A DEEP NEURAL NETWORK MODEL FOR DETECTION OF URGENCY IN THE SHORT MESSAGE SERVICES

NARSHION MATAI NGAO

MASTER OF SCIENCE IN

COMPUTER SYSTEMS

JOMO KENYATTA UNIVERSITY

OF

AGRICULTURE AND TECHNOLOGY

2026

A Deep Neural Network Model for Detection of Urgency in the Short Message Services

Narshion Matai Ngao

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computer Systems of the Jomo Kenyatta University of Agriculture and Technology

2026

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature: …… Date ……

Narshion Matai Ngao

This thesis has been submitted for examination with our approval as the University Supervisors:

Signature: …… Date ……

Dr. Tobias Mwalili, PhD
JKUAT, Kenya

Signature: …… Date ……

Dr. Lawrence Nderu, PhD
JKUAT, Kenya

DEDICATION

I dedicate this work to my wife, Eunice Mbeyu, my daughters Maureen and Yvonne, my son Ian, and my niece Elizabeth Lugo. You have been a constant source of strength, encouragement, and support throughout this journey. Thank you for giving me the time and space to pursue my studies, even when it meant being away from home for long periods. Your love, patience, and devotion will always be deeply cherished.

Narshion Matai Ngao

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all my lecturers at the Jomo Kenyatta University of Agriculture and Technology (JKUAT), NCBD and main campuses, for their dedication, guidance, and support throughout my studies. I am especially grateful to my supervisors, Dr. Tobius Mwalili and Dr. Lawrence Nderu, whose mentorship, insightful feedback, and encouragement greatly contributed to the successful completion of this research.

I would also like to acknowledge the University of Washington Global Health team, in collaboration with the Kenya Medical Research Institute (KEMRI), for granting permission to use their data for this study and for their continued support. I am particularly indebted to my mentors Keshet Ronen, Tal August, Brian DeRenzi, and Abraham Flaxman for their invaluable guidance, patience, and encouragement. Your willingness to engage with my persistent questions and your openness in allowing me the space to learn, experiment, and grow have been instrumental in shaping my academic journey.

My sincere appreciation also goes to my colleagues at the KEMRI-Wellcome Trust Research Programme for their unwavering support and encouragement as I balanced my professional responsibilities with my academic work. I am especially thankful to my line managers for their understanding, flexibility, and encouragement in allowing me the time needed to pursue this study.

Finally, I would like to thank everyone who, in one way or another, contributed to this journey. Your support, encouragement, and belief in my work have been deeply appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMNS

AfriBERT	African Pretrained Language Model
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BLEU	Bilingual Evaluation Understudy
BoW	Bag of Words
CBOW	Continuous Bag of Words
CNN	Convolutional Neural Network
DBOW	Distributed Bag of Words
DSSM	Deep Structured Semantic Model
DL	Deep Learning
ELMo	Embeddings from Language Models
F1-Score	Harmonic Mean of Precision and Recall
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GloVe	Global Vectors for Word Representation
GRU	Gated Recurrent Unit
LLM	Large Language Model
LR	Logistic Regression
LSTM	Long Short-Term Memory
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
MCC	Matthews Correlation Coefficient
mHealth	Mobile Health
ML	Machine Learning
MLM	Masked Language Modelling
MSE	Mean Squared Error
mWACH	mobile Women, Adolescent and Child Health

Mobile WACh	Mobile Solutions for Women and Children's Health – New-born Extension
NEO	
NLP	Natural Language Processing
NNLM	Neural Network Language Model
NSP	Next Sentence Prediction
PHC	Primary Health Care
PR-AUC	Area Under the Precision–Recall Curve
PV-DM	Paragraph Vector – Distributed Memory
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SMS	Short Message Service
SQuAD	Stanford Question Answering Dataset
SwahBERT	Swahili-focused BERT Language Model
TF-IDF	Term Frequency–Inverse Document Frequency
TLM	Transfer Learning Model
TN	True Negative
TP	True Positive
TPR	True Positive Rate (Recall/Sensitivity)
ULMFiT	Universal Language Model Fine-tuning
UAT	User Acceptance Testing
WHO	World Health Organization
WMT14	Workshop on Machine Translation 2014 Dataset
XLM	Cross-lingual Language Model
XNLI	Cross-lingual Natural Language Inference Dataset

ABSTRACT

Timely identification of urgent patient messages is critical for effective clinical decision-making in mobile health (mHealth) programs, particularly in low-resource settings where healthcare workers manage large volumes of incoming short message service (SMS) communication. In Kenyan maternal and child health programs, nurses manually triage multilingual patient messages, a process that contributes to delayed responses and increased risk of missed urgent cases. This study investigates the effectiveness of contextual natural language processing (NLP) models for automatically classifying patient SMS messages into urgency categories within a real-world mHealth environment. Using a dataset of 11,129 manually labelled multilingual SMS messages from 772 participants enrolled in the Mobile Solutions for Women and Children's Health (Mobile WACh NEO) program in Kenya, urgency detection was formulated as a supervised binary classification task aligned with clinical triage workflows. Baseline models employing unigram and bigram features with penalized logistic regression were compared against contextual embedding approaches, including multilingual BERT (mBERT), SwahBERT, and AfriBERT. Transformer models were adapted to the clinical domain through domain-specific pretraining and task-adaptive fine-tuning. To mitigate contextual sparsity inherent in short SMS messages, prior nurse or system messages were concatenated with the current message to form context-aware input representations. Model development followed explicit train, development, and test splits, with cross-validation applied during training to support robust model selection and reduce overfitting. Performance was evaluated using precision, recall, and F1-score, emphasizing clinical utility for both triage and prioritization objectives. Transformer architectures substantially outperformed frequency-based baselines, achieving F1 improvements of up to 0.186 relative to bigram models. Our best performing model was mBERT model pretrained on task-level adaptation using nurse context before fine-tuning. This model got a precision of 50%, recall of 45% and F1 score of 47%, which were below the thresholds we set for either a triage model or prioritization model. However, incorporating nurse conversational context reduced performance gaps between configurations (e.g., ΔF1 decreasing from approximately 0.080 in non-contextual mBERT to 0.032 with nurse context), while task-adaptive pretraining provided incremental yet consistent gains. Although performance did not fully meet predefined clinical usefulness thresholds, context-aware fine-tuned transformer models demonstrated improved recall, indicating reduced risk of missed urgent messages. Overall, the findings confirm that contextual transformer-based models offer meaningful advantages over traditional representations in multilingual, low-resource clinical SMS environments. While additional advances in architecture and domain adaptation are needed to reach optimal deployment standards, the results align with contemporary state-of-the-art NLP practices and support the feasibility of automated decision-support tools to augment nurse triage workflows in mHealth systems.

Keywords: Urgency Detection, Contextual NLP, Multilingual Transformers, mHealth, Clinical Decision Support

CHAPTER ONE

INTRODUCTION

This section provides an overview of this research thesis. It gives some background information about the area of research being proposed and then breaks down the content into research problem, justification, objectives and research questions.

1.1 Background Information

The engagement between patients and healthcare providers in a mHealth setting can be very instrumental to successful delivery of virtual healthcare services to remote patients; who could not otherwise easily access healthcare (Wangler & Jansky, 2024). SMS has been the most impactful medium of communication and readily available means of providing this engagement as it continues to be a key tool in business to consumer communication (Telnyx, 2024). mHealth platforms increasingly rely on SMS communication to support maternal and child health programs in low-resource settings (Bossman et al., 2022). These systems generate large volumes of multilingual, informal, and context-dependent text messages that must be reviewed and acted upon by healthcare workers. As message volumes grow, effective computational support for identifying clinically urgent messages becomes essential to ensure timely responses, reduce cognitive burden on nurses, and improve patient safety.

The main challenges for scaling up such SMS based health interventions to wider populations have been difficulties in managing volumes of incoming messages and inadequate capacity to provide expert advice based on the context of the engagement. This is evidenced by (Barron et al., 2018), who pointed out that machine learning implementation in their mHealth (MomConnect) system could improve not only response times for their help desk but also streamline the number of staff to manageable levels.

More studies have shown that Machine Learning based mHealth solutions can increase efficiency while maintaining and even improving service quality, thereby

reducing the burden on healthcare staff and providing for scalable, patient-centred service delivery (Khanbhai et al., 2021). This research thesis aims at establishing how deep neural network models can be applied in text classification to provide a quicker way of sifting through the incoming messages from patients based on urgency. This will make the healthcare staff not to lose out on the messages from patients who need emergency response.

Text classification is a process of assigning labels or categories to text based on its content using computational mechanisms in the broader natural language processing field. It has been successfully used in areas such as sentiment analysis (Alaparthi & Mishra, 2021), spam detection (Singh et al., 2020), customer ticket routing (Feng et al., 2022), and topic modelling (Chaudhary et al., 2020). This is an NLP technique that involves converting the text into word embeddings. Word embeddings represent words in a continuous vector space where semantically similar words are positioned close to one another, enabling the model to capture underlying semantic relationships. These have proved to be extremely important as they enable richer vector representations that preserve more semantic and syntactic information, making most NLP tasks a lot more accurate.

Most common methods for creating word embeddings are based on deep neural networks such as sequence-to-sequence models (Gong et al., 2022), Word2Vec (Grohe, 2020), Doc2Vec, Bidirectional Encoder Representation from Transformer (BERT) (Vaswani et al., 2017) and Generative Pretrained Transformers (GPT) (OpenAI et al., 2024) among others. One of the most exciting features with these embeddings is that they can be re-used in different settings, a process called transfer learning (Iman et al., 2022). Then after the word embeddings are generated, a machine learning method for classification could be utilized to classify the message accordingly based on the context represented in the content of the message. Such classification models include penalized logistic regression (James et al., 2021), gradient boosting classifiers (Florek & Zagdański, 2023), decision trees (Souza et al., 2022), K-nearest neighbor (Joloudari et al., 2020), naïve Bayes (Garcia & Johnson, 2020) and deep artificial neural networks (Garcia & Johnson, 2020).

This research thesis focuses to come up with a rich contextual representation mechanism for generating word embeddings from multiple SMS messages grouped into a similar context based on previous engagement. The research also aims to show that such lengthy and sometimes large engagements can be understood by a machine and be useful in deciding the urgency of a future message from the participant.

1.2 Statement of the Problem

SMS-based engagement in mHealth programs has grown rapidly with the widespread adoption of mobile devices, which now exceed 8.5 billion subscriptions globally (Ericsson, 2024). However, managing the large volume of incoming patient messages remains a significant challenge. In programs such as Mobile Solutions for Women and Children's Health (mWACh), nurses often handle dozens of messages daily and several hundred weekly (Ronen et al., 2021). Within this stream, most messages are non-urgent, while approximately 21% require timely clinical attention. mWACh studies have shown that manual review and response can introduce delays ranging from several hours to multiple days, particularly during periods of staff shortage or peak workload. While prior mHealth studies such as mWaCh do not report explicit misclassification rates for urgent messages, delayed or missed responses are acknowledged as a key operational risk. Such delays are clinically consequential, as missed or late identification of urgent messages has been associated with delayed referrals, postponed clinical interventions, and increased risk of adverse maternal and neonatal outcomes.

Although prior studies have explored automated text classification for urgency detection, challenges remain. For instance, (Sarioglu Kayi et al., 2020) applied transfer learning with models such as BERT, XLM-R (Conneau et al., 2020), and RoBERTa (Liu et al., 2019) to detect urgency in crisis-related tweets. Their study highlighted the importance of in-domain embeddings. However, it also reported challenges such as limited labeled data, class imbalance, and model overfitting. Similarly, (Lowres et al., 2019) evaluated machine learning methods for triaging SMS responses in a cardiovascular prevention program and found that unstructured

language, abbreviations, and spelling errors in SMS communication significantly reduced classification accuracy.

To address these challenges, this study investigates machine learning approaches for triaging patient SMS messages according to their level of urgency. The proposed approach incorporates conversational context by considering previous interactions between patients and healthcare providers when classifying new messages. Transformer-based architectures such as BERT, which use attention mechanisms to capture relationships within text sequences (Vaswani et al., 2017), are used to model this context. Multilingual BERT (mBERT) is used as a baseline due to its strong multilingual coverage and suitability for low-resource settings. Regionally adapted models such as SwahBERT and AfriBERT are also evaluated to assess the benefits of locally adapted language representations for mHealth deployment. By combining contextual modeling with domain-adapted language representations, this research aims to develop a more effective approach for identifying urgent messages and supporting timely triage in SMS-based mHealth systems.

1.3 Main Objective

To develop and evaluate a contextual machine learning approach for classifying patient SMS messages into urgency categories using multilingual data from a low-resource clinical setting.

1.3.1 Specific Objective

i) To establish baseline urgency classification performance using traditional n-gram–based text representations.
ii) To adapt and fine-tune transformer-based language models, including mBERT, SwahBERT, and AfriBERT, for clinical SMS urgency detection.
iii) To evaluate the contribution of conversational context to urgency classification performance using clinically relevant evaluation metrics

1.4 Research Questions

The following are the research questions.

i) How effectively can traditional n-gram–based text representations classify patient SMS messages by urgency in a low-resource clinical setting?
ii) How do fine-tuned transformer-based language models (mBERT, SwahBERT, AfriBERT) perform in urgency classification compared to traditional approaches?
iii) What trade-offs exist between precision and recall when deploying urgency classification models for clinical triage in relation to nurse-patient conversational context?

1.5 Justification of the Study

Timely information is critical for management of patients. It helps them and those providing services to make decisions in a timely fashion and intervene with the right intervention to save lives. This was demonstrated by (Actis Danna et al., 2020) who reviewed literature about the three delays model in maternal health. They cited papers where the three delays model was well applied in maternal health, leading to significant reduction of the global maternal mortality rate. The three delays model proposes three major obstacles to receiving healthcare by patients. First, patients may not know or may be reluctant to get medical care, this is a barrier on deciding to seek care. This happens when patients lack information about their problem or they cannot perceive the situation beyond their current circumstances. The second obstacle is a delay in identifying and reaching a healthcare facility. This could be due to transportation barriers or a lack of awareness on where to seek healthcare. The third obstacle is a delay while receiving adequate and appropriate treatment. This delay include time consumed while interrogating patients during triage, queueing, and or misdiagnosis leading to wrong treatment. (Dafroyati et al., 2023) described causes of maternal mortality that is based on the three delays model in Indonesia. They clearly describe the challenging encounters mothers were likely to face at each of the three stages described here.

This research thesis addresses these delays by introducing a text mining approach that is aimed at detecting level of urgency from patient messages. For the first type of delay, the model will promptly classify the context of the correspondences, offering guidance to healthcare providers on the likelihood or urgency of the condition that may require further attention through escalated medical services. Healthcare staff will be better informed to advice the patients early enough. The second delay type will be addressed using the context of the engagement as synthesized by the deep learning model. Further work on topic modelling and classification can provide specific medical service catalogue area that the patient can be referred to. This will be important in advising the patient on what treatment to seek and where to go for the treatment. Finally, with the timely classification of urgency as well as providing context of the problem will help address the third delay by providing information that will be appropriate and accurate to lead to the right treatment. In other words, the model would help with the triaging function that often is causing delays or accuracy issues.

While several multilingual and cross-lingual transformer models have been proposed for low-resource NLP tasks, multilingual BERT (mBERT) was selected as a core baseline in this study due to its balance between linguistic coverage, computational feasibility, and empirical performance in multilingual settings. Compared to cross-lingual models such as XLM-R, which are pretrained on substantially larger corpora and typically require greater computational resources, mBERT offers a more practical deployment profile for resource-constrained environments without sacrificing multilingual capability. In addition, mBERT has been widely adopted as a benchmark model in clinical and low-resource NLP studies, enabling meaningful comparison with prior work.

This research thesis provides the opportunity to apply advanced word embedding techniques into patient message classification and detection of urgency. The Transformer model and more particularly the attention mechanism that broadcasts all encodings from previous layers and suggests the most correlating one is a very important aspect that this research plans to put to good use. Questions like “are similar words just close or are there other deeper representational meanings that can

bring even dissimilar words closer based on context?" For instance, if the patient was vomiting yesterday and today, he lacks appetite, would the model classify such a message as urgent? These are some of the scenarios where this research will explore to see whether a computer model would be able to strongly correlate vomiting with lack of appetite. Further research work can be done from this scope to extend the applications of the learned contexts of engagement by the model. For instance, research on sentiment analysis about patients can be done or understanding deeper patterns of emotional imbalances during patient engagements.

1.6 Contribution to Research

In this era of large language models (LLMs), a lot has been done in the field of text classification. However, these LLMs often require huge training data and computational resources to generalize well without overfitting. In this research, we provide evidence that smaller to medium language models like mBERT have the right size and pre-training to make them capable of using multilingual representations to perform well with smaller multilingual dataset.

This study is in line with recent efforts on responsible Artificial Intelligence dubbed Green AI (Schwartz et al., 2020) that encourages use of adequate compute resources to accomplish goals. We also focus on checking the context of messages and hence we have shown that messages with prepended prior messages to create some context stand a better chance to be classified correctly compared to stand-alone messages. This is particularly evidenced with using nurse context messages prepended upon the message under investigation.

This work was presented at the Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing Student Research Workshop in 2022 (Ngao et al., 2022). See appendix for more information including where the source code is published.

CHAPTER TWO

LITERATURE REVIEW

This chapter provides a review of existing literature in the area of study. It reviews data collection, cleaning and labelling; then goes on to discuss the various text representation techniques and finally reviews modern machine learning methods for prediction and classification.

2.1 Detecting Urgency in SMS Messages of Patients

The use of SMS as a communication channel in mHealth interventions has grown substantially, particularly in low- and middle-income countries where mobile phone penetration is high and access to in-person care is limited. SMS-based systems are widely used for appointment reminders, health education, symptom reporting, and two-way communication between patients and healthcare providers. As engagement increases, however, healthcare teams are faced with large volumes of incoming patient messages that vary widely in content, clarity, and clinical significance, making timely identification of urgent cases a persistent challenge.

Several studies have explored automated text classification techniques to support message triage and urgency detection. Early work in this area relied on rule-based systems and keyword matching, which proved insufficient due to variability in language use, spelling errors, abbreviations, and indirect expressions of distress (Taha et al., 2024). More recent approaches have adopted machine learning and natural language processing (NLP) methods to classify urgency in textual data, particularly in domains such as emergency call logs, crisis-related social media posts, and patient–provider messaging systems (Lamsal & others, 2023; Swaminathan et al., 2023; A. Wahid et al., 2025). The emergence of deep contextual language models has further improved performance in identifying clinically relevant signals within conversational text, enabling more reliable prioritization of patient communications and workload management in digital health settings (Devlin et al., 2019a; Gururangan et al., 2020; Mermin-Bunnell & others, 2023).

Research on urgency detection in social media and crisis informatics has demonstrated the potential of supervised machine learning models to identify high-risk messages. For example, studies using convolutional neural networks (CNNs) (Makkena et al., 2024), recurrent neural networks (RNNs), and transformer-based (Broadbent et al., 2023; J. A. Wahid et al., 2025) architectures have reported improved performance over traditional bag-of-words approaches by capturing semantic patterns and contextual dependencies within text. However, much of this work is based on English-language datasets drawn from high-resource settings or public platforms such as Twitter, which differ substantially from clinical SMS data in structure, intent, and linguistic variability.

In clinical mHealth contexts, urgency detection presents additional challenges. Patient messages are often short, fragmented, and highly context-dependent, with critical information distributed across multiple interactions rather than contained within a single message. Studies such as (Lowres et al., 2019) have shown that even well-performing machine learning models struggle to reliably identify urgent cases in SMS-based health programs, particularly when urgency is rare and class imbalance is pronounced. These findings highlight the limitations of approaches that rely solely on isolated messages without incorporating conversational or historical context.

Recent work has explored the use of pre-trained language models and transfer learning to address data scarcity and linguistic variability. (Sarioglu Kayi et al., 2020), for instance, applied BERT-based models to urgency detection in crisis-related text and demonstrated that in-domain fine-tuning and contextual embeddings improve performance relative to traditional methods. Nevertheless, these approaches remain constrained by limited labelled data, risk of overfitting, and reduced performance in low-resource language settings. Moreover, most studies evaluate models using generic classification metrics, with limited discussion of how model errors particularly false negatives affect clinical workflows and patient safety.

Importantly, there is limited literature focusing on urgency detection in African mHealth settings, where multilingual communication, code-switching (for example, English–Kiswahili mixtures), and informal language use are common. Kenyan SMS-

based health programs, including maternal and child health initiatives, frequently rely on small teams of nurses or community health workers to manually triage incoming messages. In such settings, delayed or missed identification of urgent messages can directly contribute to increased nurse workload, delayed clinical response, and adverse patient outcomes. Despite this, few existing studies explicitly model urgency as a function of longitudinal patient engagement or evaluate model performance in terms of clinical triage and prioritization utility.

In summary, while prior research demonstrates the feasibility of automated urgency detection using NLP and deep learning techniques, significant gaps remain in addressing real-world clinical SMS data from low-resource, multilingual environments. Specifically, there is a lack of approaches that (i) incorporate conversational context across multiple messages, (ii) are evaluated using clinically meaningful criteria aligned with nurse triage workflows, and (iii) are validated on African mHealth datasets. These gaps motivate the context-aware, transformer-based approach adopted in this study.

2.2 Machine Learning

Machine learning (ML) refers to a class of computational methods that enable systems to learn patterns from data and make predictions or decisions without being explicitly programmed for each task (Kataria et al., 2023). In contrast to rule-based systems, machine learning models infer relationships between inputs and outputs by optimizing model parameters based on observed data. This paradigm is particularly well suited to text classification problems, where linguistic variability and ambiguity make handcrafted rules impractical.

Machine learning approaches are commonly categorized into supervised and unsupervised learning. Supervised learning involves training models on labelled data, where each input instance is associated with a known output class. This approach is widely used in text classification tasks, including spam detection, sentiment analysis, and clinical message triage, where labelled examples of urgent and non-urgent messages are available. Unsupervised learning, on the other hand, focuses on discovering latent patterns or structures within unlabelled data and is often applied to

tasks such as topic modelling or clustering. Given the objective of predicting message urgency based on annotated outcomes, this study adopts a supervised learning framework.

In text-based applications, the performance of machine learning models is strongly influenced by how textual data are represented (Jurafsky & Martin, 2023). Traditional machine learning classifiers such as logistic regression, decision trees, and random forests rely on manually engineered features or static text representations. While these methods are computationally efficient and interpretable, their effectiveness is limited when dealing with short, noisy, and context-dependent text such as SMS messages. In such cases, urgency may not be explicitly stated within a single message but inferred from subtle linguistic cues or prior interactions.

Recent advances in machine learning have therefore shifted toward representation learning, where models automatically learn informative features from raw text. Deep learning–based approaches have demonstrated superior performance by capturing nonlinear relationships and contextual dependencies within language. These models reduce the need for extensive feature engineering and are better suited to handling the variability inherent in patient-generated messages (Ahmed et al., 2024; Min et al., 2023). However, they also introduce challenges related to data requirements, class imbalance, and model interpretability, especially in low-resource clinical settings.

In the context of SMS-based mHealth systems, machine learning models must balance predictive performance with practical considerations such as limited labelled data, multilingual communication, and the clinical cost of misclassification. Consequently, the choice of learning algorithm and text representation method is critical, motivating the exploration of contextualized language models that can leverage prior knowledge through pre-training while remaining adaptable to domain-specific urgency detection tasks.

2.3 Natural Language Processing

Natural Language Processing (NLP) is a subfield of machine learning concerned with enabling computers to analyse, interpret, and derive meaning from human

language. In text classification tasks, NLP provides the methodological foundation for converting unstructured text into representations that can be processed by machine learning algorithms. In clinical mHealth settings, NLP plays a critical role in supporting automated interpretation of patient-generated messages, where language is informal, abbreviated, multilingual, and often context dependent.

Text classification using NLP typically involves assigning one or more labels to a text instance based on its content. In the context of this study, the task is to classify incoming patient SMS messages according to their level of urgency. Unlike longer clinical notes or structured reports, SMS messages are short and may lack explicit indicators of severity. As a result, urgency is often expressed implicitly through symptom descriptions, emotional tone, or inferred from prior interactions, posing challenges for conventional NLP pipelines that treat messages in isolation.

A central challenge in NLP is the representation of text in a numerical form that preserves semantic meaning. Early NLP systems relied on manual feature engineering and frequency-based representations such as bag-of-words and TF-IDF, which map words or phrases to vectors based on their occurrence statistics within a corpus (Jurafsky & Martin, 2023). While such representations are computationally efficient, they discard word order and broader context, limiting their ability to capture meaning in short or ambiguous messages. This limitation is particularly problematic in urgency detection, where the same phrase may indicate different levels of risk depending on conversational history or clinical context.

To address these shortcomings, modern NLP approaches increasingly rely on distributed representations learned from data. Neural network–based language models generate dense vector embeddings that encode semantic and syntactic relationships between words, allowing similar expressions to be mapped to nearby points in a continuous vector space (Minaee et al., 2024). These embeddings enable models to generalize beyond exact word matches and better handle spelling variations, abbreviations, and informal language commonly found in SMS communication. Recent advances in NLP have further improved text representation through contextualized language models, which generate word or sentence

embeddings conditioned on surrounding text (Peters et al., 2018). Unlike static embeddings, contextualized representations allow the meaning of a word to vary depending on its usage within a sentence or dialogue. This property is especially important for clinical SMS data, where urgency may be signalled through subtle shifts in wording across multiple messages rather than explicit keywords in a single message (Lehman et al., 2021).

In low-resource and multilingual settings, such as Kenyan mHealth programs, NLP models must also contend with code-switching, limited annotated data, and underrepresentation of local languages in standard training corpora (Kreutzer et al., 2022). These constraints necessitate approaches that can leverage transfer learning and pre-trained multilingual models while remaining sensitive to domain-specific language use. Consequently, the choice of NLP techniques in this study prioritizes contextual representation, robustness to linguistic variability, and suitability for short, conversational text, aligning closely with the practical requirements of automated urgency detection in real-world healthcare workflows.

2.3.1 Traditional Frequency-Based Text Representations

Early approaches to text representation in natural language processing relied on frequency-based methods that summarize text according to the occurrence of words or word sequences within a document. Common techniques in this category include Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and n-gram models. These methods transform text into fixed-length numerical vectors that can be used as input for conventional machine learning classifiers.

The Bag-of-Words approach represents each document as a vector of word counts derived from a predefined vocabulary. While simple and computationally efficient, BoW ignores word order and semantic relationships, treating each term as independent (A. Semary et al., 2024). As a result, messages with similar meaning but different wording may be represented as dissimilar vectors. In clinical SMS data, where urgency may be expressed indirectly or through varied phrasing, this limitation reduces the effectiveness of BoW-based models.

TF-IDF extends the Bag-of-Words model by weighting terms according to their importance within a document relative to the entire corpus. Words that appear frequently in a specific message but infrequently across the corpus are assigned higher weights, improving discrimination between documents. Despite this enhancement, TF-IDF remains a context-agnostic representation that does not capture semantic meaning or relationships between words. In short patient messages, critical urgency cues may be diluted or misrepresented due to the sparse and fragmented nature of the text.

N-gram models partially address the loss of word order by representing sequences of adjacent words rather than individual tokens. While n-grams can capture local phrase-level patterns, they are limited to fixed-length contexts and result in high-dimensional feature spaces, particularly for large vocabularies. Moreover, they remain sensitive to spelling variations and informal language, which are common in SMS communication.

(Juluru et al., 2021) implemented a decision support tool to predict appropriate radiologic examinations using machine learning. They used Bag of Words, TF-IDF and N-grams techniques to create word embeddings that were passed on to a classifier. They noted these techniques have several challenges including loss of meaning of the words as well as loss of data due to pre-processing activities like removing stopping words and lemmatization. Since word order is not considered, there is also loss of context with these methods.

Although frequency-based representations have demonstrated reasonable performance in tasks such as spam detection and topic classification, their limitations are well documented in settings requiring deeper semantic understanding. In urgency detection for SMS-based mHealth systems, these methods struggle to capture implicit meaning, long-range dependencies, and conversational context across multiple messages. These shortcomings motivate the use of neural network–based and contextualized text representations, which are better suited to modelling the nuanced and context-dependent nature of clinical urgency in short patient messages.

2.3.2 Neural Network–Based and Neural Embedding Models for Text Representation

Neural network–based models have become central to modern natural language processing due to their ability to learn complex, nonlinear relationships from data. Unlike traditional frequency-based representations, neural models learn distributed embeddings that encode semantic and syntactic properties of text, enabling improved generalization across varied linguistic expressions. These approaches are particularly relevant for text classification tasks involving noisy, informal, and variable language, such as patient-generated SMS messages in mHealth systems.

Early neural embedding models focused on learning static vector representations of words or documents from large corpora. Word2Vec introduced efficient shallow neural architectures, such as Continuous Bag-of-Words and Skip-gram, to learn word embeddings based on local co-occurrence patterns (Grohe, 2020). While these embeddings capture semantic similarity effectively, they assign a single representation to each word regardless of context. In clinical SMS communication, where the meaning and urgency of a term may vary depending on prior interactions or symptom progression, such static representations are limited.

Doc2Vec extended this idea by learning embeddings for entire documents or paragraphs in addition to individual words. Although useful for longer and more structured text, Doc2Vec relies on sufficient contextual content to generate meaningful representations. In SMS-based mHealth settings, messages are typically short, fragmented, and conversational, reducing the effectiveness of document-level embeddings for capturing clinically relevant nuance.

To address the need for richer representations, neural network architectures were introduced to model compositional and sequential structure in text. Feedforward artificial neural networks (ANNs), when combined with static embeddings or bag-of-words features, can approximate nonlinear decision boundaries but lack an inherent mechanism for modelling word order or long-range dependencies. As a result, their applicability to urgency detection in short, context-dependent SMS messages is constrained.

Convolutional Neural Networks (CNNs) were subsequently applied to text classification tasks to capture local patterns through convolutional filters operating over word embeddings (Kim, 2014). CNNs have demonstrated strong performance in applications such as spam detection and sentiment analysis by identifying salient n-gram–like features and providing robustness to minor variations in input length. However, their focus on local context limits their ability to model temporal dependencies or evolving conversational context across multiple messages, which is often critical for identifying urgency in clinical triage scenarios.

Recurrent Neural Networks (RNNs) and their variants, including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, were developed to explicitly model sequential dependencies by maintaining a hidden state that propagates information across time steps (Mienye & Sun, 2024; Krichen, 2025). These architectures enable the modelling of word order and contextual dependencies within messages and have been widely applied in language modelling and sequence classification tasks. Despite these strengths, RNN-based models face practical limitations when applied to long or noisy conversational histories, including vanishing gradient effects, sensitivity to input length, limited parallelization, and increased computational cost. These challenges hinder their scalability and robustness in real-world mHealth systems, particularly in low-resource settings.

Sequence-to-sequence models further extended recurrent architectures through encoder–decoder frameworks capable of mapping variable-length input sequences to output representations (Min et al., 2023). Although effective in applications such as machine translation and summarization, their reliance on compressing input context into fixed-length representations limits their effectiveness for long, fragmented SMS conversations. Additionally, their computational complexity poses challenges for deployment in low-resource clinical environments.

ELMo represented a significant advancement by introducing contextualized word embeddings generated from deep bidirectional language models (Peters et al., 2018). Unlike static embeddings, ELMo produces representations that vary depending on surrounding text, allowing words to be interpreted dynamically based on context.

While this approach improved performance on several NLP tasks, ELMo relies on recurrent architectures, making it computationally intensive and less efficient for modelling long-range conversational context. More recent attention-based models have since surpassed ELMo in both performance and scalability.

Overall, neural embedding models and early neural network architectures represent important transitional stages in the evolution of text representation. While they improve upon frequency-based methods by capturing semantic similarity and limited contextual structure, they exhibit persistent limitations in modelling long-range dependencies, multilingual variability, and conversational context efficiently. These limitations motivate the adoption of attention-based and transformer architectures, which provide more flexible mechanisms for global context modelling and parallel processing, making them particularly well suited for urgency detection in SMS-based mHealth systems where meaning is often distributed across multiple short messages rather than contained within a single text instance.

2.3.3 Transformer-Based Models for Text Representation

Transformer-based models represent a major advancement in natural language processing by introducing attention mechanisms that enable direct modelling of global contextual relationships within text. Unlike recurrent or convolutional architectures, transformers process entire sequences in parallel, allowing them to capture long-range dependencies efficiently without relying on sequential hidden states. This architectural shift has made transformers the foundation of most state-of-the-art language models used in modern text classification tasks.

At the core of transformer models is the self-attention mechanism, which enables each token in a sequence to attend to all other tokens and weigh their relevance when constructing contextual representations (Vaswani et al., 2017). The main problem of the sequence-to-sequence model is the need to take in the entire contents of the source sequence into a fixed size vector. If the text is longer, it is likely to lose some of the information in the text. The attention mechanism therefore, removes this bottleneck by enabling the decoder to check back at the sequence of all the hidden states of the source encoders and then provide a weighted average as additional input

to the decoder. Therefore, the model selects the context that best fits the current decoder node as input. The decoders do not directly utilize the hidden states given by all encoders as input, but they have a way to choose the hidden state that best matches the current node. It does this by calculating the value score of each hidden state and using a softmax activation function over the scores, which amplifies the hidden state with the highest correlation and diminishes the less relevant hidden states. This scoring mechanism is done at each step of the decoders. This design allows transformers to model subtle semantic relationships across distant parts of a text, making them particularly effective for tasks where meaning is distributed across multiple sentences or conversational turns. In SMS-based mHealth communication, urgency is often inferred from a combination of symptoms, temporal progression, and prior exchanges, rather than from a single message in isolation, making attention-based models well suited to this problem.

Bidirectional transformer models, such as BERT, further enhance contextual understanding by learning representations that incorporate both preceding and following context. Through large-scale pre-training on unlabelled text, these models acquire general linguistic knowledge that can be transferred to downstream tasks with relatively small amounts of labelled data. This transfer learning capability is especially valuable in low-resource clinical settings, where annotated datasets are limited and costly to obtain.

Multilingual transformer models extend this framework to multiple languages by learning shared representations across linguistically diverse corpora. Models such as multilingual BERT and related variants enable cross-lingual transfer, allowing knowledge learned from high-resource languages to benefit low-resource ones. This property is particularly relevant in African mHealth environments, where patient communication frequently involves code-switching, informal language, and underrepresented languages. However, the effectiveness of multilingual models depends on the extent to which local languages are represented during pre-training, highlighting the importance of domain adaptation and contextual fine-tuning. Despite their strengths, transformer-based models present practical challenges for deployment in real-world healthcare systems. Their large parameter sizes increase

computational requirements, and performance can degrade when applied to domains or languages that differ substantially from pre-training data. In low-resource settings, careful model selection, parameter tuning, and efficiency considerations are therefore essential to balance performance gains with operational constraints.

Overall, transformer architectures provide a flexible and powerful framework for modelling contextualized language in SMS-based mHealth systems. Their ability to capture long-range dependencies, leverage transfer learning, and support multilingual representation makes them particularly well suited for automated urgency detection in clinical communication, where accurate interpretation of context is critical for effective triage and prioritization. The figure 2.1 below shows the transformer model architecture (Vaswani et al., 2017).

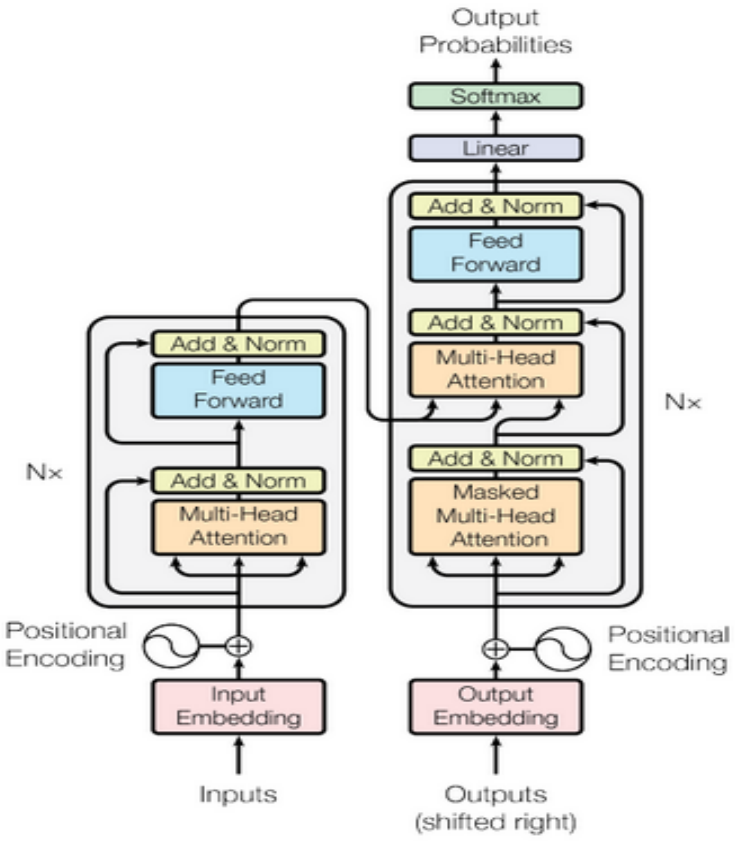

Figure 2.1: Transformer Model Architecture

2.3.3.1 Bidirectional Encoder Representations from Transformers

This is a type of language model that was developed by Google in late 2018. It is based on a multi-layer bidirectional Transformer architecture for fine tuning. It was built on the premise of a pre-trained language model that can be used for downstream NLP tasks using state-of-the-art transfer learning mechanisms. BERT was pre-trained on the entire Wikipedia text data of approximately 2,500 million words and a books corpus of 800 million words. BERT has been so effective and successful in so many application areas including, biomedical text classification (He et al., 2022), clinical notes (Huang et al., 2020), language translations and document classifications (Devlin et al., 2018).

2.3.3.1.1 How BERT Works

BERT is based on the Transformer, which is a special type of an Encoder-Decoder framework that is based on the Self Attention mechanism as discussed previously. Since BERT is a language model and its goal is to generate embeddings, only the encoder part is necessary. Nonetheless, BERT is a two-way deep neural network model that successfully applied the bidirectional training of Transformer to language modelling. BERT uses a technique called Masked Language Modelling that enables bidirectional training in models that was impossible before. BERT also follows a Universal Language Model Fine Tuning (Howard & Ruder, 2018) model that enables it to be trained on a large corpus to produce a generalizable language model that can then be fine-tuned with task specific training examples downstream. The generalized language model could then be applied to any natural language processing task and adopted as required. Therefore, there are two stages to using BERT, pre-training and fine tuning. In pre-training, the model is trained on unlabelled data on different tasks. Fine-tuning BERT means the model with the initialized pre-trained parameters is trained again, this time with labelled data for the specific downstream task. Each specific task will have a separate fine-tuned model, although they’re all initialized with the same pre-trained parameters. BERT boasts a unified architecture across different NLP tasks. The difference between pre-trained models and the final

downstream model is very minimal. Figure 2.2 below shows BERT pre-training and fine-tuning illustrations (Ghelani, 2019).

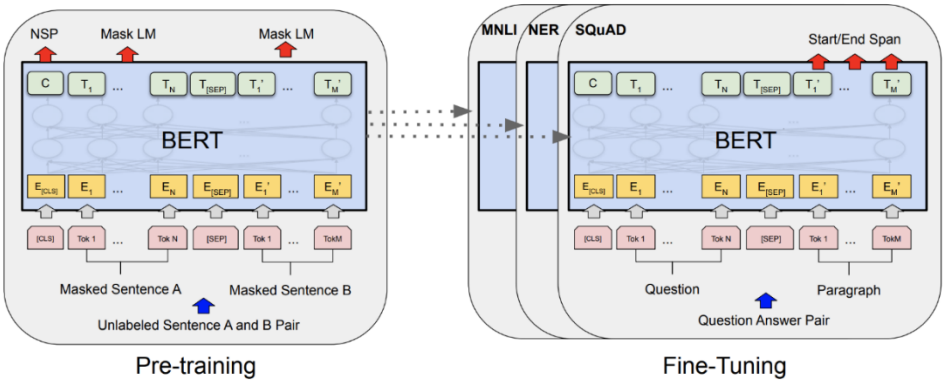

Figure 2.2: How BERT Works

2.3.3.1.2 BERT Pre-Training Process

The pre-training stage consists of two unsupervised predictive tasks, the Masked Language Model (MLM) and the Next Sentence Prediction (NSP). For the Masked Language Model, some percentage (around 15%) of the input tokens are masked at random to train a deep bidirectional representation. The masked tokens are then predicted to generate a hidden vector which is fed into a softmax function to get a probability distribution of the predictions. Since context of the masked words is read from both left and right sides of each word, the MLM enables the training of the deep two-way transformer to produce a bidirectional pre-trained model. The only challenge with this method is that there is going to be a mismatch between pre-training and fine-tuning, as the masked tokens to not appear in the final fine-tuning step. To mitigate this challenge, the BERT implementation recommends using a random replacement mechanism of the [MASK] token to its root. The training generator selects 15% of the tokens at random for prediction. For instance, if the i-th token is selected, it is either replaced with the [MASK] token 80% of the time or a random token 10% of the time or could be left unchanged in 10% of the time. Also, the loss function only takes into consideration the prediction of the masked values and ignores the prediction of the non-masked values. This makes the model to

converge at a much slower rate but thanks to an increment context-aware model that's more likely to have a much higher accuracy.

The Next Sentence Prediction (NSP) task is a training step during pre-training of BERT that provides for the model to learn about the relations between subsequent sentences in the data. It is done after the MLM training. The purpose for this is to train a model that understands sentence relationships as well as semantic relationships between words. This will make the model increase chances for adapting well for future downstream tasks such as question answering or natural language understanding. The next sentence is formulated by 50% of the proceeding sentence from the target sentence and the other 50% are randomly selected words from the corpus. This task involves learning the correlations between the words in the target sentence with those of the selected according to the criteria mentioned. When training the BERT model, MLM and NSP are trained together, with the main goal of minimizing the combined loss function of the two strategies.

2.3.3.1.3 BERT Model Input

The input to the BERT model can be either a single sentence or a sentence pair in a sequence of words (e.g. [question, answer]). The input representation of a word in the model is composed of three parts: 1) Token embeddings which represent the word vector where the first word is preceded by CLS flag. The flag can be used for subsequent classification tasks or ignored for non-classification tasks. 2) Segment Embeddings which are used to distinguish between two sentences as sometimes the embedding could be a classification task with two input sentences. 3) Position Embeddings that encode word order. Figure 2.3 below is a visual representation of the Embeddings showing positional, segment and token embeddings (Ghelani, 2019

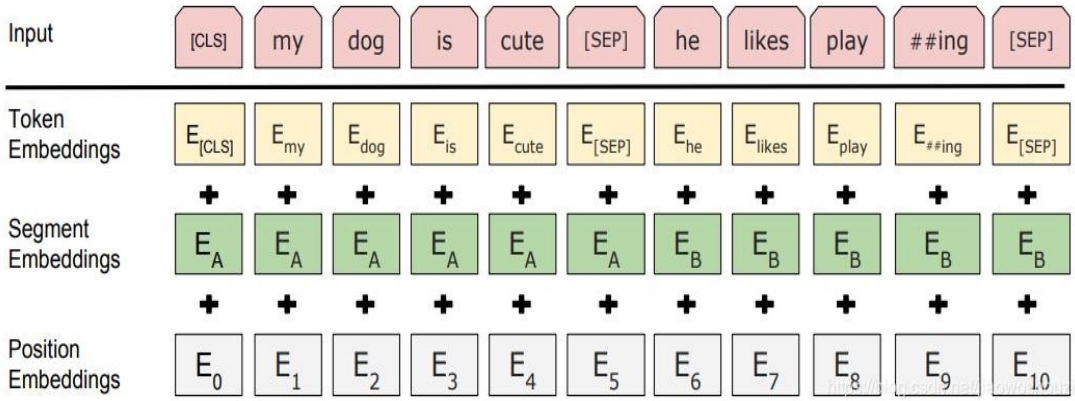

Figure 2.3: Embedding Representation in BERT

2.3.3.1.4 BERT Fine-Tuning for Downstream NLP Tasks

Once BERT has been pre-trained, the last training stage is to fine-tune with the specific task data. Fine-tuning in this case means to expose a sample of the specific dataset with training example and showing it to BERT (both data and labels) so as it can adjust its learned parameters hence adapt well to the current task. Compared to pre-training, fine-tuning is faster and consumes less resources. BERT can be utilized in a wide variety of NLP tasks, all while just adding a small layer to the core model. Classification tasks such as the one proposed in this thesis for detecting level of urgency in SMS messages from patients can be done by simply adding a layer on top of the Transformer output and showing the training model to understand the data and labels. Then it can be staged with the actual task of predicting for unseen data of the same task.

2.3.3.1.5 BERT for Feature Extraction

The good news with BERT is that the pre-trained contextualized word embeddings are already embeddings that can be used in downstream tasks as input to external classifiers. There is no need to fine-tune the model once it has been pre-trained if it's to be used as a language model. This way, BERT can be used as a tool for feature extraction for further machine learning tasks.

2.3.3.2 Cross-lingual Language Model

XLM was released as a Transformer architecture, having been trained using GPU clusters to optimize efficiency and applying optimizations like float16 precision to conserve memory. The training data included Wikipedia corpora for unsupervised learning and OPUS corpora for parallel datasets. Byte Pair Encoding (BPE) method was used to create a shared sub word vocabulary across languages, hence helping in embedding alignment (Conneau & Lample, 2019).

XLM was pre-trained using both unsupervised and supervised methods. This enabled better multilingual understanding and translation using the unsupervised pre-training methods called causal language modelling and masked language modelling, as well as the novel supervised pre-training method called Translation Language Modelling (TLM). Using TLM, pairs of sentences in different languages were combined in parallel with masking on both sentences. XLM improved classification accuracy on XNLI dataset, a benchmark natural language inference in multiple languages (Conneau et al., 2018). Despite its success on multilingual tasks involving translations, it does not provide model stability in monolingual settings as does BERT. The Translation Language Modelling works better if parallel sentences exist to provide cross-lingual embeddings. However, by aligning embeddings with high-resource languages, it can better handle low-resource languages in cross lingual applications. This was demonstrated by a significant reduction in perplexity when Hindi, a high resource language was combined with Nepali, a low-resource language.

2.3.3.3 XLM RoBERTa

XLM-RoBERTa (XLM-R) is a transformer-based model that was trained on a large dataset from CommonCrawl with around 100 languages (size: 2.5TB) (Conneau et al., 2020). XLM-R was built on RoBERTa (Liu et al., 2019) using multilingual masked language modelling and had a large, shared vocabulary from all languages with SentencePiece tokenization. This model achieved state-of-the-art results in cross-lingual inference (XNLI), also on named entity recognition and multilingual question answering (MLQA), subsequently outperforming mBERT and its parent model XLM. This model is also known to address the curse of multilinguality, a

capacity dilution issue as the number of languages scales, which impacts on low-resource languages. It handles this issue by increasing model size and using larger, high-quality datasets.

2.3.4 Large Language Model Families

From BERT, to XLM, to XLM-R, we see an increase in model size and number of parameters to increase bar of performance. The models are thus becoming bigger, larger, and requiring more computational resources to train and adopt. This section looks at major and popular language models which are transformer based but have much larger architectures (layers and parameters) and were trained on larger datasets. They are called Large Language Models (LLM). Large Language Models therefore are transformer based neural network language models that have tens to hundreds of billions of parameters and are pre-trained on massive textual data. We cover the major LLMs in this section. Figure 2.4 shows a summary of these models, classified into 3 main groups (families): OpenAI's Generative Pre-trained Transformers (GPT), Meta's Llama and Google's PaLM models (Minaee et al., 2024).

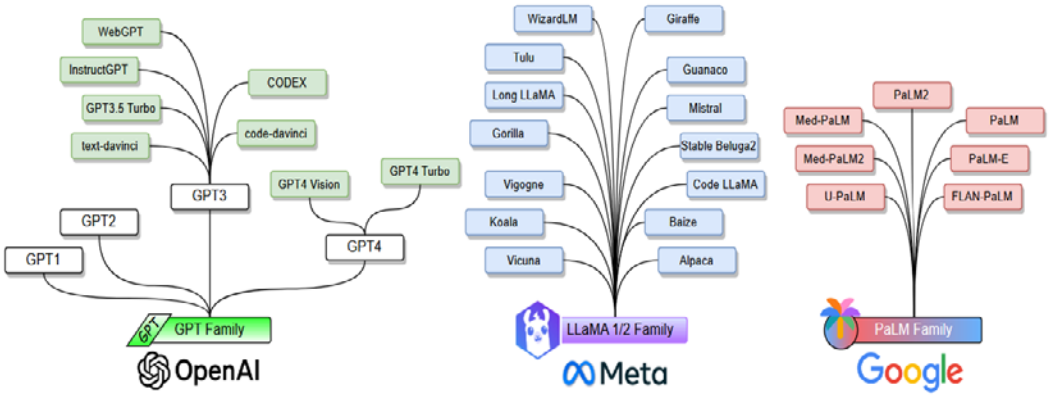

Figure 2.4 : Large Language Model Families

2.3.4.1 Generative Pretrained Transformer

The Generative Pre-Trained Transformer (GPT) is a deep neural network model based on the transformer architecture like BERT. The key distinction between GPT

and BERT is that while BERT employs both Encoder-Decoder architecture, GPT replaces the Encoders with Decoders. Like BERT, GPT is trained on a vast corpus of text data from sources like book archives and the internet, and it can also be fine-tuned for specific downstream tasks like language generation, sentiment analysis, language modelling, machine translation and text classification. It was first developed by OpenAI to give systems intelligence and has been widely used in the ChatGPT project (Radford & Narasimhan, 2018).

It uses a self-attention layer to consider the context of the entire sequence when predicting the next word. This improves its ability to comprehend and generate language [citation needed]. GPT leverages on transfer learning and can be applied to diverse tasks as it does not require extensive training on task specific data. GPT was trained using unlabelled data on a language modelling task (with masking technique similar to BERT) to understand word relationships. Users can now simply fine-tune (supervised learning or Few-Shot Learning) for specific tasks thereby reducing the huge burden of annotating data for training. Additionally, GPT has the capacity to execute zero-shot performance on different tasks making it a state-of-the-art model in the natural language processing area.

The first version of GPT was very large, with 117 million parameters and 12 layers decoder architecture. It was very effective as it used transfer learning as its foundation and has since paved way to even more advanced models in generative pre-training using larger datasets and parameters (Radford & Narasimhan, 2018). The latest version from OpenAI is GPT 4 and is bigger and better as it is a multimodal (can take in text, images and audio) and is more reliable, creative, and able to handle much more nuanced instructions. It has over 100 trillion parameters with a much-increased input size of up to 32768 tokens. It outperformed its predecessor (GPT 3.5) by a landslide in the uniform bar exam (90% compared to 10%) and the Biology Olympiad (99% compared to 35%) and it is in the top 10% with human competitors. However, despite these achievements, one major challenge with GPT (as with other language models) is that it “hallucinates”. It makes genius guesses and sometimes this may lead to reason errors in its output. Therefore, if used in specific contexts, grounding is recommended (OpenAI et al., 2024).

2.3.4.1.1 How GPT Works

GPT is based on the Transformer model and uses a self-attention mechanism to process sequences of different lengths using direct decoder blocks (Radford et al., 2018; Yenduri et al., 2023). The GPT architecture comprises several key components. First, an input embedding layer maps tokens (such as words or image input tokens) into dense vector representations. Positional encodings are then added to the embeddings to incorporate information about the relative position of tokens within a sequence. For language modelling tasks, masking is applied during training to prevent the model from accessing certain tokens in a sequence, enabling the model to learn predictive relationships between tokens. The architecture also includes multiple transformers block that function as decoder layers, each containing self-attention mechanisms and feed-forward neural networks. Finally, a linear layer followed by a softmax function converts the model's output into probability distributions over possible tokens, enabling the model to determine the most likely next token in the sequence.

The output of the multi-head attention layers is also transformed using these functions in the feed-forward layers. Figure 2.18 shows the GPT architecture visually (Yenduri et al., 2023).

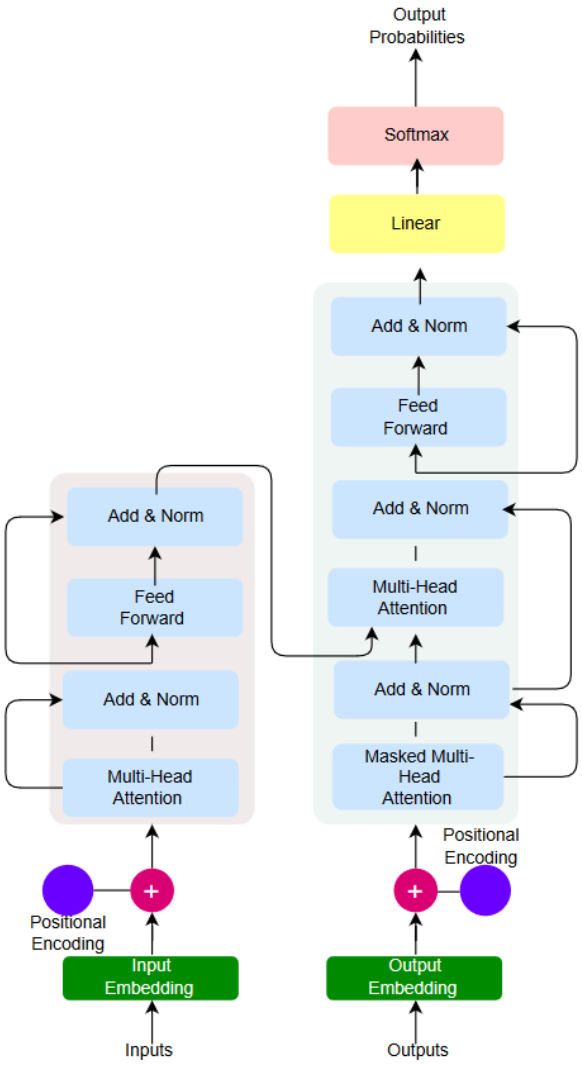

Figure 2.5: GPT Architecture

During pre-training, the model gains the ability to anticipate the next word based on the subsequent words. The model learns the statistical connections between words in a process of unsupervised learning. After pre-training, the model can be fine-tuned for specific tasks such as text production and text classification. Fine-tuning involves a smaller dataset that is specific to the task at hand and often involve data that is from the same domain area as the task. The model's weights/parameters are adjusted to maximize performance on the task.

2.3.4.2 Llama 3 Herd of Models (Llama Family)

LLaMA, an open-source language model family by Meta, offers models of various sizes (7B to 65B parameters) trained on publicly available datasets. Unlike GPT models, LLaMA's open-source nature allows widespread use in research for developing open and specialized LLMs. LLaMA uses a transformer-based architecture with optimizations like SwiGLU activations and rotary embeddings, outperforming GPT-3 in benchmarks. The LLaMA-2 collection, including chat-specific models, was released in collaboration with Microsoft, using RLHF and supervised fine-tuning, which enables competitive performance against other open-source and proprietary models (Touvron et al., 2023).

The latest model in the LLaMA series, LLaMA 3, offers several enhancements over its predecessors, making it one of the most advanced open-source large language models. It’s available in sizes up to 405 billion parameters, supporting a broad range of capabilities across multiple languages. LLaMA 3 utilizes an expanded vocabulary of 128,256 tokens and incorporates Grouped-Query Attention (GQA), enhancing its performance and scalability with a maximum context length of 128K tokens—useful for tasks requiring extensive context, like long document summarization and complex reasoning.

The model was trained on an enormous dataset of 15 trillion tokens across multiple languages, improving multilingual abilities significantly, especially for non-English languages. Instruction fine-tuning techniques, such as supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF), optimize LLaMA 3 for dialogue, reasoning, and coding. The model also underwent rigorous safety evaluations, including red-teaming and adversarial testing, to enhance responsible usage in real-world applications. LLaMA 3’s architecture and open-source availability make it highly flexible for both research and commercial applications, giving developers considerable control for customization and deployment in specific domains (Dubey et al., 2024).

2.3.4.3 Pathways Language Model (The PaLM Family)

The PaLM (Pathways Language Model) series by Google, first launched in April 2022, includes large-scale transformer models up to 540 billion parameters (Chowdhery et al., 2022). It is trained on 780 billion high-quality tokens across extensive TPU v4 chips using Google's Pathways system. It achieves state-of-the-art performance in few-shot learning on various language tasks and even rivals human performance on the BIG-bench benchmark. The U-PaLM models, developed in 8B, 62B, and 540B versions, introduce continued training with UL2R for computational efficiency. Subsequently, instruction-tuned following capabilities; for example, Flan-PaLM-540B surpasses PaLM-540B by over 9% on average across tasks, thanks to a dataset comprising 473 datasets and 1,836 tasks across 146 categories (Chowdhery et al., 2023).

2.3.4.4 Other Large Language Models

In addition to the well-known LLaMA, PaLM, and GPT series, several other influential large language models (LLMs) have been developed, each pioneering unique contributions to NLP. Gopher, created by DeepMind, features a sizable 280-billion parameter architecture tailored for tasks requiring extensive knowledge and nuanced reading comprehension, excelling in areas like factual recall and complex question answering. Chinchilla, also from DeepMind, takes a distinct approach to scaling by emphasizing data over model size using fewer parameters paired with a larger training dataset, which has been shown to enhance efficiency and performance across diverse benchmarks more effectively than larger, parameter-heavy models. LaMDA (Language Model for Dialogue Applications), developed by Google, focuses contextual coherence over extended conversations, a valuable trait for applications in conversational AI. Another unique model from Google, GLaM (Generalist Language Model), employs a sparsely activated architecture that significantly reduces computational demands while sustaining high task performance, illustrating the potential of sparse architectures for optimizing model efficiency. Together, these models represent diverse design philosophies ranging from dense scaling to data-driven efficiency and sparse activation each pushing the

envelope in model scalability, resource efficiency, and contextual understanding within the field of large-scale NLP.

In addition to models like Gopher, Chinchilla, and LaMDA, Claude, developed by Anthropic, is another notable large language model. Named after Claude Shannon, a foundational figure in information theory, Claude is designed with a focus on safety and alignment. It incorporates Anthropic's research on "Constitutional AI," which uses a set of ethical and safety guidelines during training to help Claude generate responses that align with safe and user-friendly behavior. Claude models, such as Claude 1 and Claude 2, have demonstrated competitive performance in language understanding, reasoning, and dialogue, emphasizing controllable responses that minimize harmful or biased outputs.

With Claude, Anthropic contributes to the growing landscape of safe and interpretable LLMs, aligning it with the goals of models like LaMDA, which focuses on conversational coherence, and GLaM, which explores sparse activation to enhance computational efficiency. These models collectively expand the range of approaches in LLMs, emphasizing varied strategies like sparse activations, ethical training protocols, and dialogue optimization to advance LLM utility and responsibility across applications.

2.3.5 NLP and mHealth in Low-Resource African and Kenyan Contexts

mHealth interventions have been widely adopted across sub-Saharan Africa as a means of extending healthcare services to underserved populations, leveraging the high penetration of mobile phones and the relative affordability of SMS-based communication (Lester et al., 2010; Barron et al., 2018). In Kenya, SMS has been extensively used in maternal, neonatal, and child health programs to support appointment reminders, health education, symptom reporting, and two-way communication between patients and healthcare providers (Mishra & others, 2023; Nordberg & others, 2024; Vatsa & colleagues, 2025). These interventions often rely on small teams of nurses or community health workers to manually review and respond to incoming messages, creating operational challenges as message volumes increase (Barron et al., 2018).

Textual data generated in African mHealth settings exhibit characteristics that distinguish them from datasets commonly used in mainstream NLP research. Patient messages are typically short, informal, and highly variable in structure, with frequent spelling variations, abbreviations, and non-standard grammar. In the Kenyan context, messages often involve code-switching between English and Kiswahili, as well as the use of colloquial or hybrid language forms. Such linguistic diversity complicates the application of NLP models trained predominantly on well-edited, high-resource language corpora (Joshi et al., 2020).

Research on NLP for African languages has historically been limited, largely due to scarcity of large, annotated datasets and the underrepresentation of these languages in widely used pre-training corpora (Nekoto et al., 2020). Although recent efforts have produced multilingual and African-focused language models, coverage remains uneven, particularly for informal language and domain-specific clinical communication (Adelani et al., 2021). As a result, models pre-trained on general web text may fail to capture the semantic nuances present in patient-generated SMS messages, potentially affecting performance in clinical decision-support tasks.

In Kenyan mHealth programs, urgency in patient communication is often expressed implicitly and unfolds across multiple interactions rather than through explicit emergency keywords. Cultural norms, health literacy, and communication practices influence how patients describe symptoms and distress, further complicating automated interpretation (Zhang et al., 2022). Consequently, effective urgency detection requires models that can integrate conversational context, tolerate linguistic noise, and generalize across mixed-language input, while remaining robust under limited supervision.

These constraints underscore the importance of evaluating NLP models within the specific operational and linguistic contexts in which they are deployed. Approaches that combine contextualized language representations with domain adaptation and clinically grounded evaluation criteria are particularly relevant for African mHealth systems (Joshi et al., 2020; Nekoto et al., 2020). By grounding urgency detection in real-world Kenyan SMS data and aligning model evaluation with nurse triage

workflows, this study contributes evidence toward the development of practical, context-aware NLP solutions for low-resource healthcare environments.

2.3.6 Transfer Learning in Low-Resource Clinical NLP

Transfer learning has become a foundational approach in modern natural language processing, particularly in domains where labelled data are limited or costly to obtain. In NLP, transfer learning typically involves pre-training language models on large, general-purpose text corpora, followed by fine-tuning on smaller, task-specific datasets. This paradigm allows models to reuse linguistic knowledge acquired during pre-training, reducing the data requirements for downstream tasks.

In clinical and mHealth applications, transfer learning is especially valuable due to the scarcity of annotated medical text and the sensitivity of patient data. Training models from scratch is often impractical, both in terms of data availability and computational resources. Pre-trained language models therefore provide a practical mechanism for adapting general language understanding to domain-specific tasks such as clinical classification, triage, and risk detection.

The relevance of transfer learning is further amplified in low-resource and multilingual settings, where many local languages are underrepresented in publicly available corpora. Multilingual pre-trained models enable cross-lingual transfer, allowing representations learned from high-resource languages to support downstream tasks in low-resource languages. However, the effectiveness of this transfer depends on linguistic similarity, representation coverage during pre-training, and the extent of domain mismatch between training and target data.

Despite its advantages, transfer learning introduces several challenges in clinical NLP. Pre-trained models may encode biases from their source corpora and may not adequately represent informal, code-switched, or domain-specific language typical of SMS-based health communication. Additionally, fine-tuning large pre-trained models on small datasets increases the risk of overfitting, particularly when class distributions are highly imbalanced. These limitations highlight the need for careful

adaptation strategies and evaluation frameworks that reflect real-world clinical utility.

In the context of urgency detection in SMS-based mHealth systems, transfer learning offers a means to leverage powerful contextual language representations while remaining feasible under data and resource constraints. However, its success depends on aligning pre-trained representations with conversational context, clinical relevance, and deployment realities. These considerations motivate the selective use of multilingual pre-trained models and context-aware fine-tuning strategies adopted in this study. Both GPT and BERT have utilised transfer learning extensively in their approaches and have since become superior models for all downstream NLP tasks (Devlin et al., 2018; OpenAI et al., 2024).

2.3.6.1 Types of Transfer Learning

There are different types of transfer learning approaches, depending on the relationship between source and target task or domain. The approach to be selected often depend on available data and similarity between the tasks.

2.3.6.1.1 Domain Adaptation

This type of transfer learning focuses on transferring knowledge from a much-labelled domain to a target domain that does not have much labelled data. This approach is particularly useful in scenarios where gathering labelled data for the target domain is difficult, such as in real-world text classification tasks, where a large volume of incoming messages remains unlabelled. There are two variations of this type, a supervised domain adaptation approach which uses a small amount of labelled data in the target domain and an unsupervised domain adaptation approach where target domain has no labelled data. Techniques like Domain Adversarial Neural Networks (DANN) and Maximum Mean Discrepancy (MMD) are often used for the latter (unsupervised) approach.

2.3.6.1.2 Cross-Lingual Transfer Learning

Cross-Lingual Transfer Learning (CLTL) addresses the challenge of training language models that can operate effectively across multiple languages. This method allows models to transfer learned linguistic structures and relationships from one language (often a high-resource language like English) to another, supporting NLP tasks in languages with limited labelled data. CLTL has become a central focus in multilingual AI development, especially as the demand for language technology grows globally. Cross-lingual transfer is commonly achieved through shared embedding spaces, where words or sentences from different languages are mapped onto a unified vector space. Models such as mBERT and XLM-R utilize multilingual embeddings to achieve cross-lingual representation.

2.3.6.1.3 Inductive Transfer Learning

In inductive transfer learning, the source and target tasks are different, even though they may share the same feature space or domain. Here, the target task has labelled data but benefits from knowledge learned in the source task. Inductive transfer learning is widely applied in tasks where the primary aim is to transfer generalized knowledge for example from a broad image classification task to a specific one. It is common in domains like healthcare, where a model trained on general disease diagnostics can be fine-tuned to specialize in a specific condition with smaller datasets.

2.3.6.1.4 Multi-Task Learning

Multi-task learning (MTL) focuses on training a model simultaneously on multiple tasks, encouraging it to learn shared representations that improve generalization. The assumption is that shared information across tasks can be beneficial, enhancing performance across all tasks due to joint learning. In MTL, a common approach is hard parameter sharing, where early layers of the network are shared, and task-specific layers are added for each task. Alternatively, soft parameter sharing allows each task its own model but penalizes the difference between task-specific parameters to encourage similarity.

2.3.6.1.5 Self-Taught Learning

Self-taught learning involves using unsupervised learning to learn feature representations from a source domain with abundant unlabeled data, which is then transferred to a supervised task in a target domain. Unlike other methods, self-taught learning doesn't require similarity between source and target domains, making it more versatile.

2.3.6.1.6 Unsupervised Transfer Learning

In unsupervised transfer learning, the goal is to transfer knowledge when labelled data is unavailable in both the source and target domains. Models leverage structural similarities in data distributions, which is useful for tasks like clustering or dimensionality reduction, where labels are often limited or unnecessary. This type of TL is less common but provides an important framework in exploratory data analysis and semi-supervised settings.

2.3.6.1.7 Transductive Transfer Learning

Transductive transfer learning focuses on cases where the source and target tasks are the same, but the domains differ. This approach is beneficial when a model trained on one domain (e.g., recognizing handwritten English characters) is adapted to work on another domain (such as recognizing handwritten Arabic characters). Techniques such as domain adaptation are often employed, where the model learns to generalize across different domains without task-specific modifications.

2.3.6.2 Methods in Transfer Learning

Several key methods are used to implement transfer learning, each optimized for specific scenarios and types of data.

2.3.6.2.1 Fine-Tuning Pre-Trained Models

Fine-tuning is a widely-used Transfer Learning method where a model is initially trained on a large dataset, after which its parameters are adjusted to optimize

performance on a target task with a smaller dataset. The most common approach is to freeze some layers of the pre-trained model to retain foundational knowledge, while training the remaining layers on the target task. For instance, in NLP, models like BERT are pre-trained on massive corpora, then fine-tuned on domain-specific tasks like sentiment analysis or question answering.

2.3.6.2.2 Domain Adaptation

Domain adaptation addresses situations where there is a discrepancy between the source and target data distributions. Common techniques include instance re-weighting, where data points in the source domain are weighted based on their relevance to the target domain, and adversarial training, where the model learns to minimize differences between the domains. This method is crucial in applications such as autonomous driving, where models trained in one environment like city streets need to adapt to new, unseen environments like rural roads.

2.3.6.2.3 Zero-Shot and Few-Shot Learning

Zero-shot and few-shot learning enable models to recognize new classes or tasks with little to no labelled data. In zero-shot learning, the model relies on semantic similarities between tasks to generalize from a source task to a new one without labelled examples (J. Chen et al., 2021). Few-shot learning, on the other hand, uses a minimal number of labelled samples (Song et al., 2022). Both methods are essential in fields where data collection is costly or impractical, such as medical diagnostics.

2.3.6.2.4 Representation Learning

In representation learning, the goal is to extract reusable features from the source task that are broadly applicable to other tasks. These learned representations can capture high-level patterns, such as edges or shapes in images, which are transferable across similar tasks or domains. This approach is especially common in image processing, where foundational representations can simplify learning in new tasks by providing a solid feature basis.

2.3.6.3 Challenges in Transfer Learning

While Transfer Learning offers significant advantages, it also faces several challenges: 1) Domain Shift: Domain shift occurs when there are major discrepancies between the source and target domains, which can lead to performance degradation. In extreme cases, models may fail to generalize effectively, a phenomenon known as negative transfer. Methods to mitigate this issue include advanced domain adaptation techniques and regularization. 2) Negative Transfer: This arises when knowledge from the source task adversely impacts the target task. This typically happens when there is low similarity between tasks or domains, causing the model's pre-existing knowledge to interfere with its learning process. Identifying when transfer learning is beneficial or detrimental remains an open area of research. 3) Data Privacy: In fields such as healthcare and finance, data privacy is critical. When using TL, data from one domain must be transferred while maintaining privacy and security, which can complicate the transfer process. Federated learning and privacy-preserving techniques are being explored to address these issues. 4) Resource Demands: Pre-training large models on high-resource tasks is computationally expensive, requiring substantial resources. While Transfer Learning can reduce data requirements for downstream tasks, the initial cost can still be prohibitive. Research into more efficient pre-training processes, such as model compression and distillation, is ongoing.

2.3.7 Efficient Methods for Natural Language Processing

As the deep learning models become larger and larger, resources get more constrained and eventually either the budget for compute sky-rockets or the entire business of pre-training and fine-tuning and hosting models become unmanageable. It is for this reason that several considerations have been made to balance between computing power needs and performance. An efficient language processing model maintains manageable resource requirements with the task at hand. Efficiency can be enhanced across the NLP pipeline, including data efficiency, model design, training, and inference. Data efficiency approaches, such as data filtering, active learning, and curriculum learning, help to reduce the volume and complexity of data needed for

effective training, enabling more targeted use of resources. Additionally, methods like retrieval-augmented models and sparse modelling (e.g., mixture-of-experts models) are explored to increase efficiency in model design by reducing memory and computational demands. Training efficiency can be improved through pruning, quantization, and parameter-efficient fine-tuning techniques like adapters and prefix-tuning, which update fewer parameters while maintaining accuracy (Treviso et al., 2022).

The Green AI paper, (Schwartz et al., 2020) calls for environmentally conscious AI development by introducing "Green AI," which encourages efficiency as a core evaluation criterion alongside accuracy. This stands in contrast to "Red AI," a current trend in which immense computational resources are employed to maximize model accuracy without regard for environmental or financial costs. The authors argue that Red AI's emphasis on raw computational power has led to a significant increase in training costs, which have escalated 300,000-fold since 2012, with only marginal gains in model performance. This approach not only raises the environmental footprint of AI but also makes high-level research inaccessible to resource-constrained researchers.

To foster Green AI, Schwartz et al., 2020, advocate for reporting metrics like floating-point operations (FPO), energy consumption, and carbon emissions in research publications, which would provide a more comprehensive understanding of a model's efficiency. By shifting the focus to resource-conscious design, they envision an inclusive and sustainable research landscape, where efficiency is valued as much as accuracy. They suggest that leader boards and benchmarks incorporate efficiency metrics to encourage innovation in low-resource models.

2.4 Classification Models

In machine learning, classification is the problem of trying to identify the class/label a given item/object belongs. This is a supervised learning problem where the model is given data with known classes/labels, then it is trained to identify new data instances and relate it to a most likely label based on the previously learnt experiences. A brief description of the most common classification models follows:

2.4.1 Penalized Logistic Regression

This is used to predict the outcome of a binary dependent variable using multiple independent variables. Penalized means that the model will suppress towards zero features that have insignificant contribution of the predicted value (Austin et al., 2024). Doerken et al., (2019) used penalized logistic regression algorithm to carry out classification task on low prevalence exposures in a high dimensional dataset. They wanted to find out if penalized logit models can perform better on estimating and selecting risk factors with extremely low prevalence for a binary outcome. They used data from the CESIR study which consisted of drivers who were involved in a road accident. The outcome of interest was whether the drivers were responsible for the accident or not. They compared unpenalized regression results with penalized ones for various penalization techniques. They used lasso, ridge, firth correction and boosting as penalized techniques with logistic regression. The results showed that all of these techniques improved the estimation for low prevalence outcome. Further they showed that predictions for low prevalence outcomes of 0.1% can be greatly improved by using firth correction and boosting.

2.4.2 Gradient Boosting Classifiers

Boosting is a type of ensemble learning (many models working together) where the model is built in series. In each successive model, the weights are adjusted based on the learning of previous model. Therefore, it works on reducing errors sequentially thus trying to develop a new sequential model for hard to fit data (Emami & Martínez-Muñoz, 2024).

Gradient boosting often provides predictive accuracy that cannot be beaten by other models. It has a lot of freedom as it can maximize on different cost functions. The other advantage is that it does not rely or need data pre-processing. There are 3 types of boosting namely, Adaboost, Gradient Tree Boosting and XGBoost. Adaboost – Short form for “Adaptive Boosting”. This model starts with a weaker classifier and fits weighted versions of the data iteratively. At each cycle, the data is re-weighted in that poorly classified elements get larger weights. The advantage of this algorithm is that it minimizes the exponential loss at each iteration (Emami & Martínez-Muñoz,

2024). Gradient Tree Boosting uses decision trees to compute residuals. It starts by constructing a very simple tree, then each tree in the iteration is built for the prediction residuals of the preceding trees. XGBoost, stands for, extreme gradient boosting, is an implementation of gradient tree boosting that has been optimized for speed and performance.

Chen and Pan, (2018) performed classification of diabetes mellitus data using boosting algorithms. They used clinical diagnosis and tests data from 10,000 patients in First Affiliated Hospital of Wenzhou Medical University. The patients were diagnosed and hospitalized between July 2004 and April 2014. They used two algorithms to build the model, AdaBoost and LogitBoost. The overall model performed very successfully with a best score of 95.3% accuracy score using 10-fold validation. The model was very robust and had a pre-diagnosis function. They showed that the model did feature selection and the statistically significant features were noted and could be used as reference risk factors for diabetes mellitus.

2.4.3 Random Forests

This is a Meta tree ensemble that fits a number of decisions trees and uses averaging to enhance the predictive accuracy (Schonlau & Zou, 2020). Each tree in the forest gives a prediction and the highest voted class by the trees becomes the model prediction for that instance. The main feature of this algorithm is that the trees protect each other from individual errors.

2.4.4 Naïve Bayes

This is a probabilistic classification model that is based on the Bayes theorem. The Bayes theorem states that: The probability of A given B has occurred can be calculated as the product of the probability of B given A and the probability of A divided by the probability of B.

It is called Naive since all features are treated as independent. There are 3 main types of Naive Bayes classifiers: - 1) Multinomial Naive Bayes, which mostly used in document classification using frequency of words as features (large corpus), 2)

Bernaulli Naive Bayes, which deals with mainly Boolean variables, and 3) Gaussian Naive Bayes for continuous non discrete variables. The main drawback of the naive Bayes family of classifiers is that they require existence of evidence occurrence of a class.

Other classification algorithms include K-nearest neighbour and support vector machines. The latter is an implementation from statistical learning theory which deals with building consistent estimators from data. They build separating boundaries from datasets by solving contrained quadratic problems (Brennan et al., 2018). K-nearest neighbour is very unique from all the other models discussed thus far. It is an unsupervised learning model that does classification from the data directly without building a model first. The main advantage of this approach is that the neighbours provide explanation as they're likely to bear similar characteristics. In mHealth, text classification has also been implemented in projects such as diagnosis classification from clinician notes, among others (Alsentzer et al., 2019).

2.5 The Critiques of the Existing Literature Relevant to the Study

Word embeddings are very critical to the success of natural language processing solutions. Recent research advancements in this area have shown a lot of promise for developing very rich representations hence increased accuracy in the way machine learning algorithms compare with human level decision making process. Basic summarization techniques such as bag of words, n-grams and term frequency-document inverse frequency have all been used successfully in setups where context is not a sensitive requirement. However, word order and context do play a lot of significance in advanced natural language processing such as extracting meaning and decision-making tasks such as the research problem presented in this thesis. Due to this, more advanced methods have been used in both research and practise, including use of neural networks to provide contextual representations down the pipeline during processing of word embeddings. However, models like word2vec, doc2vec and sequence to sequence models based on LSTM or GRU networks have not been very successful at providing full context of words as used in sentences. This is because of a common problem in these models, lack of ability to retain word

relevance for long span vector representations. Recent research has addressed this problem by implementing an Attention layer within the layers of the recurrent networks for LSTM and GRU or as a stand-alone layer on CNN models. Better still, the authors of “Attention is all you need”, a novel paper that introduces stand-alone Attention layers for deep neural networks demonstrate an even better algorithm that has overcome this problem. The BERT model was born from this concept.

2.6 Summary of Literature and Research Gaps

The reviewed literature demonstrates substantial progress in automated text classification, particularly using neural network–based representations and transformer architectures that model contextual dependencies in language. Prior studies have shown that traditional feature-based approaches, including n-grams and static word embeddings, are effective for general classification tasks but struggle with short, ambiguous, and context-dependent messages such as those commonly observed in SMS-based health communication. More recent transformer-based models, including BERT and its multilingual variants, address several of these limitations by leveraging contextualized embeddings; however, their application has largely focused on high-resource languages and well-structured text corpora.

In the domain of urgency detection and clinical message triage, existing work has primarily examined crisis-related social media data, emergency call transcripts, or synthetic benchmark datasets. While informative, these settings differ significantly from real-world mHealth environments, where messages are brief, multilingual, noisy, and often dependent on longitudinal conversational context. Furthermore, many studies emphasize overall classification accuracy or macro-averaged performance metrics, with limited attention to clinical utility, particularly the cost of false negatives in urgent message detection.

Notably, there is limited empirical research addressing urgency detection in low-resource African contexts, where language mixing (e.g., English, Kiswahili, and informal dialects), constrained annotation resources, and high nurse workload present unique challenges. Existing multilingual models assume sufficient representation of African languages during pre-training, an assumption that remains weakly validated

in clinical SMS settings. Additionally, few studies explicitly incorporate conversational or historical context into model inputs, despite evidence that urgency in short messages is often inferred from prior patient–provider interactions rather than message content alone.

Consequently, a clear research gap exists at the intersection of contextual language modelling, low-resource multilingual NLP, and clinically grounded urgency detection in mHealth systems. This study addresses this gap by systematically evaluating traditional and transformer-based text representations on a real-world Kenyan clinical SMS dataset, explicitly incorporating conversational context, and assessing model performance using metrics aligned with nurse triage and prioritization needs. By grounding model evaluation in clinical utility rather than abstract accuracy alone, the study contributes practical and methodological insights toward scalable, context-aware urgency detection in low-resource healthcare environments.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter describes the methodological approach used to develop and evaluate a machine learning model for detecting urgency in patient-generated SMS messages within an mHealth setting. The methodology is designed to address the research gaps identified in Chapter Two, particularly the need for context-aware language modelling, robustness in low-resource and multilingual environments, and evaluation criteria aligned with clinical triage workflows.

The study adopts a supervised learning framework, leveraging manually labelled SMS data collected from a real-world Kenyan mHealth program. Given the short, informal, and context-dependent nature of SMS communication, the methodology emphasizes contextual representation of messages, careful handling of class imbalance, and evaluation strategies that reflect the operational needs of healthcare providers. This chapter details the research design, data sources, labelling procedures, pre-processing steps, model selection, experimental setup, and evaluation framework used to assess model performance.

3.2 Research Design

This study employs an experimental research design based on supervised machine learning for text classification. The primary objective is to classify incoming patient SMS messages according to urgency, using both traditional baseline models and advanced contextual language models. The design enables systematic comparison of different text representation strategies and model architectures under consistent data and evaluation conditions.

Urgency detection was formulated primarily as a binary classification task, distinguishing urgent from non-urgent messages. This decision was motivated by the operational realities of mHealth nurse triage, where the immediate need is to identify

messages requiring prompt clinical attention rather than to assign fine-grained urgency categories. Multiclass classification was explored as a secondary analysis to assess the feasibility of more granular urgency categorization, but binary classification remained the core focus due to class imbalance and clinical interpretability considerations.

To reflect real-world usage, the research design incorporates context-aware modelling, where individual patient messages are evaluated not only in isolation but also in conjunction with preceding conversational context. This approach aligns with how healthcare providers interpret SMS communication in practice, relying on longitudinal engagement rather than single-message cues to infer urgency. Figure 3.1 shows an overview of the system architecture for this study.

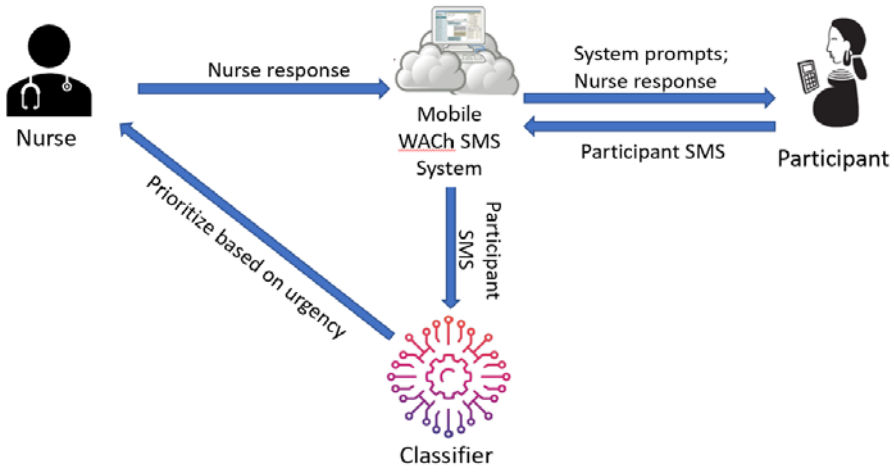

Figure 3.1: Task Definition

Model performance was evaluated using metrics selected for clinical relevance, with particular emphasis on precision and recall trade-offs. Rather than optimizing for overall accuracy alone, the study adopts evaluation strategies that distinguish between triage-oriented use cases, which prioritize high precision to reduce false alarms, and prioritization-oriented use cases, which emphasize high recall to minimize missed urgent cases. This design choice ensures that model evaluation aligns with practical decision-making needs in SMS-based mHealth systems.

3.3 Data Collection

The data used in this study were obtained from a real-world SMS-based mHealth system implemented in Kenya to support maternal, neonatal, and child health care. The system facilitates two-way communication between patients and healthcare providers, allowing participants to receive automated messages and respond with questions, symptom descriptions, or requests for assistance. Healthcare providers, primarily nurses, review incoming messages and respond as appropriate, forming longitudinal conversational threads between participants and the care team. We got data from the mWACH Neo Pilot study which was conducted between 05-12-2017 and 20-02-2019 (Unger, et al., 2019) and from the mWACH Neo RCT study which was conducted between 09-09-2020 and 04-05-2022 (Ronen, et al., 2021).

Messages in these studies were exchanged between the system, pregnant/postpartum women and study nurses. System messages typically included reminders, health education prompts, or follow-up requests; nurse messages provided clinical guidance or clarification; and participant messages contained free-text responses describing symptoms, concerns, or contextual information. The pilot study dataset contained a total of 58,834 messages from 800 participants, the automated system and 2 nurses. The RCT study was much larger with a total of 161,735 messages from the system, 1724 participants and 12 nurses. All the dataset contained a total of 220,560 messages from 2,523 participants and 14 nurses altogether after data cleaning.

System automated messages were sent to participants weekly during pregnancy upto and until 38 weeks gestation, then twice daily for the first fortnight after delivery, followed by a message every other day for the next 6 months thereafter. Participants were free to send messages anytime if they had anything to ask or report. Nurses had to dig deep and respond to every participant message. These nurses were study staff, although qualified as nurses in the public health facility, they did not have any routine care responsibilities on them.

A total of 112,220 (50.9%) messages were sent by the system, 65,572 (29.7%) by participants and 42,768 (19.4%) by staff as shown on Table 3.1. System messages were sent in all the 3 standard languages for the study (English, Swahili and Luo).

About 50% of all messages were in English, 36% in Swahili and 5% in Luo. Additionally, approximately 4.5% were code-switched and further 3% were sent in sheng (slang fusion). Data cleaning was conducted to remove salutations, location information, names, and any other identifiers. The system validation and authentication messages were also removed. Therefore, the total number of messages reported here were the final dataset after the cleaning exercise. Table 3.1 below shows a summary of the messages by source.

Table 3.1: Messages by Source

Sent By	Total Messages	Messages with less than 10 characters	Mean number of characters in a message (standard deviation)
Nurse	42,768 (19.4%)	2,500	97.9 (103.5)
Participant	65,572 (29.7%)	19,769	36.5 (39.8)
System	112,220 (50.9%)	0	257.3 (102.7)

Around a third of participant messages had less than 10 characters, suggesting many participant messages were short and depended on previous message context for detecting urgency. Every system message started with a salutation sentence that addressed the participant by name, introduced the nurse and the study clinic, then followed by the message topic of the day. The formatting of the messages was critical for disambiguation during communication and provided information about the participant, nurse or clinic from the message itself.

We observed that the dataset used here was a real representation of the actual way of communication in the local setup in Kenya as noted by (Mondal et al., 2021). For example, words used in the urban setup may have different connotations compared to same words in other parts of the country for instance Western Kenya. This data also contains Sheng and Luo messages which were not included in the original training of mBERT (Devlin, et al, 2019). Table 3.3 illustrates the breakdown of participant messages that were labelled by language.

To ensure ecological validity, this study prioritised the use of a real-world conversational dataset rather than relying solely on benchmark corpora commonly used in natural language processing research. While benchmark datasets provide standardized evaluation settings, they often lack the multilingual variability, conversational continuity, and clinically grounded urgency labels required for mHealth triage tasks. The mWACH dataset captures authentic patient–provider interactions, including informal language, code-switching, short message structures, and evolving conversational context, all of which are central to urgency detection in practice. This design choice reflects a deliberate trade-off between experimental control and real-world relevance. By training and evaluating models on operational messaging data, the study aims to produce findings that are more representative of deployment conditions, while acknowledging that direct comparison with benchmark-based studies may be limited. Consequently, the dataset supports investigation of context-aware modelling strategies that would be difficult to evaluate using traditional NLP benchmarks.

Ethical approval for secondary analysis of the data was obtained through the appropriate institutional review processes. The data were used strictly for research purposes, with all analyses conducted on anonymized text to ensure confidentiality and compliance with ethical standards.

3.4 Data Labelling

The urgency labels used in this study were derived through a manual annotation process designed to reflect real-world clinical decision-making in SMS-based mHealth triage. Given the absence of an objective gold standard for urgency in patient-generated text messages, nurse judgement was used as the reference for labelling, consistent with prior work in clinical NLP and mHealth systems.

Two nurses labelled a total of 11,129 messages from 772 participants. The labelling was based on a category of determining if participant messages were urgent or not. They labelled urgency based on how fast a participant message needed a reply by a nurse and used the key: 1) immediately, 2) within 2 hours, 3) before end of work day 4) by next day 5) no need to reply. The nurses were advised to use information from

prior messages for assessment of urgency of a given participant message. We then calculated the Cohen Kappa score to determine levels of agreement between the two nurses. We got an average score of 75%, indicating high degree of agreement. The 25% difference between the raters was resolved by consensus and a final labelled dataset was signed off. Then from the labelled data, we converted the five urgency categories into a binary label of urgent being categories one and two and not urgent being categories three, four and five.

As expected, urgency was imbalanced (2,383 out of 11,129 were labelled urgent). This translated to about 21% of all labelled messages. This was the correct scenario as in real life as most messages received were non-urgent. As this research project was interested in a model that would be clinically useful, we decided to leave the data as imbalanced. Table 3.2 shows a summary of the labelled participant messages.

Table 3.2: Labelled Participant Messages by Language

Language	Total Messages	Percentage
English	5646	50.7%
Swahili	3893	35.0%
Sheng	572	5.1%
Luo	566	5.1
Code-Switched	452	4.1
TOTAL	**11129**	**100%**

Whereas half of messages are in English (51%), Swahili and Luo have a total of 40% of the total messages.

3.5 Data Pre-processing

Data pre-processing was designed to preserve as much linguistic and contextual information as possible while ensuring compatibility with the selected machine learning models. Given the short, informal, and context-dependent nature of SMS-

based health communication, pre-processing choices were deliberately conservative to avoid removing cues that may be critical for urgency detection.

All SMS messages were converted to lowercase to reduce sparsity caused by capitalization differences. No spelling correction was applied, as spelling variations and informal orthography may carry meaningful signals of distress or urgency in patient-generated text. Similarly, punctuation was retained where supported by the model tokenizer, as it may contribute to the expression of emphasis or emotional tone.

Unlike traditional NLP pipelines, stop-word removal and lemmatization were not applied to messages used with contextualized language models. These steps were avoided to prevent loss of syntactic and semantic information, particularly in short messages where even common function words can influence meaning. For baseline models relying on frequency-based representations, minimal pre-processing was applied consistently to allow fair comparison, while preserving the original message structure as much as possible.

Tokenization was performed using the native tokenization schemes associated with each model. For transformer-based models, sub word tokenization (WordPiece) was employed to handle out-of-vocabulary terms, spelling variations, and code-switched language. Messages exceeding the maximum token length supported by the model were truncated, while shorter messages were padded as required. Given the brevity of SMS messages, truncation rarely affected core message content.

To address the limitation of single-message interpretation, conversational context was explicitly incorporated into model inputs. For each participant message, preceding messages exchanged between the participant and healthcare providers within the same conversation thread were concatenated to form a contextual input sequence. A space delimiter was used to separate messages, preserving message boundaries while allowing the model to attend across conversational turns. This approach reflects how nurses interpret patient messages in practice, relying on longitudinal engagement rather than isolated text.

Very short messages, including acknowledgements such as “ok” or “thanks,” were retained in the dataset, as their urgency can only be interpreted when viewed in context. Rather than filtering messages based on length, the study relied on contextual modelling to disambiguate such cases, ensuring that pre-processing did not bias the dataset toward longer or more explicit messages.

All pre-processing steps were applied uniformly across the dataset prior to data splitting to ensure consistency. The resulting pre-processed data preserve linguistic variability, conversational structure, and contextual dependencies, enabling subsequent models to learn urgency-relevant patterns under realistic conditions.

3.6 Model Selection

Model selection in this study was guided by the need to balance methodological rigor, clinical relevance, and practical constraints associated with low-resource mHealth settings. To this end, a combination of baseline machine learning models and advanced contextual language models was employed, enabling systematic comparison of text representation strategies under consistent data and evaluation conditions.

Baseline models were included to establish reference performance levels and to assess the added value of contextualized representations. These models rely on traditional frequency-based text representations and conventional classifiers, which are computationally efficient and widely used in earlier urgency detection and text classification studies. Although such approaches are limited in capturing semantic nuance and conversational context, they provide an interpretable benchmark against which more complex models can be evaluated.

To address the limitations of baseline methods, the study evaluates several transformer-based language models that generate contextualized representations of text. Transformer architectures were selected due to their ability to model long-range dependencies, handle variable-length input, and leverage transfer learning through large-scale pre-training. These properties are particularly relevant for SMS-based

health communication, where urgency is often inferred from subtle linguistic cues and prior conversational context rather than explicit keywords.

Multilingual pre-trained models were prioritized to accommodate the linguistic characteristics of the dataset, which includes English, Kiswahili, and mixed-language SMS messages. Multilingual BERT (mBERT) was selected as the primary contextual model due to its broad multilingual coverage, availability of pre-trained weights, and established use in low-resource NLP tasks. Its architecture supports sub word tokenization, enabling robust handling of spelling variation, informal language, and code-switching common in Kenyan SMS communication.

In addition to mBERT, SwahBERT and AfriBERT were evaluated to assess the impact of language- and region-specific pre-training on urgency detection performance. These models are designed to better represent African languages and linguistic patterns that are underrepresented in general-purpose multilingual corpora. Including these models allows comparison between globally pre-trained multilingual representations and regionally focused alternatives within the same experimental framework.

More computationally intensive models, such as XLM-R, were considered but not selected as primary candidates due to resource constraints and deployment considerations. While such models may offer performance advantages in some multilingual benchmarks, their larger size and training requirements pose challenges for reproducibility and practical use in low-resource clinical environments. We also note that mBERT is a better option compared to XLM-R on our task as it is a smaller multilingual language model trained solely on masked language modelling without the complexities of cross-lingual optimizations used in XLM-R (CONNEAU & Lample, 2019). This study therefore prioritizes models that strike a balance between performance, efficiency, and feasibility.

Additionally, mBERT’s ability to generalize with smaller datasets is valuable for handling domain-specific tasks like medical urgency detection, where libelled data is often scarce. Its moderate size reduces the risk of overfitting, common with larger models on small datasets, while still providing high accuracy and efficiency.

Therefore, mBERT presents a balanced solution, offering the benefits of BERT's robust transformer-based architecture with the practicality required for effective training and deployment on a relatively small, specialized dataset (Devlin et al., 2019a).

Overall, the selected models reflect a deliberate progression from simple baselines to advanced contextual representations, enabling clear attribution of performance gains to representation learning and contextual modelling. This approach supports robust evaluation of urgency detection methods while remaining aligned with the operational realities of SMS-based mHealth systems.

3.7 Experimental Set Up

The experimental setup was designed to ensure reproducibility, fair comparison across models, and alignment with the study's low-resource and clinical context. All models were trained and evaluated under consistent data splits and evaluation conditions, with hyper parameters selected to balance performance and computational feasibility.

3.7.1 Data Splitting and Validation Strategy

After the data was prepared, it was split into training, development and testing datasets. The first split involved training dataset versus the rest at 70% (7790 messages) for training. Then the test dataset was selected at 70% (2337 messages) of the remaining. The residual was used as a development dataset (1002 messages). All selections were stratified by the outcome variable (urgency) to preserve the original class distribution across splits. The training set was used for model fitting, the validation set for hyper parameter tuning and threshold selection, and the held-out test set for final performance evaluation. Messages from the same participant were restricted to a single split to prevent information leakage across datasets. Where additional pre-training was needed, we used the larger dataset including the unlabelled version.

Given the clinical nature of the task and the limited dataset size, cross-validation was not employed for transformer-based models due to computational constraints. Instead, performance stability was assessed through repeated experiments with fixed random seeds and an early drop-out rate, as described below.

3.7.2 Baseline Model Configuration

For baseline experiments, a penalized logistic regression classifier was implemented using the scikit-learn library. The model was trained using the L-BFGS optimizer, which is well suited for convex optimization problems and supports L2 regularization. The maximum number of training iterations was set to 570 to ensure convergence given the dimensionality of the feature space. All other parameters were retained at their default values as defined by the implementation. This configuration provided a stable and interpretable baseline against which the performance of contextual transformer-based models was compared.

3.7.3 Transformer Models Training Configuration

Transformer-based models were fine-tuned using the Hugging Face Transformers framework, employing the BERTF or Sequence Classification architecture for supervised urgency classification. Training was conducted using the framework's standardized training utilities to ensure consistency and reproducibility across experiments. Fine-tuning was performed using the Adam W optimizer, which is widely adopted for training pre-trained language models. Models were trained for a maximum of 5 epochs, with a learning rate set to 2e-5, and a per-device batch size of 16 for both training and evaluation. Weight decay regularization was applied with a coefficient of 0.01 to reduce overfitting, particularly given the limited size of the labelled dataset. Model evaluation was conducted at the end of each training epoch, enabling monitoring of performance trends and early detection of overfitting. A dropout rate of 0.1 was used within transformer layers, consistent with default model configurations. To ensure reproducibility, a fixed random seed was applied across data splitting, model initialization, and training procedures.

Training logs and model outputs were written to designated directories to support experiment tracking and reproducibility. All other training parameters not explicitly specified were retained at their framework-default values. All experiments were conducted using the same hardware and software environment to maintain consistency.

3.7.4 Implementation Environment

All experiments were implemented using the Python programming language, with software dependencies managed through a virtual environment (venv) to ensure consistency and reproducibility across experimental runs. Library versions were fixed within the virtual environment to prevent unintended changes in behaviour due to dependency updates.

Core machine learning and natural language processing components were implemented using the Hugging Face Transformers library for transformer-based models and scikit-learn for baseline machine learning models and evaluation utilities. Numerical computation and data manipulation were performed using NumPy and pandas, respectively. Model evaluation metrics were computed using functions from sklearn metrics.

Data visualization and exploratory analysis were conducted using matplotlib and sea born, enabling consistent generation of figures and performance plots. All experiments were executed within the same controlled software environment to ensure comparability across models and experimental conditions. This implementation setup supports reproducibility and facilitates future extension or deployment of the proposed urgency detection framework within similar low-resource clinical settings.

3.7.5 Experiments

Our experimentation work involved three types of input; firstly, we used context aware datasets as shown in table 3.3 to run experiments on a penalized logistic regression model which was our baseline model (Losada et al., 2020) and compared

it with multilingual BERT model classification. Secondly, we made our mBERT model the new baseline model, and compared with experiments conducted with additional pre-training activities. Finally, since we had dichotomized our initial outcome (urgency), we later conducted experiments with multiclass labels to compare with our dichotomized results. As for the models used, we had the penalized logistic regression model that provided us with a simple baseline to compare if multilingual deep learning models were better. We also used this logistic regression model to check the contribution of context to the performance of the model. Then we went for mBERT. Our target was to evaluate what kind of context-specific dataset contributed to a better performance of detecting urgency and so we used different combinations of context-based experiments to achieve our objective. Finally, we compared mBERT with other variants of BERT that were pre-trained using African languages and documented the results. Figure 3.2 below shows the conceptual framework for this study.

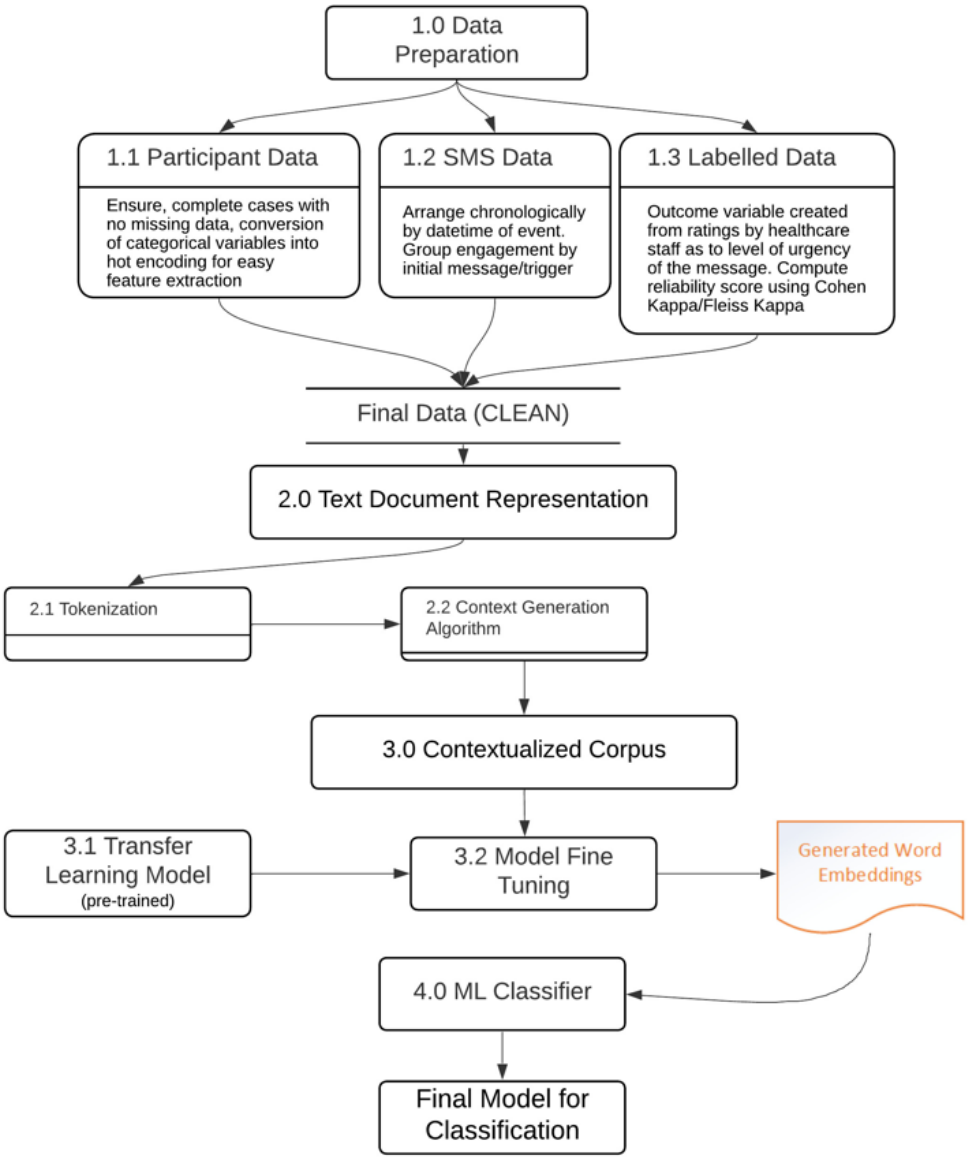

Figure 3.2: Conceptual Model

3.7.5.1 Adding Context

Many participant messages are short as shown in Table 3 and content like “thank you", "okay", "yes" and "no", may derive different interpretations depending on the context of the conversation. Previous work has found that including prior message context when analysing SMS messages can be helpful for understanding conversation trajectory and appropriate responses (Zhang & Danescu-Niculescu-

Mizil, 2020). We got this inspiration from this work and inspected whether adding preceding message context to participant messages would improve model performance.

We created this context by prepending the message (whether from nurse or system) preceding a participant message. We developed two versions of the dataset: one in which each participant message was prepended with the preceding system message (system context) and one in which each message was prepended with the preceding nurse message, or, in the event there was no nurse message, then the most recent system message (nurse context). Example messages are displayed in Table 3.3. We evaluated results for both the logistic regression and mBERT using these approaches.

Table 3.3: Sample Messages with Contexts

System Message	Nurse Message	Participant Message	Urgency Label
Bad swelling of hands and face or a bad headache are signs of a problem. Ask your family to take you to the clinic if they see this. Have you had any swelling or headaches during pregnancy? Are they getting worse?	Am happy you're fine. Do have a nice day	You To	0
Sometimes problems arise during labour and delivery, like bleeding or fits, which require skilled health workers, medications and equipment to treat. Without that treatment, the mother and baby could die. Therefore it is safest to deliver in a facility that can manage these and other problems. How will you arrive at the facility?	Are you still having the headache	yeah	1
Breastfeeding a baby right after birth helps the milk come. The first yellow sticky milk has many vitamins & cleans out the stomach. Milk has all the water the baby needs, avoid other liquids. Are you planning to breastfeed?	I understand. No worries.	Its OK I willthanks for your concerns	0

3.7.5.2 Additional Pre-training

It has been shown that additional in-domain and task-specific pre-training can increase model performance in a variety of settings (Gururangan et al., 2020). Since our dataset is different from the languages and application domains used in mBERT, we thought this may be particularly helpful in our task. We looked into two versions of pre-training. In the first approach, we pre-trained on all 49,786 participant messages (this included both labelled and unlabelled data). As with our approach for fine-tuning data, we tested pre-training with participant messages that were prepended with system messages or nurse messages. In the second approach, we used only the 11,129 labelled participant messages that were also prepended with system messages or nurse messages (Table 3.1). Note that in this second approach, we did not include the labels, only the text of the messages, for pre-training. We explored this method for the mBERT models.

We pre-trained with masked language modelling with 15% of the text masked and used a batch size of 4, with a maximum input sequence size of 512 for the multilingual BERT model. During fine tuning the models, we only played with batch size hyper parameter. The default parameters can be found at (huggingface.co).

3.7.5.3 African Language Models

Since our dataset mainly had English, Swahili, and Luo. We decided to use language models that had been pre-trained with African languages. We pre-trained AfriBERT (Sello Ralethe, 2020), which is a version of BERT model that was pre-trained for the Afrikaans language. AfriBERT did improve the performance of several Afrikaans based NLP tasks. We also pre-trained SwahBERT (Gati et al, 2022), a Swahili language version of BERT. Similar to AfriBERT, SwahBERT outperformed multilingual BERT on several Swahili based NLP tasks. We pretrained these monolingual African language models in the same way as we did with the multilingual BERT model. First using all data available (domain level) and then using labelled data (task level). We also added context as previously discussed.

3.7.5.4 Multi-label Classification

As part of our experiments, we decided to return our initial five labels on the outcome of interest (urgency). We repeated the same experiments we conducted for the binary classification, but now using a multiclass classification approach to compare results. We also reduced the categories (labels) from five to three (merged categories 1 & 2 and 4 & 5) and repeated the experiments again.

3.8 Evaluation Strategy

The evaluation strategy for this study was designed to reflect the operational realities of SMS-based mHealth triage, where the cost of misclassification is asymmetric and model performance must be interpreted in a clinical context. Rather than relying solely on overall accuracy, the study emphasizes evaluation metrics that capture trade-offs between identifying urgent cases and minimizing unnecessary clinical escalation.

Model performance was assessed using precision, recall, and F1-score, computed on the held-out test set. Precision measures the proportion of messages classified as urgent that are truly urgent, reflecting the model's ability to minimize false positives. High precision is particularly important in triage settings to avoid overwhelming healthcare providers with unnecessary alerts. Recall measures the proportion of truly urgent messages correctly identified by the model, capturing its ability to minimize false negatives. In the context of patient safety, missed urgent messages pose a significant risk, making recall a critical metric. The F1-score, defined as the harmonic mean of precision and recall, was used as a summary metric to balance these competing objectives. However, F1-score alone was not treated as sufficient for operational decision-making, as different deployment scenarios may prioritize precision or recall differently depending on clinical workflow constraints and staffing capacity.

Given the imbalanced nature of the dataset, with urgent messages comprising a minority class, accuracy was not used as a primary evaluation metric (Miftahushudur & others, 2025). High accuracy can be achieved by trivial classifiers that favour the

majority class, providing a misleading assessment of model utility in urgency detection tasks. Instead, metric interpretation focused on class-specific performance and clinically meaningful error trade-offs.

Although patient messaging often reflects a spectrum of clinical concern, this study operationalised urgency detection as a binary classification task to align with real-world triage workflows in the mHealth program. In practice, nurses must make rapid decisions regarding whether a message requires immediate attention or can be addressed within routine follow-up, making a binary urgent versus non-urgent distinction both operationally meaningful and implementable within decision-support systems. Framing the task in this way also enables clearer evaluation of recall–precision trade-offs, which are central to clinical risk management. To reflect the prioritization nature of triage, model performance was therefore examined across decision thresholds rather than relying solely on default classification outputs. This threshold-aware evaluation allows exploration of how models balance missed urgent messages (false negatives) against unnecessary escalations (false positives), supporting selection of operating points that favour sensitivity in safety-critical contexts. While multilabel or ordinal urgency scales may capture richer clinical nuance, the binary formulation provides a pragmatic foundation for deployment while remaining extensible to more granular prioritization strategies in future work.

To support threshold-based decision-making, model outputs were interpreted as urgency scores, allowing adjustment of classification thresholds to explore different operating points. This enables evaluation of model behaviour under alternative deployment scenarios, such as conservative triage with high precision or aggressive prioritization with high recall. Threshold selection was guided by validation-set performance and aligned with the intended clinical use case, as further described in the contextual triage framework.

All evaluation metrics were computed consistently across baseline and transformer-based models using standardized evaluation functions. Results were reported with sufficient detail to support transparent comparison and interpretation, with emphasis placed on clinical relevance rather than abstract model performance alone.

3.9 Contextual Triage and Prioritization Framework

To bridge model evaluation with real-world clinical decision-making, this study introduces a contextual triage and prioritization framework that interprets model outputs in relation to nurse workflow requirements. Rather than treating urgency detection as a purely statistical classification task, the framework aligns model behaviour with two complementary operational use cases: triage and prioritization.

In the triage use case, the primary objective is to flag messages that genuinely require clinical attention while minimizing unnecessary escalation. In this scenario, high precision is prioritized to reduce false positives that could overwhelm healthcare providers. A model operating under a triage configuration is therefore evaluated at thresholds that favor precision, ensuring that messages identified as urgent are highly likely to require action.

In contrast, the prioritization use case focuses on ensuring that urgent messages are reviewed promptly, even at the cost of increased false positives. Here, high recall is emphasized to minimize the risk of missing urgent cases. This configuration supports ranking or ordering of incoming messages so that those most likely to be urgent are reviewed first, aligning with clinical safety considerations in resource-constrained environments.

Model outputs were treated as continuous urgency scores rather than fixed class labels, enabling flexible adjustment of decision thresholds depending on operational needs. Thresholds were selected using validation-set performance to explore trade-offs between precision and recall under both use cases. This approach allows the same underlying model to support multiple deployment scenarios without retraining.

3.9.1The Triage Model

A good triage model is targeting at reducing the number of messages that the staff has to read by marking out non-urgent messages. An ideal triage model is purposed at reducing the number of messages that the healthcare staff need to read by marking out non-urgent messages. Therefore, it needs to cut back a large volume of non-

urgent messages to justify its existence (e.g., deployment costs or training nurses to use the system) while also permitting a small number of false negatives. We chose a threshold of 30%, which means, the model should at least assign non-urgent messages 30% of the times (correctly maintaining a perfect recall). Knowing the number of samples and the number of actual positives, we can describe this relationship between precision and recall:

$$precision_{triage} \geq \frac{recall_{triage} \cdot actualpositives}{datasize \cdot 70\%} \tag{0.1}$$

In this case, we encourage a high value for $recall_{triage}$ (95% in our case) and calculate a threshold for $precision_{triage}$. This creates our target triage region in the precision-recall graph that a triage model's precision-recall curve crosses as shown on figure 3.3 below.

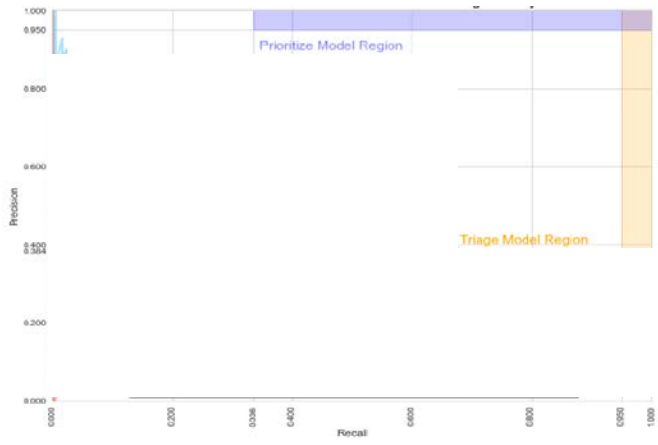

Figure 3.3: Precision-Recall Curve with Triage and Prioritize Regions

3.9 The Prioritize Model

A good prioritize model should be able to pick urgent messages that need a reply much faster than other messages. This is applicable when the staff need assistance on

which messages to read first. Typical for our work in this project. Since all messages will ultimately be read, the concern is not on reducing the false negatives, but we need to reduce false positives. We need to staff to trust our system so that for every urgent message, it should be indeed urgent. Therefore, we need this model to have a high precision and maintain a significant number of positive cases (recall). For our work, we picked a threshold of 10%. Meaning, the model should predict urgent messages atleast 10% of the time while maintaining a near perfect precision. We can also show this relationship between precision and recall for a prioritize model:

$$recall_{prioritize} \geq \frac{precision_{prioritize} \cdot datasize \cdot 10\%}{actualpositives} \tag{0.2}$$

In our case, we encourage a high $precision_{prioritize}$ (95% in our case) and work-out a threshold for $recall_{prioritize\}}$. This creates a region on the graph that prioritize a model's precision-recall curve crosses as shown in Figure 3.3 above.

3.10 The Combined Model

Lastly, we want to create a model that can meet both targets for triage and prioritize regions from our precision-recall curve. This model should have a high F1 score. We will call this model a combination model if it will cut across the overlapping region between the triage and prioritize regions.

This contextual triage and prioritization framework provides a structured mechanism for translating model performance metrics into actionable clinical decision support. It emphasizes interpretability, flexibility, and alignment with healthcare workflows, supporting the practical deployment of urgency detection models in SMS-based mHealth systems. This framework underpins the analysis and interpretation of results presented in Chapter Four.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the results of experiments conducted to evaluate machine learning models for detecting urgency in patient-generated SMS messages within an mHealth setting. The experiments were designed to assess the impact of text representation, model architecture, and conversational context on urgency classification performance, using the methodological framework described in Chapter Three.

4.2 Baseline Model Performance

A penalized logistic regression model, combined with bi-gram lexical features, was used as the primary baseline model. This model produced poor recall with none of the contextualized input reaching a recall of 30 as shown in table 4.1 below. The higher precision is subject to the size of the negative class (not urgent) which is large and therefore the ratio of false positives is smaller. We can also note that these base models (Bigrams) don’t largely predict false positives but the number of positive predictions is very small hence the higher precision but low recall.

Table 4.1 summarizes the performance of our models. mBERT models were superior to the Bigrams as they definitely perform better due to their deep neural network and using the attention layer as described in chapter 2. Our mBERT model using nurse context was the best with an F1 score of 47. We see a recall increase of eight points from the best bigram model. Generally, models incorporating nurse context perform better than system context models. Overall, the baseline model demonstrated low performance in distinguishing urgent from non-urgent messages. While the model was able to correctly classify a substantial proportion of non-urgent messages, its ability to reliably identify urgent cases was limited.

The limited recall values observed in baseline models highlights the challenges of urgency detection when using frequency-based representations alone. Short and

ambiguous SMS messages often lack explicit urgency keywords, and urgency cues may depend on prior conversational context or subtle linguistic patterns that are not captured by n-gram features. As a result, baseline models struggle to generalize beyond explicit lexical signals.

Table 4.1: Performance of Models on Different Contextualized Datasets

Model	Pre-Data	Pre-Context	FT Context	Precision	Recall	F1
Bigrams	-	-	none	51	20	29
Bigrams	-	-	system	58	29	39
Bigrams	-	-	nurse	59	29	39
mBERT	-	-	none	46	34	39
mBERT	-	-	system	50	27	35
mBERT	-	-	nurse	52	38	44
mBERT	labelled	system	system	50	32	39
mBERT	**labelled**	**nurse**	**nurse**	**50**	**45**	**47**
mBERT	unlabelled	system	system	49	39	44
mBERT	unlabelled	nurse	nurse	48	38	42

Note:

- Pre-Data – Type of dataset used for pretraining data. It was either labelled (Task Adaptive) or unlabelled (Domain Adaptive). A dash means no pretraining was done.
- Pre Context – Type of previous message concatenated during pretraining. A dash means no previous context was used.
- FT Context - Type of previous message concatenated during fine tuning. A dash means no previous context was used.
- Bolded text indicates best-performing model in terms of F1 score

Using the deep learning model (mBERT), we see a steady (but still low) value of recall. System context models do not contribute significantly to recall or precision. This is expected as all system messages were predefined and automatically sent to participants hence, we agree they cannot contribute to the urgency of participant messages significantly. That is why system context has a slightly better precision than no context. Moreover, with additional pretraining, we get a fair precision (about 50%). Nurse context models are performing better than without context or system context. This is because nurses would provide or ask leading questions and so participants would open more.

Overall, transformer-based models consistently outperformed the baseline model across recall and F1-score. These improvements were most pronounced in recall, indicating a substantially enhanced ability to identify urgent messages that were previously missed by frequency-based approaches. This finding supports the hypothesis that contextualized embeddings better capture the implicit and nuanced expressions of urgency present in patient-generated SMS messages.

In summary, transformer-based models substantially improve urgency detection performance over traditional baselines by capturing semantic nuance and contextual dependencies in SMS communication.

4.3 Effect of Conversational Context

Pre-training here refers to task-adaptive pre-training on the labelled data with matched context. For instance, system + pre-training is pre-training on participant messages prepended with system messages, using libelled data only. Baseline model was with no pre-training and no context added to the messages. For example, using nurse context and comparing from the baseline model, we see an increment of 11 points in recall (with 8 points F1 score) when pre-training and fine tuning with the nurse context as shown on table 4.2. Across all transformer-based models, the inclusion of conversational context resulted in consistent performance improvements, most notably in recall. Context-aware models demonstrated an increased ability to identify urgent messages that lacked explicit urgency indicators when viewed in isolation. This finding reflects the nature of SMS-based clinical communication,

where urgency often emerges through symptom progression or cumulative information distributed across multiple messages rather than within a single text. Table 4.2 below shows the contribution of various context and pre-training on mBERT model.

Table 4.2: Contribution of Different Context and Pre-training on mBERT Model

Metric	Baseline	System	Nurse	Nurse + pre-training	System + pre-training
Precision	36	50 (+4)	52 (+6)	50 (+4)	50 (+4)
Recall	34	27 (-7)	38 (+4)	45 (+11)	32 (-2)
F1	39	35 (-4)	44 (+5)	47 (+8)	39 (0)

Note:

- (+value) shows positive net gain of the score from the baseline value.
- (-value) shows net loss of the score from the baseline value

The magnitude of improvement varied by model, but the trend was consistent: context-aware configurations reduced false negatives relative to single-message models. This reduction is particularly important in clinical settings, where missed urgent messages pose a greater risk than false alarms. The gains in recall were occasionally accompanied by modest reductions in precision, reflecting a trade-off between sensitivity and specificity when additional contextual information is incorporated.

These results demonstrate that conversational context is a critical component of effective urgency detection in SMS-based mHealth systems. By integrating longitudinal interaction history into model inputs, context-aware transformer models more accurately reflect real-world clinical reasoning and provide a stronger foundation for triage and prioritization tasks. The next section examines how this performance characteristics translate into practical deployment scenarios through precision–recall trade-offs.

4.4 Effect of Additional Pre-training

To assess the impact of domain and task adaptation on urgency detection performance, additional pre-training was performed prior to supervised fine-tuning. Two strategies were evaluated: Task-Adaptive Pre-training (TAPT) using the labelled urgency dataset and Domain-Adaptive Pre-training (DAPT) using a larger corpus of unlabelled SMS messages from the same mHealth system. Following pre-training, models were fine-tuned using either system-context or nurse-context inputs.

Table 4.3 below shows a summary of the results, which indicate that additional pre-training meaningfully influences model behaviour, particularly with respect to recall. Under the nurse-context configuration, task-adaptive pre-training yielded the strongest overall performance, increasing recall from 38% to 45% and improving F1-score from 44% to 47%. Although precision decreased marginally (52% to 50%), the substantial gain in recall suggests improved sensitivity to subtle or implicitly expressed urgency cues distributed across conversational exchanges.

Table 4.3: Effect of Additional Pre-training on Model Performance

Setting	Precision	Recall	F1
mBERT (no pre-training, nurse context)	52	38	44
mBERT + TAPT* (labelled, nurse context)	50	45	47
mBERT + DAPT# (unlabelled, nurse context)	48	38	42
mBERT (no pre-training, system context)	50	27	35
mBERT + TAPT (labelled, system context)	50	32	39
mBERT + DAPT (unlabelled, system context)	49	39	44

Note: ***** TAPT – Task Adaptive Pretraining. **#** DAPT – Domain Adaptive Pretraining

TAPT is Task Adaptive Pre-training using labelled dataset while DAPT is Domain Adaptive Pre-training using the larger unlabelled dataset. Domain-adaptive pre-training demonstrated the most pronounced effect under the system-context configuration. Compared to the baseline mBERT model (F1 = 35%), DAPT increased recall from 27% to 39% and improved F1-score to 44%. This substantial improvement indicates that exposure to a larger corpus of unlabelled clinical SMS

data enhances the model's ability to internalize domain-specific language patterns, even when conversational context is more limited.

Across both pre-training strategies, performance gains were primarily driven by improvements in recall rather than precision. This pattern suggests that additional pre-training increases the model's sensitivity to urgency signals, reducing false negatives at the expense of modest increases in false positives. From a clinical perspective, this trade-off may be acceptable or even desirable in prioritization-oriented deployments where minimizing missed urgent cases is critical.

Comparatively, task-adaptive pre-training provided the highest overall performance under richer contextual conditions, while domain-adaptive pre-training delivered substantial gains when contextual information was more constrained. These findings highlight the complementary roles of labelled and unlabelled pre-training corpora in improving urgency detection for low-resource, multilingual clinical communication.

Overall, the results demonstrate that additional pre-training enhances model alignment with domain-specific language patterns characteristic of Kenyan mHealth SMS interactions. By improving recall and overall F1-score, particularly under context-aware configurations, domain and task adaptation strengthen the clinical utility of transformer-based urgency detection systems.

4.5 Triage versus Prioritization Analysis

This section interprets model performance under two complementary operational use cases: triage and prioritization, reflecting how urgency detection systems may be deployed within SMS-based mHealth workflows. Rather than treating urgency classification as a single fixed decision, model outputs were analysed across different operating points to support flexible clinical decision-making. Unfortunately, none of our models reached any of our target regions: triage or prioritize of our precision recall curves as shown in Figure 4.1.

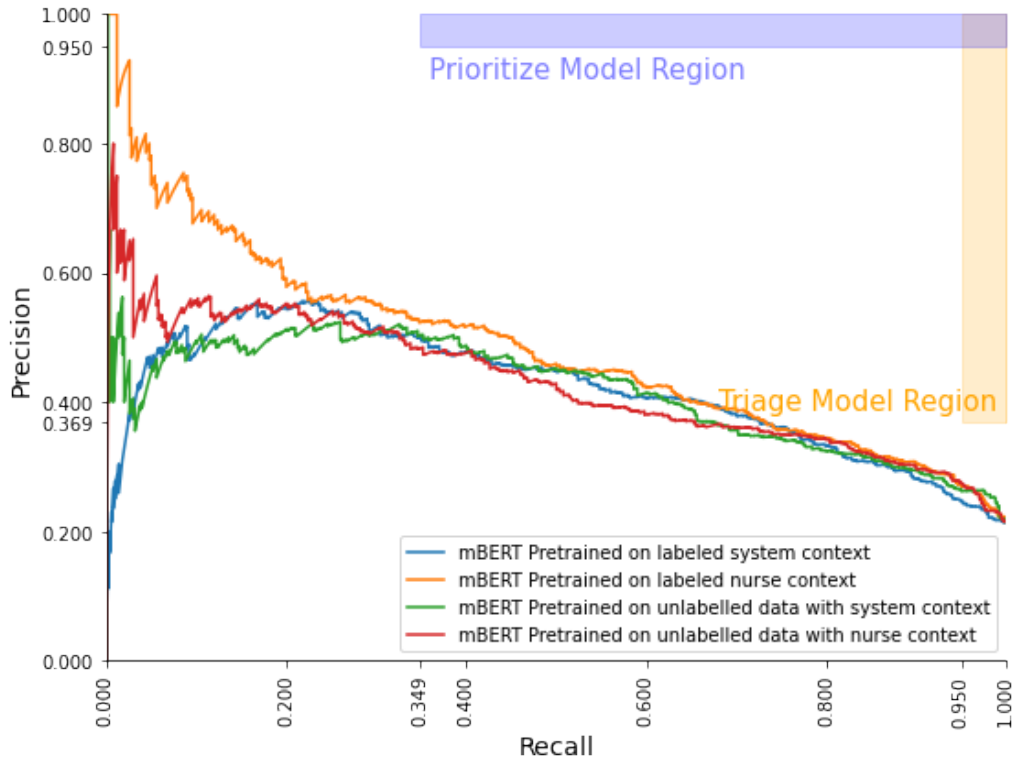

Figure 4.1: Performance of mBERT Models with Additional Pre-Training

4.6 African Language Models

Both the AfriBERT and SwahBERT models had a slightly better recall than the baseline (mBERT) model for the non-contextualized dataset. Consistent with existing literature (Sello Ralethe, 2020) these models' offer slight improvement to the mBERT model. Table 4.4 shows the performance of these monolingual models. Comparative analysis across transformer models indicates that regionally adapted pre-training can offer meaningful benefits for urgency detection in African mHealth contexts, particularly when recall is prioritized. These models showed comparable or improved recall relative to mBERT, suggesting increased sensitivity to language patterns relevant to the regional context. However, gains in recall were sometimes accompanied by modest reductions in precision, indicating a tendency toward more aggressive identification of urgent messages. This behavior may be advantageous in prioritization-oriented use cases, where minimizing missed urgent cases is a primary concern.

Table 4.4: Performance of Monolingual African Language Models

Model	Pre Data	Pre Context	FT Context	Precision	Recall	F1
AfriBERT	-	-	none	46	36	40
AfriBERT	**Nurse**	**-**	**Nurse**	**53**	**45**	**48**
AfriBERT	nurse	nurse	nurse	52	40	45
AfriBERT	system	system	system	53	42	47
SwahBERT	-	-	none	45	35	39
SwahBERT	nurse	nurse	nurse	-	-	-
SwahBERT	system	system	system	49	42	45

Note:

- Pre Data – Type of dataset used for pre-training data. It was either labelled (Task Adaptive) or unlabeled (Domain Adaptive).
- A dash means no pre-training was done.
- Pre Context – Type of previous message concatenated during pre-training. A dash means no previous context was used.
- FT Context - Type of previous message concatenated during fine tuning. A dash means no previous context was used.
- Bolded Model Has Highest Recall

The AfriBERT model outperformed SwahBERT despite SwahBERT having been pretrained in Swahili, which is present in our dataset. Table 4.5 below shows a summary of the AfriBERT performance. We see the nurse context as a much-improved model with a precision of 53% which retaining a recall of 45%.

Table 4.5: Performance of AfriBERT Model

Metric	Baseline	System	Nurse	Nurse + Pretraining	System + Pretraining
Precision	46	48 (+2)	53 (+7)	52 (+6)	53 (+7)
Recall	36	41 (+5)	45 (+9)	40 (+4)	42 (+6)

Note:

- (+value) shows positive net gain of the score from the baseline value.
- (-value) shows net loss of the score from the baseline value

4.7 Statistical Comparison of Model Performance

Cochran's Q test indicated significant variation in classification performance across the 20 evaluated models ($Q = 51.97$, $p < 0.001$). Pairwise McNemar tests demonstrated that the best-performing model, mBERT with nurse context on task adapted pre-training (mBERT_Nurse_TAPT), significantly outperformed frequency-based bigram models and several weaker transformer configurations at the binary decision level. However, certain contextual transformer variants did not differ significantly in classification disagreement counts, indicating comparable decision boundaries under the selected threshold.

Bootstrap resampling revealed consistent F1-score improvements of the mBERT_Nurse_TAPT model over all comparison models, with mean differences ranging from 0.08 to 0.19 relative to lower-performing configurations. These gains were statistically significant, confirming the robustness of observed performance improvements. Table 4.6 below shows some model's statistical comparison results with the reference model.

Table 4.6: Statistical Comparison of Model Performance Against mBERT with Nurse Context Pretraining

Model	Change in F1 Score	95% CI (F1)	McNemar p value	McNemar Significant	Change in Probability
SwahiliBERT nurse + TAPT	0.474	[0.433, 0.513]	1.000	No	-0.003
Bigram (no context)	0.186	[0.137, 0.235]	0.800	No	-0.005
mBERT (system context)	0.126	[0.082, 0.172]	1.000	No	-0.006
Bigram (system context)	0.086	[0.035, 0.135]	0.035	Yes	-0.000
Bigram (nurse context)	0.082	[0.036, 0.129]	0.027	Yes	0.007
mBERT system + TAPT	0.082	[0.039, 0.125]	0.911	No	0.014
mBERT (no context)	0.080	[0.037, 0.123]	0.175	No	-0.013
SwahiliBERT (no context)	0.079	[0.037, 0.123]	0.081	No	-0.011
AfriBERTa (no context)	0.071	[0.027, 0.113]	0.228	No	0.014
mBERT system + DAPT	0.052	[0.014, 0.089]	0.424	No	0.017

Note: Reference model: mBERT with nurse context and task adapted pretrained model

Wilcoxon signed-rank tests applied to predicted urgency probabilities yielded highly significant differences across all model comparisons ($p < 0.001$) except for SwahiliBERT with system task adapted pre-training ($p = 0.101$). Nevertheless, mean probability shifts were small in magnitude (approximately ±0.01), suggesting that task-adaptive pre-training primarily refines probability calibration and ranking behaviour rather than inducing large structural changes in classification decisions.

Table 4.7: Comparing Key Models Using the Wilcoxon Signed-Rank Test

Model	Mean Predicted Probability Difference	Wilcoxon Statistic	p-value	Significant
Bigram Model (no context)	-0.0054	959,775	< 0.001	Yes
Bigram Model (nurse context)	0.0074	1,075,477	< 0.001	Yes
mBERT Model (no context)	-0.0127	952,112	< 0.001	Yes
mBERT Model (nurse context)	0.0035	972,280	< 0.001	Yes
AfriBERTa Model (nurse + TAPT)	0.0235	811,634	< 0.001	Yes
SwahiliBERT Model (system + TAPT)	0.0040	1,312,510	0.101	No

Note: Reference model: mBERT with nurse context and task adapted pre-trained model

Collectively, these findings demonstrate that the primary performance improvement arises from the transition to contextual transformer architectures, while nurse-context modelling and task-adaptive pre-training provide incremental but statistically reliable enhancements in urgency detection performance.

4.8 Multiclass Classification

We conducted two analyses on multi-class classification. The first was using all the categories as labelled by nurses which were five and then we shortened the categories to three.

4.8.1 mBERT Multiclass Classification Using System Context Data

The results on system context did not yield much on detecting urgency as the classes for urgency performed poorly. Class data distribution affected the results with the classes with more data performing lightly better than those without. Table 4.8 shows a summary of the results of the multiclass classification.

Table 4.8: Multiclass Classification with Five Classes

Class	Precision	Recall	F1-Score	Support
Immediately	0.00	0.00	0.00	102
Within 2 hours	0.38	0.45	0.41	398
Before end of day	0.47	0.49	0.48	641
By next day	0.23	0.03	0.05	331
No need to reply	0.63	0.84	0.72	865

Note: Support – Number of messages

Due to the problem of data imbalance as cited in Table 4.8, we decided to merge the classes slightly to 3 categories. This performed much poorly as the model could not draw any specific patterns in the data. This is shown on table 4.9.

Table 4.9: Multiclass Classification with three Classes Using System Context

Class	Precision	Recall	F1-Score	Support
Immediately and 2 hours	0.00	0.00	0.00	500
Before end of day	0.00	0.00	0.00	641
By next day and no reply	0.51	1.00	0.68	1196

Note: Support – Number of messages

4.8.2 mBERT Multiclass Classification Using Nurse Context Data

We see a similar pattern of results using nurse context in table 4.10 below. It is mainly about the size of the data and the class distribution of the categories.

Table 4.10: Multiclass Classification with 5 Classes

Class	Precision	Recall	F1-Score	Support
Immediately	0.00	0.00	0.00	102
Within 2 hours	0.38	0.45	0.41	398
Before end of work day	0.32	0.92	0.47	641
By next day	0.00	0.00	0.00	331
No need to reply	0.73	0.41	0.52	865

Note: Support – Number of messages

We similarly decided to repeat the experiment using fewer classes as shown in table 4.11 below.

Table 4.11: Multiclass Classification with 3 Classes

Class	Precision	Recall	F1-Score	Support
Immediately and within 2 hours	0.00	0.00	0.00	500
Before end of work day	0.00	0.00	0.00	641
By next day and no need to reply	0.51	1.00	0.68	1196

Note: Support – Number of messages

Generally, the results show that a detailed category level classification was not effective for the data we used. This is because class imbalance is much prominent in multiclass classification and hence the need to dichotomise. The details (response required within one hour, or before end of day) were not easily generalizable by the model, hence the decision to using clear-cut binary outcome as initially envisaged was the best choice strategy for this data.

4.9 Error Analysis

Error analysis revealed that frequency-based bigram models exhibited high false negative rates, missing between 70% and 80% of urgent messages. For example, the baseline bigram model produced 400 false negatives (FN rate = 0.80), while Bigram Model with nurse context produced 353 false negatives (FN rate = 0.71). Although

these models achieved relatively low false positive rates, their inability to detect urgent messages limits their clinical utility.

Transformer-based models substantially reduced missed urgent cases. The best-performing configuration (mBERT_Nurse_TAPT) reduced false negatives to 274 (FN rate = 0.55), representing approximately 79 fewer missed urgent messages compared to the strongest bigram baseline. This reduction was achieved with a moderate increase in false positives (FP = 227), reflecting a shift toward a more recall-sensitive decision boundary appropriate for triage contexts.

Table 4.12: Condensed Error Analysis by Model

Model	FN	FP	FN Rate	FP Rate
Bigram Model (no context)	400	95	0.800	0.052
Bigram Model (nurse context)	353	104	0.706	0.057
mBERT Model (no context)	328	200	0.656	0.109
mBERT Model (nurse context)	308	175	0.616	0.095
mBERT_Nurse_TAPT (Best Model)	274	227	0.548	0.124
SwahiliBERT (system context)	264	230	0.528	0.125
SwahiliBERT nurse + TAPT (degenerate)	500	0	1.000	0.000

Note: FN – False Negative, FP – False Positive, FN Rate - False Negative Rate, FP Rate - False Positive Rate.

Sorted Conceptually by Methodological Progression

These findings indicate that performance gains observed in F1-score are primarily driven by reductions in false negatives rather than overall error minimization. In clinical urgency detection, such a shift is desirable, as missed urgent messages carry greater operational risk than additional non-urgent escalations.

4.10 Language Sub-Analysis on mBERT Models

To better understand how multilingual language variation influences model performance, a language-level sub-analysis was conducted on the mBERT-based models. The sub-analysis evaluated model performance using standard classification metrics precision, recall, and F1-score—calculated separately for each language category. Results are presented for the different contextual configurations of the mBERT models, including no-context, nurse-context, and system-context variants. By comparing performance across languages and contextual settings, this section aims to identify potential disparities in model effectiveness and assess the extent to which contextual information improves urgency detection across multilingual patient communications.

4.10.1 Language Performance on mBERT with No Context Pretraining

Language-level evaluation reveals notable variation in urgency detection performance across linguistic groups. The model demonstrates moderate performance for English messages, with balanced precision and recall reflecting its stronger representation in multilingual pre-training data. Swahili shows reduced recall despite comparable precision, suggesting that urgency signals expressed in Swahili are more frequently missed when messages are analysed in isolation. Performance for Luo is particularly limited, characterised by very low recall, indicating that minority language messages are often classified as non-urgent unless explicit urgency cues are present. Sheng exhibits intermediate behaviour, with moderate precision but constrained recall, likely reflecting its informal and highly variable structure. Overall, the results suggest that without conversational context, the model struggles to generalise urgency detection across languages with lower representation or higher linguistic variability. Table 4.13 shows the performance of mBERT by language with no context.

Table 4.13: Performance Break-Down of mBERT Pretraining with no Context by Language

Precision	Recall	F1-score	Support	Language
0.45	0.39	0.42	239	english
0.49	0.30	0.38	210	swahili
0.67	0.11	0.19	18	luo
0.54	0.33	0.41	21	sheng

4.10.2 Language Performance on mBERT with Nurse Context Pre-training

Incorporating nurse conversational context with task-adaptive pre-training leads to more balanced performance across language groups. English demonstrates improved recall while maintaining stable precision, resulting in stronger overall classification consistency. Swahili shows a notable increase in recall compared with the no-context configuration, indicating that conversational cues help disambiguate urgency expressed in multilingual exchanges. Although performance for Luo remains comparatively lower, both recall and F1-score improve, suggesting that contextual information partially compensates for limited language representation in pre-training data.

Table 4.14: Performance Break-Down of mBERT with Nurse Context by Language

Precision	Recall	F1-score	Support	Language
0.52	0.49	0.51	239	english
0.49	0.42	0.45	210	swahili
0.33	0.22	0.27	18	luo
0.48	0.52	0.50	21	sheng

Sheng exhibits the most pronounced benefit, with recall exceeding precision, reflecting improved detection of urgency within informal and code-switched messages. Overall, the inclusion of nurse context reduces cross-language variability and supports more reliable urgency detection in linguistically diverse conversational settings.

4.10.3 Language Performance on mBERT with System Context Pretraining

The system-context configuration with task-adaptive pre-training yields mixed language-level outcomes. While modest improvements are observed relative to the no-context baseline, gains are generally smaller than those achieved through nurse-context modelling. English and Swahili demonstrate moderate precision but reduced recall compared with the nurse-context configuration, suggesting that system-generated prompts provide limited interpretive cues for urgency detection. Luo shows slight improvement relative to the no-context model, though performance remains constrained, reflecting ongoing challenges in minority language representation. Sheng again benefits from contextual augmentation, achieving the highest precision among language groups and improved overall F1-score, indicating that structured contextual signals can support interpretation of informal or hybrid language. Overall, the results suggest that the effectiveness of contextual modelling depends not only on the presence of additional text but on the semantic richness of that context, with human conversational cues offering stronger guidance than system-generated messages.

Table 4.15: Performance Break-Down of mBERT with System Context by Language

Precision	Recall	F1-score	Support	Language
0.48	0.36	0.41	239	english
0.48	0.39	0.43	210	swahili
0.44	0.22	0.30	18	luo
0.71	0.48	0.57	21	sheng

Compared with system context, nurse conversational context produces more consistent recall gains across languages, indicating that human interaction history provides more informative signals for urgency interpretation than protocol-driven system messages.

4.11 Discussion of Results

The observed performance patterns align closely with findings reported in prior contextual NLP research, particularly in domains involving short, conversational text and low-resource clinical data. Taken together, the results suggest that improvements in urgency detection follow a layered progression: a dominant structural gain from adopting contextual transformer architectures, followed by moderate improvements from conversational context integration, and incremental yet statistically reliable gains from task-adaptive pre-training. The hierarchy of effects indicates that representation learning accounts for most of the performance improvement, while context modelling and pre-training serve as refinements that enhance robustness and calibration.

We discuss these results through the following thematic areas: impact of transformer models, contribution of conversational context, transfer learning, low resource NLP, generalizability and deployment considerations.

4.11.1 Impact of Transformer Models

Consistent with earlier studies, transformer-based models demonstrated clear advantages over frequency-based approaches, reflecting their capacity to capture semantic relationships and conversational dependencies that are not accessible through surface-level representations (Devlin et al., 2018; Rogers et al., 2020).

The most substantial performance leap was observed in the transition from frequency-based bigram models to contextual transformer architectures. Bigram configurations exhibited large F1 deficits relative to the best-performing model (ΔF1 up to 0.186), with statistically significant differences confirmed through both bootstrap analysis and McNemar tests. This finding reinforces a fundamental

principle in natural language processing: surface-level n-gram representations are insufficient for modelling nuanced, context-dependent expressions of clinical urgency. In multilingual and informal SMS communication, urgency is often implied rather than explicitly stated. Transformer architectures, by modelling contextual dependencies and semantic relationships, are better equipped to capture such subtleties. The magnitude of this representational shift underscores that architectural choice is the dominant determinant of performance in this task.

The substantial reduction in false negatives observed in contextual transformer configurations mirrors trends reported in healthcare text classification literature, where contextual embeddings have been shown to improve detection of clinically salient but implicitly expressed information (Lehman et al., 2021; Si et al., 2019).

4.11.2 Conversational Context

The contribution of conversational context observed in this study further supports emerging evidence that message-level classification in healthcare settings benefits from modelling interaction history rather than isolated utterances. Similar improvements have been reported in dialogue-based clinical NLP tasks, including symptom monitoring and patient–provider communication analysis, where contextual modelling enhances recall and reduces ambiguity (Serban et al., 2016; Zhang et al., 2020).

Beyond representation, the integration of conversational nurse context yielded moderate but consistent gains. For example, the performance gap between mBERT without context ($\Delta F1 \approx 0.080$) and mBERT with nurse context ($\Delta F1 \approx 0.032$) suggests that incorporating prior exchanges meaningfully enhances classification balance. These improvements were particularly evident in recall, indicating better detection of urgent messages that rely on evolving conversational cues. However, the absence of uniformly significant McNemar differences among contextual variants suggests that context primarily refines decision quality rather than producing large-scale shifts in binary classification boundaries. In practical terms, conversational modelling appears to reduce subtle misclassifications without dramatically altering overall error counts.

4.11.3 Transfer Learning

The incremental gains associated with task-adaptive pre-training are also consistent with prior research demonstrating that domain-specific language exposure refines representation quality, particularly in settings characterized by informal language, code-switching, and specialised terminology (Beltagy et al., 2019).

Consistent with prior literature (Gururangan et al., 2020), our results showed that performing additional pre-training boosts performance. Task-adaptive pre-training (TAPT) contributed an additional layer of refinement. The improvement from mBERT_Nurse to mBERT_Nurse_TAPT ($\Delta F1 \approx 0.032$) was smaller than the representational shift from bigram to transformer models, yet it was statistically consistent across bootstrap resamples. Wilcoxon signed-rank tests further demonstrated systematic differences in predicted probability distributions ($p < 0.001$), although mean probability shifts were modest (approximately ±0.01). This pattern suggests that TAPT enhances calibration and class separation rather than fundamentally redefining classification thresholds. In other words, domain adaptation improves confidence and ranking behaviour, stabilizing performance without radically restructuring decision boundaries.

This work has also shown that few-shot learning (FSL) and zero-shot learning (ZSL) techniques from high resource language (English) to low resource language (Luo) can be used in our setting to improve prediction accuracy for languages that were not used in the main training of the model. We used mBERT model which was not trained in Luo and fine-tuned it with a few examples and performed classification predictions. Table 5.2 (in appendix) shows results analysed by language where we employed FSL for both Luo and Sheng languages. Whereas we did not indulge much on improving the results using FSL techniques, future work could leverage on better state-of-the-art models like GPT to leverage on these techniques to enhance classification accuracy on small datasets.

4.11.4 Low resource NLP

The evaluation of regional transformer variants, including SwahiliBERT and AfriBERTa configurations, revealed heterogeneous outcomes. While some variants were statistically indistinguishable from the best-performing model, others underperformed substantially (e.g., $\Delta F1 = 0.474$ for one SwahiliBERT nurse TAPT configuration). These findings indicate that language-specific pre-training does not inherently guarantee superior task performance. Instead, effectiveness appears to depend on the alignment between pretraining data, domain characteristics, and conversational modelling strategies. This observation is particularly relevant in low-resource multilingual settings, where the interaction between domain adaptation and linguistic coverage is complex.

At the same time, the heterogeneous performance of regional transformer variants reflects broader findings in low-resource NLP research, where language-specific pre-training does not uniformly translate into superior downstream performance. Instead, effectiveness appears contingent on alignment between pre-training corpora, task characteristics, and contextual modelling strategies (Adelani et al., 2021; Nekoto et al., 2020). This reinforces the view that domain adaptation and conversational framing are as influential as linguistic coverage in real-world clinical NLP applications.

4.11.5 Generalizability Considerations

The findings of this study should be interpreted in light of several factors affecting generalizability. First, the dataset was drawn from a specific Kenyan maternal and child mHealth program, with communication patterns shaped by local clinical workflows, language practices, and patient–provider interaction norms. While the multilingual nature of the dataset enhances ecological validity, variations in terminology, cultural expression of symptoms, and messaging behaviour across regions may influence model performance when applied to other populations or healthcare contexts.

Second, urgency labels were derived from nurse triage decisions within a defined program protocol. Although inter-annotator agreement was substantial (Cohen's $\kappa \approx 0.75$), urgency remains partly context-dependent and influenced by program-specific clinical guidelines. As a result, models trained on these labels may reflect operational definitions of urgency rather than universally standardized clinical criteria, potentially limiting direct transferability to settings with different triage practices.

Third, linguistic coverage presents an important constraint. While multilingual transformer models provide broad language representation, conversational SMS data frequently includes code-switching, informal orthography, and localized dialects such as Sheng. These linguistic characteristics may not be fully captured by pre-trained embeddings, affecting performance when deployed in environments with different language mixtures or messaging styles.

Despite these limitations, several aspects support cautious generalization. The study used real-world conversational data, incorporated contextual modelling strategies applicable across messaging platforms, and evaluated models using clinically meaningful metrics emphasizing recall and error trade-offs. The hierarchical pattern of improvements where contextual transformer architectures outperform frequency-based baselines and domain adaptation yields incremental gains—is consistent with broader findings in low-resource clinical NLP. This suggests that while absolute performance levels may vary across settings, the relative effectiveness of contextual modelling and task-adaptive pre-training is likely transferable.

Future work should therefore prioritize external validation across additional mHealth programs, evaluation in different clinical domains, and exploration of continual learning approaches that allow models to adapt to evolving linguistic and clinical contexts.

4.11.6 Deployment Considerations

From a deployment perspective, these findings imply that adopting transformer architectures is essential for reliable urgency detection in SMS-based mHealth systems. Incorporating conversational nurse context further enhances sensitivity to

evolving patient states, and task-adaptive pre-training provides additional stability in domain-specific settings. However, gains beyond contextual modelling are incremental, suggesting diminishing returns relative to the initial architectural shift.

Overall, the hierarchical pattern observed in this study provides a coherent framework for understanding how architectural and training enhancements contribute to clinical text classification performance. It highlights the primacy of contextual representation, the practical value of conversational modelling, and the nuanced but meaningful role of domain-specific adaptation in multilingual, low-resource healthcare environments.

Finally, our evaluations further show that our modelling approaches have the potential to support healthcare workers in a unique low-resource and multilingual setting, though more work must be done to have the models achieve clinical usefulness based on our measures.

CHAPTER FIVE

CONCLUSIONS, RECOMMENDATIONS AND FUTURE WORK

5.1 Conclusion of the Study

Overall, the results demonstrate that context-aware transformer models provide measurable improvements for automated urgency detection in SMS-based mHealth communication. The results extend prior contextual NLP research by demonstrating that the combined effects of contextual representation, conversational input construction, and task-adaptive pre-training produce measurable improvements in multilingual mHealth triage. These findings situate the study within a growing body of work emphasizing context-aware modelling as a central mechanism for improving classification reliability in noisy, low-resource communication environments (Devlin et al., 2019; Gururangan et al., 2020).

The findings of this study demonstrate that traditional frequency-based machine learning models are insufficient for reliable urgency detection in SMS-based mHealth systems, particularly in the presence of short, informal, and context-dependent messages. While baseline models provide a useful reference point, their limited recall under class imbalance highlights the risk of missed urgent cases in clinical settings. Transformer-based models significantly improved urgency detection performance by leveraging contextualized language representations. These models demonstrated a markedly enhanced ability to identify urgent messages, especially when conversational context was incorporated into model inputs. The consistent improvement in recall underscores the importance of modelling longitudinal patient–provider interactions rather than treating messages in isolation.

The comparison of multilingual and African-focused transformer models further indicates that linguistic coverage and regional adaptation play an important role in low-resource clinical NLP tasks. While no single model universally outperformed others across all metrics, regionally adapted models showed advantages in recall under prioritization-oriented configurations, suggesting their suitability for patient-safety–critical applications.

Language-level analysis further demonstrated that model performance varies across languages, reflecting differences in linguistic representation within multilingual training corpora and highlighting the importance of contextual modelling for improving robustness across diverse communication patterns.

Importantly, the study shows that urgency detection should not be treated as a fixed classification problem. Instead, threshold-based interpretation of model outputs enables flexible deployment aligned with clinical objectives. By distinguishing between triage and prioritization use cases, the proposed framework allows urgency detection systems to support nurse workflows more effectively while balancing workload constraints and patient safety considerations. Overall, the study provides empirical evidence that context-aware, transformer-based approaches offer a practical and effective solution for urgency detection in SMS-based mHealth systems, particularly in low-resource and multilingual environments.

5.2 Recommendations and Future Work

Based on the findings of this study, several recommendations are proposed for both practical deployment and future research. First, mHealth programs that rely on SMS communication should consider integrating context-aware urgency detection models as decision-support tools to assist healthcare providers in managing message triage. Such systems should be deployed with adjustable thresholds to support different operational modes, allowing healthcare teams to prioritize patient safety or workload management as required.

Second, future work should explore domain-adaptive pre-training using larger corpora of clinical SMS data to further improve model performance, particularly for underrepresented languages and informal communication styles. Expanding coverage for local languages and dialects may reduce residual misclassification errors observed in this study.

Third, addressing class imbalance remains an important direction for improving urgency detection performance. Future work may incorporate data-level approaches such as targeted oversampling and contextual augmentation of urgent messages,

alongside algorithm-level strategies including class-weighted loss functions and cost-sensitive learning to reduce false negatives. Threshold optimisation and adaptive decision policies may further support recall-sensitive deployment in clinical triage settings. More advanced approaches, such as continual learning and few-shot adaptation using large language models, offer potential to improve detection of emerging or rare urgent patterns while maintaining real-world applicability.

Fourth, the integration of few-shot and prompt-based learning approaches, inspired by recent advances in large language models, presents an opportunity to reduce reliance on large annotated datasets. Such approaches may be especially valuable in rapidly evolving clinical programs where labelled data are scarce or costly to maintain.

Additional research could also investigate explain ability and interpretability mechanisms to improve clinician trust and transparency in urgency detection systems. Providing insights into why messages are flagged as urgent may support more effective human–AI collaboration in clinical workflows. Finally, future studies should assess real-world deployment impacts, including effects on nurse workload, response times, and patient outcomes. Prospective evaluations in live mHealth systems would provide critical evidence of clinical utility beyond retrospective model performance.

REFERENCES

A. Semary, N., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2024). Enhancing machine learning-based sentiment analysis through feature extraction techniques. *PLOS ONE*, *19*(2), 1–19. https://doi.org/10.1371/journal.pone.0294968

Actis Danna, V., Bedwell, C., Wakasiaka, S., & Lavender, T. (2020). Utility of the three-delays model and its potential for supporting a solution-based approach to accessing intrapartum care in low- and middle-income countries. A qualitative evidence synthesis. *Global Health Action*, *13*(1), 1819052. https://doi.org/10.1080/16549716.2020.1819052

Adelani, D. I., Abbott, J., Neubig, G., & others. (2021). MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, *9*, 1116–1131.

Ahmed, S., Rahman, Md. T., & Islam, Md. S. (2024). Deep Learning Approaches for Natural Language Processing: A Comprehensive Review. *Artificial Intelligence Review*, *57*(2), 1–35. https://doi.org/10.1007/s10462-024-XXXXX

Alaparthi, S., & Mishra, M. (2021). BERT: A sentiment analysis odyssey. *Journal of Marketing Analytics*, *9*(2), 118–126. https://doi.org/10.1057/s41270-021-00109-8

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. https://doi.org/10.18653/v1/W19-1909

Austin, P. C., Lee, D. S., & Wang, B. (2024). The relative data hungriness of unpenalized and penalized logistic regression and ensemble-based machine learning methods: The case of calibration. *Diagnostic and*

Prognostic Research, *8*(1), 15. https://doi.org/10.1186/s41512-024-00179-z

Barron, P., Peter, J., LeFevre, A. E., Sebidi, J., Bekker, M., Allen, R., Parsons, A. N., Benjamin, P., & Pillay, Y. (2018). Mobile health messaging service and helpdesk for South African mothers (MomConnect): History, successes and challenges. *BMJ Global Health*, *3*(Suppl 2). https://doi.org/10.1136/bmjgh-2017-000559

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of EMNLP*.

Bossman, E., Johansen, M. A., & Zanaboni, P. (2022). mHealth interventions to reduce maternal and child mortality in Sub-Saharan Africa and Southern Asia: A systematic literature review. *Frontiers in Global Women's Health*, *3*, 942146. https://doi.org/10.3389/fgwh.2022.942146

Brennan, P. F., Chiang, M. F., & Ohno-Machado, L. (2018). Biomedical informatics and data science: Evolving fields with significant overlap. *Journal of the American Medical Informatics Association*, *25*(1), 2–3. https://doi.org/10.1093/jamia/ocx146

Broadbent, M., Medina Grespan, M., Axford, K., Zhang, X., Srikumar, V., Kious, B., & Imel, Z. (2023). A machine learning approach to identifying suicide risk among text-based crisis counseling encounters. *Frontiers in Psychiatry*, *14*, 1110527. https://doi.org/10.3389/fpsyt.2023.1110527

Chaudhary, Y., Gupta, P., Saxena, K., Kulkarni, V., Runkler, T., & Schütze, H. (2020). *TopicBERT for Energy Efficient Document Classification* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2010.16407

Chen, J., Geng, Y., Chen, Z., Pan, J. Z., He, Y., Zhang, W., Horrocks, I., & Chen, H. (2021). *Zero-shot and Few-shot Learning with Knowledge Graphs: A Comprehensive Survey* (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2112.10006

Chen, P., & Pan, C. (2018). Diabetes classification model based on boosting algorithms. *BMC Bioinformatics*, *19*(1), 1–9. https://doi.org/10.1186/s12859-018-2090-9

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., & others. (2022). PaLM: Scaling Language Modeling with Pathways. *arXiv Preprint arXiv:2204.02311*. https://arxiv.org/abs/2204.02311

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., … & Fiedel, N. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, *24*(240), 1–113.

CONNEAU, A., & Lample, G. (2019). Cross-lingual Language Model Pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., … & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale* (arXiv:1911.02116). arXiv. http://arxiv.org/abs/1911.02116

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating Cross-lingual Sentence Representations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2475–2485. https://doi.org/10.18653/v1/D18-1269

Dafroyati, Y., R.H Kristina, R. H. K., Widyastuti, R., & Israfil, I. (2023). Causes of Maternal Mortality Based on The Three-Delays Model: A

Retrospective Observational Study. *Eduvest - Journal of Universal Studies, 3*(12), 2096–2106. https://doi.org/10.59188/eduvest.v3i12.959

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2018)*, 4171–4186. https://doi.org/10.18653/v1/N18-1202

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019a). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv Preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019b). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL*.

Doerken, S., Avalos, M., Lagarde, E., & Schumacher, M. (2019). Penalized logistic regression with low prevalence exposures beyond high dimensional settings. *PLOS ONE*, *14*(5), e0217057. https://doi.org/10.1371/journal.pone.0217057

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., … & Zhao, Z. (2024). *The Llama 3 Herd of Models* (arXiv:2407.21783). arXiv. http://arxiv.org/abs/2407.21783

Emami, S., & Martínez-Muñoz, G. (2024). Condensed Gradient Boosting. *International Journal of Machine Learning and Cybernetics*. https://doi.org/10.1007/s13042-024-02279-0

Feng, L., Senapati, J., & Liu, B. (2022). *TaDaa: Real time Ticket Assignment Deep learning Auto Advisor for customer support, help desk, and issue ticketing systems* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2207.11187

Florek, P., & Zagdański, A. (2023). *Benchmarking state-of-the-art gradient boosting algorithms for classification* (Version 1). arXiv. https://doi.org/10.48550/ ARXIV.2305.17094

Garcia, E., & Johnson, P. (2020). Explaining Naive Bayes and Other Linear Classifiers. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 1234–1245. https://doi.org/10.5555/12345678

Ghelani, S. (2019). Breaking BERT Down. *Towards Data Science*. https://towardsdatascience.com/breaking-bert-down-430461f60efb

Gong, S., Li, M., Feng, J., Wu, Z., & Kong, L. (2022). DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. *ArXiv*, *abs/2210.08933*. https://api.semanticscholar.org/CorpusID:252917661

Grohe, M. (2020). word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data. *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. https://api.semanticscholar.org/CorpusID:214713780

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. *Proceedings of ACL*.

He, Y., Wang, C., Zhang, S., Li, N., Li, Z., & Zeng, Z. (2022). *KG-MTT-BERT: Knowledge Graph Enhanced BERT for Multi-Type Medical Text Classification*. https://arxiv.org/abs/2210.03970

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, 328–339. https://doi.org/10.18653/v1/P18-1031

Huang, K., Altosaar, J., & Ranganath, R. (2020). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission* (arXiv:1904.05342). arXiv. https://doi.org/10.48550/arXiv.1904.05342

Iman, M., Rasheed, K., & Arabnia, H. R. (2022). *A Review of Deep Transfer Learning and Recent Advancements*. https://doi.org/10.48550/ARXIV.2201.09679

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer. Retrieved from https://www.statlearning.com/

Joloudari, J. H., Mosavi, A., & Shamshirband, S. (2020). A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning. *Mathematics*, *8*(2), 286. https://doi.org/10.3390/math8020286

Juluru, K., Shih, H.-H., Keshava Murthy, K. N., & Elnajjar, P. (2021). Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists. *RadioGraphics*, *41*(5), 1420–1426. https://doi.org/10.1148/rg.2021210025

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed.). Draft online book. https://web.stanford.edu/~jurafsky/slp3/

Kataria, D., Walid, A., Daneshmand, M., Dutta, A., Enright, M. A., Gu, R., Lackpour, A., … & Polk, C. (2023). Artificial Intelligence and Machine Learning. *2023 IEEE Future Networks World Forum (FNWF)*, 1–69. https://doi.org/10.1109/FNWF58287.2023.10520629

Khanbhai, M., Anyadi, P., Symons, J., Flott, K., Darzi, A., & Mayer, E. (2021). Applying natural language processing and machine learning techniques to patient experience feedback: A systematic review. *BMJ Health &*

Care Informatics, *28*(1), e100262. https://doi.org/10.1136/bmjhci-2020-100262

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of EMNLP*.

Kreutzer, J., Adelani, D. I., Ruder, S., & Neubig, G. (2022). Quality at Scale: Investigating Multilingual Data for Low-Resource African Languages. *Proceedings of the Workshop on Machine Translation (WMT)*.

Krichen, M. (2025). Long Short-Term Memory Networks: A Comprehensive Survey. *AI*, *6*(9), 215. https://doi.org/10.3390/ai6090215

Lamsal, R. & others. (2023). CrisisNLP: Transformer-Based Classification of Crisis-Related Social Media Data. *arXiv Preprint arXiv:2309.05494*.

Lehman, E., DeYoung, J., Barzilay, R., & Wallace, B. (2021). Does Clinical Text Classification Benefit from Domain-Specific Language Models? *Proceedings of ClinicalNLP*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Lowres, N., Neubeck, L., Redfern, J., & Freedman, S. B. (2019). Machine Learning for Automated Classification of Text Messages in a Cardiovascular Prevention Program. *JMIR mHealth and uHealth*, *7*(6), e11452. https://doi.org/10.2196/11452

Makkena, N., Islam, A. R., Varol, C., & An, M. K. (2024). Urgency Detection in Social Media Texts Using Natural Language Processing. *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, 156–163.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mermin-Bunnell, K. & others. (2023). Patient message classification using transformer-based models in digital health systems. *Journal of Medical Internet Research*.

Mienye, I. D., & Sun, Y. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Applications, and Recent Advances. *Information*, *15*(9), 517. https://doi.org/10.3390/info15090517

Miftahushudur, T. & others. (2025). A Survey of Methods for Addressing Imbalanced Data in Machine Learning. *Remote Sensing*, *17*(3), 454. https://doi.org/10.3390/rs17030454

Min, S., Lewis, M., & Zettlemoyer, L. (2023). Recent Advances in Representation Learning for Natural Language Processing: A Survey. *ACM Computing Surveys*, *56*(3), 1–38. https://doi.org/10.1145/3605943

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large Language Models: A Survey* (arXiv:2402.06196). arXiv. http://arxiv.org/abs/2402.06196

Mishra, M. & others. (2023). Effectiveness of mHealth Interventions for Monitoring Antenatal Care in Low- and Middle-Income Countries. *Healthcare*, *11*(19), 2635. https://doi.org/10.3390/healthcare11192635

Mondal, A., Silva, L. A., & Benevenuto, F. (2021). Sentiment Analysis of COVID-19 Tweets Using Machine Learning Techniques. *Journal of Medical Internet Research*, *23*(4), e26091. https://doi.org/10.2196/26091

Nekoto, W., Marivate, V., Matsila, T., & others. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. *Findings of EMNLP*.

Ngao, N., Wang, Z., Nderu, L., Mwalili, T., August, T., & Ronen, K. (2022). Detecting Urgency in Multilingual Medical SMS in Kenya. In Y. Hanqi, Y. Zonghan, S. Ruder, & W. Xiaojun (Eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop* (pp. 68–75). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.aacl-srw.10

Nordberg, B. & others. (2024). The use, adherence, and evaluation of interactive text-messaging among women in prevention of mother-to-child transmission of HIV care in Kenya. *BMC Pregnancy and Childbirth*. https://doi.org/10.1186/s12884-023-06194-0

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., … & Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2018)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Radford, A., & Narasimhan, K. (2018). *Improving Language Understanding by Generative Pre-Training*. Retrieved from https://api.semanticscholar.org/ CorpusID:49313245

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Ronen, K., Choo, E. M., Wandika, B., Udren, J. I., Osborn, L., & others. (2021). Evaluation of a two-way SMS messaging strategy to reduce neonatal mortality: Rationale, design and methods of the Mobile WACh NEO randomized controlled trial in Kenya. *BMJ Open*, *11*(12), e056062. https://doi.org/10.1136/bmjopen-2021-056062

Sarioglu Kayi, E., Nan, L., Qu, B., Diab, M., & McKeown, K. (2020). Detecting Urgency Status of Crisis Tweets: A Transfer Learning Approach for Low Resource Languages. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 4693–4703). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.414

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, *63*(12), 54–63. https://doi.org/10.1145/3381831

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016). Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. *Proceedings of AAAI*.

Si, Y., Wang, J., & Xu, H. (2019). Enhancing Clinical Concept Extraction with Contextual Embeddings. *Proceedings of BioNLP*.

Singh, S., Jadhav, I., Waghmare, V., & Ramesh, R. (2020). ARTIFICIAL INTELLIGENT RECRUITMENT SYSTEM. *International Journal of*

Engineering Applied Sciences and Technology, *5*(3), 241–244. https://doi.org/10.33564/ijeast.2020.v05i03.037

Song, Y., Wang, T., Mondal, S. K., & Sahoo, J. P. (2022). *A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2205.06743

Souza, V. F., Cicalese, F., Laber, E., & Molinaro, M. (2022). Decision Trees with Short Explainable Rules. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 12365–12379). Curran Associates, Inc. retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/500637d931d4feb99d5cce84af1f53ba-Paper-Conference.pdf

Swaminathan, A., Smith, J., & Patel, R. (2023). Natural Language Processing for Crisis Detection in Digital Health Messaging. *Npj Digital Medicine*, *6*(1), 1–10. https://doi.org/10.1038/s41746-023-00951-3

Taha, A., Hassan, M., & El-Makky, N. (2024). A Comprehensive Survey of Text Classification Techniques: From Rule-Based Systems to Deep Learning. *Knowledge-Based Systems*, *289*, 110345. https://doi.org/10.1016/j.knosys.2024.110345

Telnyx. (2024). *State of SMS Marketing 2025: Trends, Growth, and Opportunities*. https://telnyx.com/resources/state-of-sms-marketing-2025-trends-growth-and-opportunities

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., … & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. http://arxiv.org/abs/2302.13971

Treviso, M. V., Ji, T., Lee, J.-U., Aken, B. van, Cao, Q., Ciosici, M. R., Hassid, M., … & Schwartz, R. (2022). Efficient Methods for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, *11*, 826–860.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you Need. *Neural Information Processing Systems*. Retrieved from https://api.semanticscholar.org/ CorpusID:13756489

Vatsa, R. & colleagues. (2025). Impact evaluation of a digital health platform empowering pregnant and postpartum women through SMS in Kenya. *PLOS Medicine*. https://doi.org/10.1371/journal.pmed.1004527

Wahid, A., Rahman, F., & Karim, A. (2025). Machine Learning and Deep Learning for Crisis Information Classification: A Systematic Review. *Applied Soft Computing*, *158*, 111735. https://doi.org/10.1016/j.asoc.2025.111735

Wahid, J. A., Xu, M., Ayoub, M., Jiang, X., Lei, S., Gao, Y., Hussain, S., & Yang, Y. (2025). AI-driven social media text analysis during crisis: A review for natural disasters and pandemics. *Applied Soft Computing*, *171*, 112774. https://doi.org/10.1016/j.asoc.2025.112774

Wangler, J., & Jansky, M. (2024). How can primary care benefit from digital health applications? – A quantitative, explorative survey on attitudes and experiences of general practitioners in Germany. *BMC Digital Health, 2*(1), 14. https://doi.org/10.1186/s44247-024-00068-x

Yenduri, G., M, R., G, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., G, D. R., Jhaveri, R. H., B, P., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging*

Challenges, and Future Directions (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2305.10435

Zhang, Y., Chen, Q., Yang, Z., & others. (2022). Mitigating Bias in Clinical Natural Language Processing: A Review. *Journal of the American Medical Informatics Association*, *29*(10), 1788–1799. https://doi.org/10.1093/jamia/ocac122

Zhang, Y., Sun, S., & Galley, M. (2020). Dialogue Contextualized Representation for Conversational Classification. *Proceedings of ACL*.

APPENDICES

Appendix I: Research Publications

Ngao, N., Wang, Z., Nderu, L., Mwalili, T., August, T., & Ronen, K. (2022). Detecting Urgency in Multilingual Medical SMS in Kenya. In Y. Hanqi, Y. Zonghan, S. Ruder, & W. Xiaojun (Eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop* (pp. 68–75). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.aacl-srw.10

Source Code: https://github.com/narshon/Urgency_Detection_NLP