SOCIAL SENSING FOR PUBLIC HEALTH : TEPISENSE AI SYSTEM

ABSTRACT

The rapid proliferation of social media has created a novel data source for real-time public health surveillance, offering a potential solution to the limitations of traditional, passive reporting systems. This paper presents TepiSense, an artificial intelligence-based system designed for the real-time surveillance of epidemic-prone diseases using Twitter data. Unlike existing systems that often equate the volume of disease-related posts with outbreak intensity, TepiSense introduces a critical refinement by classifying tweets into 'general' discussions and 'indication' tweets that specifically highlight the presence of patients or active cases. The system integrates a comprehensive pipeline of pre-processing, feature extraction (TF-IDF, Word2Vec, and a hybrid approach), and classification using nine machine/deep learning models and three Large Language Models (LLMs). Evaluated on a novel, manually annotated Twitter Epidemic Surveillance Corpus (TESC) of 23.9K English and 13K Urdu tweets across six diseases, the system demonstrates that the mBERT LLM achieves superior performance with F-measure values of 0.96 and 0.83 for topic and indication classification, respectively. Crucially, correlation analysis with real-world COVID-19 case data from the WHO reveals that 'indication' tweets generated by TepiSense show a significantly stronger correlation (0.58-0.63) than raw keyword-based tweets, validating its epidemiological relevance. The development of an interactive dashboard further operationalizes TepiSense, establishing it as a powerful, scalable tool for enhancing public health intelligence and enabling timely intervention.

EXISTING SYSTEM

The existing paradigm for social media-based disease surveillance, as exemplified by systems like Epitweetr and foundational data collection frameworks such as ESS-T, primarily relies on a streamlined pipeline. This pipeline involves collecting tweets based on a predefined set of disease-related keywords and hashtags. The core metric for assessing disease intensity or public concern is typically the raw frequency or temporal trend of these keyword-matched tweets. These aggregated counts are then often visualized on dashboards, sometimes enhanced with basic geo-location filtering derived from user profile metadata or tweet content. The underlying assumption is that a surge in mentions of a disease term like "dengue" on a social platform corresponds directly to an increase in actual disease incidence in the population.

Disadvantages of the Existing System:

1. Inability to Discern Epidemiological Signal from Noise: The most significant drawback is the treatment of all disease-mentioning tweets as equal. These systems cannot differentiate between a tweet reporting "I have a high fever and tested positive for dengue" (a high-value 'indication' tweet) and one discussing "New research on dengue prevention" or "Historical dengue cases from 2010." This conflation leads to a noisy and often misleading signal that reflects public interest or media coverage more accurately than true disease prevalence.
2. Reliance on Keyword-Based Data Collection: The initial data gathering is based on simple keyword matching. This method is prone to including a substantial volume of irrelevant tweets, such as those using disease names homonymously (e.g., "corona" referring to the sun) or for promotional purposes, which dilutes the quality of the dataset and necessitates extensive post-hoc filtering.
3. Limited Linguistic and Geographic Scope: The development and application of these systems have been largely concentrated on English-language content and developed countries. This creates a significant blind spot for non-English speaking regions, such as Pakistan where Urdu is widely used on social media, despite these regions often being highly vulnerable to outbreaks of the very diseases under surveillance.

PROPOSED SYSTEM

The proposed system, TepiSense, introduces a sophisticated, AI-driven architecture that fundamentally enhances social media-based epidemic surveillance. It moves beyond simple keyword counting to a nuanced understanding of tweet content. The system is structured around four core modules: a Pre-processor for cleaning and normalizing text; a Feature Extractor that employs TF-IDF, Word2Vec, and a hybrid of both to convert text into numerical representations; a Classifier that leverages a comprehensive suite of nine machine/deep learning models and three state-of-the-art Large Language Models (mBERT, mT5, XLM-RoBERTa); and an Evaluator for performance assessment. TepiSense's innovation lies in its two-stage classification process. First, it identifies tweets relevant to a specific disease ('topic' classification), and second, it performs a critical fine-grained classification to label these tweets as either 'general' or 'indication,' with the latter being the primary signal for disease activity.

Advantages of the Proposed System:

1. High-Fidelity Signal Extraction through 'Indication' Classification: By distinguishing 'indication' tweets from general chatter, TepiSense filters out the noise that plagues existing systems. This provides public health officials with a much cleaner, more actionable signal that is directly correlated with real-world case data, as validated by a correlation of 0.58-0.63 with WHO COVID-19 statistics.
2. Superior Accuracy Powered by Advanced LLMs: The integration and benchmarking of modern Large Language Models, particularly mBERT, enable TepiSense to achieve exceptional classification accuracy (F-measure of 0.96 for topic, 0.83 for indication). These models excel at understanding context and semantic nuance, far surpassing the capabilities of simpler keyword-based or traditional ML models.
3. Multi-lingual and Clinically Validated Framework: TepiSense is designed and evaluated on both English and Urdu tweets, making it applicable to a broader and more diverse demographic. Furthermore, the system's output is not just a metric of online activity but is empirically validated against ground-truth health data, establishing its credibility and practical utility for real-world public health decision-making and resource allocation.

SYSTEM REQUIREMENTS

- **H/W System Configuration:-**

➢ Processor	-	Pentium –IV
➢ RAM	-	4 GB (min)
➢ Hard Disk	-	20 GB
➢ Key Board	-	Standard Windows Keyboard
➢ Mouse	-	Two or Three Button Mouse
➢ Monitor	-	SVGA

SOFTWARE REQUIREMENTS:

Operating system	**:**	Windows 7 Ultimate.
Coding Language	**:**	Python.
Front-End	**:**	Python.
Back-End	**:**	Django-ORM
Designing	**:**	Html, css, javascript.
Data Base	**:**	MySQL (WAMP Server).