A Novel Multi-Stage Rule-Based Information Extraction Framework for Disease Outbreak Detection with Enhanced Geographical Granularity

Manju Joy
Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India | Department of Computer Applications, Federal Institute of Science and Technology, Angamaly, Kerala, India
19phcsp010@avinuty.ac.in (corresponding author)

M. Krishnaveni
Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India
krishnaveni_cs@avinuty.ac.in

Received: 19 September 2025 | Revised: 7 October 2025, 19 October 2025, and 21 October 2025 | Accepted: 22 October 2025

 DOI: https://doi.org/10.48084/etasr.14931

ABSTRACT

A critical limitation of modern event-based bio-surveillance systems is their reliance on headline-level data rather than a systematic analysis of the underlying news narratives. Although they provide general awareness, they often lack the geographical granularity needed to generate actionable insights and support location-specific public health interventions. To address this limitation, this study presents a method for extracting epidemic-related information from unstructured text corpora on digital platforms. By transforming unstructured text into a structured format using advanced NLP techniques, this approach facilitates the visualization of outbreak information on Google Maps with enhanced spatial detail. Kerala is recognized as the "Alarm bell of India" for consistently reporting the first instance of many outbreaks that have emerged throughout the country, including the first Nipah outbreak in 2018, the first COVID-19 case in 2020, and the index Monkeypox case in 2022. This pattern highlights Kerala's pivotal role in the early detection of emerging infectious diseases, highlighting the need for a more advanced and efficient bio-surveillance system to enhance public health preparedness. This study presents a novel approach to identify entities related to outbreaks using an optimized DistilBERT model, combined with a multi-stage rule-based method for automated information extraction from unstructured text corpora, achieving an overall precision of 94.86%. Experimental results demonstrate that the proposed framework is highly effective in tracking diseases that severely impact public health and disrupt socio-economic stability.

Keywords-information extraction; named entity recognition; co-reference resolution; outbreak detection; epidemic surveillance

I. INTRODUCTION

The COVID-19 pandemic has demonstrated that health threats emerging in one country can significantly impact public health in others, regardless of geographical distance, emphasizing the need for efficient bio-surveillance systems. Event-based bio-surveillance systems complement traditional ones, offering real-time infectious disease monitoring that supports a more comprehensive assessment of disease burden [1]. Community-based participatory surveillance and web news extraction are two of the most widely adopted approaches to detect outbreaks in near real-time. In [2], a community-inclusive reporting system was presented for Foot-and-Mouth Disease (FMD) outbreaks in the Gairo district of Tanzania, allowing livestock keepers to report cases by SMS, USSD, and voice calls. This approach enabled local communities to actively participate in surveillance, reducing communication delays and improving the response to outbreaks, but faces limitations due to participants' limited digital literacy and device accessibility. In [3], the impact of Community-Based Healthcare Interventions (CBHIs) on the early detection of non-communicable diseases in Indonesia was investigated. Despite a low participation rate (3.2%), this study highlighted the importance of decentralized, community-driven health

interventions in improving preventive care, but challenges involved low coverage, gender disparities, and a lack of detailed participation data.

Most current community-based surveillance mechanisms are designed to investigate a limited number of diseases. This study focuses on outbreaks of infectious diseases that pose potential threats to human populations. Online news articles are valuable for providing data relevant to public health monitoring and outbreak detection [4]. Since unstructured text in news reports cannot be used directly for surveillance, this study presents an efficient method to transform textual disease descriptions into a structured format. This structured information is then mapped to Google Maps to facilitate real-time visualization of outbreak dynamics and improve public awareness.

Numerous surveillance systems have been proposed that focus on outbreak reporting, using unstructured textual data from online sources. HealthMap, a globally recognized bio-surveillance system, employs a dictionary-based pattern-matching approach to extract outbreak information from global media, supporting the detection of numerous diseases mentioned in different languages [5]. Eagle Eye [6], a Global Framework for Disease-Related Topic Extraction, employs a WBiLSTM-TF-IDF mechanism, a deep learning–driven method that identifies and extracts key terms from diverse Internet sources, including news reports, social networking platforms, and Web search queries, to deliver country-specific disease outbreak information. In [7], PADI-web was proposed to monitor five emerging animal diseases. This method used machine learning on a manually labeled corpus collected from Google News and RSS feeds to extract disease-related entities, focusing on animal health threats. BioPak Flasher [8] was proposed to detect outbreaks in Pakistan, collecting data from Urdu news channels, ProMED Mail, and validated World Health Organization alerts. The news articles were processed using three dictionaries to extract disease and location information—one for disease names, one for Pakistani cities, and one for latitude and longitude information. In [9], BioCaster was presented for Automatic Disease outbreak detection from RSS feeds using Deep neural NLP models to integrate text with knowledge graphs. The pre-trained PubMed BERT was utilized for relevance classification, and entity linking was performed using the SapBERT model.

In [10], a corpus was developed for Latin America, applying NER with RNNs, CRFs, and rule-based methods, and using co-occurrence for relation extraction. PEACOCK [11] is a map-based multitype infectious disease outbreak information system that collects statistical infectious disease outbreak data from the Korea CDC, news reports, and search query data for a data-driven framework to monitor contagious diseases. MedISys [12] was designed to track food and foodborne risks and categorize incoming news articles using information extraction techniques, including keyword searches and entity recognition. Its effectiveness was evaluated using a retrospective case-study approach. EPIWATCH [13] is an AI-powered epidemic intelligence and early warning system that uses NLP-based NER, ArcGIS geolocation, and expert review to process open-source data and detect outbreaks before official recognition, as seen in the Ebola outbreak (2013) and COVID-19 signals in Hubei (2019). However, this system faces challenges such as data overload, false positives, limited public health adoption, and funding constraints.

In [14], GPHIN was proposed to collect outbreak information from the global news services Factiva and Al Bawaba. This system utilizes a blend of artificial intelligence and human expertise to ensure that the information provided is both comprehensive and nuanced. Through the application of a tailored keyword taxonomy combined with a Boolean search, relevant news articles are identified, duplicates are removed, and relevancy scores are assigned based on predefined keyword values.

Existing bio-surveillance systems rely on news headlines to detect disease outbreaks. Although headlines provide quick awareness of emerging events, they lack the contextual depth necessary to obtain actionable insights. Detailed news stories, which accompany these headlines, often contain critical information such as specific locations affected by the disease and the number of cases and fatalities reported. However, these disease monitoring systems rarely exploit this rich and fine-grained information, resulting in limited geographical granularity and inadequate support for public health interventions. To address this gap, the proposed approach processes full-length news stories using a multi-stage, rule-based information extraction pipeline. This enables the identification of outbreak events with enhanced geographical precision and detailed epidemiological attributes, thereby transforming raw news content into actionable knowledge for bio-surveillance.

According to statistics released by the Directorate of Health Services (DHS) of Kerala in July 2024, a significant increase in cases of infectious diseases was observed in this region during the period from January 2024 to July 11, 2024 [15]. A total of 9,259 dengue cases were reported, with 23 fatalities. Among the 1,258 confirmed leptospirosis cases, 67 resulted in deaths. In addition, 919 cases of H1N1 (Influenza A) were documented, with 15 deaths. These data highlight Kerala's susceptibility to emerging infectious diseases, stressing the importance of sustained surveillance, timely detection, and strengthened public health measures. Kerala has a history of being the first state in India to identify and report most of the emerging infectious disease outbreaks that have recently been reported. Thus, Kerala was used as a case study to apply NLP-based methods to extract information on epidemic outbreaks.

Since traditional machine learning models require extensive training data, developing a rule-based approach is a viable alternative for specific domain-focused tasks. Rule-based approaches offer interpretability and domain-specific customization for extracting outbreak information from news reports and other text sources. Academic research often regards rule-based approaches as obsolete, yet they continue to dominate the commercial landscape. Leading technology giants such as IBM, SAP, and Microsoft rely heavily on rule-based systems for information extraction. This highlights a vast, untapped opportunity for researchers to refine and enhance rule-based methods, making them more systematic, effective, and efficient [16, 17].

This study introduces a novel framework that combines transfer learning, machine learning, and NLP techniques to process unstructured text for monitoring disease outbreaks. Since the pre-trained BERT model can be fine-tuned on task-specific small datasets to effectively address a variety of NLP tasks [18], a fine-tuned DistilBERT model, a lighter version of BERT, is used to ensure the precise identification of disease-related entities. The proposed multistage rule-based approach systematically converts unstructured information into a structured form, achieving a precision of 94.86%. Thus, the proposed framework enables efficient exploitation of online data for outbreak detection and geospatial mapping.

II. METHODOLOGY

Disease stories existing as unstructured text and available in news articles are given as input to the model. The steps of the proposed information extraction pipeline include data collection, co-reference resolution, normalization, Named Entity Recognition (NER), relation extraction, Subject-Verb-Object (SVO) triplet generation, and document construction. This method is based on the idea that by integrating a fine-tuned pre-trained language model with a rule-based approach, information extraction can be achieved with greater efficiency and accuracy. Figure 1 illustrates the information extraction pipeline used.

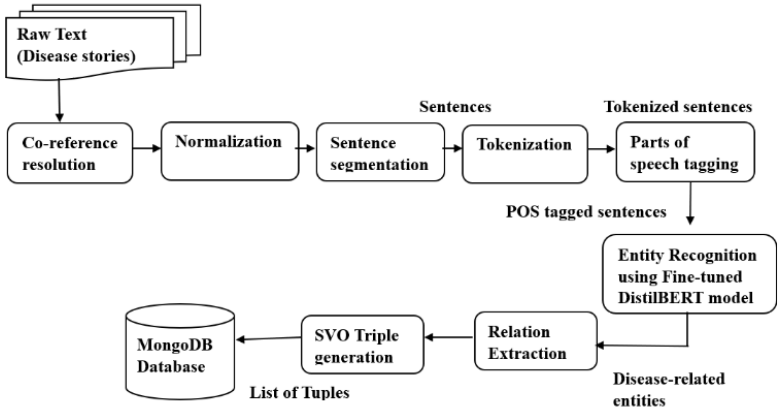

Fig. 1. Method diagram.

A. Data Collection

News articles were collected using a web scraping technique from online portals covering Kerala-specific news. The dataset comprises 1,624 outbreak-related articles covering the period from March 2003 to February 2025. Frequent epidemic-prone diseases in the region include mosquito-borne infections (such as dengue, malaria, and Japanese encephalitis), respiratory diseases (including H1N1 and COVID-19), and other communicable diseases, including hepatitis, leptospirosis, meningitis, and measles. Nipah, Monkeypox, and Amoebic Encephalitis appeared less frequently in news reports.

B. Co-Reference Resolution

Co-reference resolution was carried out to identify and link words or phrases that refer to the same entity within a text. Without resolving these references, there is a risk of discarding the document due to missing or unlinked entities. Co-references were systematically resolved by replacing generic references with their explicit counterparts. For example, occurrences of "the district" were substituted with the corresponding district name, while "the disease" and "the cases" were mapped to the specific disease mentioned in the document.

C. Normalization

Normalization is a crucial task in information extraction to resolve ambiguities that arise from the polysemy and synonymy of terms that occur in documents [19]. By applying these preprocessing techniques, a consistent representation of outbreak-related information is obtained, enhancing the reliability of downstream data extraction and analysis. Normalization was performed using 80 regular expressions. A word-to-digit mapping approach was employed to standardize numerical representations within news articles. This step is vital to ensure consistency in data processing, particularly when extracting outbreak-related numerical information, such as case counts and death counts. Word-to-Digit Mapping was implemented using a predefined dictionary to map commonly used number words (e.g., one, two, dozen) to their corresponding numeric representations (1, 2, 12). Word-based numbers were replaced with their digit equivalents. This transformation enhances numerical consistency, facilitating accurate data mining and analysis in subsequent processing steps. By applying this normalization, instances like "ten cases were reported" were converted to "10 cases were reported," ensuring uniform numerical representation across the dataset. Ordinal references in articles were replaced with structured numeric labels, such as case-1, case-2, death-1, etc. This conversion standardized outbreak records for easier storage and retrieval. For example, the "first case of dengue" was replaced with "case-1 dengue," and the "second death of monkeypox" was replaced with "death-2 monkeypox."

The date expressions were standardized to the format dd-mm-yyyy, and the disease names were normalized using a predefined list of synonyms to eliminate ambiguities. An age

pattern normalization step was used to avoid ambiguity in the extraction of the case count and ensure that the numerical values in the outbreak reports refer only to the number of affected individuals. This step eliminates age references, as they are typically insignificant in the tracking of epidemiological outbreaks. In sentences where an individual's age is mentioned alongside case counts, such as "A 22-year-old boy is among the 20 people affected by monkeypox", the age reference was removed and the sentence was modified to "A boy is among the 20 people affected by monkeypox." This transformation ensures that the numerical values extracted from the outbreak reports correspond precisely to case counts, rather than age-related information.

To ensure the relevance of the information extracted, past outbreak events reported in current stories were filtered out during preprocessing, as they fall outside the temporal scope of interest. Negation handling was performed by converting negative statements into explicit zero counts. For example, the statement "No cases of monkeypox reported this month" was transformed into "0 cases of monkeypox reported this month" to ensure that the extracted data explicitly reflects the absence of cases. Uncertainty in reported data was also handled.

D. Named Entity Recognition (NER), Subject-Verb-Object (SVO) Triplet Generation, and Document Construction

The structural and syntactic features of the textual content were considered to design rules for information extraction. The information extraction process begins with tokenization. After sentence segmentation and tokenization, two fundamental tasks required for the automatic extraction of information from text are the identification of relevant entities and the relationship between them [20-22]. A fine-tuned DistilBERT-based NER model was utilized for the efficient identification of outbreak-related entities, including the names of 14 districts and villages in Kerala, disease names, and cardinal entities representing death or case counts. False positives in NER are problematic because they reduce the precision of the model. DistilBERT is fine-tuned to achieve high precision in entity recognition using the Optuna framework. Precision refers to the percentage of items that the system identifies as locations or disease names that are actually correct. A custom NER dataset, structured in the standard CoNLL-2003 format, comprising 6,627 annotated sentences stored in a CSV file with four columns (sentence number, token, POS tag, and NER label), was used to fine-tune the DistilBERT model. DistilBERT generates contextualized embeddings that encode semantic meaning and dependency between words within the input sequence, while also offering robustness in handling out-of-vocabulary tokens. DistilBERT has demonstrated state-of-the-art performance in the entity recognition task, achieving significant improvements with an accuracy of 96% [23].

In relation extraction, a relation represents how entities appearing in a sentence are correlated. A distance-based heuristic was used to determine the relationship between entities, and SVO tuples were generated to transform unstructured text into structured representations. Since news reports are written in English, an SVO language, this method aligns well with the underlying linguistic structure [24]. An initial set of extraction rules was derived from a linguistic analysis of training documents, which capture the typical characteristics of texts encountered in real-world applications. These rules, formulated based on the features of the training data and the specific concepts targeted for extraction, were then iteratively adjusted and fine-tuned. This approach allowed for seamless incorporation of extensions or modifications to accommodate changes in the document corpus.

Predefined relations were used for linking entities. When rules were applied individually, the SVO tuples generated were not highly accurate. Therefore, a multistage rule-based approach was implemented, in which relevant pieces of information were extracted in three distinct stages, each focusing on a specific category of data, ensuring greater precision and accuracy. In the first stage, geopolitical entities and disease names were identified in sentences, and entity linking was performed. Cardinal and temporal information extraction were performed in subsequent stages. A single-stage method attempts to extract all relevant details in a single pass, increasing the probability of errors due to overlapping entity types, contextual ambiguities, and missed associations. Once structured outbreak information was extracted, it was stored in a MongoDB database, making it readily available for downstream analysis and visualization.

III. RESULTS AND DISCUSSION

A comparative analysis between the proposed multistage and single-stage information extraction approaches was conducted, and the results are shown in Table I. The information extracted using the proposed method in the form of SVO tuples was compared with a set of ground-truth SVO tuples, manually verified by domain experts. By performing a matching process, false positives, false negatives, and standard evaluation metrics such as precision, recall, and F1-score were calculated.

TABLE I. COMPARISON OF MULTI-STAGE WITH SINGLE-STAGE INFORMATION EXTRACTION

Evaluation metrics	**Single -stage**	**Multistage**			
		Stage1	**Stage2**	**Stage3**	**Proposed method**
Precision	**43.04**	98.4	92.14	100	**94.86**
Recall	88.50	99.25	93.58	93.75	**95.96**
F1-score	57.92	98.82	92.85	96.77	**95.41**
Tuple accuracy	90.37	99.66	93.32	97.21	**97.29**
FPR	0.47	0.01	0.10	0.03	0.04
FNR	0.1	0.00	0.07	0.03	**0.03**

In single-stage information extraction, a low precision of 43.04% indicates that the model extracts many incorrect tuples. In multi-stage information extraction, 94.86% of extracted tuples are correct. Concerning recall, in single-stage, 88.50% of correct tuples were extracted compared to ground truth tuples, whereas multistage information extraction extracted 95.96% of correct tuples. In single-stage, 90.37% of extracted tuples matched exactly with ground-truth tuples, whereas 97.29% of extracted tuples matched exactly with ground-truth tuples in multi-stage information extraction. In tuple accuracy, partial matches were not counted as correct. In extracted tuples, a correct subject and object linked using an incorrect verb reduced the tuple accuracy. False Positive Rates (FPR) indicate that 47% of extracted tuples were incorrect in single-stage

information extraction, whereas only 4% of extracted tuples were incorrect in multi-stage. The False Negative Rate (FNR) indicates that 10% of correct tuples were missed in single-stage information extraction, whereas only 3% correct tuples were missed in multi-stage. The results suggest that the multi-stage information extraction strategy is highly modular and interpretable, allowing better control and debugging at each stage.

To facilitate a clearer understanding of the temporal and spatial distribution of disease outbreaks, various visualization techniques can be applied to the structured information obtained using the proposed method. The choropleth map in Figure 2 shows the spatial distribution of COVID-19 cases reported across various districts of Kerala on 21/07/2021 as published in Mathrubhumi News, a leading news platform covering Kerala news. The case statistics provided are visualized using a choropleth map created using a Python library. A quick assessment of the geographic concentration of the disease was made possible, helping to identify hotspots and clusters, with case counts aggregated by color intensity. Darker shades indicate an increase in reported cases.

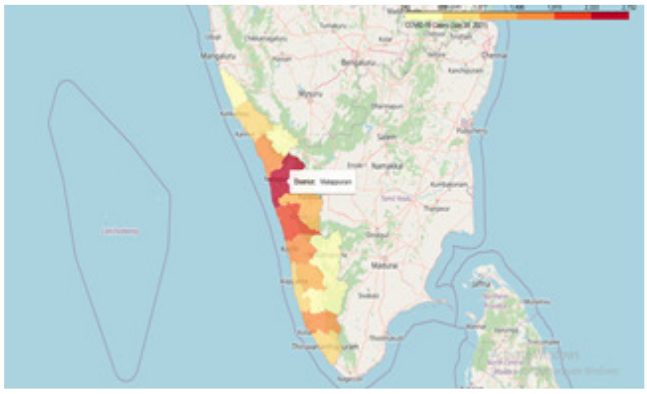

Fig. 2. Choropleth map of COVID-19 cases reported in Mathrubhumi News.

A. Case Study and Comparative Analysis

The performance of the proposed framework was evaluated by comparing the outbreak visualization of a sample news article reporting a death from Amoebic Encephalitis with the corresponding visualization produced by HealthMap, as shown in Figure 3.

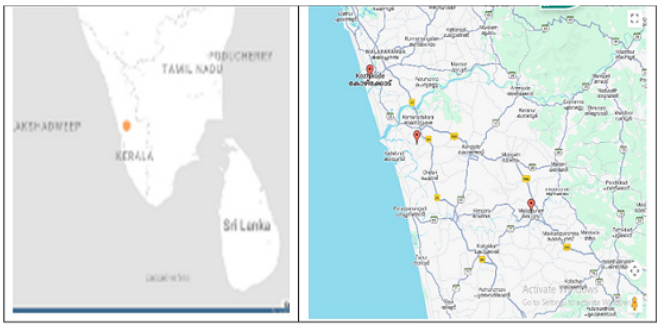

Fig. 3. Outbreak information on Health Map and the proposed framework.

HealthMap generates its visualization by processing only the news headline—"Amoebic Encephalitis Claims Another Life in Kerala, Toll at 6 in a Month"—which constrains its ability to pinpoint specific affected locations. Although this headline-based approach enables rapid aggregation and broad-scale monitoring, it suffers from limited spatial resolution, as headlines generally provide only coarse geographical references (e.g., state or national level) while omitting finer contextual details. In contrast, the proposed framework analyzes complete news narratives rather than relying solely on headlines. For example, from the detailed report "Presently, 11 people are undergoing treatment in Kozhikode for amoebic encephalitis. Shaji, from Chelambra in Malappuram district, has succumbed to amoebic encephalitis, a brain-eating infection, in Kerala," the proposed framework successfully extracts precise district and locality level information, enhancing spatial accuracy by correctly marking the three disease-affected locations on Google Maps. Processing full-text articles significantly increases the fidelity and granularity of outbreak intelligence, strengthening its utility for public health monitoring and informed decision-making.

Table II compares the proposed approach for information extraction with previous works, showing that multi-stage information extraction is superior in terms of precision, recall, and F1-score. This can be attributed to the integration of a fine-tuned DistilBERT for NER, which provides higher contextual understanding and entity detection accuracy than rule-based or dictionary-driven NER systems. The multi-stage rule-based information extraction process ensures accurate relationship identification and reduces false positives and negatives. Methods relying on regular expressions, gazetteers, or static dictionaries, such as [10, 25], are less adaptable to linguistic variations and dependencies in news reports.

TABLE II. COMPARISON WITH PREVIOUS WORKS

Study	Algorithm/Method used	Performance
Proposed method	Uses a fine-tuned DistilBERT model for NER and a multi-stage rule-based method for information extraction.	Precision: 94.86% Recall: 95.96% F1-score: 95.41%
[10]	Uses Freeling for NER. Regular expressions, gazetteers, and rule-based algorithms are used for information extraction.	Precision: 48% Recall: 46% F1-score: 47%
[25]	Disease and location names are extracted using two different dictionaries. Uses rules and regular expressions for information extraction.	Precision: 89% Recall: 94% F1 score: 92%

IV. CONCLUSION

With the advancement in Natural Language Processing (NLP) techniques and the advent of Large Language Models (LLMs), fully automated systems for disease outbreak surveillance have become feasible. The proposed rule-based information extraction framework systematically converts unstructured outbreak-related text into high-precision structured data, tailored to monitor outbreaks across different regions of Kerala. The proposed framework adopts a multi-stage processing pipeline to extract essential outbreak attributes from unstructured text corpora. This method enhances the

accuracy and reliability of outbreak surveillance, demonstrating strong applicability in real-world public health contexts.

In the dataset collected, outbreak reports are dominated by diseases such as dengue and leptospirosis, which are frequently mentioned in news articles. However, other diseases are reported only once or twice a year, leading to a significant imbalance in data availability, which makes it challenging for machine learning models to effectively predict outbreaks of less frequently reported diseases, as they lack sufficient training samples to learn meaningful patterns. Consequently, this analysis is limited to basic trends, such as the frequency of dengue or monkeypox cases reported over the years, rather than providing detailed insights into regional outbreak patterns or cross-disease interactions. However, the proposed method for converting unstructured reports into structured knowledge is broadly applicable to various public health contexts, including veterinary outbreaks. It can support integration with national surveillance systems to enable timely epidemic monitoring and geospatial visualization for future applications. If patient-level data from hospitals and health centers becomes available, more advanced analyses—such as predicting disease trends and identifying regional patterns—can be 0070erformed. In the future, the proposed method can be extended to incorporate multilingual sources, particularly Malayalam newspapers, to enhance the detection of emerging events and improve the timeliness and comprehensiveness of outbreak surveillance. Rather than replacing conventional surveillance systems, bio-surveillance systems complement them by providing early warnings, emerging disease trends, and strengthening situational awareness for timely public health responses.

REFERENCES

[1] J. O'Shea, "Digital disease detection: A systematic review of event-based internet biosurveillance systems," *International Journal of Medical Informatics*, vol. 101, pp. 15–22, May 2017, https://doi.org/10.1016/j.ijmedinf.2017.01.019.

[2] A. Kijazi, M. Kisangiri, S. Kaijage, and G. Shirima, "A Monitoring System for Transboundary Foot and Mouth Disease (FMD) considering the Demographic Characteristics in Gairo, Tanzania," *Engineering, Technology & Applied Science Research*, vol. 11, no. 4, pp. 7302–7310, Aug. 2021, https://doi.org/10.48084/etasr.4140.

[3] Sujarwoto and A. Maharani, "Participation in community-based healthcare interventions and non-communicable diseases early detection of general population in Indonesia," *SSM - Population Health*, vol. 19, Sep. 2022, Art. no. 101236, https://doi.org/10.1016/j.ssmph.2022.101236.

[4] J. Feldman, A. Thomas-Bachli, J. Forsyth, Z. H. Patel, and K. Khan, "Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1355–1359, Nov. 2019, https://doi.org/10.1093/jamia/ocz112.

[5] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports," *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, Mar. 2008, https://doi.org/10.1197/jamia.M2544.

[6] B. Jang, M. Kim, I. Kim, and J. W. Kim, "EagleEye: A Worldwide Disease-Related Topic Extraction System Using a Deep Learning Based Ranking Algorithm and Internet-Sourced Data," *Sensors*, vol. 21, no. 14, Jul. 2021, Art. no. 4665, https://doi.org/10.3390/s21144665.

[7] E. Arsevska *et al.*, "Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System," *PLOS ONE*, vol. 13, no. 8, Aug. 2018, Art. no. e0199960, https://doi.org/10.1371/journal.pone.0199960.

[8] M. Nasir, M. Bakhtyar, J. Babar, S. Lakho, B. Ahmed, and W. Noor, "BIOPAK FLASHER: Epidemic Disease Monitoring and Detection in Pakistan Using Text Mining," in *Soft Computing Applications*, vol. 1438, V. E. Balas, L. C. Jain, M. M. Balas, and D. Baleanu, Eds. Springer International Publishing, 2023, pp. 519–536.

[9] Z. Meng *et al.*, "BioCaster in 2021: automatic disease outbreaks detection from global news media," *Bioinformatics*, vol. 38, no. 18, pp. 4446–4448, Sep. 2022, https://doi.org/10.1093/bioinformatics/btac497.

[10] A. Dellanzo *et al.*, "Digital surveillance in Latin American diseases outbreaks: information extraction from a novel Spanish corpus," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, Art. no. 558, https://doi.org/10.1186/s12859-022-05094-y.

[11] B. Jang, M. Lee, and J. W. Kim, "PEACOCK: A Map-Based Multitype Infectious Disease Outbreak Information System," *IEEE Access*, vol. 7, pp. 82956–82969, 2019, https://doi.org/10.1109/ACCESS.2019.2924189.

[12] A. Rortais, J. Belyaeva, M. Gemo, E. Van Der Goot, and J. P. Linge, "MedISys: An early-warning system for the detection of (re-)emerging food- and feed-borne hazards," *Food Research International*, vol. 43, no. 5, pp. 1553–1556, Jun. 2010, https://doi.org/10.1016/j.foodres.2010.04.009.

[13] H. F. Bradford *et al.*, "Inactive disease in patients with lupus is linked to autoantibodies to type I interferons that normalize blood IFNα and B cell subsets," *Cell Reports Medicine*, vol. 4, no. 1, Jan. 2023, Art. no. 100894, https://doi.org/10.1016/j.xcrm.2022.100894.

[14] D. Carter, M. Stojanovic, and B. De Bruijn, "Revitalizing the Global Public Health Intelligence Network (GPHIN)," *Online Journal of Public Health Informatics*, vol. 10, no. 1, May 2018, https://doi.org/10.5210/ojphi.v10i1.8912.

[15] A. Jose, "Kerala in grip of epidemics, 144 deaths reported in 2024," *The New Indian Express*, Jul. 12, 2024. https://www.newindianexpress.com/states/kerala/2024/Jul/12/kerala-in-grip-of-epidemics-144-deaths-reported-in-2024.

[16] L. Chiticariu, Y. Li, and F. R. Reiss, "Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, USA, 2013, pp. 827–832, https://doi.org/10.18653/v1/D13-1079.

[17] B. Waltl, G. Bonczek, and F. Matthes, "Rule-based information extraction: Advantages, limitations, and perspectives," *Jusletter IT*, Feb. 2018.

[18] M. I. Salih, S. M. Mohammed, A. K. Ibrahim, O. M. Ahmed, and L. M. Haji, "Fine-Tuning BERT for Automated News Classification," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22953–22959, Jun. 2025, https://doi.org/10.48084/etasr.10625.

[19] T. Almeida, R. A. A. Jonker, R. Antunes, J. R. Almeida, and S. Matos, "Towards discovery: an end-to-end system for uncovering novel biomedical relations," *Database*, vol. 2024, Jul. 2024, Art. no. baae057, https://doi.org/10.1093/database/baae057.

[20] J. Jiang, "Information Extraction from Text," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Boston, MA, USA: Springer US, 2012, pp. 11–41.

[21] S. Singh, "Natural Language Processing for Information Extraction." arXiv, 2018, https://doi.org/10.48550/ARXIV.1807.02383.

[22] M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane, "Information extraction from electronic medical documents: state of the art and future research directions," *Knowledge and Information Systems*, vol. 65, no. 2, pp. 463–516, Feb. 2023, https://doi.org/10.1007/s10115-022-01779-1.

[23] M. Joy and D. M. Krishnaveni, "Enhancing Disease Outbreak Detection: Named Entity Recognition with Fine-tuned DistilBERT," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 10, pp. 4648–4660, May 2024.

[24] Zae Myung Kim, Y. S. Jeong, and Ho-Jin Choi, "Understanding news stories through SVO triplets," in *2016 International Conference on Big Data and Smart Computing (BigComp)*, Hong Kong, China, Jan. 2016, pp. 498–501, https://doi.org/10.1109/BIGCOMP.2016.7425978.

[25] M. T. Nguyen and T. T. Nguyen, "Extraction of disease events for a real-time monitoring system," in *Proceedings of the Fourth Symposium on Information and Communication Technology - SoICT '13*, Danang, Vietnam, 2013, pp. 139–147, https://doi.org/10.1145/2542050.2542084.