Vol. 5 No. 2 (2026)

Pakistan Journal of Medical & Cardiological Review

https://pakjmcr.com/index.php/1/about

Online ISSN 3007-2387

Print ISSN 3007-2379

# NLP-Based Medical Text Mining for Early Detection of Disease Outbreaks and Public Health Trends

**Hamdan Tariq**
College of Ophthalmology and Allied Vision Sciences, King Edward Medical University, Lahore & University of Chester, United Kingdom
Email: hamdantariqod07@gmail.com
**Syed Muhammad Junaid Hassan (**Corresponding Author)
Assistant Professor, Department of Information Technology, Faculty of ICT, Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS) Email: smjunaid.it@gmail.com
**Muhammad Essa Siddique**
PhD (IT) scholar at Dr. A. H. S Bukhari Postgraduate Centre of ICT, Faculty of Engineering & Technology, University of Sindh Jamshoro
Email: Essasiddique@live.com
**Wahaj Ali**
Department of Information Technology, The Islamia University of Bahawalpur
Email: wahaj.ali@iub.edu.pk
**Shehryar Irshad**
National University of Modern Languages (NUML)
Email: shehryarirshad11@gmail.com
**Talha**
University of Makran Email: talha@uomp.edu.pk

**Author Details**

**Keywords:** Natural Language Processing, Medical Text Mining, Event-Based Surveillance, Disease Outbreak Detection, Biobert, Clinicalbert, Syndromic Surveillance, Electronic Health Records, Public Health Intelligence, Deep Learning In Epidemiology

**Received on 01 May 2026**
**Accepted on 20 May 2026**
**Published on 01 Jun 2026**

**Corresponding E-mail & Author*:**

**Syed Muhammad Junaid Hassan (**Corresponding Author)
Email: smjunaid.it@gmail.com

**Abstract**

Natural Language Processing (NLP) and medical text mining have emerged as transformative tools for shifting public health surveillance from reactive, indicator-based systems to proactive, event-based intelligence. This review explores how advanced NLP techniques including Named Entity Recognition (NER), relationship extraction, text classification, and sentiment analysis enable the real-time extraction of actionable insights from unstructured data sources such as electronic health records (EHRs), clinical narratives, news reports, and social media. Domain-adapted models like BioBERT, ClinicalBERT, and BERTweet, combined with deep learning architectures (Bi-LSTM with multi-head attention achieving 98.25% accuracy), facilitate early detection of disease outbreaks, syndromic surveillance, and trend monitoring. Global frameworks such as HealthMap and WHO's EIOS demonstrate the practical impact of these technologies. While challenges including data noise, cross-lingual privacy risks, and the digital divide persist, multimodal fusion and AI-driven systems offer significant potential for improving epidemic preparedness, response speed, and public health decision-making in an increasingly interconnected world.

## Introduction

The paradigm of global health surveillance is undergoing a fundamental shift from reactive, indicator-based reporting to proactive, event-based intelligence. This transition is necessitated by the complexities of a globalized society, where the speed of pathogen transmission often outpaces the administrative and laboratory timelines of traditional public health systems (Aryffin et al., 2025a). At the heart of this evolution is Natural Language Processing (NLP) and medical text mining, disciplines within biomedical informatics that enable the extraction of actionable insights from the vast, unstructured narratives that constitute the majority of modern medical and digital data (Li et al., 2024). As the volume of electronic health records (EHRs), clinical narratives, social media interactions, and news reports continues to expand, NLP provides the scalable solutions required for real-time patient phenotyping, disease prediction, and public health decision support (Sindhushree et al., 2025).

## The Strategic Shift to Event-Based Surveillance

Historically, public health authorities have relied on indicator-based surveillance (IBS), which utilizes structured data from healthcare facilities and diagnostic laboratories. While IBS provides high specificity through laboratory-confirmed cases, it is frequently plagued by significant reporting lags due to the time required for patients to seek care and for diagnostic tests to be processed (Beyer, 2025). Furthermore, IBS often produces incomplete data concerning emerging infectious diseases, as traditional systems are optimized for known pathogens rather than novel threats (Klem et al., 2017).

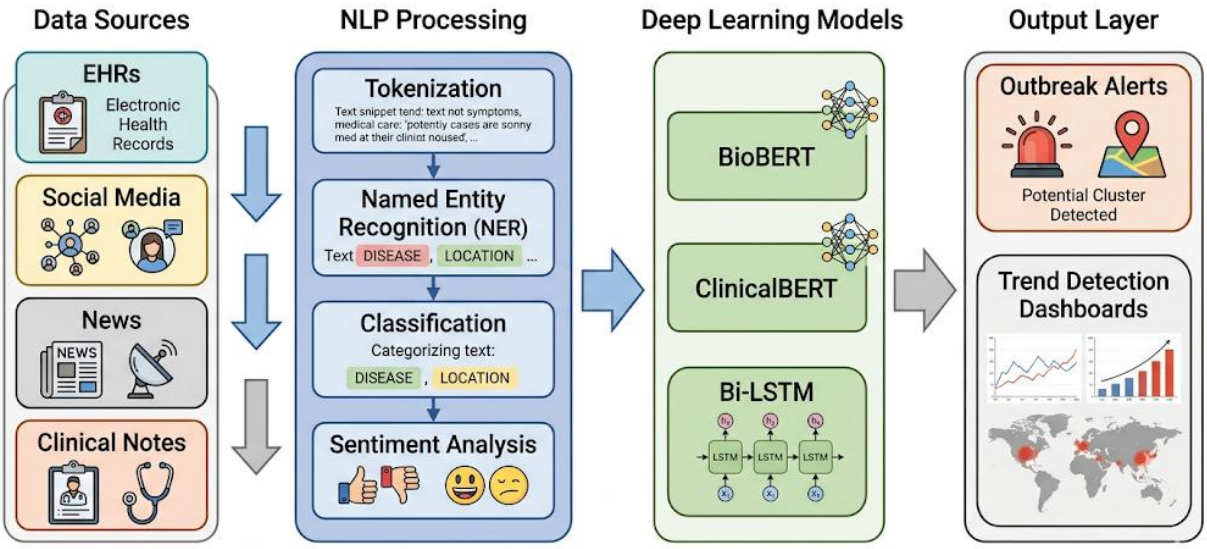
In response to these limitations, event-based surveillance (EBS) has emerged as a critical adjunct. EBS focuses on the systematic collection, analysis, and interpretation of unstructured data from diverse sources, including news media, social platforms, and clinical notes, to identify unusual health events or clusters before they are formally recognized by official channels (Abbood et al., 2020). This approach is designed to detect early epidemic signals by monitoring the "infosphere" the continuous stream of digital communication that often reflects public concerns long before a formal diagnosis is recorded (Pant et al., 2025).

The efficacy of EBS was demonstrated during the 2014 Ebola epidemic. Digital surveillance systems identified clusters of a "strange fever" nine days before the World Health Organization (WHO) released a formal confirmation (Hao et al., 2022). Similarly, during the initial stages of the COVID-19 pandemic, digital intelligence platforms flagged undiagnosed pneumonia clusters in Wuhan on December 30, 2019, providing an early warning that preceded international alerts (Wellcome Trust, 2023).

## Technical Foundations and the Evolution of Clinical NLP

The technical landscape of medical text mining has transitioned through distinct phases: rule-based systems, traditional machine learning (ML), and the current era of deep learning (DL) and large language models (LLMs). Each phase has addressed specific challenges inherent in the complexity of medical language, such as idiosyncratic acronyms and technical jargon (Kraveer, 2025). Figure 1 illustrates the complete workflow of NLP-based medical text mining systems used for outbreak detection. The integration of heterogeneous data sources with advanced NLP models enables real-time extraction of epidemiological signals.

**Figure 1. Workflow of NLP-Based Medical Text Mining for Disease Outbreak Detection**




**Rule-Based and Statistical Foundations**

Early NLP applications relied on manually curated rules. While highly interpretable, these required domain experts to curate rules for every possible linguistic expression. Rule-based systems are often rigid and fail to generalize across informal data sources like social media (Liu et al., 2024). Statistical machine learning models like Random Forests and Naive Bayes classifiers eventually gained prominence by leveraging annotated datasets to identify patterns (Aryffin et al., 2025b).

**The Transformer Revolution and Domain Adaptation**

The introduction of the Transformer architecture, based on self-attention mechanisms, allowed models to process text in parallel, enabling the capture of global context (. Bidirectional Encoder Representations from Transformers (BERT) has achieved state-of-the-art results across various public health tasks (Ghojogh & Ghodsi, 2020).

**Table 1. Domain-Specific BERT Variant Models and Use Cases**

| Model Variant | Pre-training Data | Primary Use Cases |
|---|---|---|
| BioBERT | PubMed abstracts and PMC articles | Biomedical NER, relation extraction (Kraveer, 2025; Li et al., 2024) |
| ClinicalBERT | MIMIC-III ICU clinical notes | Patient phenotyping, EHR mining (Kraveer, 2025; Sindhushree et al., 2025) |
| SciBERT | Scientific papers | Technical jargon handling (Kraveer, 2025; Sindhushree et al., 2025) |
| BERTweet | English tweets | Social media-based syndromic surveillance (Sindhushree et al., 2025; Villanueva-Miranda et al., 2025) |
| CT-BERT | COVID-19 Twitter data | Pandemic trend tracking (Sindhushree et al., 2025; Villanueva-Miranda et al., 2025) |

**Core Methodologies in Medical Text Mining**

**Named Entity Recognition and Relationship Extraction**

Named Entity Recognition (NER) identifies and categorizes medical concepts such as diseases and symptoms. Specialized models like Survice-BERT identify outbreak-related entities like CASE_COUNT and DEATH_COUNT in reports with high

accuracy. Relationship Extraction (RE) identifies connections between these entities, such as linking a pathogen to a specific exposure (Kafkas & Hoehndorf, 2019).

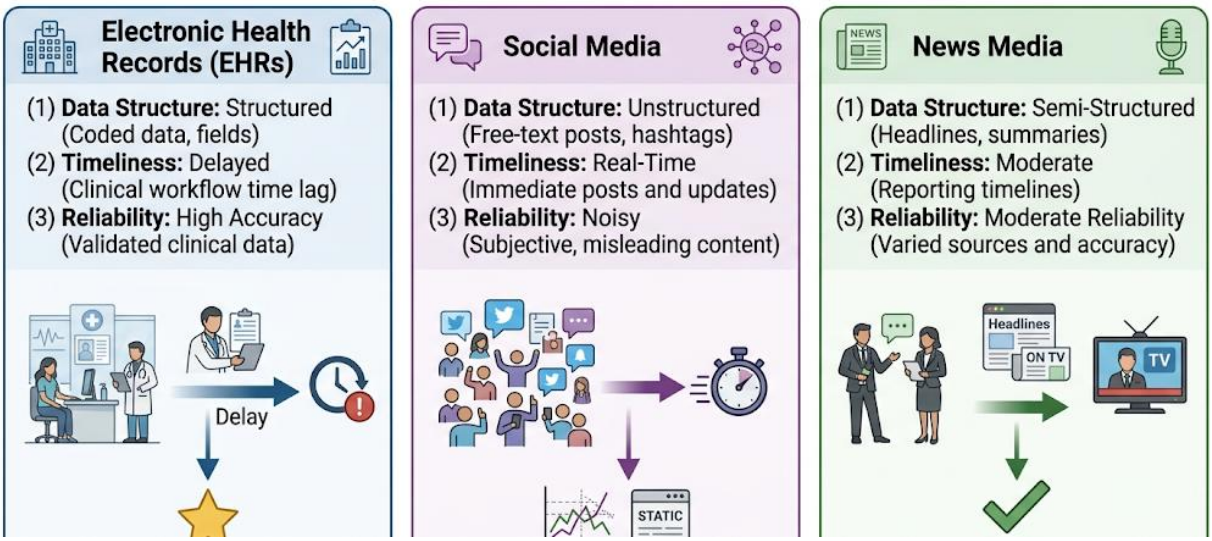
**Text Classification and Sentiment Analysis**

Text classification is used to filter digital information to identify articles relevant to public health concerns. Sentiment analysis gauges the psychological response to a crisis. Monitoring shifts in sentiment can provide early warnings of social non-compliance with health guidelines, which is critical for effective pandemic management (Abbood et al., 2020).

**Data Modalities: Electronic Health Records vs. Social Media**

Electronic Health Records (EHRs) offer objective data such as vital signs and clinical narratives. However, their use is hampered by data fragmentation and a bias toward individuals with insurance. Social media platforms provide real-time data that can precede clinical encounters (Villanueva-Miranda et al., 2025). Despite its speed, social media is subject to "noise," such as media attention causing spikes in topic frequency that do not correspond to actual incidence rates (Wellcome Trust, 2023). Figure 2 shows the different data sources contribute unique strengths and limitations to public health surveillance systems. Integrating these sources improves both the speed and accuracy of outbreak detection.

**Figure 2.** Comparison of Data Sources in Public Health Surveillance




**Advanced Predictive Architectures**

Modern disease prediction models utilize deep learning to capture temporal relationships in news data. Experimental results show that the Bi-LSTM + Multi-Head Attention architecture achieves the highest accuracy for outbreak prediction (CADTH, 2024).

**Table 2. Performance Comparison of Deep Learning Architectures for Disease Outbreak Prediction**

| Model Architecture | Accuracy (%) | Precision | Recall |
|---|---|---|---|
| LSTM (Unidirectional) | 91.60 | 0.90 | 0.89 |
| Bi-LSTM | 92.59 | 0.92 | 0.91 |
| Transformer | 92.76 | 0.93 | 0.92 |
| Bi-LSTM + Multi-Head Attention | 98.25 | 0.98 | 0.97 |
| *(Data source: Sindhushree et al., 2025).* | | | |

The performance of these models is often measured using the F1-score, calculated as follows:

F1=2·Precision·RecallPrecision+Recall*F*1=2·Precision+RecallPrecision·Recall (Li et al., 2024).

**Global Frameworks for Epidemic Intelligence**

Several frameworks operationalize NLP for real-time surveillance. HealthMap queries news aggregators and social media to extract disease mentions and location data. The Epidemic Intelligence from Open Sources (EIOS), led by the WHO, processes millions of multilingual items daily to identify trends (Aryffin et al., 2025a).

**Multimodal Fusion and Ethical Challenges**

The integration of multimodal data fusion (MMDF) consolidates heterogeneous streams like textual notes, genomics, and climate indicators (Villanueva-Miranda et al., 2025). Ethical barriers include cross-lingual privacy leakage, where LLMs reveal sensitive data when queried in a different language. Furthermore, the "digital divide" limits the effectiveness of surveillance in areas with limited internet access (Beyer, 2025).

**Conclusion**

NLP-based medical text mining represents a paradigm shift in public health intelligence, enabling the transition from slow, structured indicator-based surveillance to rapid, event-based systems capable of detecting outbreaks days or weeks before traditional confirmation. By leveraging domain-specific transformer models, deep learning architectures, and multimodal data fusion, these technologies extract critical signals from vast unstructured sources, supporting real-time phenotyping, sentiment tracking, and predictive analytics for diseases ranging from Ebola to COVID-19. Global platforms like HealthMap and EIOS have already proven their value in providing early warnings and informing timely interventions.

However, realizing the full potential requires addressing persistent challenges such as data quality, model interpretability, ethical concerns around privacy and bias, and equitable access in low-resource settings. Future advancements should focus on robust multimodal integration, explainable AI for clinical trust, standardized evaluation frameworks, and stronger collaboration between public health agencies, technology developers, and researchers. As NLP continues to evolve alongside large language models and real-time data streams, it will play a pivotal role in building resilient, proactive global health surveillance systems capable of mitigating the impact of emerging infectious diseases and safeguarding population health in an interconnected world.

**References**


Abbood, A., Ayoola, R., Bakrania, S., Elshafey, A., Gammino, V., Glass, K., Hall, I., House, T., Hughes, G., Keeling, M., Knight, G., Lopez, J., Maina, J., Nsoesie, E., Paolotti, D., Pellis, L., Polonsky, J., Riley, S., Simons, D., & Wardle, J. (2020). EventEpi: A natural language processing framework for event-based surveillance. *PLoS Computational Biology, 16*(11), e1008277. https://doi.org/10.1371/journal.pcbi.1008277

Aryffin, H. A. K., Ismail, M. A., Ahmad, N., Mohamad, M. S., & Zakaria, Z. (2025b). Advancing localized public health surveillance in Malaysia by enhancing EIOS with Google COVID-19 data integration. *Malaysian Journal of Science Health & Technology, 11*(1), 43–52. https://doi.org/10.33102/mjosht.v11i1.455

Liu, Z., Wang, Y., Sun, J., Zhang, X., Mu, Y., & Liu, Y. (2025). *Cross-lingual privacy leakage in large language models*. arXiv. https://arxiv.org/abs/2506.00759

Beyer, H. (2025). *2025 watch list: Top 5 new and emerging AI technologies in health care*. CADTH.(https://canjhealthtechnol.ca/index.php/cjht/article/view/ER0015/2417)

Crescendo AI. (2026, April 2). Google introduces TurboQuant, a memory compression breakthrough for large AI models. *Crescendo AI News*. https://www.crescendo.ai/news/latest-ai-news-and-updates

Kraveer, A. (2025, June 4). Natural language processing in clinical text mining: A review of techniques and applications. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, *2*(1), 57-70.

Li, Y., Li, J., He, J., & Tao, C. (2024). AE-GPT: Using large language models to extract adverse events from surveillance reports A use case with influenza vaccine adverse events. *PLoS ONE*, *19*(3), e0300919. https://doi.org/10.1371/journal.pone.0300919

Liu, L., Blake, V., Barman, M., Gallego, B., Churches, T., Kennedy, G., Ooi, S.-Y., & Delaney, G. P. (2024). Using natural language processing to extract information from clinical text for populating clinical registries: A systematic review. *PLoS ONE*, *19*(7), e02844598. https://doi.org/10.1371/journal.pone.02844598

Methuku, V. (2025). NLP and AI for public health intelligence: Automating disease surveillance from unstructured data. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, *2*(1), 43-56.(https://doi.org/10.62762/TETAI.2025.222799)

Pant, D., Grandhe, R. R., Agrawal, J., Kalra, J. S., Khalikar, S. V., Garg, V.,... Mathew, M. (2025). Health Sentinel: An AI pipeline for real-time disease outbreak detection. *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, 23–42. https://doi.org/10.18653/v1/2025.nlp4pi-1.3

Sindhushree, G. S., Amarnath, R., Nagabhushan, P., & Javed, M. (2025). Disease outbreak prediction using news data and Natural Language Processing (NLP) techniques. *Big Data and Cognitive Computing*, *9*(11), 291. https://doi.org/10.3390/bdcc9110291

Villanueva-Miranda, I., Xiao, J., & Xie, Y. (2025). Artificial intelligence-enabled early warning systems for national infectious disease surveillance. *Frontiers in Public Health*, *13*, 1609615. https://doi.org/10.3389/fpubh.2025.1609615

Wellcome Trust. (2023, February 2). *New digital tools use climate data to better predict and prepare for infectious diseases outbreaks*. https://wellcome.org/insights/articles/new-digital-tools-use-climate-data-better-predict-and-prepare-infectious-diseases-outbreaks

World Health Organization. (2025). *Epidemic intelligence from open sources (EIOS) strategy 2024-2026*.(https://doi.org/10.2471/B09476)

Aryffin, H. A. K., Sahbudin, M. A. B., Ali Pitchay, S., Abhalim, A. H., & Sahbudin, I. (2025a). Technological trends in epidemic intelligence for infectious disease surveillance: A systematic literature review. *PeerJ Computer Science*, *11*, e2874. https://doi.org/10.7717/peerj-cs.2874

Klem, F., Wadhwa, A., Prokop, L. J., Sundt, W. J., Farrugia, G., Camilleri, M., ... & Grover, M. (2017). Prevalence, risk factors, and outcomes of irritable bowel syndrome after infectious enteritis: a systematic review and meta-analysis. *Gastroenterology*, *152*(5), 1042-1054.

Hao, R., Liu, Y., Shen, W., Zhao, R., Jiang, B., Song, H., ... & Ma, H. (2022). Surveillance of emerging infectious diseases for biosecurity. *Science China Life Sciences*, *65*(8), 1504-1516.

Ghojogh, B., & Ghodsi, A. (2020). Attention mechanism, transformers, BERT, and GPT: tutorial and survey.

Kafkas, Ş., & Hoehndorf, R. (2019). Ontology based mining of pathogen–disease associations from literature. *Journal of biomedical semantics*, *10*(1), 15.