Digital Surveillance for Infectious Diseases through Large Language Models

Ny Haingo Miantsatiana Andry (haingo@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Dr. Joicymara Xavier
Centre for Epidemic Response and Innovation, South Africa
Universidade Federal dos Vales do Jequitinhonha e Mucuri, Brazil

Co-supervised by: Dr. Houriiyah Tegally
Centre for Epidemic Response and Innovation, South Africa
Stellenbosch University, South Africa

6 June 2024

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Scan your signature

Ny Haingo Miantsatiana Andry, 6 June 2024

Contents

1. Introduction

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. 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.

Real-time informal data sources, such as social media and news articles, have proven to be incredibly valuable for monitoring and early detection of infectious diseases. News media, for instance, can provide early warning signs of heightened disease activity, even before official sources have reported them (Brownstein et al., 2009). There is a growing body of research that supports the use of social media data for infectious disease surveillance and prediction (Wang et al., 2023). This approach, known as event-based surveillance, is much faster than traditional indicator-based surveillance, which depends on formal reports from healthcare institutions (Paquet et al., 2006). By capturing data in real time, event-based surveillance offers timely warnings. Text-based epidemic intelligence has proven its worth in various outbreak scenarios, including the early detection of the A(H1N1) pandemic and the more recent COVID-19 pandemic (Brownstein et al., 2009; Shausan et al., 2023).

1.1 Overview of text-based epidemic intelligence

We refer to the detection of epidemics using health-related textual data as text-based epidemic intelligence. This process can be viewed as a two-step procedure, as shown in Figure 1.1 illustrates the general pipeline for epidemic intelligence (Joshi et al., 2020).

- **Health mention classification**: This step refers to identifying text pertinent to epidemic intelligence. It can be broken down into a sequence of interconnected natural language processing components. Most digital surveillance systems employ a relevancy classifier as part of their health mention classification, which can be binary or more fine-grained (Valentin et al., 2021; Meng et al., 2022).
- **Health event detection**: In this step, the text concerning public health risks of interest is taken as an input to predict health events. A health event may be an outbreak over time or a geographical region. The textual units may be arranged based on their associated timestamps to predict an outbreak over time. Similarly, to predict an outbreak over geographical regions, the textual units may be arranged based on the related regions.

Figure 1.1: General pipeline of text-based epidemic intelligence

1.2 Objective

Our goal is to develop an AI system for disease surveillance. By leveraging natural language processing techniques, we aim to transform unstructured text data from news articles into structured, actionable information. This structured data will enable rapid monitoring of infectious diseases on a global scale.

To achieve this, we will harness language models, such as BERT, and explore the potential of generative

large language models. By treating the task as information extraction, we will use these models to identify critical entities within the text.

1.3 Structure of the essay

This essay is composed of five chapters. Chapter 2 provides an extensive literature review, while Chapter 3 details the data collection methods and methodology. Chapter 4 offers a presentation and analysis of the findings. Finally, Chapter 5 concludes our essay.

2. Literature review

A good deal of prior work has attempted to use natural language processing for infectious disease surveillance and forecasting. This chapter will review some of the key approaches and systems that have been developed.

2.1 Event-based surveillance systems

One notable example of an event-based surveillance system is Biocaster, which employs an ontology-enhanced approach. The system scans news feeds for health-related articles, including multilingual sources. Relevant articles are translated into English, and topic classification techniques are applied to identify those pertinent to the medical domain. Information extraction methods, such as named entity recognition, are then used to construct relationships within the BioCaster ontology. The system utilizes the derived relation tuples to predict public health events (Collier et al., 2008; Meng et al., 2022).

BlueDot, initially started as a transport network modeling tool but has since expanded to monitor and forecast infectious disease outbreaks. BlueDot utilizes both AI and human moderation and offers search capabilities in multiple languages. However, it is important to note that BlueDot is not publicly accessible and is available only to paying clients. Moreover, the system has access to closed-source information, such as government data, which is typically provided by clients (MacIntyre et al., 2023).

EBS systems are tailored to meet specific needs and priorities, such as geographical or disease-specific considerations, and employ customized surveillance definitions and data collection methods. As such, there is no universal EBS system that fits all situations; instead, the optimal choice depends on the unique context and requirements of each specific scenario.

3. Dataset and methodology

References

Brownstein, J. S., Freifeld, C. C., and Madoff, L. C. Digital Disease Detection — Harnessing the Web for Public Health Surveillance. *The New England journal of medicine*, 360(21):2153–2157, May 2009. ISSN 0028-4793. doi: 10.1056/NEJMp0900702. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917042/.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., and Taniguchi, K. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941, Dec. 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn534. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2639299/.

Joshi, A., Karimi, S., Sparks, R., Paris, C., and Macintyre, C. R. Survey of Text-based Epidemic Intelligence: A Computational Linguistics Perspective. *ACM Computing Surveys*, 52(6):1–19, Nov. 2020. ISSN 0360-0300, 1557-7341. doi: 10.1145/3361141. URL https://dl.acm.org/doi/10.1145/3361141.

MacIntyre, C. R., Chen, X., Kunasekaran, M., Quigley, A., Lim, S., Stone, H., Paik, H.-y., Yao, L., Heslop, D., Wei, W., Sarmiento, I., and Gurdasani, D. Artificial intelligence in public health: the potential of epidemic early warning systems. *Journal of International Medical Research*, 51 (3):03000605231159335, Mar. 2023. ISSN 0300-0605. doi: 10.1177/03000605231159335. URL https://doi.org/10.1177/03000605231159335. Publisher: SAGE Publications Ltd.

Meng, Z., Okhmatovskaia, A., Polleri, M., Shen, Y., Powell, G., Fu, Z., Ganser, I., Zhang, M., King, N. B., Buckeridge, D., and Collier, N. BioCaster in 2021: automatic disease outbreaks detection from global news media. *Bioinformatics*, 38(18):4446–4448, Sept. 2022. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btac497. URL https://academic.oup.com/bioinformatics/article/38/18/4446/6651060.

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.

Valentin, S., Arsevska, E., Rabatel, J., Falala, S., Mercier, A., Lancelot, R., and Roche, M. PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, 13:100357, Dec. 2021. ISSN 2352-7714. doi: 10.1016/j.onehlt.2021.100357. URL https://www.sciencedirect.com/science/article/pii/S2352771421001476.

Wang, A., Dara, R., Yousefinaghani, S., Maier, E., and Sharif, S. A Review of Social Media Data Utilization for the Prediction of Disease Outbreaks and Understanding Public Perception. *Big Data and Cognitive Computing*, 7(2):72, June 2023. ISSN 2504-2289. doi: 10.3390/bdcc7020072. URL https://www.mdpi.com/2504-2289/7/2/72. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.