets exist in the system.	Tweets are analysed and classified based on sentiment.	Sentiment data is stored for visualization and prediction.
Detect Outbreaks	Pre-processed data is available.	The LSTM model analyses trends and predicts outbreaks.	Prediction results are stored for user access.
View Visualizations	Predictions and sentiment analysis are complete.	The user interacts with charts displaying trends and insights.	The system updates insights based on real-time data.

4.4.2 Class Diagram

The class diagram depicted in Figure 4.3 is a key component of the system's object-oriented design, offering a visual representation of the classes, their attributes, and the relationships between them. In the Disease Outbreak Detection Tool, the class diagram models the core components of the system, each designed to encapsulate specific functionalities and data handling processes. The primary class, Tweet, serves as the foundational data entity, representing each individual tweet collected by the system. This class includes attributes such as the tweet's text, timestamp, sentiment (derived from sentiment analysis), and location. The DataProcessor class is responsible for handling the preprocessing of the raw tweet data. It cleans the text by removing irrelevant elements (e.g., URLs, hashtags, mentions), performs sentiment analysis, and prepares the data for further processing. The OutbreakPredictor class encapsulates the core machine learning model, specifically the Long Short-Term Memory (LSTM) model, which processes the pre-processed tweet data to identify patterns and predict potential outbreaks. This class interacts with historical tweet data and applies predictive algorithms to forecast disease trends. The UserInterface class is responsible for managing the interaction between the health official and the system. It allows the user to view collected data, explore trends, and visualize outbreak predictions through intuitive graphical representations.

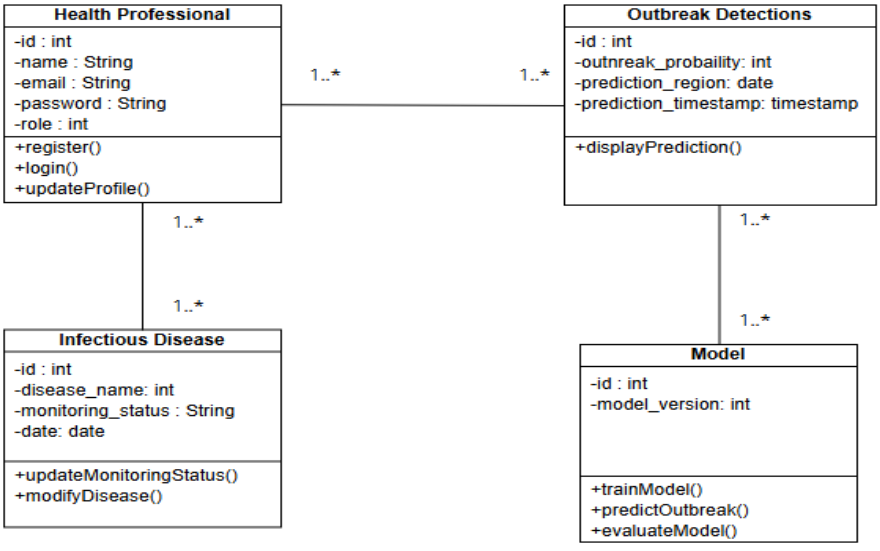


Figure 4.3: Class Diagram.

4.4.3 Sequence Diagram

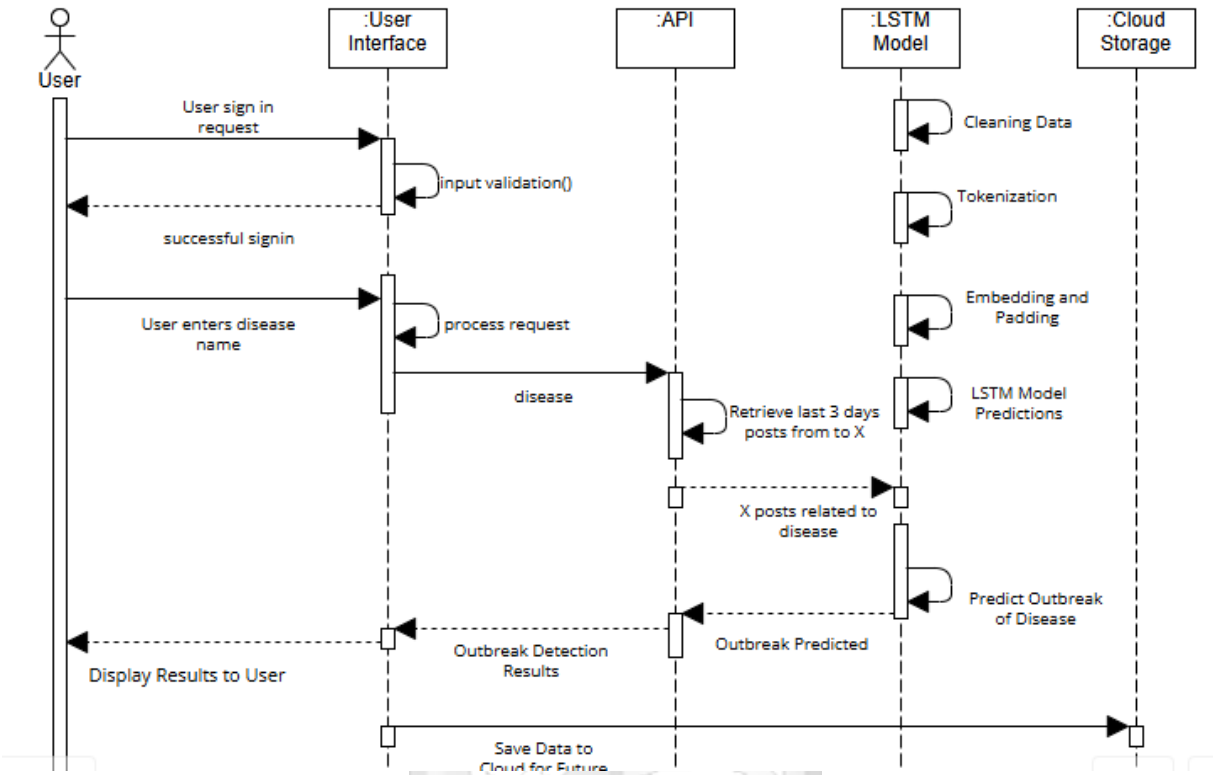


Figure 4.4: Sequence Diagram

4.4.4 Database Schema

The database schema in Figure 4.5 outlines the structure of the system's database, which is designed to efficiently store and manage the large volume of data collected from Twitter for outbreak detection. This schema is crucial for organizing the data into relevant tables, ensuring that information related to disease outbreaks, sentiment analysis, and user interactions is structured and easily accessible for processing.

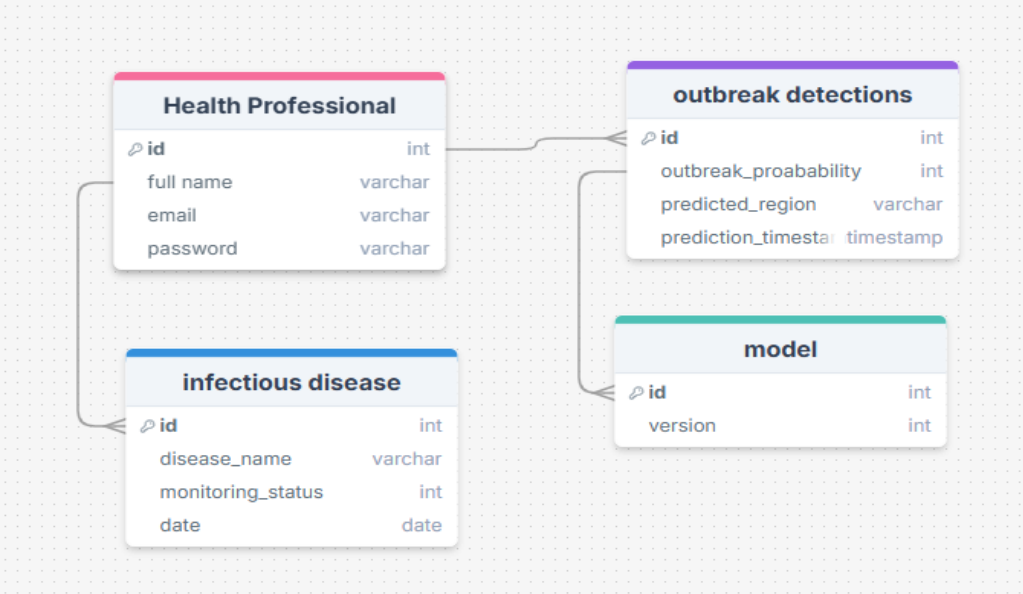


Figure 4.5: Database Schema

4.5 Wireframes

Wireframes are low-fidelity visual representations of the user interface. They serve as an essential tool in the design process, providing a blueprint for the layout and functionality of each screen in the application. In the case of the Disease Outbreak Detection system, wireframes were developed for key interfaces to ensure that the system is interactive and user-friendly.

4.5.1 Home Wireframe

This wireframe serves as the landing page of the system, providing a summary of disease outbreaks, trends, and recent activities. The wireframe includes essential components such as a navigation menu, a dashboard for monitoring outbreak statistics, and a section displaying recent disease mentions and analysis results. The user is presented with an overview of the system's current status and can easily navigate to detailed views for further exploration. The visual representation of the wireframe is shown in Figure 4.6.

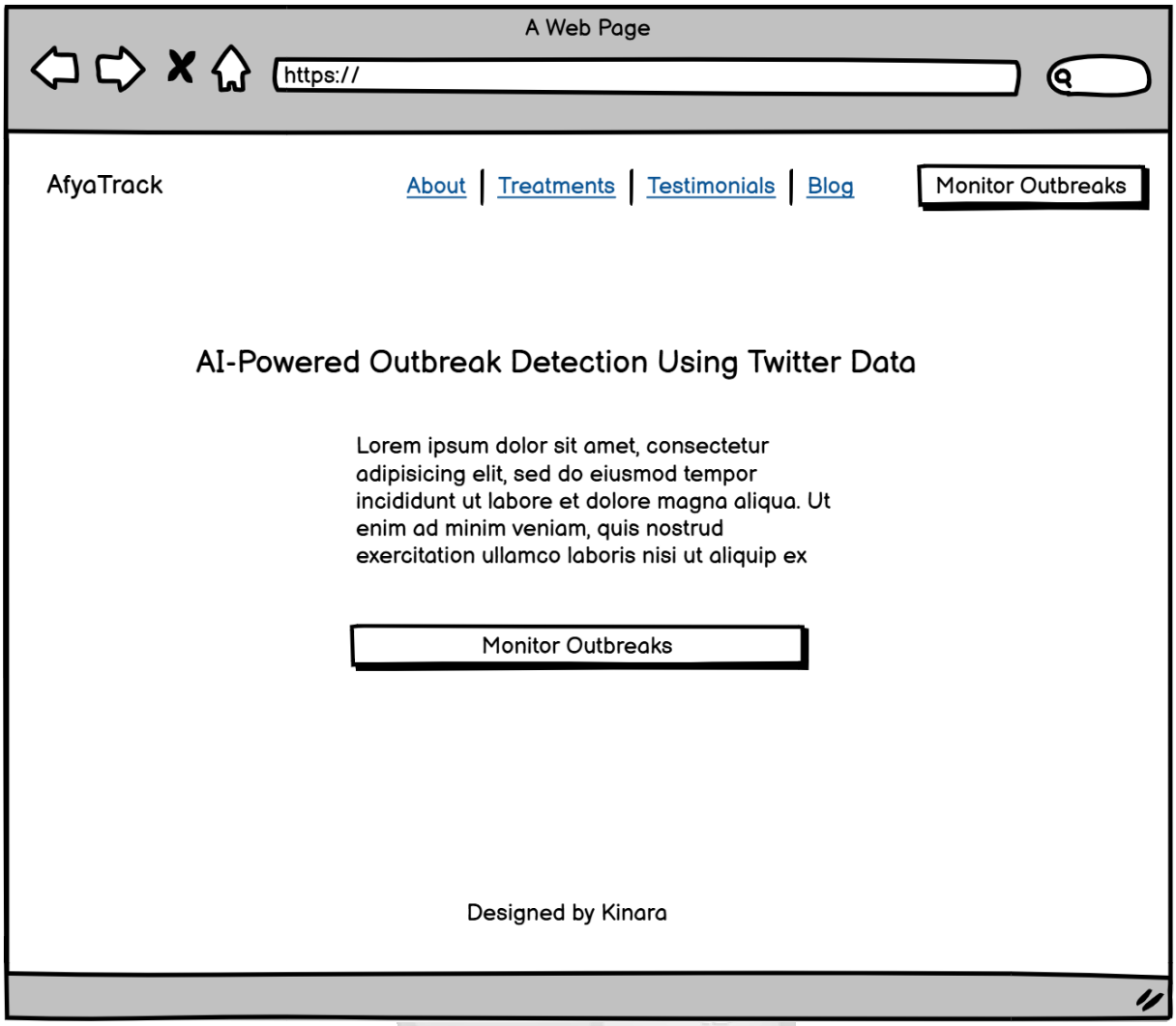


Figure 4.6: Home Page Wireframe

4.5.2 Login Wireframe

The Login wireframe in Figure 4.7 the first point of interaction for users, providing secure access to the system. The wireframe for this page includes fields for entering username and password, along with options for password recovery and new user registration. The simple design ensures that the login process is straightforward and secure, allowing only authorized health officials to access the system.

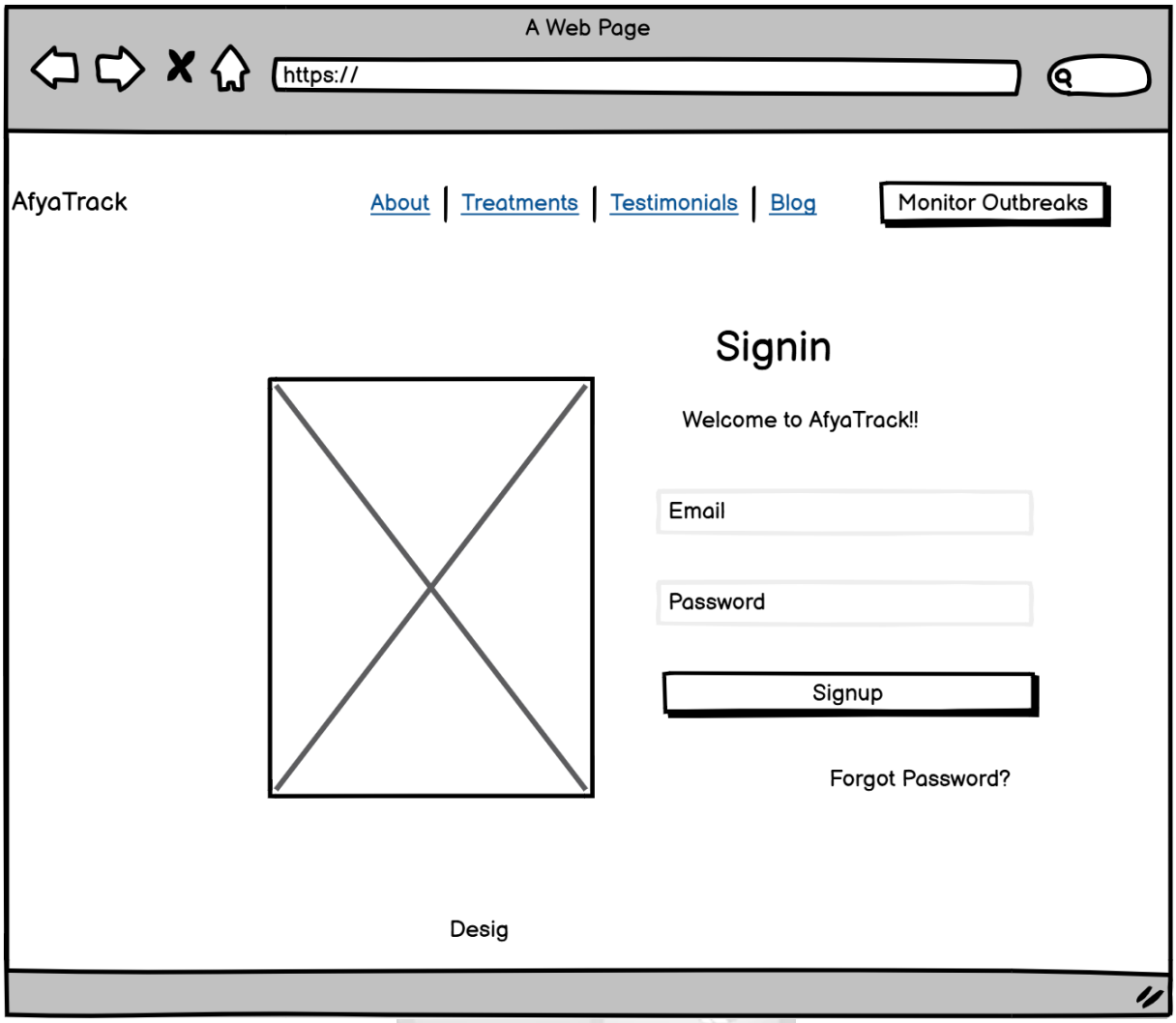


Figure 4.7: Login Wireframe

4.5.3 Register Wireframe

The wireframe allows new users, such as health officials, to create an account and gain access to the system. The wireframe includes fields for entering personal details, professional information, and login credentials. The design prioritizes simplicity and clarity, ensuring that registration is a quick and straightforward process. Figure 4.8: Register Wireframe illustrates how the registration page is structured, providing an easy-to-use interface that supports user-friendly account creation.

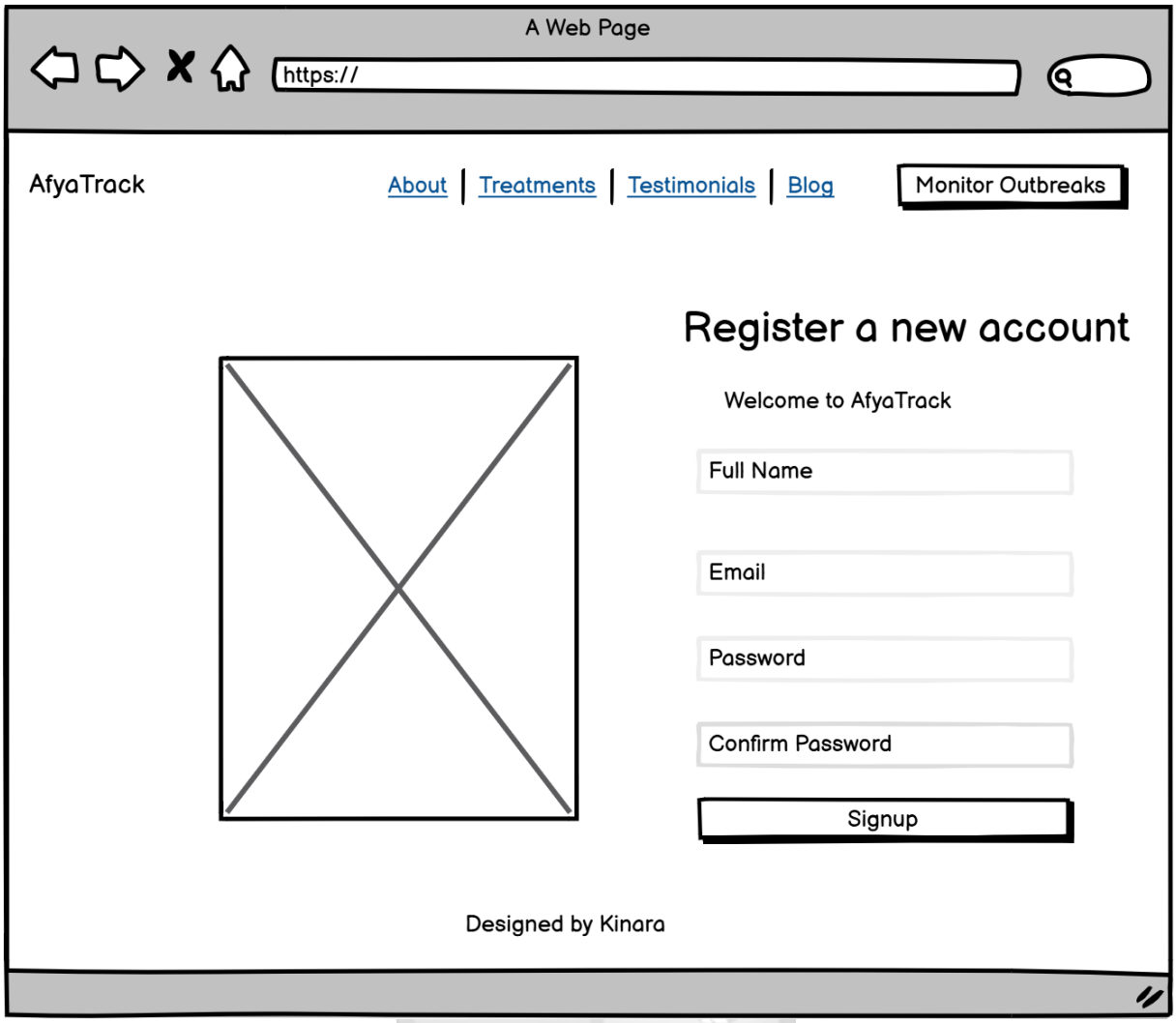


Figure 4.8: Register Wireframe

4.5.4 Monitor Outbreaks Wireframe

The monitor Outbreaks wireframe, as depicted in Figure 4.9, enables health officials to track disease outbreaks based on the data collected from Twitter. The wireframe includes charts and graphs that visualize trends in disease mentions, sentiment distributions, and predicted outbreaks. This page allows users to interact with the data, drilling down into specific disease mentions, locations, and time periods.

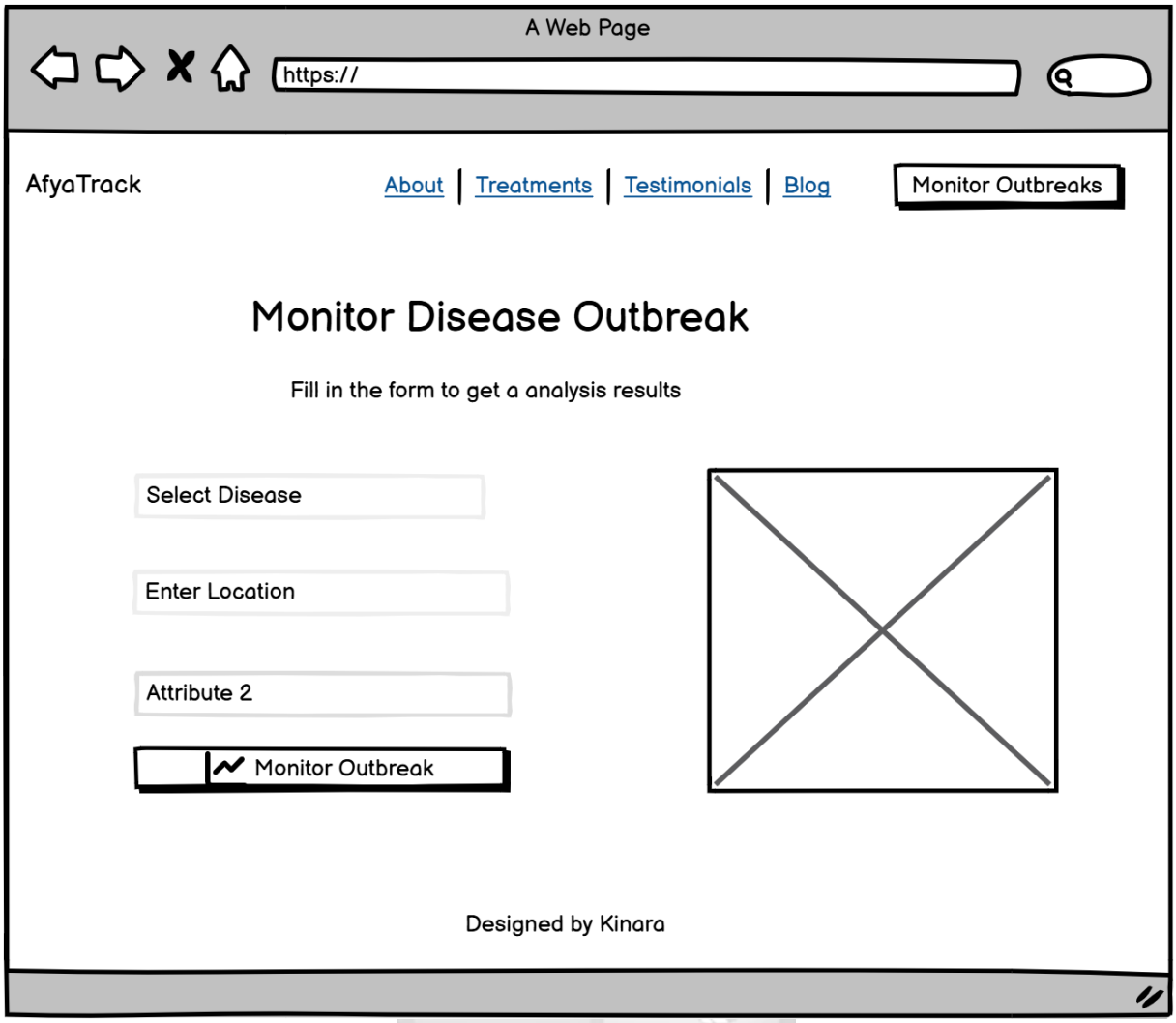


Figure 4.9: Monitor Outbreaks Wireframe

4.5.5 Results Wireframe

The results wireframe, presented in Figure 4.10, displays the outcome of the outbreak detection process, providing health officials with a detailed view of predictions, trends, and analysis. The wireframe features graphs and tables that summarize the detected outbreaks and their likelihoods, enabling users to make informed decisions about public health responses.

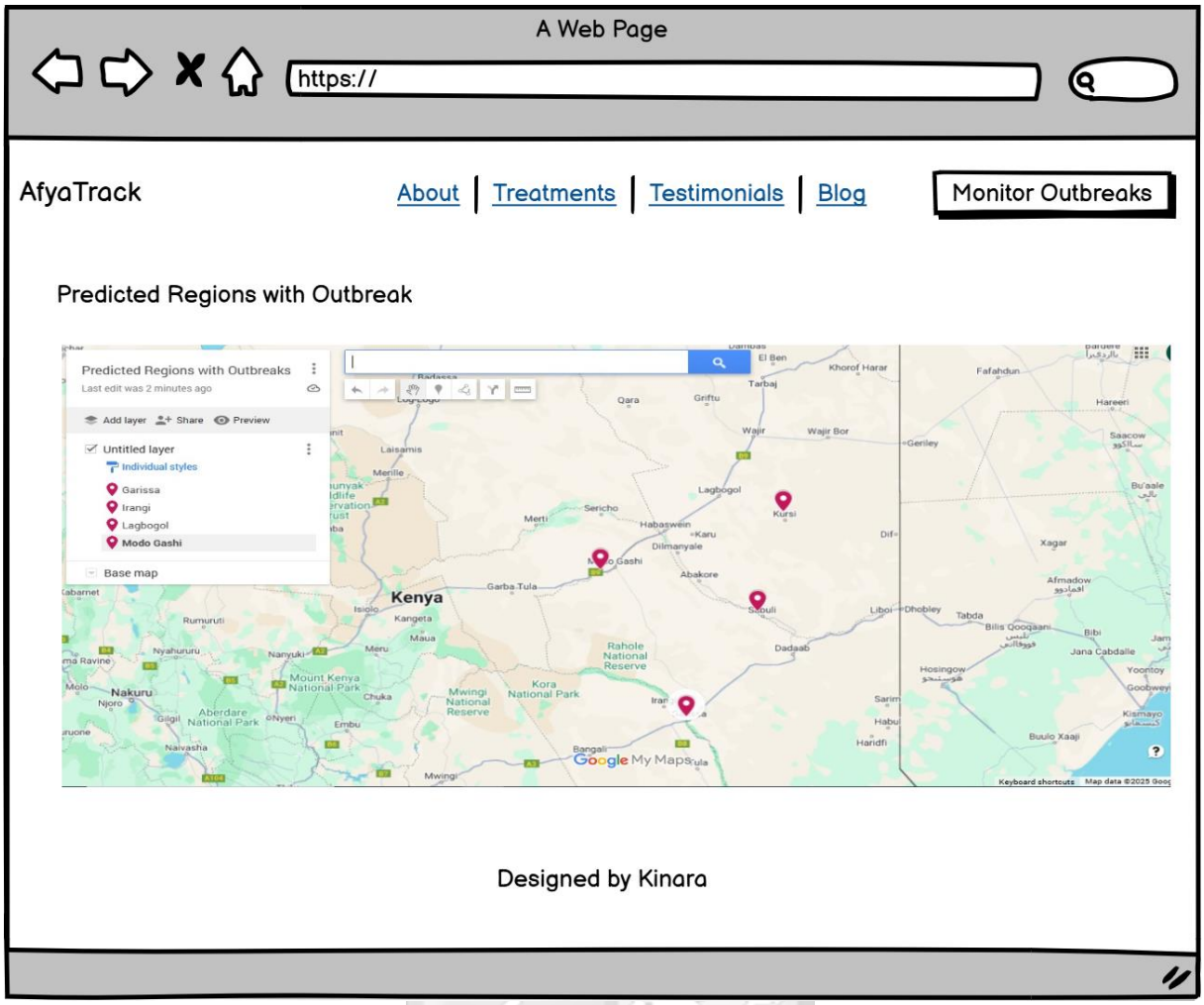


Figure 4.10: Results Wireframe

Chapter 5: System Implementation and Testing

5.1 Introduction

This phase involves the implementation of the solution guided by the design and architecture of the system. This stage involves the development, integration, and deployment of various components that were defined during the design process. For the Disease Outbreak Detection system, implementation encompasses the development of data collection pipelines, data preprocessing techniques, machine learning models for outbreak prediction, and the user interface for health officials. Testing is a critical aspect of the implementation phase, ensuring that the system functions as intended and meets the predefined requirements. The system is subjected to different testing methodologies to validate its performance, accuracy, reliability, and usability. This chapter discusses the core components of the system, outlines the development environment, and details the methodologies used in testing the system's functionalities and model performance.

5.2 Model Components

The Long Short-Term Memory (LSTM) model used in this system was designed to predict disease outbreaks by analysing historical tweet data aggregated over time. LSTM is particularly suitable for this task due to its capacity to capture long-term dependencies in sequential data, making it effective for time-series predictions such as disease outbreaks, which evolve over time. This architectural choice was based on the need to process temporal patterns in social media data, where the relationship between early signals and eventual outbreaks is typically spread across time. The model architecture comprises two primary LSTM layers, each followed by a Dropout layer to prevent overfitting. The LSTM layers enable the model to learn complex patterns inherent in the sequential nature of tweets, while the Dropout layers reduce overfitting by randomly setting some neuron outputs to zero during training, promoting generalization. Finally, a Dense output layer was used to provide the final prediction, which outputs a continuous value that correlates with the likelihood of an outbreak occurring.

5.2.1 LSTM Layers

The purpose of the LSTM layers is to process sequential data (tweets over time) to identify patterns and trends indicative of disease outbreaks. The model starts with two LSTM layers, each consisting of 50 units. The first LSTM layer returns sequences, meaning it passes its output as a sequence to the next layer. This is necessary because the LSTM needs to capture temporal

dependencies across multiple time steps. The input shape of the first layer is defined as (X_train.shape[1], 1), where X_train.shape[1] represents the number of time steps and 1 denotes a single feature per step. The second LSTM layer, also with 50 units, does not return sequences. Instead, it processes the data received from the first LSTM layer and compresses it into a single output value. This value serves as the primary feature for predicting disease outbreaks. By utilizing two LSTM layers, the model can capture both short-term and long-term dependencies within the data. Figure 5.1 visually depicts this architecture, demonstrating how the data flows through the layers.

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dropout, Dense

# Initialize the model
model = Sequential()

# First LSTM layer
model.add(LSTM(units=50, return_sequences=True, input_shape=(X_train.shape[1], 1)))
```

```
# Second LSTM layer
model.add(LSTM(units=50, return_sequences=False))
```

Figure 5.1: LSTM Layers

5.2.2 Dropout Layers

The Dropout **layers** are introduced after each LSTM layer to mitigate overfitting, which can occur when a model becomes too tailored to the training data. The Dropout rate is set to 0.2, meaning that 20% of the neurons in these layers are randomly "turned off" during training. This randomness forces the model to learn more generalized features, preventing it from overly relying on any single neuron or feature. Dropout is a regularization technique that enhances the model's ability to generalize to new, unseen data, making it more robust in real-world applications. Figure 5.2 provides a visual representation of how the Dropout layers are positioned after each LSTM layer, ensuring that overfitting is reduced while maintaining the integrity of the model's learning process.

```
# Dropout layer
model.add(Dropout(0.2))
```

```
# Dropout layer to avoid overfitting
model.add(Dropout(0.2))
```

Figure 5.2: Dropout Layers

5.2.2 Dense Output Layer

The Dense output layer is the final layer in the LSTM model, and it serves the purpose of generating the prediction. Since the task is a regression problem (predicting the likelihood of an outbreak), the output is a continuous value. This output represents the predicted magnitude of the disease outbreak based on the input data. The Dense layer consists of a single neuron, which processes the output from the previous layers and produces the final prediction. The purpose of the Dense layer is to consolidate all the learned information from the LSTM layers and produce a scalar output that indicates the likelihood of a disease outbreak. Figure **5.3** shows the structure of this layer and its role in generating the final prediction.

```
# Output layer with one neuron for regression
model.add(Dense(units=1))
```

Figure 5.3: Dense Output Layer

5.3 Disease Outbreak Detection Tool

The Disease Outbreak Detection Tool developed in this study is a real-time surveillance system designed to monitor social media platforms for early signals of infectious disease outbreaks. The tool integrates multiple components, including data collection, preprocessing, sentiment analysis, machine learning model development, and visualization of results. This section outlines the core features of the tool's user interface and its functionality for detecting disease outbreaks.

5.3.1 Home Interface

The Home Page serves as the entry point to the system, presenting health officials with an overview of the most recent disease mentions, sentiment trends, and predictive analysis results.

Figure 5.4: Home Page illustrates the design of the home interface.

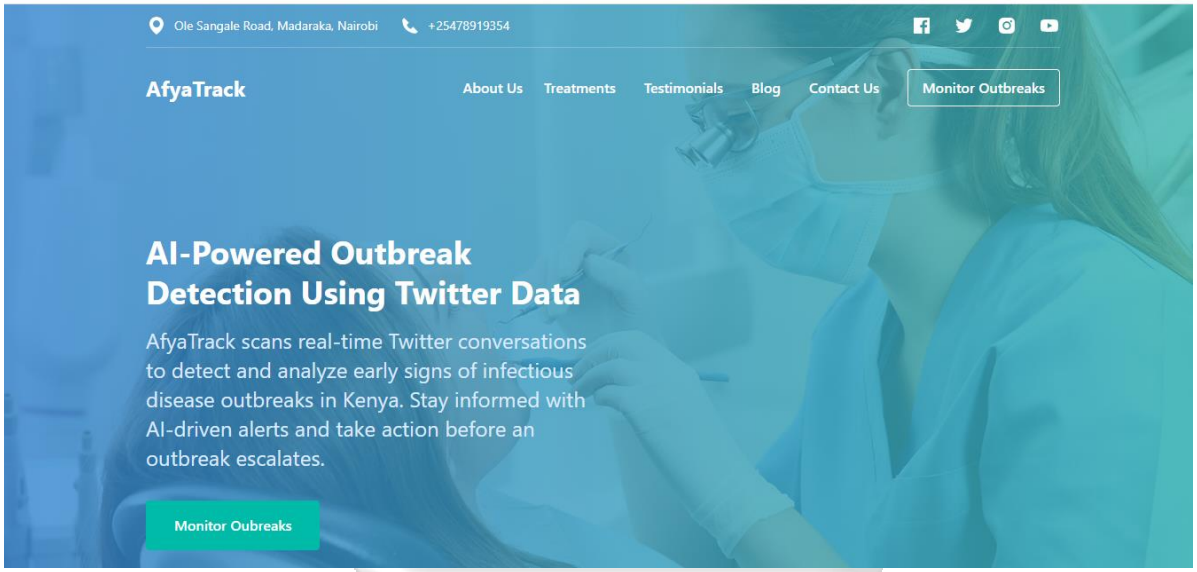


Figure 5.4: Home Page

5.3.2 Monitor Outbreak Interface

The Monitor Outbreak Page is the central tool for tracking disease outbreaks in real-time. This page displays a form that allows users to track an outbreak of the infectious diseases. The Monitor Outbreak Page also allows users to zoom in on specific time periods, regions, or diseases to investigate further. Figure 5.5: Monitor Outbreaks Interface provides a visual representation of this feature.

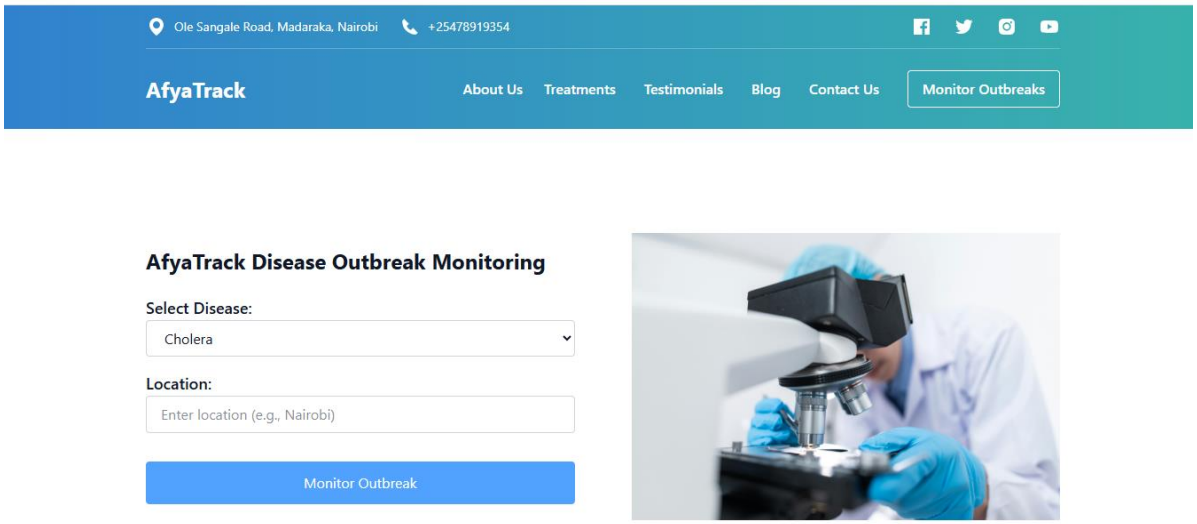


Figure 5.5: Monitor Outbreaks Interface

5.3.3 Detection Results Interface

The Results Page is where health officials can view the model's predictions and outbreak likelihoods n given geographical locations and the disease, they queried. This page presents a clear overview of the predicted outbreaks, including a confidence score that indicates the probability of an outbreak. Figure 5.6: Detection Results Interface shows how these results are displayed to users.

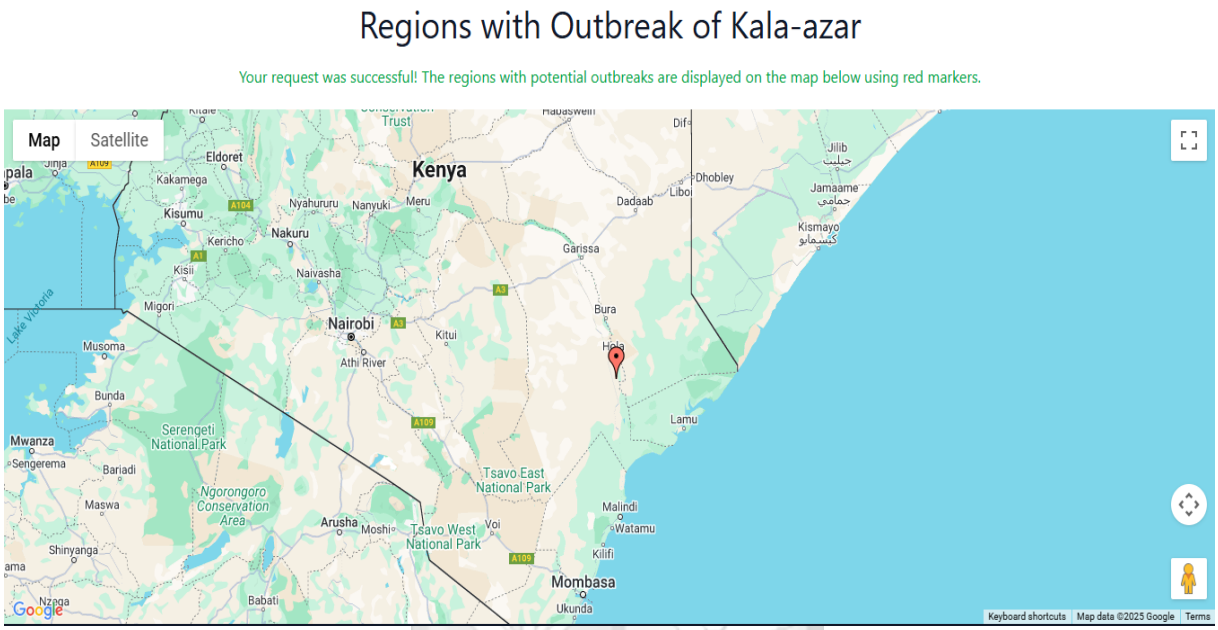


Figure 5.6: Detection Results Interface

5.4 System Implementation

The process of system implementation is vital in transforming the theoretical design and architecture into a fully functioning solution. In this phase, various system components—such as data collection, preprocessing, model training, and the creation of the user interface—are integrated to form a cohesive system. The Disease Outbreak Detection Tool leverages a modular approach, where each component operates independently, yet contributes to the overall task of monitoring and predicting potential disease outbreaks. By using tools such as Python, Google Colab, Flask, and HTML, the system has been developed to efficiently process real-time data, apply machine learning models, and provide actionable insights for health officials.

5.4.1 Development Environment

The development environment was chosen with great care to ensure efficient coding, testing, and deployment. The following tools were utilized:

i). Python: Used for data preprocessing, feature engineering, and machine learning model development.

ii). Google Colab: Provided a cloud-based platform for training the LSTM model and conducting experiments.

iii). Flask: Used to create APIs connecting the backend and frontend.

iv). HTML and Tailwind CSS: Employed to design a responsive and user-friendly interface.

v). Windows OS: Served as the primary platform for testing and running the application.

5.4.2 Data Collection

The data collection process, as depicted in Figure 5.7, entailed querying Twitter for tweets related to infectious diseases and their geographical locations. The Twitter API is used to collect tweets in real-time, which are then stored for further analysis. This module retrieves data based on predefined disease-related keywords, ensuring the system stays up to date with current health trends. A significant challenge in working with social media data is the informal language commonly used by users. Tweets often contain slang, abbreviations, and informal syntax, which traditional models may fail to process effectively. To address this challenge, the study incorporated preprocessing steps such as tokenization, lowercasing, and lemmatization, which helped the model better understand the context of words used informally. Additionally, the model was trained on a large, labelled dataset to ensure robustness to these variations in language. Another limitation addressed was sample bias. Social media data often exhibits skewed representation, where certain regions, demographics, or health-related topics are overrepresented, while others are underrepresented. This bias can negatively impact model predictions, resulting in poor generalization. To mitigate this issue, ta augmentation techniques were utilized to generate a more diverse sample set, thereby reducing the effect of underrepresented categories.

```python
def collect_kenya_tweets():
    # Define the search terms for infectious diseases and locations in Kenya
    search_terms = ['cholera', 'measles', 'malaria', 'polio', 'Kenya', 'Garissa', 'Mandera','Wajir']

    # Set up the tweet collection parameters
    geo_location = '1.2921,36.8219,1000km'  # Garissa coordinates and a 1000km radius
    language = 'en'  # Only English tweets

    # Collect tweets
    tweets = tweepy.Cursor(api.search_tweets,
                           q=" OR ".join(search_terms),
                           geocode=geo_location,
                           lang=language,
                           tweet_mode='extended').items(1000)

    # Store the tweets in a list
    tweet_data = []
    for tweet in tweets:
        tweet_data.append([tweet.created_at, tweet.full_text, tweet.user.screen_name, tweet.user.location])
```

Figure 5.7: Data Collection

5.4.3 Data Pre-processing

Data preprocessing stands as a vital process which converts raw data into suitable material for analysis. The machine learning model requires preprocessing of collected tweets through multiple processing stages.

5.4.3.1 Data Cleaning

The tweet text is cleaned by removing irrelevant elements, including URLs, user mentions, and hashtags. This ensures that only meaningful content is used for further analysis. Figure 5.8 provides a clear visual representation of how the cleaning process enhances the quality of the data for further analysis.

```python
def clean_text(text):
    # Remove URLs
    text = re.sub(r'http\S+', '', text)
    # Remove mentions (@usernames)
    text = re.sub(r'@\S+', '', text)
    # Remove hashtags (#topic)
    text = re.sub(r'#\S+', '', text)
    # Remove non-alphabetical characters and convert to lowercase
    text = re.sub(r'\W', ' ', text)
    text = text.lower()
    # Tokenize the text (split by spaces)
    words = text.split()
    # Remove stopwords (common words that don't add much meaning)
    stop_words = set(nltk.corpus.stopwords.words('english'))
    words = [word for word in words if word not in stop_words]
    # Rejoin the words into a cleaned text
    return ' '.join(words)
```

Figure 5.8: Data Cleaning

5.4.3.2 Sentiment Analysis

Sentiment analysis is applied to classify each tweet as positive, negative, or neutral based on the text's sentiment. This classification provides valuable insights into the public's perception of disease outbreaks. Figure 5.9: shows how the sentiment analysis process categorizes tweets based on the emotional tone conveyed.

```python
def get_sentiment(text):
    analysis = TextBlob(text)
    if analysis.sentiment.polarity > 0:
        return 'positive'
    elif analysis.sentiment.polarity < 0:
        return 'negative'
    else:
        return 'neutral'
```

```python
# Apply sentiment analysis to each tweet
tweets_df['sentiment'] = tweets_df['cleaned_text'].apply(get_sentiment)
```

Figure 5.9: Sentiment Analysis

5.4.3.3 Temporal Aggregation

Tweets are aggregated into daily time windows, creating a time-series dataset that allows the system to analyse trends in disease mentions over time. Figure 5.10 illustrates how tweets are aggregated into daily windows for trend analysis.

```
# **Step 5: Create a Time-Series of Disease Mentions**
tweets_df['date'] = pd.to_datetime(tweets_df['date'])  # Convert the date column to datetime
tweets_df.set_index('date', inplace=True)

# Count the number of disease-related mentions per day
daily_mentions = tweets_df.resample('D').size()
```

Figure 5.10: Temporal Aggregation

5.4.3.4 Feature Engineering

Key features are extracted from the cleaned and processed data, including sentiment scores, counts of disease mentions, and temporal features, all of which are crucial for training the prediction model. Figure 5.11 demonstrates how these features are derived and organized for use in the machine learning model.

```
# Create the dataset for LSTM: We need to create sequences of past data to predict the future.
def create_dataset(data, time_step=1):
    X, y = [], []
    # Adjusted loop condition to handle smaller test_data
    for i in range(len(data) - time_step - 1):
        X.append(data[i:(i + time_step), 0])
        y.append(data[i + time_step, 0])
    return np.array(X), np.array(y)

# Define time step (how many previous days to use for prediction)
time_step = 3  # Use the past 7 days to predict the next day
X_train, y_train = create_dataset(train_data, time_step)
X_test, y_test = create_dataset(test_data, time_step)

# Check if X_test is empty before reshaping
if X_test.size > 0:  # Only reshape if X_test has data
    # Reshape the input to be suitable for LSTM [samples, time steps, features]
    X_train = X_train.reshape(X_train.shape[0], X_train.shape[1], 1)
    X_test = X_test.reshape(X_test.shape[0], X_test.shape[1], 1)
else:
    print("X_test is empty. Adjust data splitting or time_step.")
```

Figure 5.11: Feature Engineering

5.4.4.4 Data Scaling

The application of feature scaling normalizes numerical data to exist within a single range. The application of this step becomes essential for LSTM models to achieve training convergence. Figure 5.12 illustrates the application of scaling to the dataset.

```
# **Step 6: Prepare Data for LSTM**
scaler = MinMaxScaler(feature_range=(0, 1))
scaled_data = scaler.fit_transform(daily_mentions.values.reshape(-1, 1))
```

Figure 5.12: Data Scaling

5.4.4.5 Splitting Data

The data is divided into training and testing sections where 80% of the data serves for training purposes and 20% remains for testing purposes. The model benefits from this data arrangement because it trains using the provided information while testing occurs on fresh data to measure its ability to make predictions. Figure 5.13 represents how the data is split for training and testing.

```
# Split the data into training and testing sets (80% training, 20% testing)
train_size = int(len(scaled_data) * 0.8)
train_data, test_data = scaled_data[:train_size], scaled_data[train_size:]
```

Figure 5.13: Splitting Data

5.4.4 Training Model

The trained LSTM model receives preprocessed data to learn about disease mention patterns within sequential data. Training weights of the model follows a process that minimizes prediction errors. **Figure 5.14** shows the steps involved in the training process.

```
# Compile the model
model.compile(optimizer='adam', loss='mean_squared_error')
```

```
# Train the model
model.fit(X_train, y_train, epochs=20, batch_size=32, validation_data=(X_test, y_test))
```

```
# **Step 8: Predict Disease Trends**
predicted_mentions = model.predict(X_test)

# Inverse transform the predictions to get the original scale
predicted_mentions = scaler.inverse_transform(predicted_mentions)
```

Figure 5.14: Model Training

5.4.5 API

A Flask API is used to connect the backend machine learning model with the frontend user interface. The API handles user requests, triggers the model's prediction process, and sends the results back to the user interface for visualization.

5.5 Testing

System testing is essential for validating the functionality and performance of the Disease Outbreak Detection system. Once all components are integrated, confirming that the system functions as expected, accurately processes input data, and provides reliable predictions is crucial. The testing process includes various methods to assess the model's prediction accuracy, system responsiveness, and overall robustness.

5.5.1 Model Accuracy

For this tool, model accuracy is measured using mean squared error. The system's ability to identify disease outbreaks in real-time is thoroughly evaluated by applying these testing methods. Testing the trained model on new, unseen data assessed the model's performance. The root mean squared error (RMSE) is computed to gauge the accuracy of the model's predictions. The model achieved a mean squared error of 0.092. Figure 5.15 shows how the model's performance is evaluated.

Figure 5.15: Model Accuracy

Chapter 6: Discussions

6.1 Introduction

In recent years, detecting infectious disease outbreaks through social media has garnered significant attention, fuelled by the increasing availability of digital data and the urgent need for rapid public health interventions. Social media platforms, particularly Twitter, have become vital sources of real-time information, providing valuable insights that complement traditional epidemiological methods. The use of Natural Language Processing (NLP) models enables the extraction of relevant health data from large volumes of unstructured text. These techniques support early detection of disease outbreaks and allow for real-time monitoring, potentially improving public health responses.

6.2 Challenges Faced in Outbreak Detection of Infectious Diseases

Disease outbreak detection from social media data is an inherently challenging task due to several factors. First, the noisy nature of social media data complicates the detection process. Social media posts often contain irrelevant or misleading information, such as personal opinions or off-topic discussions, which can obscure true outbreak signals. Ordun et al. (2020) highlighted how users may tweet about symptoms that are not related to infectious diseases, which can lead to false positives in the detection models. For example, posts about "headaches" or "fatigue" might be associated with various non-epidemic causes, potentially skewing the results. Additionally, the problem of data imbalance persists. Disease-related posts may constitute a small fraction of the total volume of social media activity, making it difficult for models to identify outbreaks in the early stages (Chakraborty et al., 2020). In many cases, only a small fraction of users engages in discussions about disease-related topics, while the majority of social media content focuses on unrelated subjects. This imbalance impacts the model's ability to accurately identify and predict outbreaks. To address this challenge and improve model performance, advanced data preprocessing and feature engineering are essential. Another challenge in outbreak detection is the multilingual nature of social media data. As highlighted by Li et al. (2020), social media users communicate in a variety of languages and dialects, adding complexity to data processing. In regions like Kenya, where multiple languages are spoken, it is crucial for NLP models to handle different languages effectively to ensure comprehensive surveillance. Although models such as XLM-R have been developed to process multilingual data, they require large, diverse datasets for effective training, which can be a significant limitation in resource-limited environments.

6.4 Existing Models and Frameworks in Detecting Outbreaks of Infectious Diseases

Existing frameworks for disease outbreak detection often combine traditional epidemiological methods with modern machine learning and NLP techniques. These hybrid models aim to bridge the gap between real-time data from social media and the structured data typically used in public health surveillance. For example, Krauer & Schmid (2022) used geospatial analysis combined with NLP to track the spread of plague outbreaks, demonstrating the importance of integrating location-based data into outbreak detection systems. Additionally, Müller et al. (2023) introduced COVID-X-BERT, a transformer model specifically designed to analyse social media posts related to the COVID-19 pandemic. This model outperformed traditional NLP models, achieving better accuracy in sentiment classification and thematic analysis. The incorporation of sentiment analysis allows for more accurate detection of public health concerns, such as fear or panic, which can signal an impending outbreak. However, as noted in Bengesi et al. (2023), the application of these models is often restricted by the quality of the data available, especially in low-resource settings. Despite the progress made, the integration of geospatial data and disease-specific keywords into these frameworks is still in its infancy. This limitation highlights the need for continued research to refine models that can handle geospatial tagging, which is essential for localized outbreak detection. As Mavragani (2020) pointed out, the lack of location-based data in social media posts poses a significant barrier to accurately pinpointing the areas most affected by disease outbreaks.

6.5 Disease Outbreak Detection Models

The Disease Outbreak Detection system developed in this study effectively utilizes a Long Short-Term Memory (LSTM) model to analyse disease-related tweets over time. The system processes historical tweet data that includes mentions of diseases like cholera, malaria, and typhoid, and the model uses these mentions to predict potential outbreaks. By tracking the frequency and sentiment of disease-related posts, the system identifies early trends in public concern, which can be indicative of emerging outbreaks. The LSTM model demonstrated high accuracy in tracking disease mentions and predicting outbreaks, as it was able to capture both short-term and long-term dependencies within the tweet data. One of the key findings of this research was the correlation between negative sentiment and the onset of disease outbreaks. The sentiment analysis performed by the LSTM model classified tweets as positive, negative, or neutral. An increase in negative sentiment, often linked to fear or concern about a disease, frequently preceded a rise in disease mentions, suggesting that public distress could serve as an

early indicator of an outbreak. This finding aligns with the work of Mansoor et al. (2020), who found that negative sentiment on social media could signal emerging health crises. However, despite the model's success in identifying trends and making predictions, it faced certain limitations. The model's ability to predict outbreaks early enough for effective intervention was constrained by the delay in social media reactions to disease events. Social media posts tend to emerge after outbreaks have started to spread, making it challenging for the model to detect them in their very early stages. Additionally, distinguishing between sporadic mentions of disease and an actual outbreak can be difficult, especially in cases where users might mention diseases in passing or in the context of past outbreaks. Incorporating additional data sources, such as geospatial information and real-time health reports, could enhance the model's capacity for early detection and localized outbreak prediction, improving its effectiveness in public health response efforts.

6.6 Model Validation

Model validation is a crucial step in assessing the accuracy and reliability of the Disease Outbreak Detection system. In this study, the performance of the LSTM model was evaluated using the Root Mean Squared Error (RMSE) metric, which provided a measure of the model's prediction accuracy. The model achieved a RMSE of 0.092, indicating a high level of accuracy in identifying disease outbreaks. This result suggests that the LSTM model is capable of making reliable predictions, accurately tracking disease mentions, and correlating them with outbreak likelihood. To ensure the model's robustness, cross-validation techniques were applied. These techniques help evaluate the model's performance on different subsets of the dataset, ensuring that the model does not overfit to the training data and generalizes well to new, unseen data. The results confirmed that the LSTM model maintained consistent performance across various subsets of the dataset, reinforcing its ability to predict outbreaks in real-time with accuracy. While the validation results were promising, there are areas where the model could be improved. For instance, the model's performance could be enhanced by incorporating geospatial tagging, which would enable more accurate localized outbreak predictions. By tracking the geographic distribution of disease-related posts, the model could identify outbreak hotspots and help public health officials allocate resources more efficiently. Furthermore, diversifying the training data could improve the model's ability to recognize a broader range of disease patterns and public health concerns, enhancing its capability to predict outbreaks across different regions and disease types. The need for continuous refinement and incorporation of diverse data sources is essential to improve the system's effectiveness in outbreak detection.

Chapter 7: Conclusion and Recommendation

7.1 Conclusion

This study demonstrates that Natural Language Processing (NLP) techniques can effectively detect infectious disease outbreaks early using social media data. Through the integration of sentiment analysis, topic modelling, and the Long Short-Term Memory (LSTM) model, the study successfully identified trends and potential outbreak signals from publicly available social media content. The results indicate that social media platforms, particularly X, hold significant potential as real-time sources of health-related data, which could complement and enhance traditional public health surveillance systems. The LSTM model employed in this research proved effective in detecting disease-related discussions and predicting potential outbreaks by analysing temporal trends and sentiment in social media posts. Notably, the study highlighted that negative sentiment, often expressed in public posts, correlates with the onset of disease outbreaks. However, despite the promising results, the study also identified limitations in the model's capacity for early detection, suggesting that further improvements are necessary to achieve optimal real-time detection. The findings of this study underscore the importance of incorporating social media data into the existing health surveillance frameworks, particularly in regions such as Kenya, where infectious diseases frequently occur. While the model demonstrated potential, the study also pointed to areas for enhancement, particularly in geospatial data integration and the handling of multilingual content, which are critical for accurate and localized outbreak detection.

7.2 Recommendations

This study provides valuable insights into the potential of utilizing social media data and Natural Language Processing (NLP) techniques for the early detection of infectious disease outbreaks. The following recommendations are offered to policymakers, government agencies, researchers, and public health organizations on how to best utilize and adopt this tool for enhancing public health surveillance:

i). Policymakers and government agencies should prioritize the integration of social media data into existing public health surveillance systems. By recognizing the value of real-time data from platforms like X, public health authorities can enhance their capacity to detect outbreaks at early stages. This tool can serve as a supplementary resource alongside traditional surveillance methods, improving the timeliness of responses.

ii). Researchers and public health professionals should focus on developing and implementing NLP models capable of processing multilingual data. In regions with diverse linguistic populations, such as Kenya, it is critical that surveillance tools be equipped to handle various languages and dialects. This will ensure comprehensive monitoring of disease-related discussions and more accurate outbreak detection.

iii). Policymakers and public health agencies should advocate for the inclusion of geospatial data in social media-based surveillance tools. By incorporating geolocation tagging, the model can more accurately pinpoint the geographical areas most affected by outbreaks. This will enable more targeted, region-specific responses and better resource allocation during public health emergencies.

iv). It is crucial that government bodies and health organizations collaborate with academic researchers to refine and validate NLP-based outbreak detection models. By working together, these stakeholders can ensure that the models are continuously updated with relevant data, remain accurate, and meet the evolving needs of public health surveillance.

v). Public health professionals can utilize sentiment analysis from social media platforms to assess public concern and response to emerging health threats. This tool can provide real-time insights into public perceptions, enabling health authorities to adjust communication strategies, dispel misinformation, and address public anxiety promptly.

vi). Governments and policymakers should invest in capacity building by providing training for public health professionals on how to effectively use and interpret data from social media surveillance tools. Understanding how to analyse sentiment, identify trends, and validate the results of such models will empower health officials to make data-driven decisions and improve public health outcomes.

vii). To maximize the effectiveness of social media-based disease detection, it is important that the general public be encouraged to participate in health-related discussions online. Raising awareness about the role of social media in outbreak detection can help ensure a more robust data stream and facilitate the identification of health concerns in real time. Public engagement can also contribute to improving the quality of data shared, allowing for more accurate predictions and responses.

7.3 Future work

Future work should build upon the findings of this study by addressing the following areas:

i). Future iterations of the model should integrate real-time data collection from multiple social media platforms to enhance responsiveness. A continuous stream of data would provide more timely alerts and contribute to more effective outbreak management.

ii). Researchers could explore hybrid models that combine NLP techniques with traditional epidemiological models. This would provide a more comprehensive approach to disease surveillance, allowing for the integration of both real-time data and epidemiological factors.

iii). Conducting longitudinal studies to assess the long-term effectiveness of early warnings generated by the model could offer valuable insights into the real-world impact of such systems on public health responses and resource allocation.

iv). Future research could also focus on the development of real-time intervention strategies, based on model predictions, to assist health authorities in making swift decisions regarding disease outbreaks.

7.4 Limitations

Despite the promising results, several limitations were identified in this study:

i). The research focused specifically on Kenya, and while the methodology developed could be adapted to other regions, further studies are required to evaluate its applicability in different geographic and healthcare contexts.

ii). Social media data is often unstructured and noisy, which can lead to challenges in filtering out irrelevant or misleading content. The accuracy of outbreak predictions may be compromised by false positives or extraneous data.

iii). This study was restricted to using data from X, and other platforms that feature multimedia content, such as Instagram, were excluded. Future research should expand the range of platforms included to provide a more holistic view of public health discussions.

7.5 Research Contribution

This research makes a significant empirical contribution to the field of health informatics by demonstrating the effectiveness of social media, particularly Twitter, as a tool for the real-time detection of infectious disease outbreaks. The study not only advances the application of

Natural Language Processing (NLP) techniques in health surveillance but also presents a novel approach to monitoring disease trends through social media platforms, which are increasingly used to share health-related information. By employing sentiment analysis and topic modelling on social media data, the research provides a concrete example of how NLP can be used to detect signals of disease outbreaks early, offering valuable insights into public health monitoring practices. The results of this study suggest that integrating social media data with traditional public health surveillance systems in Kenya can significantly enhance early outbreak detection capabilities. This integration can aid health authorities in making informed, real-time decisions that can improve resource allocation and expedite public health interventions. For example, during an outbreak of diseases like malaria or cholera, health authorities could leverage the findings from this research to better identify at-risk regions based on real-time discussions on social media, thus facilitating quicker responses. This methodology also offers an opportunity to augment existing surveillance systems, which often face challenges related to reporting delays and data incompleteness. Furthermore, the findings of this study lay a solid foundation for future research in public health surveillance, particularly in the context of Kenyan health authorities. The model developed can be tailored to local languages and specific regional health concerns, allowing for the creation of a robust early warning system that can be implemented nationwide. Kenyan health authorities could operationalize this approach by incorporating it into their existing systems, ensuring that real-time social media monitoring becomes an integral p art of their disease surveillance strategy. This would not only improve outbreak prediction but also enhance communication strategies during public health emergencies. The contributions of this research extend beyond the Kenyan context, offering valuable insights for health authorities in other regions with similar challenges in disease surveillance. By showcasing the potential of NLP techniques in identifying disease outbreaks through social media, this study contributes to the growing body of literature at the intersection of NLP and public health. Future applications of this model can be expanded to monitor other diseases or regions, making it adaptable and scalable across various public health sectors.

References

Abbood, A., Ullrich, A., Busche, R., & Ghozzi, S. (2020). EventEpi—A natural language processing framework for event-based surveillance. *PLOS Computational Biology, 16*(11), e1008277. https://doi.org/10.1371/journal.pcbi.1008277

Al-Garadi, M. A., Yang, Y.-C., & Sarker, A. (2022). The Role of Natural Language Processing during the COVID-19 Pandemic: Health Applications, Opportunities, and Challenges. *Healthcare*, *10*(11), 2270. https://doi.org/10.3390/healthcare10112270

Ali, Ahmed., & Sameh, Hesham. (2024). Developing a theoretical framework for enhancing green project approaches via Agile methodology. *Scientific Reports, 14*(1). https://doi.org/10.1038/s41598-024-78613-x

Alsudias, L., & Rayson, P. (2021). Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic: Adapting for Informal Language in Arabic X (Preprint). *JMIR Medical Informatics*. https://doi.org/10.2196/27670

Amin, S., Muhammad Ali Zeb, Hani Alshahrani, Hamdi, M., Alsulami, M., & Shaikh, A. (2023). Social Media-Based Surveillance Systems for Health Informatics Using Machine and Deep Learning Techniques: A Comprehensive Review and Open Challenges. *Computer Modeling in Engineering & Sciences*, *0*(0), 1–10. https://doi.org/10.32604/cmes.2023.043921

Bengesi, S., Oladunni, T., Olusegun, R., & Audu, H. (2023). A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From X Tweets. *IEEE Access*, *11*, 11811–11826. https://doi.org/10.1109/ACCESS.2023.3242290

Fallatah, D. I., & Adekola, H. A. (2024). Digital epidemiology: harnessing big data for early detection and monitoring of viral outbreaks. *Infection Prevention in Practice*, *6*(3), 100382–100382. https://doi.org/10.1016/j.infpip.2024.100382

Field, A., Park, C. Y., Theophilo, A., Watson-Daniels, J., & Tsvetkov, Y. (2022). An analysis of emotions and the prominence of positivity in #BlackLivesMatter tweets. *Proceedings of the National Academy of Sciences*, *119*(35). https://doi.org/10.1073/pnas.2205767119

Fu, K.-W., Liang, H., Saroha, N., Tse, Z. T. H., Ip, P., & Fung, I. C.-H. (2016). How people react to Zika virus outbreaks on X? A computational content analysis. *American Journal of Infection Control*, *44*(12), 1700–1702. https://doi.org/10.1016/j.ajic.2016.04.253

Gautam, A. S., & Raza, Z. (2024). Disease outbreak prediction using natural language processing: a review. *Knowledge and Information Systems*, *66*(11), 6561–6595. https://doi.org/10.1007/s10115-024-02192-6

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets. *IEEE Access*, *8*, 181074–181090. https://doi.org/10.1109/access.2020.3027350

Kadam, V., Kumar, S., Bongale, A., Wazarkar, S., Kamat, P., & Patil, S. (2021). Enhancing Surface Fault Detection Using Machine Learning for 3D Printed Products. *Applied System Innovation*, *4*(2), 34. https://doi.org/10.3390/asi4020034

khan, Z. F., & Alotaibi, S. R. (2020). Applications of Artificial Intelligence and Big Data Analytics in m-Health: A Healthcare System Perspective. *Journal of Healthcare Engineering*, *2020*, 1–15. https://doi.org/10.1155/2020/8894694

Krauer, F., & Schmid, B. V. (2022). Mapping the plague through natural language processing. *Epidemics*, *41*, 100656. https://doi.org/10.1016/j.epidem.2022.100656

Landi, F., Baraldi, L., Cornia, M., & Cucchiara, R. (2021). Working Memory Connections for LSTM. *Neural Networks*, *144*, 334–341. https://doi.org/10.1016/j.neunet.2021.08.030

Leelawat, N., Jariyapongpaiboon, S., Promjun, A., Boonyarak, S., Saengtabtim, K., Laosunthara, A., Yudha, A. K., & Tang, J. (2022). X data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning. *Heliyon*, *8*(10), e10894. https://doi.org/10.1016/j.heliyon.2022.e10894

Liang, H., Fung, I. C.-H., Tse, Z. T. H., Yin, J., Chan, C.-H., Pechta, L. E., Smith, B. J., Marquez-Lameda, R. D., Meltzer, M. I., Lubell, K. M., & Fu, K.-W. (2019). How did Ebola information spread on X: broadcasting or viral spreading? *BMC Public Health*, *19*(1). https://doi.org/10.1186/s12889-019-6747-8

Liu, Y., Whitfield, C., Zhang, T., Hauser, A., Reynolds, T., & Anwar, M. (2021). Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems*, *9*(1). https://doi.org/10.1007/s13755-021-00158-4

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. Springer Us.

Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., & Borodulin, A. (2024). Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review. *Polymers*, *16*(18), 2607–2607. https://doi.org/10.3390/polym16182607

Mansoor, M., Gurumurthy, K., U, A. R., & Prasad, V. R. B. (2020). Global Sentiment Analysis Of COVID-19 Tweets Over Time. *ArXiv:2010.14234 [Cs]*. https://arxiv.org/abs/2010.14234

Marhon, S. A., Cameron, C. J. F., & Kremer, S. C. (2013). Recurrent Neural Networks. *Intelligent Systems Reference Library*, 29–65. https://doi.org/10.1007/978-3-642-36657-4_2

Martín-Corral, D., García-Herranz, M., Cebrian, M., & Moro, E. (2024). Social media sensors as early signals of influenza outbreaks at scale. *EPJ Data Science*, *13*(1). https://doi.org/10.1140/epjds/s13688-024-00474-1

Martinez, L. S., Savage, M. W., Jones, E., Mikita, E., Yadav, V., & Tsou, M.-H. (2022). Examining Vaccine Sentiment on X and Local Vaccine Deployment during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, *20*(1), 354–354. https://doi.org/10.3390/ijerph20010354

Mirugwe, A., Ashaba, C., Namale, A., Akello, E., Bichetero, E., Kansiime, E., & Juwa Nyirenda. (2024). Sentiment Analysis of Social Media Data on Ebola Outbreak Using Deep Learning Classifiers. *Life*, *14*(6), 708–708. https://doi.org/10.3390/life14060708

Müller, M., Salathé, M., & Kummervold, P. E. (2023). COVID-X-BERT: A natural language processing model to analyse COVID-19 content on X. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1023281

Olusegun, R., Oladunni, T., Halima Audu, YAO Houkpati, & Staphord Bengesi. (2023). *Text Mining and Emotion Classification on Monkeypox X Dataset: A Deep Learning-Natural Language Processing (NLP) Approach*. *11*, 49882–49894. https://doi.org/10.1109/access.2023.3277868

Omar, M., Brin, D., Glicksberg, B., & Klang, E. (2024). Utilising Natural Language Processing and Large Language Models in the Diagnosis and Prediction of Infectious Diseases: A Systematic Review. *American Journal of Infection Control*. https://doi.org/10.1016/j.ajic.2024.03.016

Oyebode, O., Ndulue, C., Mulchandani, D., Suruliraj, B., Adib, A., Orji, F. A., Milios, E., Matwin, S., & Orji, R. (2022). COVID-19 Pandemic: Identifying Key Issues Using Social Media and Natural Language Processing. *Journal of Healthcare Informatics Research*. https://doi.org/10.1007/s41666-021-00111-w

Pilipiec, P., Samsten, I., & Bota, A. (2023). Surveillance of infectious diseases using social media: A systematic review. *PLOS ONE*, *18*(2), e0282101. https://doi.org/10.1371/journal.pone.0282101

Qin, Z., & Ronchieri, E. (2022). Exploring Pandemics Events on X by Using Sentiment Analysis and Topic Modelling. *Applied Sciences*, *12*(23), 11924. https://doi.org/10.3390/app122311924

Raza, S., & Schwartz, B. (2023). Constructing a disease database and using natural language processing to capture and standardise free text clinical information. *Scientific Reports*, *13*(1), 8591. https://doi.org/10.1038/s41598-023-35482-0

Sarantopoulos, A., Kourmpani, C. M., Atshaya Lily Yokarasa, Chiedza Makamanzi, Antoniou, P., Nikolaos Spernovasilis, & Constantinos Tsioutis. (2024). Artificial Intelligence in Infectious Disease Clinical Practice: An Overview of Gaps, Opportunities, and Limitations. *Tropical Medicine and Infectious Disease*, *9*(10), 228–228. https://doi.org/10.3390/tropicalmed9100228

Shen, C., Chen, A., Luo, C., Zhang, J., Feng, B., & Liao, W. (2020). Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. *Journal of Medical Internet Research*, *22*(5), e19421. https://doi.org/10.2196/19421

Song, C., Wang, X.-K., Cheng, P., Wang, J., & Li, L. (2020). SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis. *Knowledge-Based Systems*, *194*, 105572. https://doi.org/10.1016/j.knosys.2020.105572

Sv, P., Lorenz, J. M., Ittamalla, R., Dhama, K., Chakraborty, C., Kumar, D. V. S., & Mohan, T. (2022). X-Based Sentiment Analysis and Topic Modeling of Social Media Posts Using Natural Language Processing, to Understand People's Perspectives Regarding COVID-19 Booster Vaccine Shots in India: Crucial to Expanding Vaccination Coverage. *Vaccines*, *10*(11), 1929. https://doi.org/10.3390/vaccines10111929

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, *53*(8). https://doi.org/10.1007/s10462-020-09838-1

Wang, Y., Willis, E., Vijaya Kumari Yeruva, Ho, D. H., & Lee, Y. (2023). A case study of using natural language processing to extract consumer insights from tweets in American cities for public health crises. *BMC Public Health*, *23*(1). https://doi.org/10.1186/s12889-023-15882-7

Wilson, A. E., Lehmann, C. U., Saleh, S. N., Hanna, J., & Medford, R. J. (2021). Social media: A new tool for outbreak surveillance. *Antimicrobial Stewardship & Healthcare Epidemiology*, *1*(1). https://doi.org/10.1017/ash.2021.225

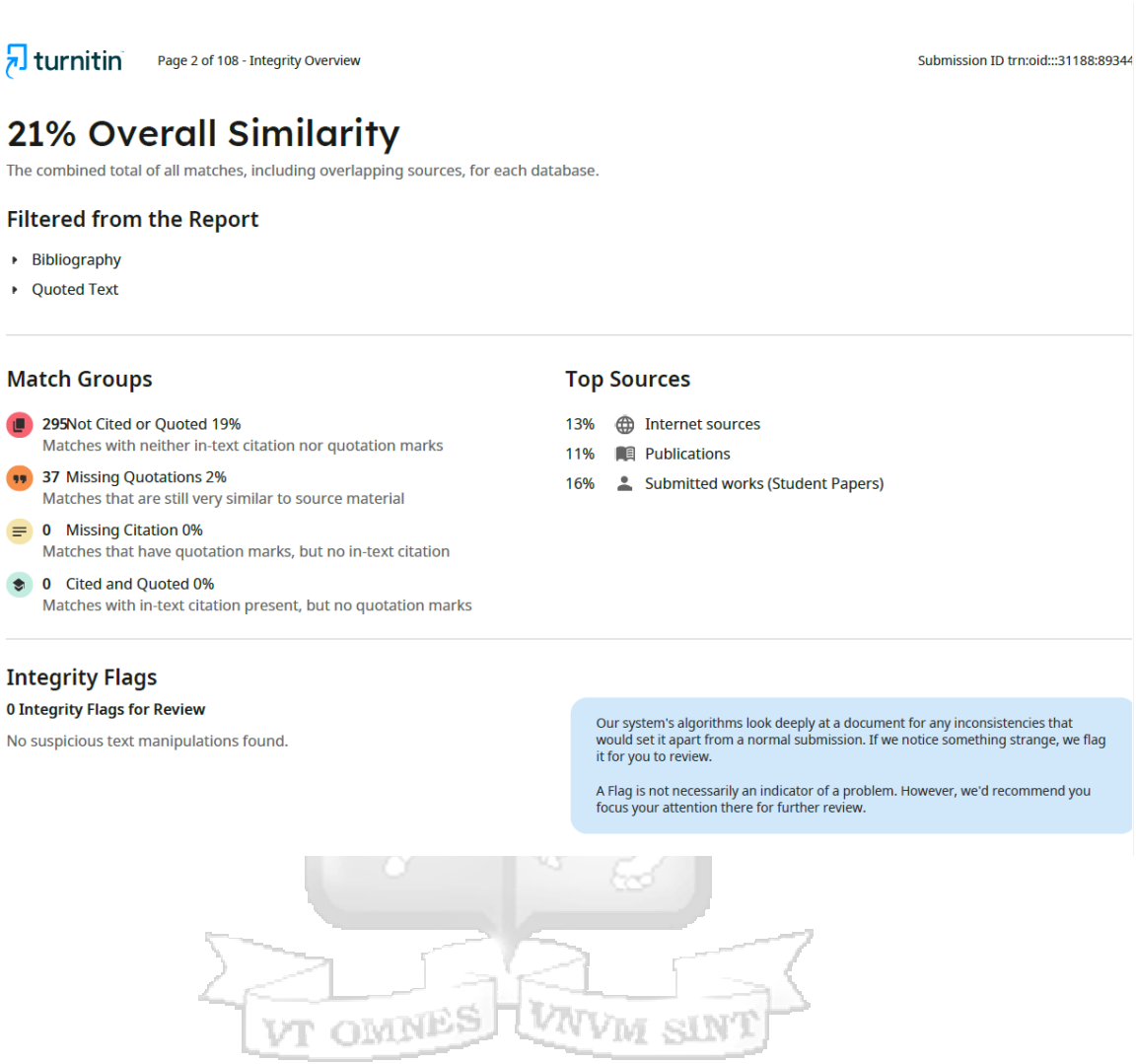
Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional Neural networks: an Overview and Application in Radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yin, H., Yang, S., & Li, J. (2020). Detecting Topic and Sentiment Dynamics Due to COVID-19 Pandemic Using Social Media. *Advanced Data Mining and Applications*, 610–623. https://doi.org/10.1007/978-3-030-65390-3_46

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures: Neural Computation: Vol 31, No 7. *Neural Computation*. https://doi.org/10.5555/3370494;website:website:dl-site;csubtype:string:Periodical;taxonomy:taxonomy:acm-pubtype;pageGroup:string:Publication

Appendices

Appendix A: Plagiarism Report



turnitin Page 2 of 108 - Integrity Overview Submission ID trn:oid:::31188:89344

21% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

295 Not Cited or Quoted 19%
Matches with neither in-text citation nor quotation marks

37 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

13% Internet sources
11% Publications
16% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix B: Ethical Approval

Strathmore UNIVERSITY

25th February 2025

Ms Kinara Calista,
calista.kinara@strathmore.edu

Dear Ms Kinara,

RE: A Natural Language Processing Model to Detect Outbreaks of Infectious Diseases in Kenya

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2645/25.** The approval period is from **25th February 2025 to 24th February 2026.**

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

Appendix C: Excerpts from the Kenya Health Policy 2014–2030 touching on the objectives of the Health Policy

1.3. Health Policy and the National Development Agenda

Over the years, Kenya has strived to overcome development obstacles and improve the socioeconomic status of her citizens, including health. Some of the initiatives include the development and implementation of the Kenya Health Policy Framework (KHPF 1994–2010), Vision 2030, the promulgation of the Kenya Constitution 2010, and fast-tracking actions to achieve the Millennium Development Goals (MDGs) by 2015. The Government of Kenya (GOK) also upholds the fundamental right to health access for every Kenyan as envisaged in Vision 2030.

Key objectives of the Kenya Health Policy 2014–2030

- ✓ Eliminate communicable conditions
- ✓ Halt and reverse the rising burden of non-communicable conditions
- ✓ Reduce the burden of violence and injuries
- ✓ Provide essential healthcare
- ✓ Minimize exposure to health risk factors
- ✓ Strengthen collaboration with private and other health-related sectors

The implementation of KHPF 1994–2010 led to significant investment in public health programmes and minimal investment in medical services, resulting to improvement of health indicators such as infectious diseases and child health. However, the emerging increase of non-communicable diseases is a threat to the gains made so far. This policy aims at consolidating the gains attained so far, while guiding achievement of further gains in an equitable, responsive, and efficient manner. It is envisioned that the

Appendix D: Mini-Budget for Thesis Project

A Natural Language Processing Model to Detect Outbreaks of Infectious Diseases in Kenya

Expense Item	Description	Estimated Cost (KES)
Twitter API Subscription	For accessing real-time and historical tweets	43,000
Printing and Binding	Thesis document printing and final binding	7,000
Internet and Data Costs	High-speed internet for research and model training	5,000
Cloud Computing Services	For hosting and training the NLP model (e.g., AWS, Google Cloud)	12,000
Software Tools and Libraries	Specialized tools like TensorFlow, PyTorch (if paid), or API integration costs	6,000
Data Preprocessing Tools	Cleaning, formatting, and annotating data	4,000
Transport Costs	For consultations with experts or data collection support	3,000
Laptop Purchase	Mid-range laptop for coding and research work	70,000
Miscellaneous Expenses	Contingencies (stationery, snacks during work, unexpected costs)	2,000

Total Estimated Cost: 152,000 KES