MAKERERE UNIVERSITY

COLLEGE OF HEALTH SCIENCES
SCHOOL OF PUPLIC HEALTH

ENHANCING OUTBREAK SURVEILLANCE THROUGH INTEGRATION OF NATURAL LANGUAGE PROCESSING IN UGANDA'S ELECTRONIC INTEGRATED DISEASE SURVEILLANCE AND RESPONSE SYSTEM.

NAKITANDWE REBECCA MELISA

2023/HD07/3084U

A DISSERTATION SUBMITTED TO THE SCHOOL OF PUBLIC HEALTH IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF HEALTH INFORMATICS AT MAKERERE UNIVERSITY KAMPALA

JANUARY 2026

DECLARATION

I, **Nakitandwe Rebecca Melisa**, hereby declare that this dissertation titled "*Enhancing Outbreak Surveillance through integration of Natural Language Processing in Uganda's Electronic Integrated Disease Surveillance and Response System."* is my original work and has not been submitted for any academic award at any other institution. All sources of information have been duly acknowledged.

NAME: NAKITANDWE REBECCA MELISA	Sign:  ……

APPROVAL

This is to certify that this dissertation was developed under our supervision and is now ready for submission.

Dr. Victoria Nankabirwa

Lecturer, Department of Epidemiology/Biostatistics,

School of Public Health,

Makerere University

DocuSigned by:
Victoria Nankabirwa
8471090DAF2C428...

……..…… Date. January 13, 2026 ……

Mr. Haron Gichuhi

Health Informatics Researcher, Department of Epidemiology/Biostatistics,

School of Public Health,

Makerere University

…….……. Date…. January 12, 2026 …….

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BIO	Beginning, Inside, Outside (tagging format)
CNN	Convolutional Neural Network
CSV	Comma-Separated Values
DHIS2	District Health Information Software 2
DSR	Design Science Research
EBS	Event-Based Surveillance
eIDSR	Electronic Integrated Disease Surveillance and Response
GPU	Graphics Processing Unit
HIV	Human Immunodeficiency Virus
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IDSR	Integrated Disease Surveillance and Response
IRB	Institutional Review Board
JSON	JavaScript Object Notation
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
NER	Named Entity Recognition
NGO	Non-Governmental Organization
PHEOC	Public Health Emergency Operations Centre
PII	Personally Identifiable Information
PPV	Positive Predictive Value
RNN	Recurrent Neural Network
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SQL	Structured Query Language
SMS	Short Message Service
SVM	Support Vector Machine
TSV	Tab-Separated Values
WHO	World Health Organization
XML	eXtensible Markup Language
XLM-R	Cross-lingual Language Model – RoBERTa

OPERATIONAL DEFINITIONS

Timeliness: The speed at which key information is processed and acted upon in the outbreak surveillance system. In this context, it refers to the time from when an SMS alert is received to when it is processed and an alert is sent to the surveillance personnel. In alignment with the 7-1-7 framework, we target a maximum turnaround time of 24 hours for processing and notification to ensure prompt detection and response.

Epidemiologic Indicators: Quantitative measures used to assess the effectiveness of disease surveillance and outbreak response, including sensitivity, specificity, time to detection, time to response, and case fatality rates.

Cost-Effectiveness: The ratio of resources spent (e.g., time, money, and labor) to the benefits achieved in outbreak detection and response when using the proposed automated system compared to the manual process.

Natural Language Processing (NLP): A field of artificial intelligence focused on enabling computers to understand, interpret, and process human language. In this study, NLP is applied to extract key information such as symptoms, location, and disease type from SMS messages.

eIDSR System: The electronic Integrated Disease Surveillance and Response system implemented in Uganda to collect and manage health data. It facilitates real-time reporting of suspected disease outbreaks through SMS.

Model: A computational representation of a system or process that is trained on data to recognize patterns and make predictions or classifications.

Sensitivity: The ability of the system to correctly identify true outbreaks (true positive rate).

Specificity: The ability of the system to correctly exclude non-outbreak situations (true negative rate).

Positive Predictive Value (PPV): The proportion of alerts generated by the system that corresponds to actual outbreaks.

F1 Score: A balanced metric combining sensitivity and PPV, providing a single measure of the system's performance in extracting and analyzing SMS messages.

Integration: The process of linking the trained NLP model with the eIDSR application to enable seamless automated processing of SMS data.

Training Dataset: The subset of data used to teach the NLP model to identify and extract key entities from SMS messages.

Validation Dataset: A subset of data used during model training to evaluate its performance and tune hyperparameters.

Test Dataset: A separate dataset used to assess the final performance of the NLP model after training is complete.

Annotation: The process of manually labeling SMS messages with key information, such as disease type, location, and symptoms, to create a training dataset for the NLP model.

Anonymization: The process of removing personally identifiable information from SMS messages to protect patient confidentiality during analysis.

ABSTRACT

Introduction: Early detection of diseases or infections is essential to prevent infectious diseases from escalating into large outbreaks. In Uganda, the Electronic Integrated Disease Surveillance and Response (eIDSR) system enables community-level reporting of suspected cases via SMS. However, manual processing of these unstructured messages often delays outbreak detection and response, particularly during high-volume reporting periods. The manual processing of incoming SMS messages within the eIDSR system creates a bottleneck that hinders timely outbreak detection and response. This delay has the potential to increase morbidity and mortality, especially in resource-limited settings. This study aimed to integrate Natural Language Processing (NLP) to automate the extraction of key information, such as disease type, location, and symptoms, from SMS alerts submitted to the eIDSR system. It also sought to understand the contextual factors that influenced model accuracy and performance.

Methods: A retrospective design was employed using historical SMS data submitted to the eIDSR system in 2024. A Bidirectional Encoder Representations from Transformers (BERT)-uncased model was fine-tuned on a manually annotated dataset to support named entity recognition. The model was evaluated using precision, recall, F1-score, and processing speed, and its performance was compared with manual extraction. McNemar's test was used to assess the statistical significance of differences between the two methods.

Results: The model achieved an F1-score of 92.6%, with recall of 94.2% and precision of 91.1%, processing approximately 48 messages per second. It extracted high-value entities such as disease, age, gender, and location, with near-perfect accuracy. Errors were concentrated around symptom span boundaries and ambiguous entries. Interviews confirmed the value of automation for reducing analyst workload and outlined key limitations of the current manual workflow, including handling of ambiguous or duplicate messages.

Conclusion: This study demonstrated the feasibility of applying NLP to automate SMS-based disease surveillance within Uganda's eIDSR system. Although human review remains necessary for edge cases, the model showed strong potential to accelerate processing, eliminate backlog, and support timely response under frameworks like 7-1-7. With targeted improvements especially in symptom handling and multilingual input. The model would be suitable for pilot integration under a human-in-the-loop deployment model.

CHAPTER ONE
INTRODUCTION AND BACKGROUND

1.1 Introduction

Timely detection and notification of infectious disease outbreaks are critical for effective public health response, particularly in resource-constrained settings where delays often lead to widespread morbidity and mortality. In Uganda, the Electronic Integrated Disease Surveillance and Response (eIDSR) system played a key role in early warning by enabling healthcare workers and community members to report suspected cases via Short Message Service (SMS). This approach aligned with the World Health Organization's (WHO) 7-1-7 framework, which outlines targets for detecting outbreaks within seven days, notifying and investigating within one day, and initiating a response within seven days (Kuehne et al., 2019). However, the system relied heavily on manual review, where personnel at the Public Health Emergency Operations Centre (PHEOC) read, interpreted, and re-entered each SMS into the Event-Based Surveillance (EBS) module. This labour-intensive process introduced delays in alert generation and slowed down response coordination.

Early notification was especially critical in the context of zoonotic diseases—those transmitted between animals and humans—which now constitute the majority of emerging infectious disease threats globally. Uganda's ecological diversity, livestock practices, and frequent human–animal interaction increased the risk of zoonotic spillovers, including diseases such as Ebola, Anthrax, Rabies, and Rift Valley Fever. These outbreaks often began with vague or non-specific symptoms, making timely and accurate extraction of information from community-reported messages essential for effective escalation and containment.

While several countries adopted digital tools to enhance surveillance, the manual processing of text-based alerts remained a bottleneck. Kenya's mobile-based reporting system (mSOS), for instance, demonstrated the utility of SMS-based alerts in improving outbreak detection (Toda et al., 2016), but like Uganda, it still depended on human interpretation and entry. During high-volume emergencies—such as the COVID-19 pandemic, manual workflows became overwhelmed, resulting in delayed verification and reduced responsiveness (Budd et al., 2020).

Natural Language Processing (NLP), a subfield of artificial intelligence, emerged as a promising solution by automating the extraction of structured information from unstructured text data like SMS.

NLP had been applied in various public health contexts globally, including the analysis of social media trends, clinical text processing, and digital epidemiology (Baclic et al., 2020). However, its integration into Uganda's surveillance systems remained limited.

This study addressed this gap by developing and evaluating an NLP-powered system to automatically extract key information, such as disease type, symptoms, location, and date of onset, from SMS messages submitted to the eIDSR system. By reducing reliance on manual review and accelerating the notification workflow, the system improved the timeliness and accuracy of outbreak alerts and response efforts, particularly for high-priority zoonotic diseases.

1.2 Background

Effective outbreak surveillance forms the backbone of a resilient public health system, enabling early detection and timely response to infectious disease threats. Globally, the nature of disease emergence is changing; zoonotic diseases, those transmitted between animals and humans, now account for over 60% of known infectious diseases and more than 75% of emerging pathogens (Sekamatte et al., 2018). According to the World Health Organization (WHO), over 60% of global outbreak detections now originate from informal sources, including SMS, phone calls, and social media posts. While these unstructured data streams offer rich early warning signals, processing them at scale remains a significant challenge for public health authorities worldwide.

Across Africa, nations have made strides in standardizing surveillance through the Integrated Disease Surveillance and Response (IDSR) strategy. To bridge the gap between community detection and national response, several African countries have piloted mobile and digital technologies. For example, Kenya's mSOS system successfully enabled real-time facility-level reporting of notifiable diseases through SMS (Toda et al., 2016), while Senegal integrated community event-based surveillance into DHIS2 during the COVID-19 pandemic to improve early detection (Seck et al., 2023). Furthermore, the application of Artificial Intelligence (AI) in African public health is growing; South Africa has employed Natural Language Processing (NLP) for optimizing workforce planning, and Nigerian innovators have utilized machine learning for neonatal diagnostics. However, despite these innovations, many regional systems remain constrained by a reliance on manual data processing.

In Uganda, the Ministry of Health has operationalized these strategies through the electronic Integrated Disease Surveillance and Response (eIDSR) platform (Preparedness, 2019). This system facilitates real-time data flow, allowing frontline health workers and community members to report suspected cases via SMS. This capability is critical given Uganda's high biodiversity and extensive livestock

interactions, which increase the risk of zoonotic spill overs. A 2017 One Health prioritization exercise identified Anthrax, Zoonotic Influenza, Viral Haemorrhagic Fevers, Brucellosis, Plague, and Rabies as top priorities for intensified surveillance.

Despite the robust reporting infrastructure in Uganda, the backend processing remains a bottleneck. Currently, incoming SMS alerts are manually reviewed at the Public Health Emergency Operations Centre (PHEOC) before being escalated. This manual process is labor-intensive and unsustainable during high-volume reporting periods or outbreak surges. While the eIDSR system captures the data, the lack of automated interpretation slows the transition from "alert" to "action." This study leveraged Natural Language Processing (NLP) to address this gap, aiming to automate the extraction of epidemiological data from these messages and align Uganda's surveillance capabilities with the 7-1-7 framework targets. The ultimate goal was to demonstrate the feasibility of using NLP to meet the targets of the 7-1-7 framework (Resolve to Save Lives, 2024), which emphasizes the importance of detecting a public health threat within 7 days, initiating investigation within 1 day, and launching a response within another 7 days.

By contributing an automated, scalable solution for SMS-based disease surveillance, this study aimed to improve Uganda's readiness for zoonotic outbreaks and enhance the country's broader public health surveillance capacity. The findings are expected to inform future deployments of NLP in similar low-resource settings and provide a foundation for digital surveillance innovations across sub-Saharan Africa.

CHAPTER TWO
LITERATURE REVIEW

Effective disease surveillance is fundamental to early outbreak detection and timely public health response, particularly in low-resource settings where manual processes often delay critical action. The 7-1-7 surveillance framework underscores the need to detect potential threats within seven days and initiate a response within one day (Resolve to Save Lives, 2024). This urgency is heightened in the case of zoonotic diseases, which account for over 60% of known infectious diseases and 75% of emerging pathogens globally (Sekamatte et al., 2018). Many of these such as Anthrax, Rabies, and Viral Hemorrhagic Fevers, have high epidemic potential in Uganda and are usually first flagged at the community level via informal, event-based reporting channels, including SMS alerts.

While SMS-based surveillance has proven valuable for capturing real-time community data, systems like Uganda's eIDSR still rely heavily on manual triage and interpretation of free-text messages at the Public Health Emergency Operations Centre (PHEOC). This manual extraction creates bottlenecks that compromise the speed and consistency of alert classification and onward escalation (Mremi et al., 2021).

To address this gap, this study explored the use of Natural Language Processing (NLP) to automate the extraction of outbreak-related entities such as disease type, symptoms, and location from incoming SMS alerts. This literature review therefore examined the broader potential of NLP in digital disease surveillance, with particular attention to its application in low-resource settings. It explored existing research on event-based surveillance systems, the limitations of manual data handling, and recent advancements in NLP and digital epidemiology.

By synthesizing this body of evidence, the review provided context for the development, training, and evaluation of an NLP-powered SMS processing model integrated with Uganda's eIDSR system. This approach aimed to improve timeliness, reduce workload at the PHEOC, and enhance the accuracy of surveillance for high-priority zoonotic diseases (Feng et al., 2018; Ibrahim et al., 2021; Nakiire et al., 2019; Masiira et al., 2019).

2.1 Reportable Diseases in Uganda

Uganda bears a significant burden of infectious diseases, many of which are classified as reportable due to their potential for rapid spread, high case fatality, and socio-economic impact. These include endemic conditions such as malaria, cholera, typhoid, and measles, as well as viral haemorrhagic fevers

like Ebola and Marburg, which have caused repeated outbreaks in the country (Masiira et al., 2019). The urgency of surveillance for these diseases is underscored by the adoption of the 7-1-7 framework, which sets a target of detecting potential threats within seven days and responding within one (Resolve to Save Lives, 2024).

Among the reportable conditions, zoonotic diseases present a particularly complex and growing challenge. Infections such as Anthrax, Rabies, Rift Valley Fever, and Brucellosis, which are transmitted between animals and humans, now constitute a substantial share of Uganda's priority health threats. In recognition of this, the Ministry of Health—alongside key stakeholders—conducted a One Health Zoonotic Disease Prioritization workshop in 2017 to rank zoonoses based on their outbreak potential, disease burden, and socio-economic impact (Uganda One Health Zoonotic Disease Prioritization Report, 2017). The resulting list included Anthrax, Zoonotic Influenza, Viral Haemorrhagic Fevers (including Ebola, Marburg, Rift Valley Fever, and Crimean Congo Haemorrhagic Fever), Brucellosis, Plague, Trypanosomiasis, and Rabies.

These diseases often emerge at the community level, particularly in rural or pastoral areas where human–animal interactions are frequent and where access to clinical surveillance is limited. Because early symptoms are typically vague—such as fever, headache, or rash—and often described using local expressions, delays in detection are common. This makes frontline SMS alerts from health workers and communities a critical entry point for early warning. However, under current workflows, these alerts must be manually read, interpreted, and classified by staff at the Public Health Emergency Operations Centre (PHEOC), creating a bottleneck that slows down response time and increases the risk of missed or misclassified events (Mremi et al., 2021; Nakiire et al., 2019).

To address this, the present study explored the use of Natural Language Processing (NLP) to automate the extraction of key outbreak-related entities such as disease name, symptoms, age, gender, location, and onset date, from unstructured SMS alerts. NLP, particularly through Named Entity Recognition (NER), can support faster, scalable analysis of community-generated data and enable real-time structuring of critical information from otherwise ambiguous or code-switched messages (Feng et al., 2018).

This capability is especially relevant for zoonotic diseases, where early warning depends on parsing scattered symptom descriptions and geographic references embedded in free-text reports. By reducing reliance on manual review, an NLP-enhanced surveillance system offers the potential to close the gap between message receipt and actionable notification. This can significantly strengthen Uganda's ability

to meet 7-1-7 goals, mitigate epidemic spread, and better allocate resources for outbreak control (Ibrahim et al., 2021; Resolve to Save Lives, 2024).

2.2 Event-Based Surveillance and 7-1-7 Framework

Event-Based Surveillance (EBS) plays a vital role in detecting and responding to public health threats in real time, particularly in low-resource settings where traditional indicator-based surveillance may be slower or less sensitive. In Uganda, EBS is a core pillar of the national disease surveillance architecture, supporting the early identification of unusual health events based on SMS alerts submitted by frontline health workers and community informants (Bochner et al., 2023).

To strengthen the timeliness and effectiveness of such systems, the 7-1-7 framework was introduced as a global benchmark: detect potential outbreaks within 7 days, notify and begin investigation within 1 day, and launch an appropriate response within the following 7 days (Frieden et al., 2021). This framework has been formally endorsed by Uganda and promoted through collaborations with the World Health Organization and Resolve to Save Lives as a means to improve national outbreak preparedness and response capacity.

Despite this commitment, practical challenges persist in meeting the 7-1-7 targets, particularly the first two: detection and notification. A multi-country study by Bochner et al. (2023) highlighted mixed progress across African countries, including Uganda. While there were gains in early detection and laboratory capacity, limitations in data processing, communication infrastructure, and manual workflows consistently undermined response timeliness.

One of the most persistent bottlenecks lies in the manual review of SMS alerts at Uganda's Public Health Emergency Operations Centre (PHEOC). Each incoming message must be read, interpreted, and re-entered into the system, a process that not only delays detection and notification but also risks inconsistencies due to human error or message ambiguity. During periods of high-volume reporting, this manual triage becomes unsustainable, contributing to delays that compromise Uganda's ability to act swiftly in line with the 7-1-7 framework.

Kenya's experience with the mSOS (mobile SMS) system offers a relevant parallel. By enabling health workers to submit real-time SMS alerts for suspected cases of notifiable diseases, the system improved early detection and supported timely response through integration with a web-based portal (Toda et al., 2016). This model demonstrated how mobile-based tools could be used to meet the 7-day detection and 1-day notification targets—provided that efficient backend processing systems are in place.

Building on this insight, Natural Language Processing (NLP) can help address this challenge: the manual handling of unstructured SMS data. By automating the extraction of key entities from the SMS alerts, NLP can significantly reduce processing delays and standardize data interpretation. This capability directly supports more timely detection, more consistent notification, and ultimately quicker outbreak response.

The integration of NLP into Uganda's SMS-based EBS system presents a practical pathway to meet the 7-1-7 performance targets. It introduces a scalable, technology-driven approach that addresses current system bottlenecks and strengthens national surveillance capacity to respond to fast-moving infectious threats more effectively.

2.3 Outbreak Surveillance and Digital Health Systems

Outbreak surveillance remains a cornerstone of public health, serving as the primary mechanism for detecting and responding to infectious disease threats. Traditional surveillance systems, however, often rely on manual reporting and data entry, which can delay critical public health responses and compromise early containment efforts (Simonsen et al., 2016). In recognition of these limitations, digital health technologies have emerged as powerful tools for modernizing disease surveillance by enabling real-time data collection, rapid analysis, and seamless communication across the health system.

In Uganda, the electronic Integrated Disease Surveillance and Response (eIDSR) system exemplifies this shift toward digital surveillance. It leverages Short Message Service (SMS) alerts from frontline health workers and community members to capture outbreak signals directly from the field (Lester et al., 2016). This SMS-based architecture offers broad geographic reach and real-time reporting, especially in low-resource or rural settings where other infrastructure may be limited. However, while it improves data flow and timeliness, the system still relies on manual review and transcription of incoming messages at the Public Health Emergency Operations Centre (PHEOC), creating a processing bottleneck that undermines rapid decision-making.

Persistent challenges such as human resource limitations, data entry inconsistencies, and the inability to handle high message volumes during outbreak surges continue to hamper the effectiveness of digital surveillance platforms (Lamorde et al., 2018). Similar experiences have been documented in Kenya, where the mobile-based mSOS system enabled health workers to report suspected cases of notifiable diseases via SMS (Toda et al., 2016). While these tools have improved community-based surveillance

and early warning capabilities, they often fall short in automating downstream data processing and extraction of structured information.

More recently, efforts in Uganda have focused on integrating disease-specific digital tools. For instance, the mobile application developed for real-time reporting of decentralized SARS-CoV-2 testing data demonstrated the potential of digital innovation to support faster case tracking and response (Nansumba et al., 2023). These targeted interventions illustrate how mobile technologies can be adapted to meet the needs of specific outbreaks.

Despite these advancements, a critical gap remains: most digital platforms are still unable to fully interpret unstructured or semi-structured data like free-text SMS alerts without manual input. As a result, valuable information may be delayed, overlooked, or inconsistently processed.

This study addressed that gap by introducing Natural Language Processing (NLP) as a layer of intelligence within the SMS-based disease reporting workflow. By automating the extraction of key entities, such as disease names, symptoms, locations, and age, from informal and varied SMS messages, the NLP-enhanced system reduces reliance on manual review and allows structured, actionable insights to be generated in real time. Integrating NLP into existing surveillance frameworks like eIDSR or mSOS allows for high-volume, rapid triage of alerts even when message formats vary or include local idioms and code-switching.

Ultimately, NLP-powered surveillance systems go beyond simply digitizing data, they enable dynamic interpretation of unstructured inputs, accelerate outbreak detection, and strengthen response capabilities. In resource-limited settings like Uganda, this can make the difference between early containment and uncontrolled spread, offering a scalable, cost-effective solution for future epidemic preparedness.

2.4 Automated Alert Systems in Disease Surveillance

Traditional disease surveillance systems have long relied on manual data processing, where incoming reports, whether through phone calls, SMS, or paper forms, are reviewed and interpreted by human analysts. This manual workflow typically involves extracting relevant information, entering it into centralized databases, and then conducting trend analysis. While this process can ensure contextual judgment, it is time-consuming, labor-intensive, and error-prone, especially during outbreak surges when timeliness is critical for effective response (Van Hoek et al., 2024). These inefficiencies are

amplified in low-resource settings like Uganda, where surges in reporting can quickly overwhelm personnel and delay critical public health actions.

To overcome these limitations, automation through Natural Language Processing (NLP) has emerged as a promising strategy. NLP enables systems to automatically interpret and extract structured data from unstructured text, such as SMS messages submitted to surveillance platforms. This capability is particularly relevant for Uganda's eIDSR system, where frontline health workers report suspected cases through free-text SMS alerts. Leveraging NLP not only accelerates the triage process but also improves accuracy by standardizing the extraction of entities such as disease type, symptoms, location, and patient demographics.

Jamal et al. (2021) demonstrated how deep learning-based NLP models can successfully classify sentiment and emotion in large-scale Twitter data, underscoring the broader capacity of NLP to interpret short-form, informal, and noisy text. This aligns well with the structure of SMS messages typically submitted to surveillance systems in Uganda. Similarly, Randriamiarana et al. (2018) emphasized the potential of digital tools in improving timeliness and consistency in outbreak reporting—particularly when integrated with structured workflows for automated data interpretation.

Several countries have piloted or implemented automated alert systems with notable success. In Senegal, Seck et al. (2023) documented the integration of alert systems with the national DHIS2 platform to monitor COVID-19 cases in real time, emphasizing the importance of ensuring rural connectivity for equitable coverage. Kenya's mSOS platform, evaluated by Toda et al. (2016), enabled health workers to send immediate alerts via SMS, resulting in improved detection and faster communication between frontline providers and district managers. Similarly, Namuye et al. (2015) highlighted the role of mobile and service-oriented technologies in supporting early warning systems capable of initiating timely interventions.

In Uganda, Bosa et al. (2016) showed that community-based reporting—through phone calls and SMS—can improve yellow fever surveillance, although challenges such as mobile network access and lack of incentives persist. While these initiatives laid the foundation for mobile-driven surveillance, the backend systems have remained largely manual. As a result, the speed and scale at which SMS data can be processed are still limited, particularly during health emergencies.

This study extended these prior efforts by building and demonstrating an NLP-enabled API capable of automatically extracting key outbreak-related information from SMS alerts. By integrating this model into a working prototype of the eIDSR alert pipeline, the study showed how automation could reduce

processing delays, increase consistency, and help district health teams receive timely, structured data for action.

Ultimately, the use of NLP for automated text analysis represents a critical advancement in digital epidemiology. It complements existing mobile-based reporting systems and aligns with global efforts to strengthen early warning and response capacity. When combined with systems like eIDSR, NLP-powered tools can significantly enhance the accuracy, speed, and scalability of disease surveillance in Uganda and other resource-constrained settings.

2.5 Performance of Manual and Automated Processing of Health Surveillance Data

A growing body of research has examined the comparative performance of manual and automated approaches to extracting key information from health surveillance data. These studies consistently highlight the time-saving and scalability benefits of automation, especially when large volumes of unstructured reports must be processed rapidly. Elhadad & Demner-Fushman (2016), evaluated the use of NLP for extracting clinical concepts from emergency department triage notes and found that NLP systems significantly reduced the time required for data abstraction while achieving comparable accuracy to trained human reviewers. Similarly, Zuccon et al. (2015) demonstrated that automated classifiers could match or exceed human accuracy in identifying disease categories from clinical free-text, suggesting that AI-powered systems can support or even replace manual triage in high-volume settings (Raza & Schwartz, 2023).

In public health surveillance specifically, studies showed that NLP-based systems were able to identify outbreak-related terms in chief complaints with a sensitivity and specificity approaching that of epidemiologists manually reviewing the same data. More recently, (Gupta & Katarya, 2020) used BERT-based models to extract adverse drug events from social media and demonstrated that the automated system outperformed human coders in consistency and speed, particularly when dealing with large datasets.

These findings reinforce the notion that, while manual review provides valuable contextual interpretation, automated NLP solutions offer significant advantages in terms of speed, reproducibility, and scalability, particularly when rapid decision-making is critical, such as during an outbreak.

However, some studies caution that automation should be viewed as a complement rather than a full replacement for human judgment. Ruis et al. (2020) emphasize the importance of human-in-the-loop validation in systems where public health implications are high. In such hybrid models, NLP systems

handle the initial extraction and flagging, while human experts focus on final validation and interpretation. This dual approach is especially valuable in settings like Uganda's eIDSR system, where the SMS inputs are highly variable, and local language use adds complexity. Overall, comparative studies support the integration of NLP to reduce the manual burden while preserving expert oversight in critical decision points.

2.6 Natural Language Processing in Public Health - Roles, Opportunities, and Challenges.

Natural Language Processing (NLP), a subfield of artificial intelligence, has emerged as a transformative tool in public health, particularly for analyzing unstructured text data such as SMS alerts, clinical notes, and digital communications. In low-resource settings like Uganda, where much of the health information is collected through informal or free-text formats, NLP provides a pathway to reduce reliance on manual data processing, improve timeliness, and enhance consistency in interpreting health-related messages (Young et al., 2019; Li et al., 2022).

Across the African continent, several initiatives demonstrate the practical applications of NLP and related technologies in healthcare. For example, in South Africa, NLP has been used to support public health workforce management, reducing administrative errors and improving efficiency (Chilunjika et al., 2022). In Nigeria, machine learning and signal processing technologies have enabled automated detection of neonatal asphyxia, improving early diagnosis and clinical decision-making (Sachin et al., 2017). These examples highlight the broad potential of NLP to enhance health systems by automating complex and labor-intensive tasks.

However, realizing this potential in low-resource settings presents unique challenges. One of the foremost issues is the limited availability of annotated datasets especially those in local languages, which are essential for training accurate and context-aware NLP models. Most African languages remain underrepresented in existing language models, and the widespread use of informal expressions, code-switching, and local idioms in SMS health reports complicates standardization and analysis (Babirye et al., 2022; Akera et al., 2022; Adebara & Abdul-Mageed, 2022). Additionally, the lack of computational infrastructure, constrained internet access, and limited technical expertise in data science and NLP further restrict deployment in resource-limited environments.

To address these barriers, researchers are increasingly leveraging transfer learning and domain adaptation. Pre-trained models such as multilingual BERT (mBERT) and XLM-RoBERTa, trained on

dozens of global languages, offer promise for processing multilingual and code-mixed text, a common feature of community-based surveillance data in Uganda. These models are particularly useful where text includes both English and local languages. In contrast, English-only models can still perform effectively when English dominates the data or when high-quality annotations exist only in English. The choice of model architecture depends on the linguistic characteristics of the data and the goals of the system being developed.

In the context of disease surveillance, NLP techniques such as Named Entity Recognition (NER) are commonly employed to extract structured information such as disease type, symptoms, location, and date of onset, from raw messages. When integrated into digital health systems, such as Uganda's SMS-based eIDSR platform, NLP can facilitate earlier detection of outbreak signals and reduce the human effort required to triage alerts (Durango et al., 2023; Jerfy et al., 2024). Advances in low-resource NLP including synthetic data generation, cross-lingual embeddings, and multilingual fine-tuning, continue to expand the feasibility of applying NLP in these settings.

Overall, NLP holds significant promise for transforming how disease surveillance systems in low-resource countries manage unstructured data. By automating the interpretation of community-level SMS alerts, NLP can enhance early warning capacity, reduce delays in outbreak response, and improve the efficiency of health surveillance systems, objectives that align directly with this study's aim of enhancing the eIDSR platform using NLP-based approaches.

2.7 Natural Language Processing Models

Natural Language Processing (NLP) has seen rapid evolution in recent years, with significant improvements in the performance of models designed for tasks such as text classification, Named Entity Recognition (NER), and information extraction. Traditional machine learning models like Support Vector Machines (SVM) and Conditional Random Fields (CRF) laid the foundation for early NLP applications but often fell short in understanding context, particularly in complex and informal text formats.

The introduction of deep learning models, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), marked an improvement in handling sequential and spatial features of text. However, these models still struggled with long-range dependencies and lacked the ability to fully capture bidirectional context.

Transformer-based models, most notably Bidirectional Encoder Representations from Transformers (BERT), have since emerged as state-of-the-art solutions. BERT was pre-trained on large English corpora (such as BooksCorpus and Wikipedia) using a masked language modeling objective, allowing it to learn deep contextual representations of language. Unlike previous models, BERT processes text in both directions simultaneously, enabling it to better capture word meanings based on context before and after each token (Devlin et al., 2019).

BERT has shown outstanding performance in a range of NLP tasks, particularly in Named Entity Recognition, where it excels at identifying complex entities within noisy or unstructured data. Its ability to model context at the sentence and subword level makes it highly effective for extracting entities such as disease names, symptoms, and locations from short and informal messages, exactly the kind of text found in Uganda's SMS-based disease surveillance system.

Fine-tuned versions of BERT have consistently outperformed previous architectures across public health-related tasks. For instance, studies have demonstrated that BERT-based models significantly improve the extraction of clinical concepts from unstructured medical records, outperforming RNN and CRF baselines in terms of precision, recall, and F1-score (Peng et al., 2019; Si et al., 2021). Its flexibility also makes it suitable for adaptation to domain-specific data through transfer learning, requiring relatively small annotated datasets for fine-tuning.

Although multilingual models like mBERT offer advantages in multi-language contexts, English-based BERT models remain highly effective and more accessible in scenarios where English dominates the available data, or where high-quality annotations are only available in English. In Uganda's case, SMS disease alerts are often composed in English or use English as a base language with local idiomatic variations. This makes an English-based BERT model a practical choice, especially in the absence of large, annotated corpora in local languages.

By fine-tuning BERT on a manually annotated dataset of historical SMS alerts, it is possible to achieve high levels of performance in extracting outbreak-related information. The model's strong contextual understanding, coupled with its ability to generalize well from limited training examples, makes it well-suited for automating disease surveillance tasks in resource-limited environments.

BERT's bidirectional architecture, deep contextual modeling, and proven success in health-related NLP tasks make it an ideal foundation for building robust, scalable solutions for automating information extraction in Uganda's eIDSR system. Its integration into the SMS processing pipeline

has the potential to substantially improve the speed, accuracy, and consistency of disease outbreak detection.

2.8 Factors Influencing Accuracy of Extracted Information

The accuracy of Natural Language Processing models in public health surveillance is heavily dependent on several structural and linguistic factors. Literature suggests that while transformer-based models like BERT have set new performance benchmarks, their real-world application is often constrained by data quality, linguistic diversity, and domain complexity. (Bagla et al., 2023)

One of the most significant predictors of NLP model performance is the quality of the input text. In clinical and community reporting, text is often "telegraphic," characterized by omitted subjects, missing punctuation, and non-standard grammar, which disrupts standard dependency parsing. Studies evaluating Named Entity Recognition (NER) robustness have shown that performance declines significantly as textual noise, such as misspellings and capitalization errors, increases (Bhadauria et al., 2024). In health surveillance specifically, the pervasive use of ad-hoc abbreviations (e.g., using "vom" for "vomiting") presents a major challenge, with some studies showing that standard NLP systems achieve suboptimal F1-scores when handling unnormalized clinical abbreviations (Luo et al., 2020).

In multilingual settings like Uganda, "code-switching"—the practice of alternating between languages within a single message—poses a unique challenge for automated systems. Most pre-trained language models are optimized for monolingual data and struggle to capture the grammatical transitions inherent in mixed-language text. Research indicates that while multilingual models (like mBERT) offer a foundation, they often require specific fine-tuning on code-switched datasets to accurately identify entities that span across English and local dialects (Xie et al., 2025). Failure to account for these linguistic shifts can lead to high rates of false negatives in disease detection.

The choice of model architecture plays a critical role in extraction accuracy. Systematic reviews have demonstrated that Bidirectional Transformer (BT) models, such as BERT, generally outperform rule-based or traditional machine learning approaches (like Conditional Random Fields) by better capturing context (Dahl et al., 2025). However, these models are data-hungry; their accuracy is directly linked to the volume and representativeness of the annotated training data. In low-resource settings, the lack of large, annotated medical corpora often forces researchers to rely on smaller datasets, which can limit the model's ability to generalize to unseen data (Pais et al., 2024).

Finally, the accuracy of any supervised learning model is capped by the quality of its ground truth annotations. Inconsistencies in how human annotators label complex entities, such as long symptom descriptions versus short keywords, introduce "label noise" that degrades model performance (Plank, 2022). Ensuring rigorous inter-annotator agreement is therefore cited as a prerequisite for building reliable surveillance systems.

2.9 Literature review conclusion

The literature reviewed highlighted a critical need to strengthen disease surveillance systems in resource-limited settings such as Uganda, where reliance on manual processing of SMS-based reports created bottlenecks that delayed timely outbreak detection and response. Although event-based surveillance (EBS) was shown to be a valuable approach for capturing community-level alerts, its effectiveness was often undermined by the challenges of manually reviewing large volumes of unstructured messages.

Natural Language Processing (NLP) emerged as a promising solution for automating the extraction of key information, such as symptoms, disease names, locations, and dates of onset, from free-text reports. Existing studies demonstrated that NLP can reduce processing time, improve data accuracy, and support real-time decision-making. The review also found that advances in transformer-based models, particularly BERT, expanded the feasibility of deploying NLP in low-resource environments, despite challenges such as linguistic diversity, limited annotated datasets, and infrastructure constraints.

However, the review revealed a notable gap in the practical integration of NLP into national surveillance platforms, especially in African contexts. This gap informed the current study, which aimed to explore how an NLP-powered system could be developed and aligned with Uganda’s eIDSR platform. By addressing these identified challenges, the study sought to demonstrate the potential of NLP to reduce manual workload, enhance the accuracy of extracted disease information, and improve the overall timeliness and effectiveness of outbreak response in Uganda.

CHAPTER THREE
STATEMENT OF THE PROBLEM, JUSTIFICATION, CONCEPTUAL FRAMEWORK

3.1 Statement of the Problem

Uganda's electronic Integrated Disease Surveillance and Response (eIDSR) platform serves as the country's main gateway for community-level outbreak intelligence, relying on Short Message Service (SMS) alerts submitted by frontline health workers and the public. While SMS is well-suited to low-resource settings, the platform still depends on personnel at the Public Health Emergency Operations Centre (PHEOC) to manually read, triage, and re-type each free-text message. This manual extraction process creates a structural bottleneck in the national surveillance workflow. During periods of reporting surges, caused by seasonal zoonotic activity, population mobility, or climate-driven disease trends, the daily volume of incoming messages often multiplies, exceeding the PHEOC's processing capacity. These backlogs delay the progression from detection to investigation and notification, compromising Uganda's 7-1-7 target of detecting a threat within seven days, initiating a field investigation within one day, and launching control measures within the following seven days (Bochner et al., 2023; Frieden & Lee, 2021).

The initial descriptions of such events, typically captured in SMS form, are often vague, locally worded, or code-switched, making them difficult to interpret quickly and consistently. As a result, during high-volume surges, the likelihood of misclassified or missed alerts increases, extending community exposure and further straining limited public health response resources (Budd et al., 2020; Randriamiarana et al., 2018).This study addressed that dual challenge: the operational bottleneck caused by manual SMS processing and the urgent requirement for rapid detection imposed by the growing threat of zoonotic outbreaks. Therefore, this study aimed to bridge this operational gap by developing and evaluating a natural language processing solution to automate the extraction of epidemiological entities from SMS alerts, thereby enhancing the efficiency and timeliness of Uganda's eIDSR system.

. 3.2 Justification for the Proposed Solution

This study addressed a critical inefficiency in Uganda's electronic Integrated Disease Surveillance and Response (eIDSR) system by integrating Natural Language Processing (NLP) to automate the extraction of outbreak-related information from SMS messages. Previously, the system depended on

manual verification and transcription of incoming alerts by personnel at the Public Health Emergency Operations Center (PHEOC) before alerts could be escalated for investigation. This manual process formed a significant bottleneck, especially during periods of increased reporting caused by seasonal trends, population movement, or disease outbreaks.

By automating the extraction of key surveillance elements including disease type, symptoms, location, age, gender, and date of symptom onset, the NLP-enhanced workflow demonstrated the potential to reduce processing time, minimize human error, and support real-time analysis of SMS-based alerts. This directly contributed to faster triage and timely identification of potential outbreaks, supporting the 7-1-7 framework's goal of rapid detection and response (Frieden & Lee, 2021).

The use of NLP, particularly named entity recognition (NER), allowed the system to handle large volumes of unstructured messages with consistent speed and accuracy. This proved especially beneficial during high-volume reporting periods, where manual capacity is often overwhelmed. The approach aligns with national priorities and international public health strategies, including the WHO's IDSR framework, by enhancing timeliness, standardization, and responsiveness in the surveillance process.

Overall, this study provided practical evidence that integrating NLP into eIDSR workflows can improve efficiency and reliability in disease report processing, thereby strengthening Uganda's capacity to respond to public health threats in a scalable and sustainable manner.

3.3 Conceptual Framework

This study adopted a Design Science Research (DSR) framework to address inefficiencies in the manual processing of SMS-based disease alerts within Uganda's electronic Integrated Disease Surveillance and Response (eIDSR) system. DSR provided a structured, iterative approach for developing and evaluating a practical solution, an NLP model and supporting API, to extract structured data from unstructured SMS alerts.

The DSR approach guided the study through continuous cycles of design, implementation, demonstration, and evaluation, ensuring that the solution was both technically sound and contextually appropriate. This made it well-suited for the resource-constrained public health setting in which the study was conducted.

The conceptual framework was composed of the following components:

Problem Identification and Motivation: The manual extraction of disease-related information from incoming SMS alerts in the eIDSR system leads to delays, inconsistent data quality, and resource inefficiencies. These bottlenecks can hinder early detection and timely public health response. Recognizing these challenges, the study is motivated to automate the processing of SMS messages using NLP to improve data extraction accuracy and operational speed.

Objectives of the Solution: The goal is to improve the efficiency and reliability of SMS-based alert processing within eIDSR. Specifically, the solution aims to:

- Automatically extract key outbreak-related information (disease type, location, symptoms, age, gender, and onset date) from unstructured SMS messages;
- Reduce the time from alert receipt to usable structured data;
- Achieve high model performance based on precision, recall, and F1-score.

Design and Development of the Artifact: A transformer-based NLP model (BERT Uncased model) was selected and fine-tuned on a manually annotated dataset of historical SMS alerts. Tokenization and entity alignment were performed to prepare the data for training. In parallel, a custom API was developed in Python to serve the trained model and receive SMS input for processing. This pipeline formed the basis for a functional prototype capable of real-time entity extraction.

Demonstration: The model and API were integrated into a demo web interface that simulated the process of submitting an SMS alert and receiving structured outputs. This demonstration showed how the trained model could be embedded into a real-world workflow and used to process incoming messages in near real-time.

Evaluation: The artifact was evaluated using a reserved test dataset. Model performance was assessed using standard Named Entity Recognition (NER) metrics: precision, recall, and F1-score. Timeliness of processing was also measured by comparing model inference time against manual processing. A confusion matrix was used to assess classification accuracy across entity types, and McNemar's test was considered for comparing model predictions to ground truth annotations for statistical significance.

Communication: The development process, evaluation findings, and integration pathway were documented and communicated to relevant stakeholders. This included public health practitioners, system developers, and academic audiences. The prototype serves as a blueprint for future deployment,

demonstrating a viable pathway for scaling NLP-based automation within national surveillance systems.

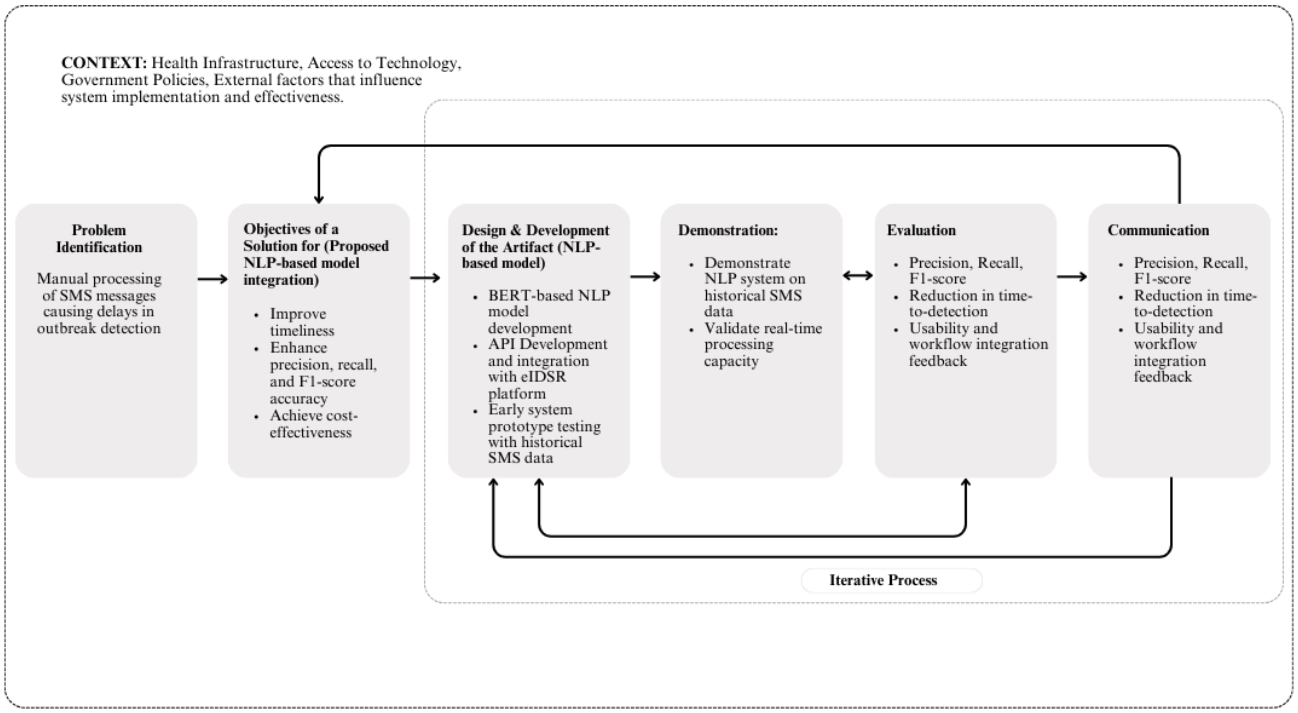

Figure 1: Figure showing conceptual framework

How it worked

The design science framework followed in this study was inherently iterative, beginning with a clear understanding of the problem and the formulation of measurable objectives. As illustrated in Figure 1, the process began with problem identification, which informed the design and development of the initial version of the NLP artefact.

This initial model was trained on a manually annotated dataset of SMS messages and embedded within a custom-built API. The prototype was then demonstrated through a web-based demo interface, simulating how SMS alerts could be routed to the model and returned as structured outbreak data.

Following this, the model underwent rigorous evaluation using a reserved test set and standard performance metrics, including precision, recall, F1-score, and processing time. Insights from this evaluation guided further refinements, including hyperparameter tuning and adjustments to the training dataset to improve generalizability. This feedback loop allowed for continuous improvement of the system's accuracy and efficiency.

By applying this iterative process, the study ensured that the final artefact was both technically robust and practically viable within the context of Uganda's health surveillance system. The resulting NLP-powered system demonstrated a feasible method for automating the extraction of structured data from SMS alerts, contributing to more timely and consistent disease reporting.

The design science framework enabled this research to effectively bridge the gap between theoretical modelling and real-world application, providing a scalable and context-appropriate solution for public health information processing.

CHAPTER FOUR
RESEARCH QUESTIONS AND STUDY OBJECTIVES

This study aimed to develop and evaluate the performance of a Natural Language Processing (NLP) model for the automated extraction of epidemiological entities from unstructured Short Message Service (SMS) alerts in Uganda's electronic Integrated Disease Surveillance and Response (eIDSR) system.

4.1 Research Questions

The following research questions guided this project:

1. What factors influence the accuracy of extracting key surveillance information, such as disease type, location, and symptoms, from SMS messages submitted to the eIDSR system?
2. How can an NLP model be developed and trained to improve the extraction of this key information, considering the identified influencing factors?
3. How does the NLP-powered system perform compared to the current manual approach, in terms of processing speed and key evaluation metrics (precision, recall, F1-score, and time-to-detection) for SMS-based disease surveillance?

4.2 Objectives

4.2.1 Main Objective

To enhance outbreak surveillance through integration of Natural Language Processing in Uganda's electronic Integrated Disease Surveillance and Response System

4.2.2 Specific objectives

1. To identify factors influencing the accuracy of key information extraction from eIDSR SMS messages, including disease type, location, and symptoms.
2. To develop a Natural Language Processing (NLP) model based on these identified factors, using a manually annotated dataset of historical SMS alerts.
3. To evaluate the performance of the NLP-powered system by comparing its processing speed and accuracy to the current manual approach.

CHAPTER FIVE
METHODS

This section describes the methods used in developing, training, and integrating a Natural Language Processing (NLP) model for automated disease report processing within the Ugandan eIDSR system. The process followed a structured approach, encompassing data preparation, model development, training, evaluation, and demonstration of integration with the eIDSR platform.

5.1 Study Area and Setting

This study utilized data from Uganda's electronic Integrated Disease Surveillance and Response (eIDSR) system, a Ministry of Health–managed platform that facilitates real-time disease surveillance through SMS reporting. The eIDSR system serves as the primary mechanism for community health workers and the public to transmit unstructured alerts to central surveillance units. The system is hosted on Ministry of Health servers and receives messages containing vital epidemiological details, including suspected disease types, symptoms, locations, and dates of onset.

This study utilized historical, de-identified SMS data collected by the eIDSR system, which represents real-world, high-stakes communication from the field. The analysis was conducted using SMS data collected over a one-year period, from 1 January 2024 to 31 December 2024. The dataset is characterized by messages containing vital epidemiological details (suspected disease types, symptoms, locations, dates of onset) that are expressed in informal, abbreviated, and code-switched language (mixing English with local Ugandan languages). These characteristics define the specific technical challenge this study addresses.

5.2 Study Design

A retrospective design was adopted, using historical SMS data from the eIDSR system to train and evaluate the performance of the NLP model. The design enabled objective assessment of the model's extraction accuracy and operational feasibility.

To complement the quantitative evaluation, a qualitative component identified contextual and operational factors that could influence model accuracy. This included informal interviews and key informant discussions with eIDSR focal persons, district surveillance officers, and data managers, as well as system walkthroughs to observe how SMS alerts were composed and submitted. These insights

guided data preprocessing steps to address common inconsistencies such as varied abbreviations, misspellings, and non-standard message structures.

5.3 Study Sample

The study used a one-year archive of SMS alerts submitted to the eIDSR system, comprising approximately 8,029 messages. These alerts were generated by community health workers and members of the public, containing unstructured text with key epidemiological details such as the suspected disease, symptoms, location, and relevant dates.

Each message contained a mix of structured and unstructured information, including:

- **Reported symptoms such as** rash, body aches, oral ulcers, and other relevant clinical signs.
- **Suspected disease names:** The specific disease mentioned in the message.
- **Patient demographics:** Age and gender of the reported case.
- **Location details:** Geographic information such as district, sub-county, or village.
- **Symptom onset and reporting dates:** Timestamps indicating when symptoms began and when the report was sent.
- Date of Onset:

The entire archive of 8,029 SMS messages was utilized for exploratory data analysis and linguistic profiling to identify the structural characteristics and variations that influence the difficulty of key a subsample of 1,331 unique SMS messages was manually annotated to form the Gold Standard dataset. This final annotated corpus was strictly partitioned into a Training Set (70%) used to fine-tune the BERT model, a Validation Set (15%) used for hyperparameter tuning, and a Held-Out Test Set (15%) used exclusively for the final, unbiased evaluation of model accuracy and speed against the manual baseline.

The aspect of time was captured using the system-generated metadata from the eIDSR platform, which automatically assigns a unique received timestamp to every SMS alert upon arrival at the Ministry of Health servers. This metadata allowed for the precise temporal anchoring of the SMS alerts within the study period.

5.3.1 Inclusion and exclusion criteria

The inclusion and exclusion criteria were essential for ensuring the quality and relevance of the data used to evaluate the NLP model's ability to process and extract information from SMS-based disease notifications. They ensured that only complete, high-quality SMS alerts, containing key details such as disease type, location, symptoms, and timestamps, were included in the analysis to enable accurate extraction of key surveillance indicators.

Inclusion Criteria

The following criteria was applied to determine the eligibility of SMS messages for inclusion in the analysis.

- **Timeframe:** Only SMS messages received within the defined one-year study period (January 1, 2024 to December 30, 2024) were included to ensure the dataset reflected the current operational context of the eIDSR system and provided a relevant basis for evaluating the NLP model's performance under present-day conditions.

- **Disease focus:** Messages referencing high-priority zoonotic conditions identified by the Ugandan Ministry of Health ("One Health Zoonotic Disease Prioritization for Multi-Sectoral Engagement in Uganda," 2017) such as Anthrax, Zoonotic Influenza Viruses, Viral Haemorrhagic Fevers (Ebola, Marburg, Rift Valley Fever, Crimean Congo Haemorrhagic Fever), Brucellosis, Trypanosomiasis, Plague, and Rabies.**Context:** Messages were only included if they contained the suspected disease or condition, patient location (at least at the district level), age, sex, and date of symptom onset.

Exclusion Criteria

The following criteria were applied to remove ineligible messages from the dataset.

- **Duplicate alerts:** SMS messages referring to the same case were identified and removed to prevent dataset inflation and analytical bias. De-duplication was performed using message content, reporter identification, and the date/time of submission.
- **Non-operational messages:** Test messages for system maintenance or training, as well as those generated due to system errors, were excluded to avoid contamination by irrelevant data.

5.4 Qualitative Study Participants and Sampling

The qualitative component of this study employed a purposive maximum variation sampling strategy to select the interview participants. This non-probability technique was used to ensure representation from all relevant levels of the eIDSR operational hierarchy, thereby capturing a diverse range of perspectives on system feasibility, potential integration barriers, and desired functionalities. The final sample size for the interviews included 10 individuals, comprising four Surveillance Officers/Data Clerks (PHEOC and District level Data Analysts), two District Surveillance Officers, and four Senior Ministry of Health Managers/Technical Advisors.

The study population was specifically segmented into two functional groups: Frontline Users (Surveillance Officers and Data Clerks) who provided critical insight into the inherent challenges of the current manual workflow, and Public Health Managers and Decision-Makers (senior staff from the MoH and relevant partner organizations) who offered essential perspectives on operational oversight, technical feasibility, resource allocation, and the ultimate institutional approval required for integrating the new system.

To ensure the relevance and authority of the expert feedback, participants were selected based on three specific criteria: they required direct involvement with the eIDSR system and SMS data for a minimum of one year; they needed to demonstrate established knowledge of Uganda’s disease surveillance strategy (IDSR); and they had to confirm their willingness and availability to participate in a structured interview.

The data collected consisted of rich, contextualized textual narratives sourced from these Key Informants via semi-structured interviews. This information comprehensively addressed the operational and institutional factors critical to the NLP model's successful integration, providing necessary context for the quantitative findings. Specifically, the content focused on three main thematic areas.

First, it explored the operational workflow and bottlenecks, detailing the manual processing challenges, data quality issues stemming from informal SMS language, and the impact of the current system on public health timeliness targets. Second, the data captured perceived usefulness and usability, documenting informants' opinions on the value proposition of automation, their level of trust in a machine learning system to triage alerts reliably, and their requirements for the system's user interface and validation mechanisms.

Finally, the data addressed integration barriers and policy implications, providing expert insights into existing technical readiness, potential institutional resistance to workflow changes, and the long-term

sustainability requirements for maintaining the NLP solution within the Ministry of Health's eIDSR system. This collection of expert feedback allows the study to move beyond technical performance metrics to assess real-world viability.

5.4.2 Ethical Review and Consent

Before commencing the interviews, the following steps were taken:

Approval: Ethical approval for the study, including the qualitative component, was secured from the Makerere University School of Public Health Research and Ethics Committee.

Informed Consent: Each prospective participant received a clear explanation of the study's purpose, interview procedures, voluntary nature of participation, and their right to withdraw at any time without penalty. Written or recorded verbal informed consent was obtained from every individual prior to the start of their interview.

Confidentiality: Participants were assured that their identities would remain confidential and that all collected data would be de-identified and reported using anonymized codes (e.g., KI-01, KI-02) to protect their privacy.

5.4.3 Interview Instrument Development

A semi-structured interview guide was systematically developed to ensure alignment with the study's qualitative objectives, particularly those related to operational feasibility and system integration. The guide was organized around several key thematic areas. These themes included an exploration of the current manual SMS processing workflow and bottlenecks experienced by frontline staff, assessment of the perceived usefulness and usability of an automated NLP system, identification of potential technical and institutional barriers that could impede integration, and collection of feedback on the necessary functionalities and design requirements for the NLP tool. Finally, the instrument addressed the expected impact of the NLP system on existing public health surveillance policies, such as the WHO's 7-1-7 framework for timely notification.

5.4.4 Interview Logistics and Execution

The interviews were executed in a systematic manner:

Scheduling: Participants were contacted via email or phone to schedule a suitable date, time, and mode for the interview (mostly conducted virtually via Zoom/Google Meet or in-person, depending on the informant's preference).

Mode: The majority of interviews were conducted virtually and in-person and lasted approximately 45 to 60 minutes.

Recording: With the explicit permission of the participant, all interviews were audio-recorded to ensure accurate capture of the data. Comprehensive field notes were simultaneously taken to document non-verbal cues and key contextual details.

5.4.5 Transcription and Preparation for Analysis

Following the completion of data collection, a rigorous process was undertaken to prepare the audio recordings for qualitative analysis. Initially, all audio recordings were professionally transcribed verbatim to convert spoken data into textual format. To ensure data fidelity, the transcripts were then meticulously verified by cross-checking them against the original audio recordings for accuracy. The crucial final step involved anonymization: all identifiable information, such as participants' names or specific dates that could trace an individual, was immediately removed or systematically replaced with the assigned unique Key Informant IDs (KI-01 to KI-10).

5.5 Study variables

5.5.1 Dependent variable/Outcome variable

The dependent variables measured the performance and effectiveness of the NLP-powered system in processing SMS-based disease notifications compared to manual extraction. The effectiveness is conceptually defined as the system's ability to successfully automate the initial triage function of the Public Health Emergency Operations Centre (PHEOC) surveillance officer by simultaneously achieving high extraction accuracy and superior operational timeliness when processing unstructured SMS alerts, as compared to the current manual extraction process.

The dependant variables included:

Time to Detection (hours): The elapsed time between receipt of an SMS report and completion of the NLP model's extraction process. This was measured continuously in seconds to assess system processing speed and responsiveness, which directly addresses the WHO 7-1-7 framework for timely notification.

Manual Processing Lag (Operational Benchmark): To provide a comparative baseline, the Date of Signal Follow-up and Date of Signal Status Update are extracted from the historical dataset. These variables are used to calculate the duration of the existing manual workflow. This calculation is intended to establish a benchmark for current responsiveness, allowing for a direct comparison between traditional human-led data entry and the automated NLP approach.

Precision (Positive Predictive Value): The proportion of extracted entities that were correct, calculated as the number of true positives divided by the sum of true positives and false positives. This reflects the reliability of the system's output.

Recall (Sensitivity): The proportion of actual relevant entities that were correctly extracted by the model, calculated as the number of true positives divided by the sum of true positives and false negatives. This indicates the model's ability to capture all relevant information.

F1 Score: The harmonic mean of precision and recall, providing a single measure of the system's ability to balance accurate detection with completeness of extraction.

5.5.2 Independent variables/Exposures

The independent variables represent the key information elements extracted from SMS-based disease notifications received through the eIDSR system. These served as inputs for the NLP model and influenced its accuracy, processing speed, and overall performance. They included:

- **Disease type:** The suspected disease or condition mentioned in the SMS (e.g., anthrax, measles, malaria, Ebola).
- **Symptoms:** Signs and symptoms described in the message, such as fever, rash, diarrhoea, vomiting, or difficulty breathing.
- **Location:** The geographical area of the suspected case, at least at the district level, with finer details such as sub-county or village where available.
- **Gender:** The sex of the suspected case, typically recorded as male or female.
- **Age:** The age of the suspected case, recorded numerically or categorically (e.g., child, adult).
- **Reporting Date (System Metadata):** The automatically generated timestamp recording when the SMS reached the Ministry of Health servers
- **Date of Onset:** The specific date or time reference mentioned by the reporter indicating when symptoms first appeared. This is critical for calculating the reporting lag (the delay between illness and notification).

These independent variables were analysed in relation to the dependent variables by comparing the model's extracted outputs against manually validated records. The evaluation focused on processing time, precision, recall, and F1-score to determine the system's ability to accurately and efficiently extract key details from real-world SMS notifications.

5.6 Model Development and Finetuning

The development of the NLP model followed a structured process encompassing model selection, training, fine-tuning, and iterative refinement. The workflow began with raw SMS data collected from the eIDSR system, which underwent pre-processing steps including cleaning, tokenization, and normalization. Messages were then annotated using the BIO tagging scheme to define key epidemiological entities of interest. A pre-trained BERT model was selected and fine-tuned on the annotated dataset, with performance evaluated using precision, recall, F1-score, and loss. Based on these results, the model was refined and improved before being operationalized into a prototype API and demo that incorporated a human-in-the-loop validation layer to handle uncertain cases. The overall workflow is summarized in Figure 3.

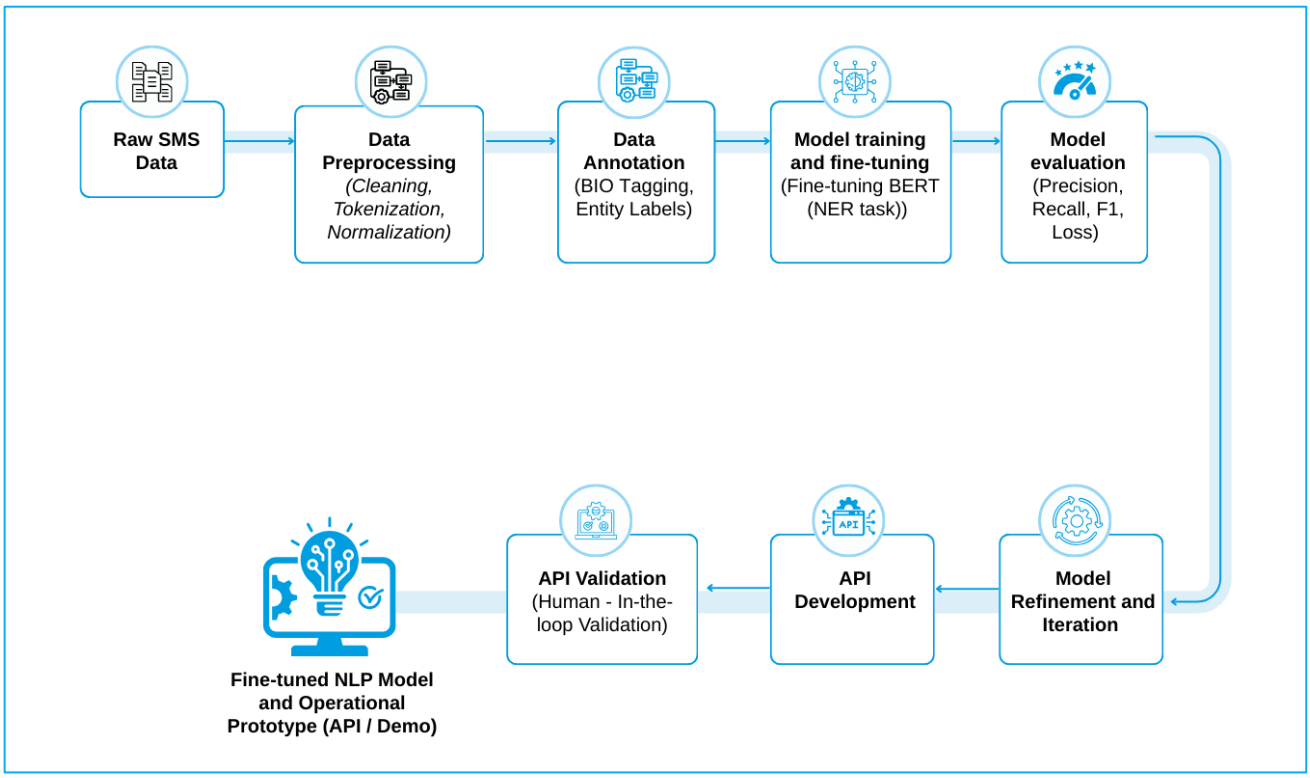

Figure 2 : model training and finetuning workflow

5.6.1 Data Acquisition

A comprehensive dataset of historical SMS messages related to disease reports was obtained from the Ugandan eIDSR system. This dataset served as the foundation for training and evaluating the Natural Language Processing (NLP) model. Given the data requirements of BERT-based NLP models, 1,331 annotated SMS messages were analysed, ensuring sufficient training, validation, and testing to optimize model performance.

5.6.2 Data Cleaning and Annotation

In this phase, comprehensive data cleaning was carried out to enhance the accuracy and consistency of the input text. This involved removing irrelevant content such as greetings, signatures, and unrelated characters, as well as standardizing message formats. Common spelling errors, abbreviations, and variations in disease reporting terminology were corrected using a domain-specific lexicon. All messages were converted to lowercase, and unnecessary punctuation and non-standard symbols were removed for consistency.

Messages were then prepared for BERT input by tokenizing them with the model's WordPiece tokenizer, which split the text into subword units compatible with transformer-based architectures. Tokenized sequences were padded to a uniform length or truncated where necessary, with the maximum sequence length set to 128 tokens. This setting balanced processing efficiency with the retention of critical context, given the typically short nature of SMS alerts.

Following pre-processing, the dataset of SMS messages was manually annotated for supervised learning. Each message was labelled with key entities relevant to public health notification processing, including disease type, reported symptoms, location, age, gender, and date of symptom onset. The annotation process was guided by a structured codebook , which specified the exact token-level tagging scheme for entities (e.g., B- for beginning of entity, I- for inside, and O for outside). This ensured that the annotations were not only consistent for human review but also directly usable for model training. To enhance reliability, the codebook was iteratively refined by consulting relevant public health literature and sample message reviews. A subset of the annotated data was re-examined after an interval of time to check for intra-annotator consistency, and discrepancies were corrected. This ensured that the final annotated dataset was consistent for human review and directly usable for model training.

The finalized annotated dataset was split into training (70%), validation (15%), and test (15%) subsets using a stratified approach to preserve the distribution of entity classes. The training set was used to fine-tune the model, the validation set supported hyperparameter tuning and early stopping, and the test set provided an unbiased evaluation of the model's performance. This annotated dataset of SMS alerts served as the foundation for developing and evaluating the NLP model's ability to automatically extract structured disease notification data from unstructured SMS text.

5.6.3 Feature extraction

The dataset was prepared in BIO tagging format to mark entity boundaries for key variables, including disease type, symptoms, location (district), age, gender, and date of symptom onset. This domain-specific customization allowed the model to recognize local health terminology, non-standard spellings, abbreviations, and variations in disease and symptom reporting commonly found in community-generated SMS alerts.

Additional pre-processing steps such as text normalization, tokenization using the model's WordPiece tokenizer, and sequence length adjustment, were implemented to align with the model's input requirements.

5.6.4 Model Selection

This study adapted the BERT Uncased model, a transformer-based architecture developed by Google and pre-trained on large-scale English text corpora (Bhutda et al., 2024). The model was selected for its suitability in handling short, unstructured, and informal text, making it well-adapted for processing SMS-based disease notification alerts such as those from Uganda's eIDSR system (Jarashanth & Nawarathna, 2022). The model's relatively compact size provides a balance between high performance and computational efficiency, enabling faster training and inference compared to larger transformer variants.

To adapt the model to the study context, it was fine-tuned on a manually annotated Named Entity Recognition (NER) dataset created from historical SMS messages.

5.6.5 Model training and fine-tuning

The model training task involved adapting its pre-trained capabilities to extract structured entities—such as disease type, location (district), symptoms, age, gender, and date of symptom onset—from unstructured SMS messages.

The model architecture comprises of 8 transformer encoder layers, each incorporating 8 multi-head self-attention mechanisms. The hidden dimension size of 512 per token allow the model to learn rich contextual representations from the short, informal, and code-mixed SMS text typical of community health reporting.

For fine-tuning, a task-specific token classification head was appended to the pre-trained BERT model to perform Named Entity Recognition (NER) using the BIO tagging format. This head assigned a label

to each token in the input sequence, allowing the model to detect and classify entity boundaries in SMS messages. Messages were tokenized using the model's WordPiece tokenizer, with padding or truncation applied to a fixed sequence length of 128 tokens for computational efficiency.

Hyperparameters were selected through a combination of established literature guidance and empirical experimentation with the annotated dataset:

Learning Rate (3e-5): Fine-tuning transformer models requires relatively small learning rates to avoid catastrophic forgetting of pre-trained language representations. Prior studies recommend values in the range of 2e-5 to 5e-5 for BERT-based models (Devlin et al., 2019; Howard & Ruder, 2018). Higher rates, such as 2e-2, typically destabilize optimization and degrade performance, as observed in preliminary trials of this study. Empirically, 3e-5 provided stable convergence and improved validation F1-scores compared to the initial baseline of 5e-5.

Batch Size (8 per device): Batch size was constrained by GPU memory availability. Larger sizes (>16) exceeded memory limits, while smaller sizes (<4) led to noisy gradients and unstable learning. A batch size of 8 struck the balance between computational feasibility and training stability, consistent with recommendations in resource-constrained fine-tuning scenarios (Sun et al., 2019).

Number of Epochs (12): Although BERT models are often fine-tuned within 2–5 epochs on general NLP tasks (Devlin et al., 2019), experiments on this dataset showed that shorter runs underfit rare classes (e.g., I-SYMPTOM). Validation performance improved steadily until around 10–12 epochs, after which F1-scores plateaued and only marginal (<0.5%) gains were observed. Thus, 12 epochs were adopted to maximize recall and generalization without significant overfitting.

Weight Decay (0.01): This regularization parameter, standard in BERT fine-tuning (Loshchilov & Hutter, 2019), was applied to penalize large weights and mitigate overfitting. Empirical tests confirmed that removing weight decay slightly increased validation loss, supporting its inclusion.

Training initially began with baseline hyperparameters from the BERT literature (learning rate = 5e-5, epochs = 4). These settings resulted in unstable convergence and underfitting of rare entity classes. Iterative adjustments, guided by validation trends, led to the refined configuration of learning rate = 3e-5, batch size = 8, epochs = 12, and weight decay = 0.01. This combination yielded the most consistent improvements in precision, recall, and F1-score across all entity categories, particularly for symptoms, which were the most variable in SMS reporting.

The cross-entropy loss function was used to optimize the model by comparing predicted labels against the annotated gold-standard labels. Performance was monitored on the validation set after each epoch, with evaluation metrics including precision (Positive Predictive Value), recall (sensitivity), and F1-score computed using the seqeval library for sequence labelling tasks.

Upon completion of training, the final model was evaluated against the test dataset to provide an unbiased measure of real-world performance. The best-performing model was selected based on the highest F1-score on the validation set, which balances precision and recall and is widely recommended for Named Entity Recognition tasks where both false positives and false negatives are critical. Precision, recall, and accuracy were also monitored, but F1-score was used as the primary criterion for model selection. The corresponding tokenizer was saved alongside the model for deployment and integration into the eIDSR system.

This fine-tuning process yielded a domain-adapted model capable of automatically extracting structured public health notification data from incoming SMS messages, enabling faster and more consistent processing compared to manual review.

5.7 Model Evaluation

5.7.1 Model evaluation

The evaluation of the fine-tuned BERT Uncased model was carried out systematically to assess its ability to accurately extract relevant disease notification information from unstructured SMS messages. The evaluation used the annotated SMS dataset that had been split into training (70%), validation (15%), and test (15%) subsets. The test set, which consisted of data unseen during training, provided an unbiased assessment of model performance.

Predictions generated by the fine-tuned model on the test set were compared against expert-generated annotations (the ground truth) to determine accuracy. The evaluation focused on standard metrics for Named Entity Recognition (NER) tasks, precision, recall, and F1-score.

Precision measured the proportion of correctly identified entities out of all entities predicted by the model (Afzal et al., 2016). High precision would indicate that the model generated few false alarms and extracted relevant entities accurately and vice versa. This was calculated as:

$$Precision = \frac{true\ Positives}{True\ Positives + False\ Positives}$$

Recall measured the proportion of disease-related entities present in the SMS messages that were correctly identified by the NLP model. In this study, entities refered specifically to annotated disease names, symptoms, and location mentions within the SMS text. A high recall indicated that the model successfully captured most of the relevant information from the messages without omitting key entities. This was computed as:

$$Recall = \frac{True\ Positives}{True\ Positive + False\ Negatives}$$

The **F1-score** is the harmonic mean of precision and recall and provides a single, balanced measure of the NLP model's entity extraction performance. In this study, the F1-score captures the trade-off between correctly extracting relevant disease-related entities (precision) and successfully identifying all relevant entities present in SMS messages (recall).

$$F1 - Score = 2 * (\frac{Precision * Recall}{Precision + Recall})$$

This metric provided a balanced measure of the model's performance by considering both accuracy in identifying relevant entities (precision) and its ability to capture all important information (recall).

Loss measured the discrepancy between the model's predicted probabilities and the actual ground-truth labels during training and evaluation. In this study, the token-level cross-entropy loss was used, which is standard in sequence labeling tasks (Goodfellow et al., 2016). Lower loss values indicate closer alignment between predicted and true labels, while higher values reflect greater divergence. This metric was useful for monitoring convergence across epochs and assessing overall learning stability.

The model achieved validation loss values ranging between 0.137 and 0.215 across training epochs, with a final test loss of 0.165, confirming stable convergence without evidence of overfitting. The cross-entropy loss was computed as:

$$\frac{1}{N} * \sum_{I=1}^{N} \sum_{C=1}^{C} y_{i,c}.\log(\mathbb{y}_{i,c})$$

where N is the number of tokens, CCC the number of classes, $y_{i,c}$ the true label (1 if class ccc is correct for token iii, otherwise 0), and $\mathbb{y}_{i,c}$ the predicted probability for class c.

A high F1-score signified that the model was both accurate and comprehensive, effectively minimizing false positives and false negatives to ensure reliable extraction of outbreak-related data from SMS messages.

In this study,

- **True Positives (TP)** represented correctly identified entities matching the ground truth labels.
- **False Positives (FP)** were entities predicted by the model but absent in the ground truth.
- **False Negatives (FN)** were entities present in the ground truth but missed by the model.

To assess operational efficiency, the average processing time per SMS was measured, calculating the time taken from receiving an SMS to completing entity extraction. This helped determine the model's potential to improve real-time disease notification processing compared to manual review.

Further validation was conducted by running the trained model on historical raw SMS data and comparing extracted entities to structured information documented in archived eIDSR surveillance reports. This retrospective analysis assessed the model's ability to replicate human-extracted data in real-world conditions.

To evaluate the alignment in classification performance between the NLP model and the manual baseline approach, a comparative error analysis was conducted using the framework of McNemar's test (Smith & Ruxton, 2020). While automation is inherently expected to provide superior operational efficiency in terms of speed, this analysis was necessary to empirically assess the accuracy gap between the two methods. In high-stakes disease surveillance, it is critical to determine the extent to which a model's performance deviates from the human expert standard rather than assuming parity through automation alone.

This approach is specifically designed for paired nominal data and is used to compare the performance of two classification systems on the same dataset by focusing on discordant pairs—cases where one method succeeded while the other failed. In this study, the annotated subset of 1,331 SMS messages was used, where each message was evaluated to determine if the model and the manual method reached the same conclusion in extracting key entities, including disease, symptoms, age, gender, and district. By quantifying these disagreements, the study identified the specific linguistic complexities that require human oversight.

Through this multi-faceted evaluation, the model's accuracy, comprehensiveness, and practical efficiency were quantified. The resulting distribution of concordant and discordant cases provides the empirical evidence required to justify a Human-in-the-Loop workflow, where the NLP system handles routine high-volume notifications while flagging complex cases for expert review within the eIDSR system.

5.7.2 Data analysis

Data analysis was conducted in two primary phases, quantitative (for the NLP model's development and accuracy) and qualitative (for operational feasibility), to align with the specific objectives of the study.

A. Identifying Influencing Factors

The initial analysis focused on rigorously establishing the Gold Standard dataset and identifying the linguistic and structural factors that complicate information extraction.

Annotation Process and Quality Control: A subsample of the 1,331 SMS records was manually annotated by two independent public health experts trained in Named Entity Recognition (NER) using the BIO (Beginning, Inside, Outside) tagging scheme. The six target entities—Disease, Symptom, Location, Age, Gender, and Date—were tagged according to detailed annotation guidelines, which were developed to address the specific complexities of the code-switched and informal SMS messages.

Consensus Building for Gold Standard: To ensure the highest level of data fidelity, the inter-rater reliability of the initial annotations was assessed. While a formal Kappa statistic was not calculated, discrepancies between the two annotators were meticulously tracked and resolved through weekly consensus meetings with a third, senior epidemiologist. This structured reconciliation process guaranteed that all ambiguous entities and challenging linguistic structures were harmonized, resulting in a single, verified Gold Standard dataset required for model training. Analysis of these recurring discrepancies helped precisely define the factors (e.g., abbreviations, multilingualism) that contribute to low extraction accuracy.

B. Model Development and Training

This objective involved computational analysis to build and optimize the entity extraction model.

Data Partitioning: The Gold Standard annotated dataset was systematically partitioned into distinct sets: Training (70%), Validation (15%), and a final Hold-out Test (15%) set to ensure unbiased evaluation.

Model Selection and Optimization: A transformer-based model, specifically a BERT-NER architecture, was selected as the base. The model was trained iteratively on the Training Set and optimized using the Validation Set to fine-tune hyperparameters and minimize the loss function.

C. Performance Evaluation and Comparison

This objective combined quantitative metrics to assess performance and qualitative analysis to assess operational feasibility, completing the comparison to the current manual approach.

Quantitative Evaluation: The final, trained model was run against the unseen Hold-out Test Set, and its performance was assessed using the following metrics:

Extraction Accuracy: The model's output was compared to the Gold Standard using three standard performance metrics from Named Entity Recognition (NER). Precision was used to reflect the model's reliability in avoiding false alarms (False Positives), ensuring that the information extracted was correct. Recall (Sensitivity) was used to reflect the model's ability to capture all critical information present in the message (minimizing False Negatives), which is paramount in an outbreak detection system. Finally, the F1-Score, calculated as the harmonic mean of Precision and Recall, served as the primary measure of overall model accuracy and robustness.

Statistical Significance: The statistical significance of differences in classification performance between the NLP model and the baseline (manual extraction) was tested using McNemar’s test, a paired non-parametric test suitable for evaluating the difference in classification outcomes on the same set of records.

Timeliness Comparison: The computational processing time (inference speed) of the NLP model per SMS was measured and directly compared to the documented baseline time taken for the current manual reading, triage, and data entry process, quantifying the gain in operational speed.

Qualitative Analysis (Operational Feasibility)

The data collected from the 10 Key Informant Interviews (KIIs) was subjected to Thematic Analysis, following the rigorous six-phase approach proposed by Braun and Clarke:

Familiarization and Coding: Transcripts were read thoroughly, and relevant textual segments were systematically coded line-by-line using qualitative analysis software (e.g., NVivo).

Theme Development: Codes were iteratively sorted into potential themes and sub-themes that captured patterns related to the operational feasibility questions (e.g., usability, barriers, trust, policy impact).

Refinement and Reporting: Themes were reviewed for coherence, defined, and used to construct a narrative argument, which was substantiated by direct quotes from the Key Informants to describe the operational environment and assess the viability of integrating the NLP-powered system.

The findings from both the quantitative metrics and the qualitative themes were ultimately triangulated in the discussion to provide a holistic conclusion on the overall effectiveness and readiness of the NLP-powered system.

5.7.3 Model Refinement and Iteration

Following initial evaluation with baseline hyperparameters (learning rate = 2e-4, 4 training epochs, batch size = 16), the model showed unstable convergence and signs of underfitting, particularly on longer SMS messages and rare entity classes. To address this, a systematic series of experiments was conducted, guided by validation trends and standard practices in transformer fine-tuning.

The learning rate was reduced to 3e-5, which is commonly recommended in prior BERT fine-tuning studies for sequence labeling tasks and empirically provided more stable optimization on the validation set. The number of epochs was increased to 12, based on observed gains in F1-score up to this point, beyond which additional training yielded diminishing returns. The batch size was set to 8, a compromise between available computational resources and gradient stability, consistent with settings used in similar low-resource NER studies. These final hyperparameter values were retained throughout subsequent refinement because they consistently produced stable convergence and balanced performance across entity classes.

At this stage, refinements focused not on further hyperparameter tuning but on data handling and annotation quality. Tokenization rules were adjusted to better preserve entity boundaries, and sentence-level annotations were re-checked for consistency. Where performance bottlenecks were identified, particularly in symptom continuation (I-SYMPTOM) tokens, additional annotated examples were incorporated to strengthen underrepresented entity types.

Model refinement was guided by evaluation on the validation and test datasets, with F1-score and recall prioritized as the main stopping criteria. F1-score was selected because it provides a balanced measure of overall performance by jointly considering precision and recall, which is the standard practice in NER evaluation (Devlin et al., 2019). Recall was given additional weight in this study because, in a public health surveillance context, false negatives (missed outbreak signals) carry greater risk than false positives (extra alerts) (WHO, 2022).

Specifically, training was considered converged when improvements in validation F1-score or recall were smaller than 0.5–1.0% across two successive epochs, as suggested by common early-stopping practices for transformer models (Howard & Ruder, 2018; Sun et al., 2019). This threshold balances avoiding premature termination (which risks underfitting) with preventing overfitting and wasted computation. In addition, both metrics were required to remain above 90%, a level that is widely considered indicative of reliable extraction performance in applied NER tasks (Li et al., 2020).

The final model version was selected once it consistently achieved strong performance across entity types, demonstrated stable generalization, and met the prioritized evaluation criteria. It was therefore deemed sufficiently optimized for deployment in an operational disease surveillance setting.

5.8 Application Programming Interface Development and Simulation

5.8.1 Application Programming Interface Development

To demonstrate how the trained NLP model could be integrated with the eIDSR system, a custom web-based prototype was developed using HTML, JavaScript (jQuery), and Bootstrap for the frontend, and Python (Flask) on the backend. The application programming interface (API) accepts POST requests containing free-text SMS messages and processes them through the trained model to extract structured entities, including disease type, age, gender, symptoms, and location.

The user interface allows manual input of sample SMS alerts and displays the extracted entities in formatted JSON for easy interpretation. Upon clicking the submit button, the message is sent to the /extract_entities API endpoint via AJAX. The backend processes the input using the trained NER model and returns the result in real-time. Error handling was implemented to provide feedback if processing fails. This setup served as a functional demonstration of how NLP outputs could be integrated into the existing eIDSR SMS workflow.

5.8.2 System Integration

While direct integration into the national eIDSR system was not implemented due to restricted access to its infrastructure, the developed API was designed to be easily deployable within the existing SMS processing pipeline. The prototype closely simulated how real-time alerts could be automatically routed through the API to the trained NLP model and returned as structured disease notification data.

To support future integration efforts, the system was developed following standard practices for modular design and interoperability. API endpoint specifications, input/output formats, and sample use cases were clearly defined and tested through the demo interface. This ensured that the integration approach remains compatible with existing workflows while offering a scalable and efficient method for entity extraction from incoming SMS alerts.

Comprehensive documentation was maintained throughout, including the API request structure, response formatting, and guidance for deploying the model in a production environment. This foundation enables smooth integration with the eIDSR platform when system access becomes available, facilitating automation of SMS triage and improving the speed and accuracy of public health response.

5.8.3 Validation

A series of validation tests were conducted to ensure the end-to-end functionality of the integrated NLP system within the eIDSR platform. Although full integration into the national eIDSR system was not possible during this study, a demo interface was developed to simulate how the trained NLP model could be used to process incoming SMS alerts in real time. Through this prototype, end-to-end validation was carried out to assess whether the model could extract relevant public health information accurately and consistently.

Test SMS alerts were submitted via a web-based interface, which routed each message through the API. The NLP model returned structured data—including disease type, symptoms, age, gender, location, and date of onset—in JSON format. The results were displayed in real time, demonstrating the system's readiness for integration into existing surveillance workflows.

Two key validation strategies were employed. **(1) Content Validity:** The system's extracted outputs were compared against manually annotated ground truth data to assess whether all relevant disease

entities were correctly identified. **(2) Concurrent Validity:** Outputs from the model were also compared with historical surveillance summaries from previously processed alerts.

5.9 Quality Assurance and Quality Control (QA/QC)

A comprehensive Quality assurance and quality control (QA/QC) plan was implemented throughout the study to ensure the reliability and validity of the results. In this context, quality assurance encompassed proactive processes designed to prevent errors, while quality control focused on procedures used to detect and correct errors.

5.9.1 Data Quality Assurance

All SMS messages were subjected to standardized pre-processing procedures to address inconsistencies, remove noise (e.g. greetings and extraneous characters), and ensure uniform formatting. A custom codebook guided the manual annotation process to maintain consistency in labelling key entities such as disease, symptoms, location, age, gender, and date of onset.

5.9.2 Model QA/QC

Model development followed a rigorous training and evaluation protocol. Training progress was continuously monitored using metrics such as precision, recall, and F1-score. Epoch-based evaluation on the validation set ensured early detection of overfitting and guided performance tuning.

The model evaluation followed a 70/15/15 split for training, validation, and test sets, which provided reliable insight into generalization performance. The training process included checkpointing, allowing restoration and comparison of multiple model versions to identify and preserve the best-performing configuration.

All scripts were version-controlled, and evaluation metrics were logged and visualized to support transparent model selection and reproducibility.

5.9.3 Data Management

Effective data management was essential to ensure the integrity, confidentiality, and reproducibility of this study. The SMS dataset, which formed the basis for model training and evaluation, was handled using secure and structured procedures throughout the project lifecycle.

To protect the privacy of individuals, all messages were anonymized by removing personally identifiable information (PII) such as names or contact details. Each SMS was assigned a unique identifier, allowing for accurate tracking and reference during annotation and analysis without compromising confidentiality.

Data cleaning, transformation, and annotation processes were version-controlled using Git software. Commit messages were used to document each update, indicating the nature of the changes (e.g. data corrections, annotation revisions, or model configuration adjustments) along with timestamps. File naming conventions and directory structures were standardized to support clarity and consistency across data and code repositories.

Datasets were stored in the following interoperable formats:

- **TSV (Tab-Separated Values)** was used for the BIO-annotated dataset to support sequence labelling.
- **CSV (Comma-separated values)** files were used for structured tabular records.
- **JSON** was used for storing API responses and model outputs.

At the end of the study, all resources were carefully archived in line with institutional data governance standards. While the annotated datasets remain protected due to confidentiality and ethical considerations, the trained models and associated scripts were uploaded to a Git-based repository. This ensured that the wider research community can access, reuse, and further develop the model, in the same spirit as the publicly available pre-trained models (e.g., Google BERT) that enabled this work. Providing open access to the trained system promotes transparency, reproducibility, and collective progress, while upholding ethical standards for data protection.

5.10 Ethical considerations

Ethical principles, particularly those related to data privacy, anonymization, and responsible use, were upheld throughout the study. Prior to any data processing, a formal data-sharing agreement was obtained from the Ministry of Health. This agreement clearly outlined the conditions for data access, use limitations, data protection protocols, and compliance with ethical research standards.

Approval to conduct this study was obtained from the Makerere University School of Public Health Institutional Review Board (IRB). The IRB reviewed the study's objectives, methodology, and ethical safeguards, including data anonymization and management procedures, to ensure adherence to national

and institutional ethical guidelines. To protect the confidentiality of individuals, all SMS data used in the study were thoroughly anonymized. Direct identifiers such as names, phone numbers, and reporter metadata were removed prior to analysis. Indirect identifiers, including specific location names or exact timestamps that could allow re-identification, were generalized or masked. Each message was assigned a unique anonymized code to enable structured analysis while preserving privacy.

The anonymized dataset was used exclusively for research purposes. Access was restricted to authorized members of the research team, and all data were stored in encrypted environments with access logs maintained to track usage. No data were shared with unauthorized third parties.

Upon conclusion of the study, the dataset was securely archived on a restricted-access Google Drive folder in line with Makerere University and Ministry of Health data retention policies. These safeguards ensured that the study complied with ethical best practices in digital health research, data protection, and public health surveillance.

CHAPTER SIX
RESULTS

This chapter presents the findings from the development, evaluation, and analysis of the NLP model designed to extract key entities from community-generated eIDSR SMS messages. Quantitative results from model training and testing are reported alongside a comparative analysis with manual extraction using McNemar's test. Additional insights were drawn from interviews with key informants including surveillance personnel, data analysts and informatics experts to contextualize the results and assess real-world applicability.

6.1 Factors Influencing the Accuracy of Key Information Extraction

This section presents findings related to the factors affecting how accurately disease type, location, and symptoms were extracted from SMS messages in the eIDSR system. The key results demonstrated that extraction accuracy is significantly challenged by a dual problem: high linguistic complexity within the messages (due to informal language, abbreviations, and code-mixing) and critical inconsistencies in the current manual workflow and operational context. These findings established the necessity of a fine-tuned, robust NLP model complemented by a Human-in-the-Loop verification system.

6.1.1 Qualitative Findings from Key Informants Interviews

Insights from ten key informants, including eIDSR focal persons, district surveillance officers, and data analysts, revealed several contextual and operational factors affecting the accuracy of information extraction from SMS messages in the eIDSR system. Table 1 summarizes their demographic and professional characteristics.

Informant ID	Age	Sex	Cadre/Role	Institution Type	Experience with eIDSR
KI-01	28	Female	Data Analyst	Ministry of Health	2 years
KI-02	30	Male	Data Analyst	Implementing Partner (NGO)	3 years
KI-03	35	Female	Data Analyst	Ministry of Health	4 years
KI-04	32	Male	Data Analyst	District Health Office	3 years
KI-05	29	Female	District Surveillance Officer	District Health Office	2 years

KI-06	38	Male	District Surveillance Officer	District Health Office	6 years
KI-07	36	Female	Technical Advisor – Surveillance	Implementing Partner (NGO)	4 years
KI-08	40	Male	Epidemiologist	Ministry of Health	7 years
KI-09	33	Female	Public Health Emergency Officer	Ministry of Health	4 years
KI-10	37	Male	Health Informatics Specialist	Implementing Partner (NGO)	5 years

Table 1: showing key informants summary

From the surveillance perspective, the current workflow depends heavily on human verification due to the inconsistent quality of incoming SMS alerts. Many messages lack essential details such as location or symptoms, contain ambiguous place names, or are sent during simulations rather than actual outbreak events. One surveillance officer remarked, *"Some messages come with just the disease name and no location, you cannot forward that to the district without first confirming,"* underscoring the burden placed on manual triage processes.

Several informants emphasized the lack of a standardized structure in SMS submissions. Reporters, often community health workers or members of the public, used free-form text, frequently including abbreviations, local language, or slang. This made automatic extraction of disease names, symptoms, and locations difficult. A data analyst observed, *"People use short forms like 'fev, vom' instead of 'fever and vomiting' which the system may miss."* Others noted that age and gender, while expected, were not always included, leading to incomplete reports.

In addition to message structure, informants pointed out broader operational challenges. These included a lack of enforced guidelines, varied literacy levels among reporters, and inconsistent mobile network coverage affecting real-time reporting. Despite these obstacles, there was cautious optimism about the potential for automation. One participant suggested,

"If a message has all the key details - disease, symptoms, location, age, gender, the model can pick and forward it fast. The ones missing information can be flagged for review."

This reinforced the value of a hybrid triage approach, where NLP accelerates the processing of complete reports while flagging incomplete ones for human review.

On the technical side Informatics experts highlighted system-level limitations within the DHIS2-based eIDSR platform, specifically Cross-Origin Resource Sharing (CORS) restrictions and limited external

API access. The system also uses specific metadata conventions, which external models would need to adhere to. Given these constraints, informants recommended building the NLP model externally using Python and exposing its functionality via a lightweight API. This would allow the model to process SMS data, structure it, and return usable outputs in a format compatible with the existing system workflows. As one informant explained, *"DHIS2 can't just receive raw Python outputs, you need to convert and adapt based on how it names and organizes its data."*

These perspectives validated the study's focus and confirmed the real-world relevance of the model's design. The findings emphasized that while automated systems hold great promise, their success depends on both improving message structure and ensuring compatibility with existing surveillance infrastructure.

6.1.2 Entity Distribution in Annotated SMS Messages

From an initial 8,029 SMS alerts recorded in the eIDSR system, 3,586 test messages, 31 blank messages, and 252 non-human case reports were excluded. Of the remaining alerts, 2,829 were removed for not referencing priority diseases identified in the study criteria. This process resulted in a final sample of 1,331 eligible SMS alerts, which formed the basis for model training and evaluation.

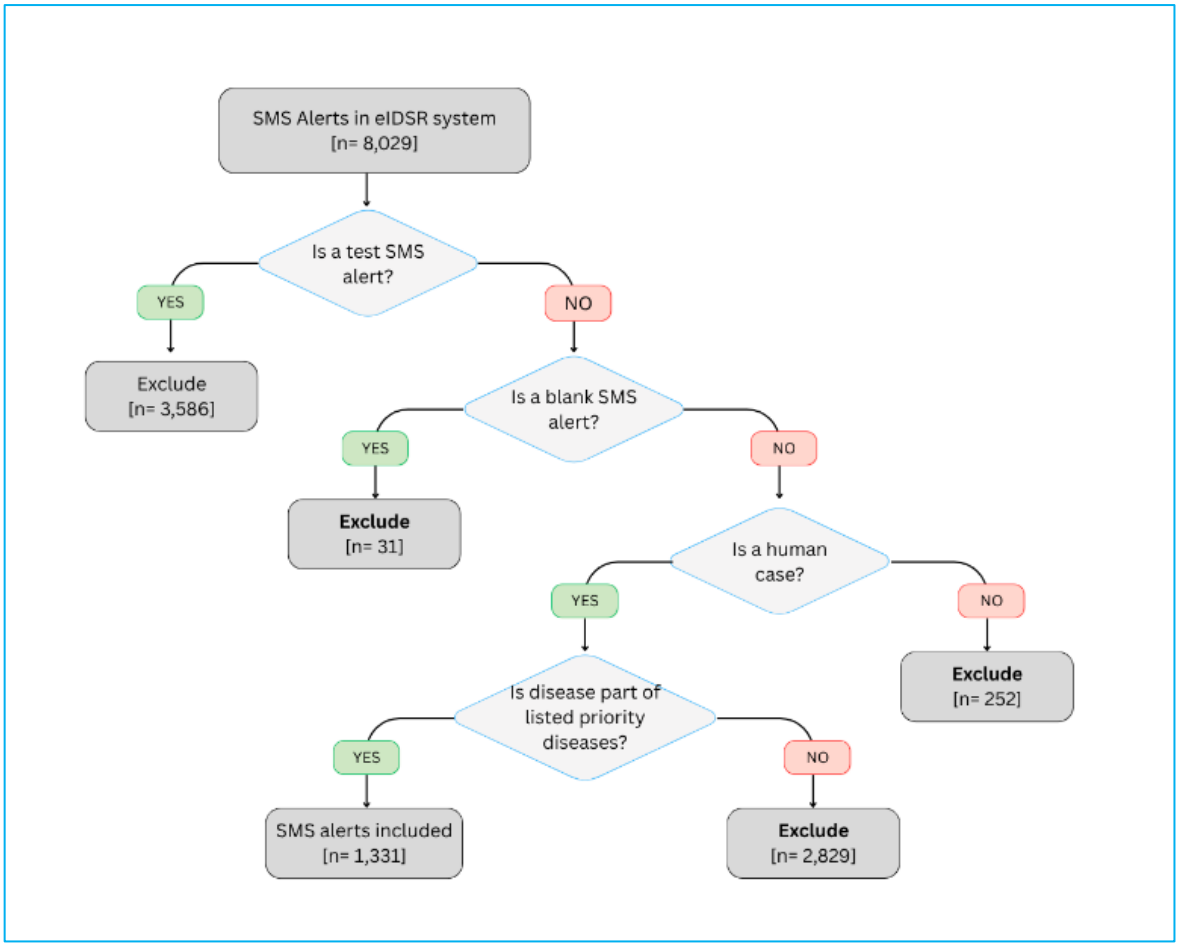

Figure 3 stepwise selection process used to derive the final dataset of SMS alerts included in the study

A total of 3,259 tokens were annotated across 1,331 SMS messages after pre-processing. Of these, 2,483 tokens (76.20%) were labeled as non-entity (O). Among the entity classes, symptoms (B-SYMPTOM and I-SYMPTOM) accounted for 324 tokens (9.94%), followed by diseases (B-DISEASE and I-DISEASE) with 188 tokens (5.76%). Mentions of districts totaled 103 tokens (3.16%), gender 115 tokens (3.53%), and age 46 tokens (1.41%).

Entity Class	Token Count	Percentage of Total Tokens
B-AGE	46	1.46%
B-DISEASE	187	5.93%
B-DISTRICT	103	3.26%
B-GENDER	115	3.65%
B-SYMPTOM	248	7.86%
I-DISEASE	1	0.03%
I-SYMPTOM	76	2.41%

O	2483	78.40%

Table 2: summarizing the total number of entities annotated in the dataset per class.

Beyond entity counts, qualitative inspection of the annotated SMS dataset revealed linguistic challenges that complicate automated processing including informal phrasing, inconsistent structure, and code-mixed language—making it a strong benchmark for evaluating the NLP model's ability to generalize in low-resource surveillance environments.

Challenge	**Example from annotated SMS data**
Informal phrasing	"person very sick wit h fever and bad cough" "Alert 42 suspected with MPOX with skin rash started 2weeks ,***difficuilty*** in swallowing and ***lympnhnodes***"
Inconsistent structure	Some messages list disease first ("cholera case in Arua"), others symptoms first ("vomiting and diarrhea in 2 children"), while others mix both
Code-mixed language	"***omwana alina fever ne cough***" (Luganda + English)

Table 3: examples of linguistic challenges

6.2: To develop a Natural Language Processing (NLP) model based on these identified factors

The second objective was to develop a Natural Language Processing (NLP) model for entity extraction based on the linguistic factors identified in the annotated dataset. The linguistic challenges identified in the data necessitated the use of a BERT-NER architecture. The model was trained for a fixed 12 epochs, and the entire training process lasted approximately 12 minutes and 17 seconds.

6.2.1 Model Training and Convergence

Validation metrics (precision, recall, F1-score, and loss) were tracked at the end of each epoch to assess the model's progression and generalization ability.

Convergence Metrics

The model exhibited strong early learning behavior, with substantial improvements occurring within the first three to five epochs. Training loss dropped sharply from 0.52 at epoch 1 to less than 0.01 by epoch 12, confirming effective optimization. Validation loss also decreased overall, stabilizing between 0.15 and 0.21 in later epochs, which suggested effective generalization without signs of severe overfitting.

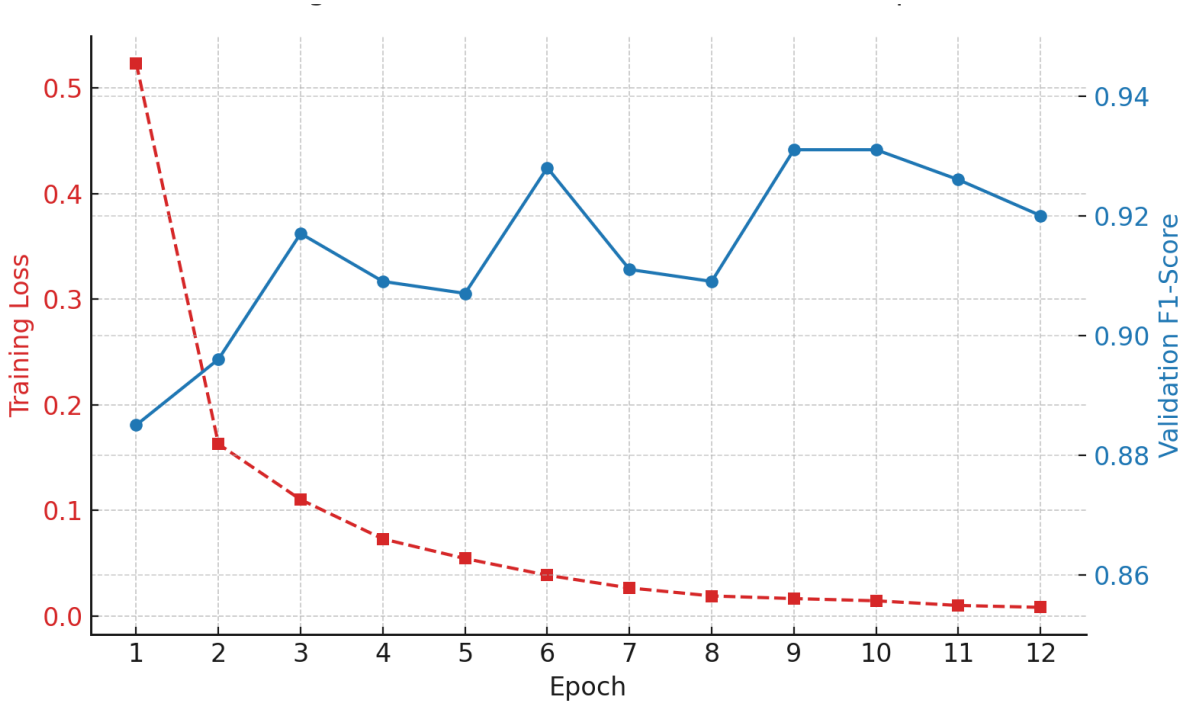

Figure 4 showing training loss and validation F1-score across epochs

6.1.3 Temporal Distribution of Alerts

The results indicated that while 746 messages (56%) were updated on the same day as the follow-up, the remaining alerts experienced significant delays. The average time for an alert to transition to a final status update was 10 days and 11 hours. A small number of outlier records showed delays of up to 350 days, typically occurring when alerts from previous periods were closed during administrative reviews. This 10-day average lag underscored the "time" gap in manual surveillance, providing a clear benchmark for the NLP model's performance, which can process similar volumes in seconds.

Temporal Metric	Value
Total Annotated Sample Size	1,331
Same-Day Updates	746 (56%)
Mean Time to Status Update	10 Days, 11 Hours
Minimum Lag (Fastest Response)	< 1 Day
Maximum Lag (System Outlier)	350 Days

6.2.2 Training Progress and Checkpoint Selection

The validation F1-scores improved rapidly in the early stages, reflecting a quick balance between precision and recall. The highest validation F1-score was observed at epoch 9 (0.931), with subsequent

strong performance at epoch 10 (0.931) and epoch 11 (0.926). The model was trained for a fixed 12 epochs without early stopping, as performance stabilized rather than degraded. The checkpoint corresponding to epoch 9 was ultimately selected for the final test-set evaluation due to its optimal balance of performance and generalization.

Epoch	Validation Loss	Precision (PPV)	Recall	F1-Score
1	0.183	0.846	0.928	0.885
2	0.152	0.863	0.932	0.896
3	0.137	0.896	0.939	0.917
4	0.144	0.878	0.943	0.909
5	0.175	0.894	0.920	0.907
6	0.152	0.918	0.938	0.928
7	0.191	0.903	0.919	0.911
8	0.215	0.896	0.923	0.909
9	0.189	0.917	0.947	0.931
10	0.189	0.920	0.943	0.931
11	0.201	0.912	0.940	0.926
12	0.211	0.906	0.935	0.920

Table 4 Validation performance across training epochs.

Early stopping was not applied during model training. Instead, the model was trained for a fixed 12 epochs. This decision was informed by the consistent performance trends observed throughout training. The validation F1-score improved steadily in the early stages, rising from 0.885 at epoch 1 to a peak of 0.931 at epochs 9 and 10. Performance then stabilized, with only minor fluctuations between 0.920 and 0.931 across the remaining epochs. Since there was no indication of overfitting, completing the full training schedule was justified.

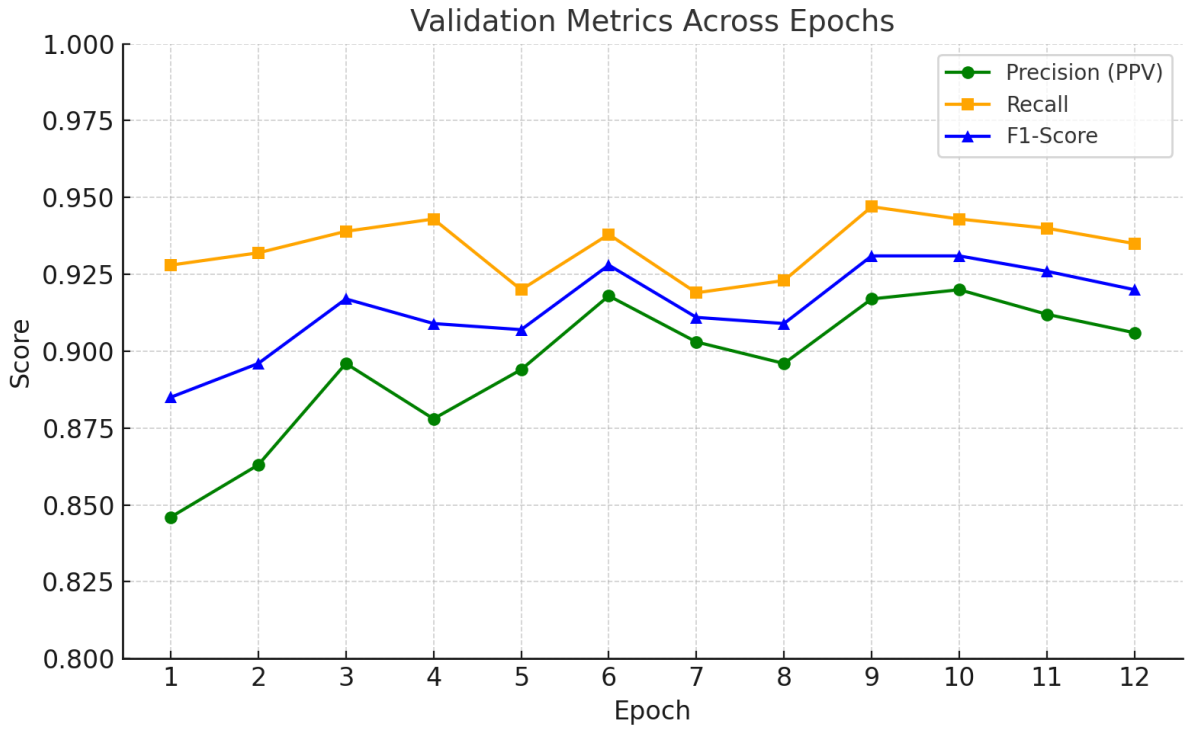

Figure 5 showing precision, recall, and F1-score trends across epochs

Figure above presents the precision, recall, and F1-score trends across epochs, highlighting early rapid gains and stable high performance in later epochs. These results confirmed smooth convergence, absence of overfitting, and consistent generalization across the training process.

6.2.3 Comparative Results of Adjusted Model States

To identify the most robust configuration for operational use, the model's performance was analyzed across different training states (adjusted epochs). This allowed for the identification of the point where the model was best adjusted to handle the linguistic nuances of the SMS data without suffering from overfitting.

Model State (Adjustment)	Precision	Recall	F1-Score	Status
Initial Adjustment (Epoch 1)	0.846	0.928	0.885	Underfitted rare classes
Intermediate Adjustment (Epoch 5)	0.894	0.920	0.907	Improving stability
Optimal Adjustment (Epoch 9)	**0.917**	**0.947**	**0.931**	**Selected for Test set**
Final Adjustment (Epoch 12)	0.906	0.935	0.920	Diminishing returns

Table 5Comparison of Model Performance across Training Adjustments

The comparative results showed that early adjustments (Epoch 1) resulted in high recall but lower precision, suggesting the model was identifying broad patterns but generating false positives. As the model parameters were adjusted through Epoch 9, precision improved by 7.1% while maintaining peak recall at 94.7%, effectively bridging the performance gap for rare entity classes such as diseases and locations.

6.3 To evaluate the performance of the NLP-powered system compared to manual approach

6.3.1 Final Test Set Performance

The final model checkpoint selected at epoch 9 (validation F1 = 0.931) was deployed on the independent test dataset. This epoch was chosen because it represented the peak validation F1-score, with stable precision and recall, ensuring the best balance between sensitivity and specificity. The test-set results confirmed robust generalization to unseen data, achieving an overall F1-score of 0.926 and a precision/recall balance that prioritized high recall (sensitivity).

Metric	Score
Test Loss	0.165
Precision (PPV)	0.911 (91.08%)
Recall	0.942 (94.21%)
F1-Score	0.926 (92.62%)
Runtime (s)	3.34
Samples/sec	49.74

Table 6: Final model performance on the test set.6.3.2 Per-Class Performance MetricsThe fine-tuned model achieved strong performance across most entity classes. Overall, the model reached a weighted average precision of 0.96, recall of 0.96, and F1-score of 0.96, demonstrating consistent accuracy across the majority of categories.

Class	Precision	Recall	F1-Score	Support
B-AGE	0.958	1.000	0.979	46
B-DISEASE	0.989	0.995	0.992	188
B-DISTRICT	0.936	1.000	0.967	103
B-GENDER	1.000	0.966	0.983	119
B-SYMPTOM	0.886	0.889	0.887	279

I-DISEASE	1.000	1.000	1.000	1
I-SYMPTOM	0.650	0.576	0.610	132
O	0.983	0.986	0.984	2,519

Table 7: Per-Class Performance Metrics

6.3.3 Token-Level Performance and Error Analysis

The token-level confusion matrix offered valuable insights into how well the model distinguished between different named entity classes. Overall, the model performed exceptionally well on structured and easily identifiable entities such as AGE, GENDER, DISEASE, and DISTRICT, while facing more challenges with the more variable SYMPTOM and I-SYMPTOM classes.

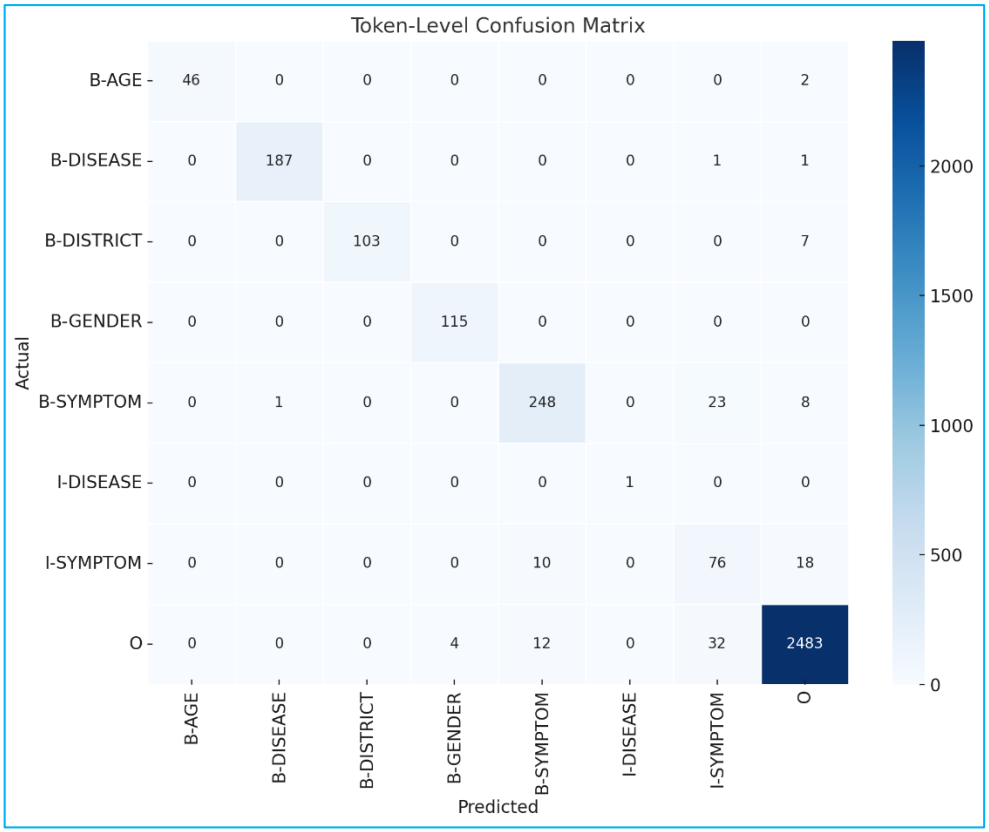

Figure 6: Token-Level Confusion Matrix

B-AGE tokens were perfectly classified (F1 = 0.98), and B-DISEASE performed very well (F1 = 0.99), with minimal confusion.

I-SYMPTOM showed the lowest performance F1= 0.61 due to higher confusion with other classes. Specifically, 32 tokens were misclassified as non-entity (O) and 23 as B-SYMPTOM. This confirms that accurately segmenting multi-word symptom phrases remains the model's greatest challenge, consistent with the linguistic factors identified

6.3.4 Comparative Evaluation and Operational Efficiency

Inference Speed and Operational Efficiency

In real-time deployment scenarios, response time is critical. A comparison of processing speeds between manual review and the NLP model is presented below.

Method	Average Time per Message	Throughput (messages/sec)
Manual Review	2–3 minutes	~0.006–0.008
NLP Model	0.021 seconds	48

Table 8: Comparison of processing efficiency between manual review and NLP model.

While manual review of an SMS message typically required 2–3 minutes per message, the NLP model processed an average of 48 messages per second. This demonstrates a substantial improvement in operational efficiency.

McNemar Test Results

To evaluate the model's performance relative to human accuracy, a head-to-head comparison was conducted on a subset of 1,331 SMS messages containing valid epidemiological information. For each message, we recorded whether the model and the manual (baseline) method correctly extracted all key entities (age, disease, symptoms, and district).

The results are summarized as follows:

Correct by both model and manual method	1,227 messages	92.2%
Correct by manual only	104 messages	7.8%
Correct by model only	0 messages	0.0%

Table 9 showing McNemar Test Results

The analysis of the performance gap between the two methods focused on the 104 discordant cases where the manual approach successfully extracted entities that the model missed. While this indicates that human review remains superior for the most complex or irregular messages, the model achieved direct agreement with the manual method in 92% of the cases. This high level of concordance demonstrates that the model can reliably automate the vast majority of SMS-based health reports, while the 8% disagreement rate provides the empirical basis for triaging specific "discordant-style" alerts for human oversight.

6.3.3 Real-World Demonstration and System Deployment

To validate the practical feasibility of this model in a real-world public health context, a demonstration system was developed to simulate its integration into Uganda's electronic Integrated Disease Surveillance and Response (eIDSR) framework. This system includes a functional backend API, a

user-friendly interface for testing SMS inputs, and a proposed architecture for mapping model outputs to the national Event-Based Surveillance (EBS) system.

Distribution of Diseases Extracted by the NLP Model

To assess the practical utility of the model in a surveillance context, the distribution of diseases extracted from the SMS dataset was analyzed. The model successfully identified a range of priority zoonotic and infectious diseases, with Mpox emerging as the most frequently detected condition (n = 1,198), reflecting the specific epidemiological landscape of Uganda during the 2024 study period. Other identified threats included Rabies (n = 47), Anthrax (n = 18), and Yellow Fever (n = 7), as well as rare but high-consequence pathogens such as Brucellosis (n = 4) and Ebola (n = 1).

Disease	Extracted
Mpox	1198
Yellow fever	7
Rabies	47
Anthrax	18
Brucellosis	4
Ebola	1
Forwarded for Human Involvement	56

Table 10: Frequency of Disease Entities Extracted by the NLP Model

Crucially, the system flagged 56 alerts as "Forwarded for Human Involvement." These cases represent messages where the model identified high ambiguity or missing critical entities (such as location or symptoms), automatically triggering the human-in-the-loop protocol. This distribution confirmed that the model is capable of both high-volume routine extraction and the selective triaging of complex cases for expert review.

Web-Based NLP API Demo

A lightweight backend inference API was built using Python and Flask to expose the trained BERT-based NLP model.

For example, when the following SMS message was entered:

"ALERT feMALE 24 YEARs OLD mpox WITH GENERALIZED SKIN RASH, IN MAKERERE UNIVERSITY HOSPITAL, KAMPALA CENTRAL, BY MOSES."

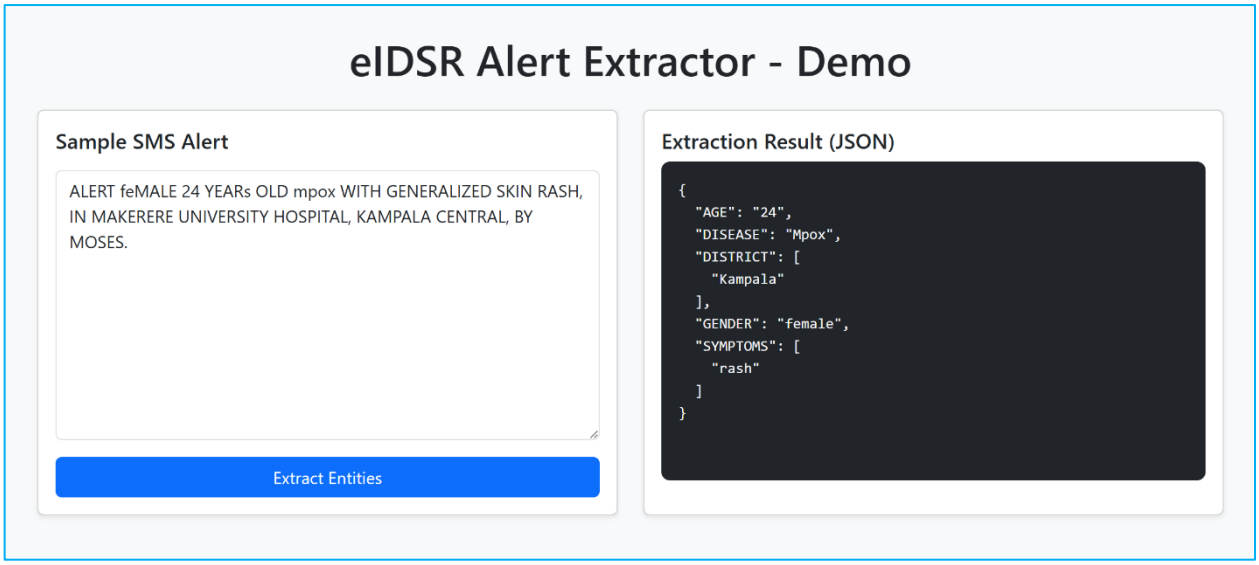

Figure 7: showing the results of the model extraction

This successful inference demonstrated the model's robustness in handling informal, unstructured, and code-mixed input. It also validates the model's readiness for operational deployment, highlighting how it can facilitate faster triaging and reduce analyst workload in the current eIDSR workflow.

This demo showcased the model's readiness for deployment, with accurate predictions even on unstructured, informal, and mixed-format input, typical of community-submitted health alerts.

The demo of the eIDSR Alert Extractor illustrated how the human-in-the-loop approach can be operationalized. In cases where the model successfully identified entities such as disease, symptoms, age, and gender but failed to detect a valid district, the system automatically generated a structured alert prompting manual review (e.g., *"District not identified – Forward for manual follow-up"*).

Figure 8: Prototype demo of the eIDSR Alert Extractor showing a human-in-the-loop workflow

This ensured that critical epidemiological information is not lost due to incomplete extraction and that uncertain outputs are escalated to health analysts for verification. By combining automated entity recognition with targeted human oversight, the workflow preserved timeliness while maintaining accountability and accuracy in surveillance reporting.

CHAPTER SEVEN
DISCUSSION

This study was designed to address a pressing operational challenge in Uganda's electronic Integrated Disease Surveillance and Response (eIDSR) system, the slow and manual extraction of critical epidemiologic information from SMS-based community alerts. The objective was to develop and evaluate a lightweight Named Entity Recognition (NER) model that could automatically extract key entities from noisy, unstructured text. This work bridges the gap between digital epidemiology and operational surveillance, with the aim of accelerating outbreak detection and improving alignment with the WHO's 7-1-7 timeliness goals.

7.1 Interpretation of Key Results

A fine-tuned BERT-base uncased model was trained on 1,331 annotated SMS messages. On the held-out test set, the model achieved a precision of 91.1%, recall of 94.2%, and an F1-score of 92.6%, with an average inference speed of 48 SMS per second. These findings are consistent with prior research in health informatics, where transformer-based models have demonstrated strong performance on short clinical or public-health texts (Cho et al., 2024), (Denecke et al., 2024). Importantly, this performance was achieved despite working with messages that were linguistically inconsistent, sparsely punctuated, and highly variable in length and structure.

7.1.1 Discussion on Factors Influencing the Accuracy of Key Information Extraction

Entity types that were lexically stable, such as AGE, GENDER, DISTRICT, and DISEASE, were identified with high reliability (F1 approx 0.95 or higher). This result confirmed that high-accuracy automated extraction is readily achievable for standardized information. These stable entities benefited significantly from either having a fixed, small vocabulary (e.g., gender, district names) or a highly predictable semantic structure (e.g., age tokens like '4yrs', '2 years'). This structural regularity enabled the contextual model to generalize effectively, even when faced with minor orthographic differences. Furthermore, the near-perfect accuracy achieved for AGE and GENDER directly validated the strategy of applying lightweight pre-processing and normalization rules (e.g., converting "4yrs," "four years," and "4yo" to a common form).

This essential pre-processing step successfully reduced superficial variability, allowing the deep learning model to focus its sophisticated contextual decoding capacity on identifying the underlying semantic intent rather than being confused by surface-level token variations. The rarity of I-DISEASE

tokens, indicating that most diseases in SMS alerts are mentioned as single words, further simplified the extraction task for this stable entity class.

By contrast, the Symptoms entity emerged as the most error-prone. In particular, I-SYMPTOM tokens (representing the continuation of a multi-word symptom phrase) recorded the lowest performance, with an F1-score of 0.61. This low score is consistent with established literature, which shows that entities requiring the extraction of nuanced meaning from unstructured, colloquial, or patient-generated text are inherently challenging (Babaian & Xu, 2024).

Studies focusing on the extraction of medical concepts from noisy data often report symptom-related F1-scores in the lower range of 0.60 to 0.85. The model's 0.61 score therefore falls within the expected lower-bound range for this difficult task in the field of NER from noisy clinical text (Shafran et al., 2020). The specific confusion observed in the error analysis, where I-SYMPTOM tokens were often misclassified as O (non-entity) or the start of a new symptom (B-SYMPTOM), revealed the challenge was one of span boundary detection. The flexible, multi-word, and often abbreviated nature of symptom expressions in SMS lacked the clear, predictable syntactic boundaries necessary for the model to precisely demarcate the entity boundaries.

This type of boundary error is a known limitation of current NER approaches when domain-specific data is highly irregular. Nonetheless, the high rate of error was concentrated on a small set of recurring span errors and common misspellings, suggesting that the model’s performance can be substantially improved by implementing targeted post-processing heuristics.

The varying performance across entity types, particularly the error-proneness of symptoms, provided critical empirical evidence that informs the design of the final surveillance system. The inability of the purely automated model to achieve perfect accuracy, especially in boundary detection and handling incomplete messages, validates the necessity of a Human-in-the-Loop workflow.

Interviews with surveillance staff confirmed that this hybrid approach aligns with operational realities. One informant highlighted the necessity of human verification when core entities are missing:

"Some messages come with just the disease name and no location, you cannot forward that to the district without first confirming."

This underscores the need for human judgment to manage messages that are structurally incomplete or ambiguous. Despite these challenges, surveillance staff recognized the system's potential to accelerate their workload, stating:

"If a message has all the key details—disease, symptoms, location, age, gender—the model can pick and forward it fast. The ones missing information can be flagged for review."

This expert validation suggested that the NER model's role is not to replace human oversight, but rather to serve as a high-speed pre-processor. By extracting stable entities with high accuracy and flagging variable or incomplete messages for review, the system accelerates routine triage while ensuring critical human verification is applied to the small subset of cases where model accuracy may be compromised.

The discovery of an average 10-day and 11-hour lag between signal follow-up and status update in the manual workflow revealed a significant bottleneck in the current surveillance timeline. While 56% of the 1,331 messages were updated on the same day, the presence of a "long tail" of delays, some reaching up to 350 days, indicated that manual tracking is inconsistent and prone to administrative delays. By integrating an NLP model capable of near-instantaneous extraction and triage, the time required to move from a raw SMS "signal" to an organized "data point" is reduced to seconds. This improvement is critical for meeting the WHO 7-1-7 framework targets, as reducing the processing lag directly accelerates the time to detection and official notification of potential outbreaks.

7.2 Model Training and Finetuning

Several design choices supported the model's effectiveness. First, the scope was deliberately limited to five epidemiologically relevant entities—disease, symptoms, age, gender, and district—which prevented the model from being stretched too thin and allowed specialization on the most operationally important variables. Second, lightweight normalization rules (e.g., standardizing shorthand forms such as *"yrs," "yo," "f," "m"*) reduced superficial variability and improved the model's ability to capture underlying patterns. Third, training on authentic SMS data, rather than synthetic examples, provided ecological validity by ensuring the model learned from the same challenges, misspellings, abbreviations, code-mixing, and irregular syntax, that characterize actual alerts.

The final model's design incorporated three critical choices to directly mitigate the identified data challenges:

First, the scope was deliberately limited to five epidemiologically relevant entities, disease, symptoms, age, gender, and district. This focused approach allowed the model to specialize its training capacity on the variables most critical to the surveillance mission, rather than dispersing effort on less relevant or more complex linguistic structures. These methodological choices echo findings from previous studies in health informatics, where transformer-based models have achieved high performance on noisy, short-form clinical or public health texts (Cho et al., 2024; Denecke et al., 2024). By tailoring model development to local reporting realities, this study demonstrated the feasibility of adapting state-of-the-art NLP techniques to a low-resource, SMS-driven surveillance setting.

Second, lightweight normalization rules (e.g., standardizing shorthand forms such as “yrs,” “yo,” “f,” “m”) were integrated. This intervention was a direct response to the high variability observed in AGE and GENDER entities, and the model's high performance on these classes later confirmed that this pre-processing step successfully guided the model toward underlying semantic patterns. The rarity of I-DISEASE tokens indicated that most diseases in SMS alerts are mentioned as single words, simplifying the extraction task and shifting the main challenge toward improving symptom and location recognition.

Third, the decision to train exclusively on authentic SMS data provided crucial ecological validity. This ensured the model learned to process the exact challenges, misspellings, abbreviations, code-mixing, and irregular syntax, that characterize actual alerts, a methodological approach vital for achieving reliable performance in low-resource surveillance settings. This tailored development, which specialized the model on the local reporting realities and minimized the impact of predictable noise, ultimately demonstrated the feasibility of adapting state-of-the-art NLP techniques to this challenging, SMS-driven environment.

Additional insights strengthened confidence in the model’s robustness. Age and gender entities were recognized with near-perfect accuracy despite inconsistent formats, confirming the effectiveness of lightweight pre-processing.

7.3 Model Performance

On the independent test set, the final model achieved a precision of 91.1%, recall of 94.2%, and an F1-score of 92.6%. This performance is highly competitive and consistent with existing research in health informatics, where transformer-based models have demonstrated strong performance of (often F1 > 0.90) when applied to short, domain-specific clinical or public-health texts (Cho et al., 2024; Denecke et al., 2024). Achieving this level of accuracy, despite working with the unique challenges of

linguistically inconsistent, noisy, and unstructured SMS data from a low-resource setting, validates the adaptability of state-of-the-art NLP techniques.

The overall performance of the model, which achieved a 92.6% F1-score, must be contextualized against the mission requirements of the electronic Integrated Disease Surveillance and Response (eIDSR) system. In the context of Event-Based Surveillance (EBS) and the globally recognized WHO 7-1-7 framework, the primary operational standard is the maximization of system sensitivity to minimize the risk of missing critical outbreak alerts (False Negatives). This means that Recall is the dominant metric, with an acceptable operational threshold typically set conservatively at 90% or higher to ensure robust detection. The model's Recall of 94.2% therefore successfully exceeds this operational threshold, confirming the system's high ability to capture the vast majority of disease signals.

The Precision of 91.1% indicated that only 8.9% of the entities identified were False Positives, and the small 3.1 percentage point gap between Recall and Precision confirms a deliberate design bias toward prioritizing sensitivity over absolute specificity—a strategic trade-off justified by the critical need to avoid missed outbreaks. Consequently, the 92.6% F1-score represents a robust and operationally reliable balance between the two metrics. The high scores validate the system's capacity to serve as a high-throughput, initial triage mechanism, allowing surveillance staff to shift focus from tedious manual extraction to critical verification and rapid response.

Comparison of Model Performance vs. Manual Baseline

A direct comparison between the NLP model's output and the manual gold-standard baseline was conducted to contextualize the model's real-world accuracy. Within the dataset of 1,331 messages, a high level of concordance was observed in 92% of cases, where the model and the manual method reached identical extraction results.

The analysis specifically examined the 104 discordant cases (8%) where the methods diverged. In these instances, the manual approach correctly identified entities that the model either missed or misclassified, while there were no cases where the model outperformed the human expert. Rather than indicating a system failure, this 8% margin identifies the specific threshold of linguistic complexity where human intervention is required.

However, this result does not diminish the model's operational value; rather, it provides the empirical evidence to justify the Human-in-the-Loop workflow. The model's high recall and speed suggested that it can reliably support outbreak detection when embedded in a HITL workflow. While the manual approach maintains an edge in handling the most complex inputs, the model's consistent accuracy

across most cases positions it as a promising supportive tool. In this role, the system accelerates routine triage while leaving uncertain cases to analysts, complementing rather than replacing human judgment. With further tuning, particularly around symptom expression variability and linguistic nuances—the primary driver of error—the model could narrow the performance gap, ultimately offering comparable accuracy with significantly faster processing.

Beyond accuracy, operational efficiency was a major strength. The model processed approximately 48 SMS messages per second, in sharp contrast to the 2–3 minutes per message required for manual review. This represents a >99% time savings and positions the system as a real-time pre-processing engine for national surveillance. Such efficiency gains are consistent with prior evaluations of NLP-based surveillance systems, which have shown that automation can reduce reporting delays and improve timeliness of outbreak detection (Baclic et al., 2020, Methuku, 2025).

Interpretation of Disease Distribution and the Role of Human-in-the-Loop Triaging

The distribution of diseases identified by the model underscored its operational relevance as a real-time situational awareness tool. The high frequency of Mpox (n = 1,198) extractions was significant because it demonstrated the model's ability to remain stable and performant during a public health emergency. In a manual system, a sudden surge of over a thousand alerts can lead to "notification fatigue" and data entry backlogs; however, the model's ability to categorize these high-volume alerts instantaneously ensured that the epidemiological curve was updated in real-time. This suggests that the NLP system is particularly valuable for scaling surveillance efforts during active outbreaks without requiring a proportional increase in human data-entry staff.

Furthermore, the successful extraction of rare but high-consequence pathogens such as Ebola, Anthrax, and Yellow Fever, represents the "needle in the haystack" capability required for effective Early Warning and Response. The importance here lies in the model's sensitivity; even a single missed case of Ebola can lead to a catastrophic outbreak. By accurately flagging these rare diseases, the model proved it is not merely a tool for routine data processing but a critical component of national health security.

The 56 alerts forwarded for Human Involvement represented the most vital safety feature of the proposed architecture. This result indicates that the model is "risk-aware", recognizing its own limitations when faced with highly ambiguous or incomplete SMS data. In a public health context, an incorrect automated decision (a false negative) is more dangerous than a delay for human review. By triaging these 56 cases, the system ensured that human experts can focus their limited time on the most complex 4% of alerts, while the model handles the 96% of routine extractions. This hybrid approach

addresses the ethical and safety concerns of using AI in healthcare, establishing a "Human-in-the-Loop" standard that preserves human oversight while maximizing automated efficiency.

7.3 Strengths and limitations

This study demonstrated several important strengths, both methodological and practical, that enhance the validity and real-world applicability of its findings.

7.3.1 Study Strengths

The study was grounded in real operational needs and used real-world data from Uganda's national eIDSR system, ensuring high ecological validity. The model was trained on actual community-submitted SMS messages, rather than synthetically constructed examples, which gave it exposure to the kind of linguistic noise, variation, and irregularity it would face in production. This training approach increased the likelihood of robust performance in a real-world setting.

The model achieved high performance across multiple metrics, most notably an F1-score of 92.6%, recall of 94.2%, and precision of 91.1%, demonstrating its capacity to accurately identify key entities even in short and unstructured text. Importantly, inference speeds averaged 48 messages per second, far exceeding the minimum threshold of 10 messages per second as it ensures scalability to thousands of SMS alerts per day (well within typical national surveillance workloads).

The full pipeline, from annotation and pre-processing to training, evaluation, and deployment via a web-based API, was designed to align with realistic implementation constraints. The model's architecture (BERT-base uncased) and tooling (Flask backend) were selected for ease of deployment in low-resource settings. This made the solution both technically sound and practically feasible.

Furthermore, the study contributed a custom, domain-specific annotated corpus of 1,331 messages, which fills a critical gap in training data for public health NLP in resource-constrained environments. This dataset, along with the token-level confusion matrix and qualitative error analysis, provided a rich foundation for understanding model behavior and planning future improvements.

7.3.2 Study Limitations

The robustness and generalizability of the NLP pipeline were influenced by several methodological constraints.

The dataset used for training, while carefully annotated, was relatively small and reflected an uneven distribution of entity types. Common symptoms and diseases were well represented, but rarer presentations occurred infrequently, increasing the risk of bias toward more typical patterns. This

imbalance may have limited the model's ability to generalize to less frequent but equally important scenarios.

Labeling quality, though supported by detailed guidelines, was also subject to occasional inconsistency. Annotating informal, unstructured text is inherently subjective, especially for symptoms that span multiple tokens or use ambiguous phrasing. As a result, some inter-annotator disagreement introduced label noise, potentially affecting model performance. Strengthening annotation through clearer guidance, periodic calibration, and consensus resolution processes would help improve consistency in future efforts.

A significant limitation of the study was that the NLP model was developed for static entity extraction rather than temporal predictive analysis. While the research established a processing lag baseline of 10.5 days through metadata analysis, the model architecture itself treats each SMS alert as an independent event. Consequently, the system is capable of identifying the Date of Onset and Reporting Date within a message, but it does not currently possess the capacity to analyze the chronological sequence of alerts to perform automated trend forecasting or outbreak cluster detection. The transition from extracting "time" as an entity to utilizing "time" as a predictive feature remains an area for future development.Another long-term concern is concept drift, the way language and reporting norms evolve over time. Changes in abbreviations, phrasing, or symptom descriptions can gradually erode the model's accuracy if not addressed. Sustaining model performance will require periodic fine-tuning using recent, manually validated data, along with updates to any associated dictionaries and rules.

Although this study strongly advocates for a Human-in-the-Loop workflow as a safeguard for high-stakes surveillance, the model development phase (training and fine-tuning) was conducted using a static, pre-annotated dataset. In this research, human experts provided the "gold standard" labels upfront rather than interacting with the model in real-time to correct misclassifications during training (Active Learning). This means the model did not benefit from an iterative feedback loop where its performance on ambiguous cases could be improved through immediate human intervention. These methodological limitations reflect the realities of working with noisy, real-world data in low-resource settings. Nonetheless, they are manageable through iterative refinement, data augmentation, and practical safeguards that can support the pipeline's long-term adaptation and operational use.

7.4 Conclusion

This study demonstrated that automating the extraction of key epidemiologic entities (disease, symptom, district, age, and gender) from Uganda's eIDSR SMS stream using a fine-tuned BERT-base model is both technically feasible and operationally transformative.

The analysis of results confirmed that the accuracy of key information extraction is highly contingent upon the lexical stability of the target entity. Entity categories with stable, standardized vocabularies such as age, gender, district, and disease names were extracted with consistently high reliability. By contrast, symptom entities proved to be the primary source of error, particularly continuation tokens (I-SYMPTOM), due to the flexible and unstructured nature of symptom descriptions in SMS alerts. This key takeaway highlights that while transformer models are robust, entity stability and reporting consistency strongly influence model performance, indicating that highly variable categories require additional post-processing strategies for refinement.

The development of the Natural Language Processing model was critically informed by these identified linguistic factors. The study demonstrated that careful design choices—including limiting the scope to five epidemiologically relevant entities, applying lightweight rules to normalize shorthand forms, and training exclusively on authentic, noisy SMS data, were critical to the model's success. These methodological considerations enabled the system to effectively adapt to Uganda's challenging, real-world surveillance environment, proving the feasibility of translating state-of-the-art NLP techniques into a viable public health tool.

The NLP-powered system's performance was evaluated based on both accuracy and efficiency compared to the manual approach. The model achieved high overall accuracy with a strong emphasis on system sensitivity (Recall), successfully meeting the operational threshold set by the WHO 7-1-7 timeliness goals. Critically, the system sustained a real-time processing speed of approximately 48 messages per second, virtually eliminating intake bottlenecks and positioning it as a time-saving mechanism for surveillance staff. While direct comparison showed that manual analysts still maintained an advantage in handling a small subset of ambiguous or complex cases (confirmed by McNemar's test), the system's high reliability across the majority of inputs and its remarkable speed underscore its value. This work establishes that by blending the scalability of machine learning with the oversight and contextual judgment of human analysts, the system is best positioned as an effective Human-in-the-Loop tool, accelerating disease surveillance and setting a precedent for adaptive epidemic preparedness across low-resource settings.

Overall, this study establishes a concrete foundation for smarter, faster, and more responsive disease surveillance in Uganda. It demonstrates that NLP can automate a previously manual bottleneck, ando that the factors influencing extraction accuracy must be carefully considered when designing models for real-world surveillance. By blending the precision and scalability of machine learning with the oversight and contextual judgment of human analysts, Uganda can move closer to achieving the

WHO's 7-1-7 surveillance targets, while setting a precedent for adaptive, technology-driven epidemic preparedness across low-resource settings.

7.5 Recommendations

This study should be understood as a proof of concept that demonstrates the feasibility and promise of automated entity extraction from SMS-based disease surveillance data. The model consistently met performance thresholds for precision, recall, and F1-score during validation tests. These outcomes support the viability of using the model to automate the extraction of disease indicators from unstructured SMS data. With access to the national system, future work can build on this foundation to evaluate real-time deployment and measure operational impact at scale.

While the results were encouraging, they also revealed limitations that can be addressed before operational deployment. Based on these findings, the following recommendations are proposed.

Human-in-the-Loop Pilot Evaluation within eIDSR: Before full deployment, the model should be evaluated in a controlled, human-in-the-loop pilot. Running behind the existing SMS gateway, the model can generate structured outputs for review without affecting live decisions. Health analysts would verify or edit outputs, allowing a real-world comparison with the current manual workflow on accuracy, speed, and triage value. This step would provide operational validation, uncover hidden edge cases, and build trust among users and decision-makers.

Symptom Extraction Refinement through Post-Processing: Given that symptoms accounted for most classification errors, the pipeline should integrate a lightweight post-processing layer to refine extraction. This could include the use of a curated symptom dictionary covering common abbreviations and misspellings, basic spell-correction routines, and rules to merge overlapping or redundant symptom spans. Incorporating these enhancements would immediately improve output quality without the need for full model retraining, particularly for SMS messages with informal or inconsistent phrasing.

Expansion and Calibration of Training Data: To further boost performance and model resilience, the training dataset should be expanded and re-annotated to increase representation of rare disease types and diverse districts, while also improving coverage of the wide variations in phrasing, spelling, and SMS formats. At the same time, existing annotation noise should be addressed through double labeling and clearer guidelines, supported by periodic consensus resolution. Routine calibration sessions and regular measurement of inter-annotator agreement will be essential to ensure quality and

consistency in future annotations, thereby strengthening the reliability of the training data and the model built on it.

A critical next step is the development of multilingual extensions of the model to support SMS messages written in local languages and code-switched formats. While this study focused on English-language messages, Uganda's linguistic diversity means that a significant portion of surveillance data may be composed in other languages or mixed-language styles. Leveraging multilingual pre-trained models like mBERT or XLM-RoBERTa, combined with targeted annotation of non-English messages, would expand the system's reach and inclusivity (Choudhury & Deshpande, 2021). This represents a key direction for building a more inclusive and comprehensive national surveillance system.

REFERENCES

Adebara, I., & Abdul-Mageed, M. (2022, January 1). Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go. Cornell University. https://doi.org/10.48550/arXiv.2203.

Adelani, D. I. (2025). Natural language processing for African languages. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2507.00297

Afzal, N., Sohn, S., Abram, S., Liu, H., Kullo, I. J., & Arruda-Olson, A. M. (2016). Identifying peripheral arterial disease cases using natural language processing of clinical notes. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI . . .)*. https://doi.org/10.1109/bhi.2016.7455851

Akera, B., Mukiibi, J., Naggayi, L. S., Babirye, C., Owomugisha, I., Nsumba, S., Nakatumba-Nabende, J., Mwebaze, E., & Quinn, J. (2022). *Machine Translation For African Languages: Community Creation Of Datasets and Models In Uganda*. https://www.semanticscholar.org/paper/M-ACHINE-T-RANSLATION-FOR-A-FRICAN-L-ANGUAGES-%3A-C-C-Akera-Mukiibi/dbca0b5d8fe1c906066951d6935a86dca3de8a50

Babirye, C., Nakatumba-Nabende, J., Jeremy, T. F., Mukiibi, J., Katumba, A., & Ogwang, R. (2022). *Building Text And Speech Datasets For Low Resourced Languages : A Case OF Languages IN East Africa*. https://www.semanticscholar.org/paper/B-UILDING-T-EXT-AND-S-PEECH-D-ATASETS-FOR-L-OW-R-L-Babirye-Nakatumba%E2%80%90Nabende/2dcc000cf43830167fbf91349b6b2399454c8858

Baclic, O., Tunis, M., Young, K., Doan, C., & Swerdfeger, H. (2020a). Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*, 161–168. https://doi.org/10.14745/ccdr.v46i06a02

Bhadauria, D., Sierra-Múnera, A., & Krestel, R. (2024). *The effects of data quality on named entity recognition*. https://www.semanticscholar.org/paper/The-Effects-of-Data-Quality-on-Named-Entity-Bhadauria-Sierra-M%C3%BAnera/b51290b6785f1f65b140c39ead705abe5deac0a3

Bagla, K., Gupta, S., Kumar, A., & Gupta, A. (2023). Noisy Text Data: foible of popular Transformer based NLP models. *ACM Digital Library*, 1–5. https://doi.org/10.1145/3639856.3639889

Biemba, G., Chiluba, B., Yeboah-Antwi, K., Silavwe, V., Lunze, K., Mwale, R. K., Russpatrick, S., & Hamer, D. H. (2017). A Mobile-Based Community Health Management information system for community health workers and their supervisors in 2 districts of Zambia. *Global Health Science and Practice*, *5*(3), 486–494. https://doi.org/10.9745/ghsp-d-16-00275

Bochner, A F., Makumbi, I., Aderinola, O., Abayneh, A., Jetoh, R., Yemanaberhan, R L., Danjuma, J S., Lazaro, F T., Mahmoud, H J., Yeabah, T O., Nakiire, L., Yahaya, A K., Teixeira, R A., Lamorde, M., Nabukenya, I., Oladejo, J., Adetifa, I., Oliveira, W K D., McClelland, A., & Lee, C T. (2023, April 13). Implementation of the 7-1-7 target for detection, notification, and response to public health threats in five countries: a retrospective, observational study. Elsevier BV, 11(6), e871-e879. https://doi.org/10.1016/s2214-109x(23)00133-x

Bosa, H. K., Majwala, R., Nakiire, L., Ario, A., Kiwanuka, N., Kibuuka, H., Downing, R., & Lutwama, J. (2016). Missed opportunities for Yellow Fever Surveillance in Uganda, July 2015 - May 2016. *International Journal of Infectious Diseases*, *53*, 116–117. https://doi.org/10.1016/j.ijid.2016.11.291

Budd, J., Miller, B. S., Manning, E. M., Lampos, V., Zhuang, M., Edelstein, M., Rees, G., Emery, V. C., Stevens, M. M., Keegan, N., Short, M. J., Pillay, D., Manley, E., Cox, I. J., Heymann, D., Johnson, A. M., & McKendry, R. A. (2020). Digital technologies in the public-health response to COVID-19. *Nature Medicine*, *26*(8), 1183–1192. https://doi.org/10.1038/s41591-020-1011-4

Bhutda, A., Sakarkar, G., Shelke, N., Paithankar, K., & Panchal, R. (2024). An AI-based approach to train a language model on a wide range corpus using BERT. *2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)*, 1–6. https://doi.org/10.1109/sceecs61402.2024.10482037

Dahl, S., Bøgsted, M., Sagi, T., & Vesteghem, C. (2025). Performance of natural language processing for information extraction from electronic health records within Cancer: Systematic review. *JMIR Medical Informatics*, *13*, e68707. https://doi.org/10.2196/68707

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North {a}Merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, *1*. https://doi.org/10.18653/v1/n19-1423

DHIS2. (2025, May 7). *DHIS2 software Releases - DHIS2*. https://dhis2.org/releases/

Durango, M. C., Torres-Silva, E. A., & Orozco-Duque, A. (2023). Named Entity Recognition in Electronic Health Records: A Methodological review. *Healthcare Informatics Research*, *29*(4), 286–300. https://doi.org/10.4258/hir.2023.29.4.286

Elhadad, N., & Demner-Fushman, D. (2016). Aspiring to Unintended Consequences of Natural Language Processing: A review of recent developments in Clinical and Consumer-Generated Text Processing. *Yearbook of Medical Informatics*, *25*(01), 224–233. https://doi.org/10.15265/iy-2016-017

Fall, I. S., Rajatonirina, S., Yahaya, A. A., Zabulon, Y., Nsubuga, P., Nanyunja, M., Wamala, J., Njuguna, C., Lukoya, C. O., Alemu, W., Kasolo, F. C., & Talisuna, A. O. (2019). Integrated Disease Surveillance and Response (IDSR) strategy: current status, challenges and perspectives for the future in Africa. *BMJ Global Health*, *4*(4), e001427. https://doi.org/10.1136/bmjgh-2019-001427

Feng, S., Grépin, K. A., & Chunara, R. (2018). Tracking health seeking behavior during an Ebola outbreak via mobile phones and SMS. *Npj Digital Medicine*, *1*(1). https://doi.org/10.1038/s41746-018-0055-z

Frieden, T R., Lee, C T., Bochner, A F., Buissonnière, M., & McClelland, A. (2021, July 6). 7-1-7: an organising principle, target, and accountability metric to make the world safer from pandemics. Elsevier BV, 398(10300), 638-640. https://doi.org/10.1016/s0140-6736(21)01250-2

Gupta, A., & Katarya, R. (2020). Social media based surveillance systems for healthcare using machine learning: A systematic review. *Journal of Biomedical Informatics*, *108*, 103500. https://doi.org/10.1016/j.jbi.2020.103500

Hahn, U., & Oleynik, M. (2020). Medical information extraction in the age of deep learning. *Yearbook of Medical Informatics*, *29*(01), 208–220. https://doi.org/10.1055/s-0040-1702001

Howard, J., & Ruder, S. (2018, January 18). *Universal Language model fine-tuning for text classification*. arXiv.org. https://arxiv.org/abs/1801.06146

Ibrahim, L M., Okudo, I., Stephen, M., Ogundiran, O., Pantuvo, J S., Oyaole, D R., Tegegne, S G., Khalid, A., Ilori, E., Ojo, O., Ihekweazu, C., Baraka, F., Mulombo, W K., Lasuba, C L P., Nsubuga, P., & Alemu, W. (2021, May 13). Electronic reporting of integrated disease surveillance and response: lessons learned from northeast, Nigeria, 2019. BioMed Central, 21(1). https://doi.org/10.1186/s12889-021-10957-9

Jamal, N., Chen, X., Al-Turjman, F., & Ullah, F. (2021, March 15). A Deep Learning–based Approach for Emotions Classification in Big Corpus of Imbalanced Tweets. Association for Computing Machinery, 20(3), 1-16. https://doi.org/10.1145/3410570

Jarashanth, S., & Nawarathna, R. (2022). Applying Transformer Models for Disease Named Entity Recognition. *IEEE.org*, 272–277. https://doi.org/10.1109/icarc54489.2022.9754023

Jerfy, A., Selden, O., & Balkrishnan, R. (2024). The growing impact of natural language processing in healthcare and public health. *INQUIRY the Journal of Health Care Organization Provision and Financing*, *61*. https://doi.org/10.1177/00469580241290095

Kambalame, D., Yelewa, M., Iversen, B., Khunga, N., Macdonald, E., Nordstrand, K., Mwale, A., Muula, A., Banda, E. C., Phuka, J., & Arnesen, T. (2024). Factors influencing operationalization of Integrated Disease Surveillance in Malawi. *Public Health*, *228*, 100–104. https://doi.org/10.1016/j.puhe.2023.12.030

Keller, M., Blench, M., Tolentino, H., Freifeld, C C., Mandl, K D., Mawudeku, A., Eysenbach, G., & Brownstein, J S. (2009, May 1). Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. Centers for Disease Control and Prevention, 15(5), 689-695. https://doi.org/10.3201/eid1505.081114

Kimera, R., Rim, D. N., & Choi, H. (2023). Fine-Tuning BERT on Twitter and Reddit Data in Luganda and English. *ACM Digital Library*, 63–70. https://doi.org/10.1145/3639233.3639344

Kuehne, A., Keating, P., Polonsky, J A., Haskew, C., Schenkel, K., Waroux, O L P D., & Ratnayake, R. (2019, December 1). Event-based surveillance at health facility and community level in low-income and middle-income countries: a systematic review. BMJ, 4(6), e001878-e001878. https://doi.org/10.1136/bmjgh-2019-001878

Lamorde, M., Mpimbaza, A., Walwema, R., Kamya, M., Kapisi, J., Kajumbula, H., Sserwanga, A., Namuganga, J. F., Kusemererwa, A., Tasimwa, H., Makumbi, I., Kayiwa, J., Lutwama, J., Behumbiize, P., Tagoola, A., Nanteza, J. F., Aniku, G., Workneh, M., Manabe, Y., . . . Kugeler, K. J. (2018). A Cross-Cutting Approach to Surveillance and Laboratory Capacity as a Platform to Improve Health Security in Uganda. *Health Security*, *16*(S1), S-86. https://doi.org/10.1089/hs.2018.0051

Lester, J., Paige, S B., Chapman, C A., Gibson, M A., Jones, J H., Switzer, W M., Ting, N., Goldberg, T L., & Frost, S D W. (2016, June 9). Assessing Commitment and Reporting Fidelity to a Text Message-Based Participatory Surveillance in Rural Western Uganda. Public Library of Science, 11(6), e0155971-e0155971. https://doi.org/10.1371/journal.pone.0155971

Loshchilov, I., & Hutter, F. (2017, November 14). *Decoupled weight decay regularization*. arXiv.org. https://arxiv.org/abs/1711.05101

Luo, Y., Henry, S., Wang, Y., Shen, F., Uzuner, O., & Rumshisky, A. (2020). The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, *27*(10), 1529-e1. https://doi.org/10.1093/jamia/ocaa106

Masiira, B., Nakiire, L., Kihembo, C., Katushabe, E., Natseri, N., Nabukenya, I., Komakech, I., Makumbi, I., Charles, O., Adatu, F., Nanyunja, M., Woldetsadik, S. F., Fall, I. S., Tusiime, P., Wondimagegnehu, A., & Nsubuga, P. (2019). Evaluation of integrated disease surveillance and response (IDSR) core and support functions after the revitalisation of IDSR in Uganda from 2012 to 2016. *BMC Public Health*, *19*(1). https://doi.org/10.1186/s12889-018-6336-2

Methuku, V. (2025). NLP and AI for Public Health Intelligence: Automating Disease Surveillance from Unstructured Data. *ICCK - Institute of Central Computation and Knowledge*, *2*(1), 43–56. https://doi.org/10.62762/tetai.2025.222799

Miftahutdinov, Z., Sakhovskiy, A., & Tutubalina, E. (2020, December 1). *KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions*. ACL Anthology. https://aclanthology.org/2020.smm4h-1.8/

Mremi, I R., George, J., Rumisha, S F., Sindato, C., Kimera, S I., & Mboera, L E G. (2021, November 9). Twenty years of integrated disease surveillance and response in Sub-Saharan Africa: challenges and opportunities for effective management of infectious disease epidemics. BioMed Central, 3(1). https://doi.org/10.1186/s42522-021-00052-9

Namuye, S., Platz, M., Okanda, P., & Mutanu, L. (2015, May 1). Leveraging health through early warning systems using mobile and service oriented technology. https://doi.org/10.1109/istafrica.2015.7190572

Nakiire, L., Masiira, B., Kihembo, C., Katushabe, E., Natseri, N., Nabukenya, I., Komakech, I., Makumbi, I., Charles, O., Adatu, F., Nanyunja, M., Nsubuga, P., Woldetsadik, S. F., Tusiime, P., Yahaya, A. A., Fall, I. S., & Wondimagegnehu, A. (2019). Healthcare workers' experiences regarding scaling up of training on integrated disease surveillance and response (IDSR) in Uganda, 2016: cross sectional qualitative study. *BMC Health Services Research*, *19*(1). https://doi.org/10.1186/s12913-019-3923-6

Nansumba, H., Nambuya, P., Wafula, J., Laiton, N., Kadam, R., Akinwusi, O., Suleiman, K., Akugizibwe, P., & Ssewanyana, I. (2023). Uptake and effectiveness of a mobile application for real-time reporting and quality assurance of decentralized SARS-CoV-2 testing in Uganda. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1053544

Ndisha, M., Hassan, A S., Ngari, F., Munene, E., Gikura, M., Kimutai, K., Muthoka, K., Murie, L A., Tolentino, H., Odhiambo, J., Mwele, P., Odero, L., Mbaire, K., Omoro, G., & Kimanga, D. (2023, September 15). Leveraging electronic medical records for HIV testing, care, and treatment programming in Kenya—the national data warehouse project. BioMed Central, 23(1). https://doi.org/10.1186/s12911-023-02265-6

Pais, D., Brás, S., & Sebastião, R. (2024). Overcoming the Small Dataset Challenge in Healthcare. *IEEE.org*, 497–502. https://doi.org/10.1109/melecon56669.2024.10608708

Plank, B. (2022). The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *ACL Anthology*, 10671–10682. https://doi.org/10.18653/v1/2022.emnlp-main.731

Preparedness, E. (2019, January 1). *Integrated disease surveillance and response technical guidelines: Booklet Two: sections 1, 2, and 3, 3rd ed*. https://www.who.int/publications/i/item/WHO-AF-WHE-CPI-01-2019

Ranade, M., & Deshpande, A. (2021). A QUALITATIVE LITERATURE REVIEW OF MACHINE LEARNING TECHNIQUES USED FOR DIAGNOSIS OF NEONATAL DISEASES IN HEALTHCARE. *International Journal of Scientific Research*, 4–7. https://doi.org/10.36106/ijsr/8529147

Randriamiarana, R., Raminosoa, G., Vonjitsara, N., Randrianasolo, R., Rasamoelina, H., Razafimandimby, H., Rakotonjanabelo, A., Lepec, R., Flachet, L., & Halm, A. (2018, April 10). Evaluation of the reinforced integrated disease surveillance and response strategy using short message service data transmission in two southern regions of Madagascar, 2014–15. BioMed Central, 18(1). https://doi.org/10.1186/s12913-018-3081-2

Raza, S., & Schwartz, B. (2023). Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02117-3

Resolve to Save Lives. (2024, December 17). *7-1-7: an organising principle, target, and accountability metric to make the world safer from pandemics - Resolve to Save Lives*. https://resolvetosavelives.org/resources/7-1-7-an-organising-principle-target-and-accountability-metric-to-make-the-world-safer-from-pandemics/#:~:text=For%20The%20Lancet's%20Viewpoint%2C%20experts,early%20response%20within%207%20days.

Ruis, F., Pathak, S., Geerdink, J., Hegeman, J. H., Seifert, C., & Van Keulen, M. (2020). Human-in-the-loop Language-agnostic Extraction of Medication Data from Highly Unstructured Electronic Health Records. *2021 International Conference on Data Mining Workshops (ICDMW)*, 644–650. https://doi.org/10.1109/icdmw51313.2020.00091

Sachin, M. U., Nagaraj, R., Samiksha, M., Rao, S., & Moharir, M. (2017). GPU based Deep Learning to Detect Asphyxia in Neonates. *Indian Journal of Science and Technology*, *10*(3). https://doi.org/10.17485/ijst/2017/v10i3/110617

Seck, O., Roka, J L., N'Diaye, M., Namageyo-Funa, A., Abdoulaye, S., Mangane, A., Dieye, N L., Ndoye, B., Diop, B., Ting, J., & Pasi, O. (2023, June 1). SARS-CoV-2 case detection using community event-based surveillance system—February–September 2020: lessons learned from Senegal. BMJ, 8(6), e012300-e012300

Sekamatte, M., Krishnasamy, V., Bulage, L., Kihembo, C., Nantima, N., Monje, F., Ndumu, D., Sentumbwe, J., Mbolanyi, B., Aruho, R., Kaboyo, W., Mutonga, D., Basler, C., Paige, S., & Behravesh, C. B. (2018). Multisectoral prioritization of zoonotic diseases in Uganda, 2017: A One Health perspective. *PLoS ONE*, *13*(5), e0196799. https://doi.org/10.1371/journal.pone.0196799

Shafran, I., Du, N., Tran, L., Perry, A. N., Keyes, L., Knichel, M., Domin, A., Huang, L., Chen, Y., Li, G., Wang, M., Shafey, L. E., Soltau, H., & Paul, J. S. (2020). *The Medical Scribe: Corpus Development and Model Performance Analyses*. https://www.semanticscholar.org/paper/The-Medical-Scribe%3A-Corpus-Development-and-Model-Shafran-Du/447249d74104d68f2b9fd2cc0fc01b43da6e7566

Simonsen, L., Gog, J R., Olson, D., & Viboud, C. (2016, October 5). Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems. Oxford University Press, 214(suppl 4), S380-S385. https://doi.org/10.1093/infdis/jiw376

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, May 14). *How to Fine-Tune BERT for text classification?* arXiv.org. https://arxiv.org/abs/1905.05583

Toda, M., Njeru, I., Zurovac, D., Tipo, S. O., Kareko, D., Mwau, M., & Morita, K. (2016). Effectiveness of a Mobile Short-Message-Service–Based Disease Outbreak Alert System in Kenya. *Emerging Infectious Diseases*, *22*(4), 711–715. https://doi.org/10.3201/eid2204.151459

Van Hoek, A. J., Funk, S., Flasche, S., Quilty, B. J., Van Kleef, E., Camacho, A., & Kucharski, A. J. (2024). Importance of investing time and money in integrating large language model-based agents into outbreak analytics pipelines. *The Lancet Microbe*, *5*(8), 100881. https://doi.org/10.1016/s2666-5247(24)00104-6

Xie, K., Babbar, N., Chen, V., & Turura, Y. (2025). Enhancing multilingual language models for Code-Switched input data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2503.07990

Young, I J B., Luz, S., & Lone, N. (2019, October 5). A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. Elsevier BV, 132, 103971-103971. https://doi.org/10.1016/j.ijmedinf.2019.103971

APPENDICES

APPENDIX 1: Informed Consent Form for Key Informant Interview

Title of the proposed study:

Enhancing Outbreak Surveillance through Integration of Natural Language Processing in Uganda's Electronic Integrated Disease Surveillance and Response System

Investigators:

Nakitandwe Rebecca Melisa, Makerere University School of Public Health.
Contact: +256 (0) 787 655 505

Background and rationale for the study:

Uganda faces a high burden of infectious diseases including high-priority zoonoses. Disease surveillance is largely SMS-based under the eIDSR system, but manual processing creates delays. NLP offers a solution by automating SMS message interpretation to improve early outbreak detection.

A description of sponsors of the research project and the organizational affiliation of the researchers:

The research is self-sponsored as part of a Master's requirement under Makerere University School of Public Health.

Purpose:

This study seeks to understand operational and contextual factors affecting SMS message composition in disease surveillance. It involves interviews with surveillance officers. The experimental component relates to using NLP for automation. Participation lasts 30–45 minutes.

Procedures:

You will be asked to participate in a one-on-one interview discussing your experience with SMS disease reporting. This will take 30–45 minutes and may be audio recorded with your permission.

Who will participate in the study:

eIDSR focal persons, district surveillance officers, and data managers. About 10–15 participants will be interviewed.

Risks/Discomforts:

There are no physical risks. However, participants may experience minor discomfort when discussing operational challenges.

Benefits:

There are no direct personal benefits. However, your input may improve national surveillance practices. No alternative procedures are required.

Confidentiality:

Your responses will be kept confidential. No names will be used in reports. Only the research team, SPH-REC, and UNCST may access anonymized data.

Alternatives:

Participation is voluntary. You may choose not to participate without any consequences.

Cost:

There is no cost for participating in this study.

Compensation for participation in the study:

Participants will receive UGX 5,000 as a token of appreciation for their time. The study carries minimal risk

Reimbursement:

Participation in this study is entirely voluntary. No reimbursement or compensation will be provided. You may choose to opt out at any time if you are not comfortable.

Questions about the study:

You may contact Nakitandwe Rebecca Melisa at beckynakitandwe@gmail.com for any questions.

Questions about participants rights:

If you have questions about your rights as a research participant, contact Dr. Kagaayi Joseph, Chairperson SPH-REC at 0773785333 or email jkagaayi@musph.ac.ug

Statement of voluntariness:

Your participation is entirely voluntary. You can withdraw at any point without any consequences.

Dissemination of results:

A summary of the study results will be shared with participants. Any significant findings relevant to participants will be communicated.

Ethical approval:

This study has been reviewed and approved by the Makerere University School of Public Health Research and Ethics Committee (SPH-REC).

STATEMENT OF CONSENT

The study has been explained to me. I understand the procedures, risks, and benefits. I know my participation is voluntary and I can withdraw at any time. By signing below, I consent to participate in this study. I will receive a copy of this form.

Name of Participant: _____

Signature/Thumbprint: _____ Date: _____

Signature of Interviewer: _____ Date: _____

Signature of Witness (if applicable): _____ Date: _____

APPENDIX 2: Key Informant Interview Guide

Study Title:

Enhancing Outbreak Surveillance through Integration of Natural Language Processing in Uganda's Electronic Integrated Disease Surveillance and Response System

Target Participants:

- eIDSR focal persons
- District Surveillance Officers
- Data Managers involved in SMS-based reporting
- Epidemiologist

Estimated Duration: 30–45 minutes

Introduction

1. Briefly describe your current role in relation to disease surveillance and SMS reporting.
2. How long have you been involved in this role?

Experience with SMS Reporting

3. Can you describe how SMS disease reports are typically submitted to the system?
4. Who usually sends these messages (community health workers, public, others)?
5. What challenges do you encounter when interpreting SMS messages?

Content and Structure of SMS Messages

6. How are disease names, symptoms, and locations usually presented in SMS reports?
7. Are abbreviations, local languages, or slang commonly used? Please give examples.
8. Do the formats of messages vary between different reporters? How?

Operational and Contextual Factors

9. Are there any guidelines or templates used to guide message composition?
10. What operational challenges affect how quickly and accurately SMS reports are processed?

Relevance to NLP and System Improvement

11. What do you think would help an automated system better understand and extract accurate information from SMS reports?
12. What kinds of errors or misunderstandings might such a system encounter?
13. What recommendations would you give to improve SMS data quality?

APPENDIX 3: NER Codebook

This NER model is designed to extract critical epidemiological information from unstructured SMS messages sent to Uganda's eIDSR (electronic Integrated Disease Surveillance and Response) system. The extracted entities enable faster detection of disease outbreaks and improve public health response.

Entity Types and Descriptions

Entity Tag	Description	Example(s)	Annotation Rules
B-DISEASE / I-DISEASE	Name of the suspected disease being reported	rabies, plague, brucellosis	Use as provided in the disease list. Use B- for the first token and I- for continuation.
B-SYMPTOM / I-SYMPTOM	Symptoms experienced or observed in the patient	itching rashes, blisters, fever	Tag multi-word symptoms as one entity using B/I tags. Use fuzzy matching for common typos.
B-AGE / I-AGE	Age of the suspected patient	4yr, 10yrs, 2 years	Include both number and unit (yr, yrs, months). Only tag if an age term is detected.
B-GENDER	Gender or sex of the patient	male, female, girl, boy	Tag single word only.
B-DISTRICT / I-DISTRICT	District or location name within Uganda	Tororo, Oyam District	Use district list. Tag multiple tokens as B/I if needed.
O	Outside – any token that does not belong to a defined entity	All other words	Default for all untagged tokens.

Annotation Guidelines

Tokenization: SMS messages are tokenized using whitespace and punctuation-aware logic (e.g., 'fever.' becomes 'fever'). Multi-word entities must be tagged as sequences (B- followed by I-).

Multi-Entity Matching: A single SMS may contain multiple entities of the same type. Each should be tagged individually.

Fuzzy Matching Rules: For SYMPTOM entities, fuzzy matching is used (e.g., 'lymph node' vs. 'lymp nodes') with ≥90% similarity threshold.

Handling Missing Info: If an SMS does not contain one of the expected entity types (e.g. no age or district), no tag is applied.

Normalization (Optional Post-Processing): After extraction, entities may be normalized to standard values (e.g., 'yrs' → 'years', 'rbies' → 'rabies').

Example Annotated Sentence (BIO Format)

Token	Label
girl	B-GENDER
10yr	B-AGE
with	O
itching	B-SYMPTOM
rashes	I-SYMPTOM
from	O
tororo	B-DISTRICT
suspected	O
rabies	B-DISEASE