Contents

1. Introduction

Timely detection of infectious diseases is crucial for preventing outbreaks and saving lives. Traditional disease surveillance methods often depend on manual reporting, which can be slow and inefficient. The rise of artificial intelligence has sparked growing interest in automating disease surveillance for early detection. In the current era of globalization, there is a significant need for epidemic intelligence due to the potential impact that epidemics can have on society and the economy. While many developed countries have already implemented automated solutions to address this challenge, it is essential to explore the possibilities for Africa, where the need is equally urgent.

Monitoring real-time informal data sources, such as social media and news articles, has shown great promise in the early detection of infectious diseases. This approach, known as event-based surveillance, is faster than traditional indicator-based surveillance, which relies on formal reports from healthcare centers (Paquet et al., 2006). Event-based surveillance captures real-time data and provides timely warnings. Informal sources of information, such as online news and social media, come in various forms; despite their differences, they all have one thing in common: they convey information through written content. Text-based epidemic intelligence has proven useful in several outbreak instances, including the early detection of the COVID-19 pandemic (Shausan et al., 2023). By leveraging AI-powered natural language processing techniques, we can extract relevant information and gain insights from unstructured text data, such as social media posts and news articles, to identify potential disease outbreaks.

We aim to develop an AI-driven system for disease surveillance. The primary goal is to transform unstructured text data, primarily obtained from news articles, into structured formats in order to rapidly detect infectious diseases. To accomplish this, we will employ Named Entity Recognition (NER) techniques to extract pertinent entities from the text data, such as geographic locations, disease names, and symptom descriptions. Our approach involves treating the issue as a named entity recognition task and constructing a dataset by utilizing news articles to train language models like BERT. Additionally, we will address the challenges associated with working with low-resource languages and the need for machine translation.

References

Paquet, C., Coulombier, D., Kaiser, R., and Ciotti, M. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Eurosurveillance*, 11(12):5–6, Dec. 2006. ISSN 1560-7917. doi: 10.2807/esm.11.12.00665-en. URL https://www.eurosurveillance.org/content/10.2807/esm.11.12.00665-en. Publisher: European Centre for Disease Prevention and Control.

Shausan, A., Nazarathy, Y., and Dyda, A. Emerging data inputs for infectious diseases surveillance and decision making. *Frontiers in Digital Health*, 5, Apr. 2023. ISSN 2673-253X. doi: 10.3389/fdgth.2023.1131731. URL https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2023.1131731/full. Publisher: Frontiers.