PHS-LLM: Real-time Multilingual Multiplatform Public Health Surveillance Pipeline using Large Language Models

- Specific Aims:

- Specific Aim 1: Finetuning Large Language Models (LLMs) for generalizable public health surveillance

Public health surveillance aims to monitor public health attitudes and behaviors with crucial population health consequences. For example, during the COVID-19 pandemic, vaccination were employed by policymakers worldwide and averted millions of preventable deaths. Vaccination highly relies on public acceptance and perspective, which can be monitored using social media data. Traditionally, surveys are employed to understand vaccine hesitancy. Recently, more and more researchers have turned to large-scale social media data and deep learning as a supplement to surveys, since social media data can be acquired instantaneously to achieve real-time public surveillance. Thanks to the development of Large Language Models (LLM), which make up for the drawback of traditional deep learning models on model generalizability (the capability of undertaking different tasks using the same model), it is possible to conduct real-time public health sentiment surveillance on social media using LLM to monitor a wide array of public health topics in real-time.

- Specific Aim 2: Designing a benchmark for multilingual multiplatform public health inforveillance

During this research, we are going to fine-tune BLOOMZ-7B, a state-of-the-art multilingual LLM, for generalizable public health surveillance that can be applied easily in low-resource settings. We will design a benchmark for measuring LLM's capability in multilingual and multiplatform social media inforveillance. The benchmark will cover a number of crucial topics in public health surveillance, such as adherence to public health interventions, trust in policymakers, and mental health. Data from the benchmark comes from social media platforms like X, Weibo, Xiaohongshu, and Reddit, covering multiple languages from major language families. The benchmark would be open to the public to evaluate future models.

- Specific Aim 3: Evaluating LLMs in public health surveillance settings

At the final stage of our research, evaluate the performance of the LLM we finetuned, along with existing LLMs, regarding their performance in public health surveillance settings. We will present the result in both numerical and graphical form.

- Research Strategy ~3 pages

- Significance

The effectiveness of many public health interventions (such as social distancing, COVID-19 testing, and vaccination) relies on the public's adherence [1, 2, 3]. Based on social media data, previous studies trained machine learning models to learn about sentiments toward public health interventions. For example, trajectories of sentiments can reveal how public adherence fluctuates due to policy changes. [1, 2, 3, 4] Also statistical analysis, such as fixed effect models and multiple regression, can reveal linkages between policy and sentiment [1, 3]. In addition, topic modeling can show the public's top concerns [2]. These studies aspire to use streaming social media data to inform policymakers close to real-time [1, 2, 3, 4]. However, researchers needed to annotate training data and train machine-learning models before implementing social media-based public health surveillance, because traditional machine-learning models are lacking in generalizability [1, 2, 3]. Currently, there's a lack of real-time surveillance based on social media data [4]. Such a gap leads to delays in evidence-based policy adjustments and may lead to population-level health consequences, such as failing to avert millions of preventable deaths [5].

- Innovation

The innovation in this project lies in developing Large Language Models (LLMs) specifically tailored for public health surveillance, an area previously unexplored. By finetuning BLOOMZ-7B with state-of-the-art techniques such as Flash attention-2, DeepSpeed ZeRO Stage3, and QLoRA, this research pioneers a scalable, efficient approach for real-time, multilingual, and multiplatform health monitoring. This strategy represents a significant evolution from traditional machine learning approaches, providing a dynamic tool capable of adapting to various languages and social media platforms. It facilitates the integration of LLMs-driven surveillance into existing public health systems, enhancing evidence-based policymaking through timely insights into public sentiments and behaviors worldwide. This innovative methodology is expected to significantly improve the speed and accuracy of public health interventions, leveraging

the generalizability of LLMs to offer a groundbreaking solution in the domain of global health surveillance.

- Research Plan

Jan 15 - Feb 20 Collect Datasets for finetuning and developing the benchmark
Feb 20 - March 20 Instruction-tune LLM
March 20 - April 20 Evaluate and compare model
April 20 - present the findings

- Specific Aim

- **Specific Aim 1: Fine Tuning Large Language Models (LLMs) for generalizable public health surveillance**

- Hypothesis: By fine tuning LLMs with a focus on public health surveillance, these models can efficiently process and analyze multilingual social media data in real-time, which improve model performance while lowering the cost.

- Rationale: Traditional methods of public health surveillance, such as surveys and questionnaires, have been instrumental in understanding public attitudes and behaviors towards health interventions, like vaccination. However, these methods are limited by their inability to provide real-time insights, a gap increasingly being filled by the use of social media data. Social media platforms offer a rich, real-time source of public sentiment and behavior but pose significant challenges in terms of data volume, velocity, and variety.

LLMs have the potential to significantly enhance public health surveillance. The inherent capabilities of LLMs make them particularly well-suited for sifting through vast amounts of data to identify genuine public health concerns. By fine tuning these models with data specifically curated for public health surveillance, they can be optimized to efficiently process and analyze content on social media. This process is critical for enhancing the speed and accuracy of detecting public health trends, sentiments towards interventions, and misinformation, which are essential for informed decision-making and timely intervention in public health crises and evidence-based policymaking.

- Experimental Approach

Data Preparation: We begin by collecting a wide array of social media data related to public health discussions, ensuring a diverse representation of languages and topics, such as vaccination debates and mental health awareness. This data is sourced

through publicly available data sources, collaborations with public health organizations, and direct access via social media platform APIs.

Model Fine-Tuning: With BLOOMZ-7B models as our foundation, we employ instruction-based fine-tuning techniques, utilizing advanced tools like Flash attention-2, DeepSpeed ZeRO Stage3, and QLoRA. This process adapts the models to better understand and analyze public health discourse, leveraging their multilingual capabilities to ensure wide applicability.

Validation and Iteration: The fine-tuned models are evaluated using a separate validation set, focusing on accuracy, precision, recall, and F1 score across various public health topics and languages. This iterative evaluation process allows us to identify and address any deficiencies, refining the models to improve performance continually.

Integration into Surveillance Pipeline: Finally, the optimized models are integrated into a real-time public health surveillance pipeline. This integration involves developing an interface for the models to process live data streams from social media platforms, ensuring the system is scalable and can handle high-volume data processing. This approach not only enhances the real-time monitoring capabilities of public health surveillance systems but also ensures adaptability to new challenges and discussions in the public health domain.

- Interpretation of Results We address this with aim 2.

- Potential Problems and Alternative Approaches Challenges such as ensuring the representativeness of social media data, addressing biases and hallucination in LLMs, and navigating the regulatory landscape regarding data use and privacy will be important to address. Additionally, the continual evolution of social media platforms and user behavior necessitates ongoing adjustments to the surveillance pipeline to maintain its effectiveness and relevance.

- **Specific Aim 2: Designing a benchmark for multilingual multiplatform public health inforveillance**

- Hypothesis: A comprehensive benchmark designed for evaluating LLMs capability and flexibility for multilingual multiplatform social media inforveillance will be established. The effective design and implementation of a benchmark will significantly enhance the accuracy, timeliness, and cultural sensitivity of health information monitoring, by providing a uniform standard for measuring LLM performance in social media inforveillance.

- Rationale: Public health inforveillance, or the monitoring and analysis of health information from various sources, is crucial for early detection of outbreaks, understanding public perceptions of health risks, and guiding public health policy and response strategies. Establishing a benchmark facilitates the standardization of methods and metrics for evaluating public health information across languages and platforms. This standardization is essential for comparing LLM performance while taking into account the cost of LLM, enhancing real-world inforveillance efforts. It also encourages the development of tools and algorithms that meet these standards, fostering innovation in the field.

- Experimental Approach We'll construct a comprehensive benchmark encompassing diverse public health scenarios, languages, and social media platforms. This involves curating datasets that reflect a wide array of public health discussions, ensuring inclusivity of underrepresented languages and regions. A suite of evaluation metrics will be designed to measure LLM performance across accuracy, timeliness, cultural sensitivity, and adaptability, with an additional focus on cost-effectiveness for deployment in varied settings. These metrics aim to provide a holistic view of an LLM's capabilities in public health inforveillance.

- Interpretation of Results The benchmark will undergo rigorous validation using both general-purpose and public health-specific LLMs to ensure its effectiveness and identify any evaluation biases. This phase is crucial for refining the benchmark and ensuring it accurately assesses LLM performance. Feedback from a broad spectrum of stakeholders, including public health experts, data scientists, and social media platform representatives, will be incorporated to fine-tune the benchmark. This step ensures the benchmark's relevance and utility across the public health inforveillance community.

- Potential Problems and Alternative Approaches

Ensuring comprehensive coverage of public health topics, languages, and cultural contexts is critical to avoid biases in model evaluation. Collaborating with international

health organizations and employing synthetic data generation can enhance dataset diversity, filling gaps in underrepresented areas and languages.

The resource-intensive nature of developing and maintaining a multilingual and multiplatform benchmark could limit access for researchers in low-resource settings. Optimizing computational efficiency and providing cloud-based benchmarking tools, along with forming partnerships with tech companies, could offer sustainable solutions, making the benchmark accessible to a wider research community. These combined efforts will pave the way for a robust, ethical, and inclusive benchmarking framework, enhancing LLM evaluation in public health surveillance.

- Specific Aim 3: Evaluating LLMs in public health surveillance settings

- Hypothesis: Based on the benchmark we created, we can compare the model we finetune with existing off-shelf models, enabling the analysis of public health surveillance data concerning previously unseen topics and datasets with a high degree of accuracy and cost-effectiveness.

- Rationale: LLMs have shown great promise in various domains, including public health surveillance, due to their ability to achieve general-purpose language generation and understanding. However, their performance in specialized tasks such as public health surveillance—especially when dealing with multilingual data, diverse topics, and dynamic, real-time health-related information—remains underexplored. Evaluating LLMs using our newly developed benchmark can systematically assess LLM’s capabilities and limitations in this context. This evaluation aims to identify the most effective models and fine-tuning approaches that can handle the complexities of public health data, ensuring accurate, timely, and culturally sensitive surveillance outcomes.

- Experimental Approach and Interpretation of Results

Utilizing the benchmark, we'll conduct comprehensive performance evaluations of both our fine-tuned models and selected off-the-shelf models. This involves assessing their ability to accurately classify, interpret, and predict public health trends from social media data, spanning multiple languages and platforms. A crucial part of the evaluation will focus on the models' performance in dealing with unseen topics and datasets, testing their generalizability and adaptability to new public health challenges. Alongside accuracy and performance metrics, we'll analyze the cost-effectiveness of deploying these models in real-world public health surveillance settings. This includes considerations of computational resources and the feasibility of scaling these solutions. Evaluating the models' capacity to provide culturally sensitive insights and timely

analysis will be integral. This ensures the surveillance outcomes are not only accurate but also relevant and respectful of diverse global perspectives.

- Potential Problems and Alternative Approaches

Evaluating Large Language Models (LLMs) for public health surveillance introduces critical challenges, notably in data quality and bias, language and cultural nuances, and privacy and ethical considerations. Ensuring models are trained on diverse, representative datasets is essential to mitigate biases and improve the accuracy of health-related predictions across global populations. Moreover, the complexity of language and cultural contexts demands sophisticated natural language processing capabilities to ensure culturally sensitive and accurate surveillance outcomes. Privacy and ethical concerns are paramount, as the use of social media data for health surveillance necessitates stringent data handling protocols to protect individual rights and maintain public trust. Addressing these issues is crucial for developing effective, ethical, and globally applicable LLM-based public health surveillance systems.

References

[1] Zhou, Xinyu, et al. "Spatiotemporal trends in COVID-19 vaccine sentiments on a social media platform and correlations with reported vaccine coverage." Bulletin of the World Health Organization 102.1 (2024): 32.
[2] Zhou, Xinyu, et al. "Comparison of public responses to containment measures during the initial outbreak and resurgence of COVID-19 in China: infodemiology study." Journal of medical Internet research 23.4 (2021): e26518.
[3] Zhou, Xinyu, et al. "Deep Learning Analysis of COVID-19 Vaccine Hesitancy and Confidence Expressed on Twitter in 6 High-Income Countries: Longitudinal Observational Study." Journal of Medical Internet Research 25 (2023): e49753.
[4] Tsao, Shu-Feng, et al. "What social media told us in the time of COVID-19: a scoping review." The Lancet Digital Health 3.3 (2021): e175-e194.
[5] Cai, Jun, et al. "Modeling transmission of SARS-CoV-2 omicron in China." Nature medicine 28.7 (2022): 1468-1475.