EpiGator: An Event-based Surveillance System for Infectious Disease Outbreaks

Yiheng Wu, Jue Hou, Lidia Pivovarova,
Sathianpong Trangcasanchai, Roman Yangarber
University of Helsinki, Finland
first.last@helsinki.fi

Abstract

We present EPIGATOR, an event-based system for global surveillance of outbreaks of infectious epidemics. The EPIGATOR pipeline integrates keyword filtering, relevance classification via fine-tuned LLMs, zero-shot information extraction for event feature normalization, event-based clustering, and multi-document summarization using instruction-tuned models. Evaluated against reports by public health specialists, EPIGATOR demonstrates 75% recall in outbreak detection. The system also provides real-time monitoring through a user-friendly interface, supporting public health surveillance and decision-making.

Keywords: Information extraction, LLMs, Public Health

1. Introduction

A primary objective of Public Health surveillance is the timely identification of infectious disease outbreaks to mitigate their impact and reduce the burden on Public Health systems. Traditional *indicator-based surveillance* (IBS) relies on official sources of information that are monitored by Public Health authorities. In contrast, *event-based surveillance* (EBS) monitors dynamic, fast-paced information streams from a wide range of sources, such as news media and social networks. This approach significantly reduces the lag between the occurrence of an event and its detection, compared to IBS (Abbood et al., 2020; Crawley et al., 2024).

EBS relies on unstructured data streams that may contain early signals of events significant for public health. The approach requires robust mechanisms capable of processing highly heterogeneous information. Given the volume and rate of delivery of open-source information, manual expert analysis is impractical for early warning detection. This highlights the need for scalable automated solutions (Bhatia et al., 2021; Hartley et al., 2013, 2010).

Information extraction from news in the context of surveillance of epidemic outrbeaks has been under development over the last two decades, with a number of extraction systems, including including MedISys (Linge et al., 2010), BioCaster (Collier et al., 2008), PULS, which is in use by European public health agencies (Barboza et al., 2013). This work, conducted in collaboration with specialists in European public health agencies, provides insights on how the model output should be organized, and, specifically, what are the relevant fields in the extracted events.

Since the epidemic surveillance domain is relatively narrow, systems based on ontologies and rules can work with high accuracy. However, the cost of *maintenance* of such systems is high. Further, providing multilingual coverage is difficult with rule-based systems, since each new language to be covered requires extensive adaptations in the rule bases and ontologies (Pivovarova et al., 2013).

In contrast, modern LLMs provide powerful multilingual abilities, as well as background knowledge, although it is fragmented and not explicit. The goal of our work is to combine experience in outbreak surveillance with state-of-the-art LLM approaches to achieve a reduction in the overall cost of system development and maintenance. We reuse as much of the previous work as possible—knowledge bases, outputs of the human-curated and rule-base systems, etc.—to guide the development and validation of the LLM-based systems.

We introduce EPIGATOR, an EBS system designed to monitor media streams from a wide range of on-line sources, detect emerging outbreak events, and generate epidemiological reports. The system operates via a pipeline of modules: (1) identify relevant articles; (2) extract specific outbreak-related information; (3) cluster articles based on events related to the same outbreak; (4) perform multi-document summarization to produce a comprehensive report for each identified cluster.

The key challenges we address in EPIGATOR include:

- **Timely and High-Volume Data**. An effective EBS system must offer both timeliness and broad data coverage (Hartley et al., 2010). The system should process a wide range of data from diverse sources in real time. The system must be able to process a massive volume of data and accurately identify important instances. On the other hand, *most* news articles are *not* relevant for outbreak surveillance, so the system must not trigger false positives.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7732–7743
11–16 May 2026.

- **Event-Based Clustering**. Prior work on EBS has typically focused on *document-level* analysis (Abbood et al., 2020; Steinberger et al., 2008). However, our goal is to *aggregate* information across multiple documents, sources and languages to yield a comprehensive view of an emerging outbreak.
- **Multi-Document Summarization**. Beyond event extraction, our goal is to generate coherent and informative summary reports that capture the main aspects of the progression of events in each given outbreak. We frame this as a Multi-Document Summarization (MDS) task.
- **Evaluation**. Automatically measuring the quality of reports generated by LLMs poses a significant challenge in its own right, since traditional lexically-based metrics, such as ROUGE (Lin, 2004) have been shown to correlate poorly with human judgments.

The paper is structured as follows. Section 2 reviews related work. Section 3 describes the collection of input data and construction of ground truth datasets for evaluation. Section 4 presents an overview of our pipeline architecture, detailing the functionality of each component. Section 5 outlines the methodology used to evaluate the effectiveness of the pipeline. Section 6 concludes the paper and outlines directions for future work.

2. Related Work

Event-based Surveillance (EBS): MedISys[1] and PULS,[2] (Linge et al., 2010; Steinberger et al., 2008), are examples of earlier media surveillance systems. As a foundation, they use the Europe Media Monitor (EMM) (Steinberger et al., 2013), which continually gathers multilingual news articles potentially relevant for public health surveillance. The information extraction module in PULS relies on rule-based pattern matching that requires an extensive manual effort for maintenance and adaptation to additional languages. We improve on this by using LLMs to go beyond document-level information extraction and synthesize coherent reports on active outbreaks.

An example of an open-source EBS system is EventEpi (Abbood et al., 2020), which follows a framework similar to MedISys, combining relevance filtering and information extraction. EventEpi's relevance filtering is based on a trained binary classifier with Word2vec embeddings serving as input features. For information extraction, EventEpi utilizes EpiTator,[3] a Python library built on the SpaCy framework, to identify epidemiological entities within text. However, EventEpi is restricted to a limited set of sources, including WHO Disease Outbreak News (WHO DONs),[4] ProMED reports,[5] and optional user-provided data. EPIGATOR exploits EMM to access articles from a wide range of global providers for broad coverage.

Information Extraction (IE): is an established technology for text understanding, actively researched over the past three decades (Piskorski and Yangarber, 2013). LLMs have demonstrated significant potential across a diverse range of NLP tasks, including IE. Conventional non-LLM models rely on task-specific architectures and costly training data, which makes IE highly resource-intensive (Atkinson et al., 2011; Huttunen et al., 2002). LLMs enable zero-shot IE where generalist LLMs are instructed to perform IE without task-specific training (Agrawal et al., 2022). Prior research has shown that zero-shot IE using LLMs can yield promising results in epidemic IE (Consoli et al., 2024), in some cases surpassing the performance of models trained from scratch with full supervision (Wei et al., 2023).

The need for multi-document or cross-document extraction is recognized in real-world applications of IE. The chronological sequence of news reports offers additional insights into the progression and evolution of the outbreak, which cannot be captured when considering the documents in isolation. Therefore, a key challenge is to accurately cluster documents from diverse sources that refer to the same sequence of events (Escoter et al., 2017; Yangarber, 2006).

Multi-document summarization (MDS): presents complex challenges compared to single-document summarization. While modern LLMs have been shown to produce summaries whose quality often surpasses that of human-written summaries for single documents (Pu et al., 2023), MDS remains relatively underexplored. Notably, hallucination—the fabrication of inaccurate information—is a persistent problem in the LLM setting (Belem et al., 2025). MDS addresses multiple problems with single-document summarization, by generating a coherent summary from *multiple* related documents by identifying salient content, resolving redundancies, and maintaining consistency across sources (Ma et al., 2022).

Early work in MDS relied on encoder-decoder architectures to process concatenated source documents to generate an abstractive summary. For example, PEGASUS (Zhang et al., 2020) introduced a novel pre-training task called "gap-sentences generation," where the model learns to generate masked sentences. PRIMERA (Xiao et al., 2021)

[1] medisys.newsbrief.eu
[2] puls.cs.helsinki.fi
[3] github.com/ecohealthalliance/EpiTator
[4] who.int/emergencies/disease-outbreak-news
[5] promedmail.org

was explicitly designed for multi-document inputs and used a "Pyramid-based Masked Sentence" pre-training strategy to encourage cross-document information aggregation. Later approaches infuse information extraction as part of summarization, marginally improving the overall performance (Zhang et al., 2023). While these approaches are effective, they require fine-tuning for specific tasks and data sets.

Instruction-tuned LLMs, like GPT-4, are trained to follow natural-language instructions. They can perform MDS in a zero-shot or few-shot fashion. The user can control the summary's length, style, and focus with prompt engineering, making them flexible. This shift from task-specific fine-tuning to instruction-based prompting has made LLMs a versatile and scalable approach to MDS.

Turining to the challenge of **evaluation** of information extraction in the multi-document context, prior work has shown that lexically-based metrics, such as ROUGE (Lin, 2004) correlate poorly with human judgment (Fabbri et al., 2021). Recently introduced metrics based on semantic similarity, such as BERTScore (Zhang et al., 2019), have been found to be insufficiently sensitive for tasks where factual consistency is a critical requirement (Maynez et al., 2020), such as ours. Previous research has also shown that both ROUGE and BERTScore are unreliable for evaluating LLM-generated summaries in news summarization, as these metrics tend to assign low scores to summaries actually preferred by human evaluators (Goyal et al., 2022). In our task, the factual consistency between the output summary and the input documents is of paramount importance. An additional difficulty is the lack of gold-standard annotations. We therefore turn to using public-health reports created by human experts as a reference, and perform manual validation.

3. Data

Our input data include articles from news agencies and specialized public-health websites worldwide, published in multiple languages. On average, these sources provide approximately 10,000 articles per day. The articles are collected using a battery of keyword queries tilted toward high coverage—hence the vast majority of the documents carry no information on infectious diseases, while important, newsworthy outbreaks are typically reported in multiple sources.

In order to *approximate* ground truth as part of our verification and system evaluation strategy, we use the Communicable Disease Threats Reports (CDTR)[6] published weekly by the European Centre for Disease Prevention and Control (ECDC). CDTR is a weekly compendium of approximately 10 topics requiring most urgent attention in the European public health context. Each CDTR contains several *sub-reports*; each sub-report is an expert-curated description of an epidemic outbreak of a communicable disease of concern for EU and global surveillance. An example of the report is shown in Appendix 3.

We use these reports to validate our findings and support the development of our surveillance process. CDTRs do not make explicit which data sources the ECDC specialists used to produce each report: they use news sources, as well as dedicated IBS sources. In addition, events in CDTR do not always concern outbreaks happening at the time of the report: some may refer to events that occurred weeks or months prior to the report. Some updates on certain diseases may be *negative*—in the form "no new cases were observed this week." Thus, although ECDC reports are an invaluable source for understanding which events are of importance for public health and medical authorities, using them *directly* as ground truth for news surveillance is a challenging problem in itself.

[6] www.ecdc.europa.eu/en/publications-and-data

4. EpiGator pipeline

Figure 1 shows our 4-stage pipeline: data filtering, relevance filtering, event extraction, and summarization. The Figure also shows 3 check procedures: relevance check, semantic check and fact check. The checks are not part of the production pipeline, they are additional semi-manual operations that we perform for evaluation. Although CDTRs (ECDC reports) are not used in the main pipeline, they are used for building a training set and for evaluation, as described in the following sections.

4.1. Stage I: Data Filtering and De-duplication

The first stage focuses on analyzing the stream of input documents to retain unique and potentially relevant articles. We detect exact- and near-duplicates and remove them using a hash-based algorithm. Then a keyword-based filter based on a large curated ontology of communicable diseases removes articles that mention no disease terms, narrowing the scope to disease-related content. This ontology includes a disease names (currently, 5227) and their synonyms or alternative forms (2415), which will be used for normalization in event-based clustering. This step prunes articles that are unlikely to discuss disease-related events, significantly reducing the computational load for

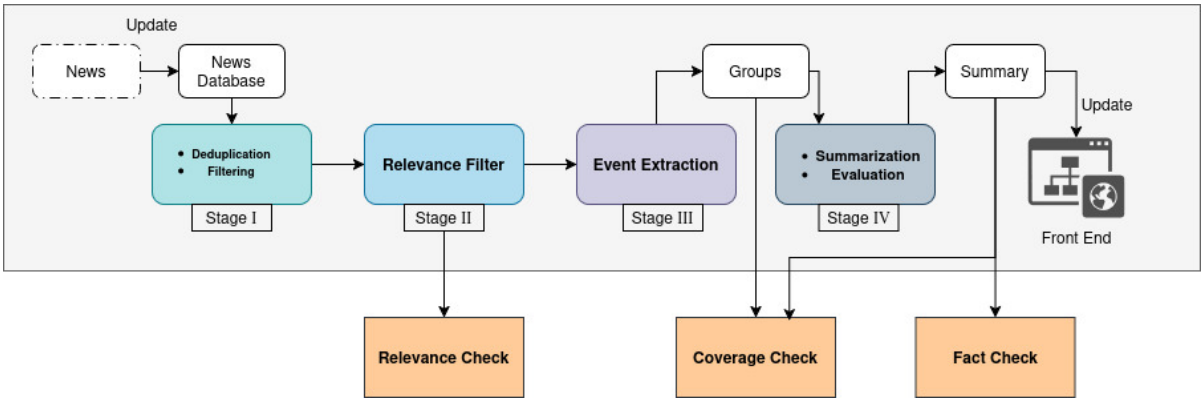


Figure 1: Stages of the processing pipeline

the subsequent, more computation-intensive processing stages.

In this stage, the input for each iteration consists of news articles from the latest week. In addition, we maintain a one-month window of news data for de-duplication purposes, to prevent the repeated use of old news.

4.2. Stage II: Relevance Filter

Keyword-based filtering cannot differentiate between various contexts in which a disease might be mentioned, e.g., an active outbreak vs. a description of a historical event. To address this issue, we formulate relevance filtering as a binary classification task, similar to EventEpi (Abbood et al., 2020). This classifier, which is based on a fine-tuned LLM, predicts whether an article is relevant to an *active* outbreak. This allows the system to distinguish between articles discussing an ongoing outbreak from those focusing on historical information, general health recommendations, research papers, etc.

Data: To train the relevance classifier, we construct a *training set* consisting of 2,000 instances that are equally balanced between positive and negative samples.

We start with a collection of all articles that mention any disease name. To find high-quality *positive* samples, we compare them against the weekly CDTR from ECDC. We extract the textual content from the CDTR (which is usually provided in PDF format), and convert it into structured data, comprising a list of outbreak events for each report. Each outbreak event is introduced by a header—e.g., *"Yellow fever—South America—2024-2025"*—and a summary text that provides details about the reported event. From each event, we extract the disease and location names to form an *outbreak key*, which serves as a reference point for linking articles from news streams to the CDTRs.

Articles whose extracted disease and location match a CDTR key within a $\pm$1-week temporal window are considered to be externally validated, high-quality positive instances.

To find high-quality negative instances, we instruct `Qwen/QwQ-32B` (Team, 2025) to detect negative instances, denoted as `Qwen-neg`. We tune a prompt on a held-out development data set. Evaluated on a manually annotated set of 300 documents, `Qwen` achieves 98% precision in the negative class, allowing us to find negative instances confidently.

Model: We fine-tune Llama-3.1-8B-Instruct[7] (Grattafiori et al., 2024) using Low-Rank Adaptation (LoRA) (Hu et al., 2022) in a supervised learning setting—to train the relevance classifier on the training set, by instructing the model to decide whether each input article is relevant. The resulting fine-tuned model is denoted as `Llama-ft`. The prompt is shown below:

```
You are a media monitor
for a health organization.
Your task is to detect and
remove irrelevant documents
that are not related to
current emergent disease
outbreaks.  A document is
considered irrelevant if
its content fits one of
the following criteria:
(1) vaccine development or
general medical articles,
(2) political or economic
topics, (3) historical events
of past disease outbreaks,
or (4) content that contains
little useful information for
further processing.
```

[7] huggingface.co/meta-llama/Llama-3.1-8B

Disease	Article	Start Date	Report Date	Location	Deaths	Cases	Status
cholera	1	September 2024	Jan 5, 2025	Malawi	13	over 200	Ongoing
	2	October 2024	Jan 4, 2025	Ghana	37	unknown	Ongoing
	3	October 2024	Jan 4, 2025	South Sudan	over 100	over 6,000	Ongoing
monkeypox	1	unknown	Jan 6, 2025	Brittany, France	1	unknown	Ongoing
	2	October 5, 2024	Jan 6, 2025	Brittany, France	1	unknown	Ongoing

Table 1: Clustered articles by disease and their extracted outbreak attributes.

model	Acc	Prec	Rec	F1
`Llama-ft`	**90.00**	**94.59**	86.42	**90.32**
`Qwen-zs`	86.67	80.80	**98.76**	88.88
`Llama-zs`	86.33	82.01	95.67	88.31

Table 2: Relevance Check: accuracy, precision, recall, and F1 scores on the positive class of the test set.

The fine-tuning was done with HuggingFace Transformers and PEFT (LoRA (Hu et al., 2022)) libraries. The configuration of hyper-parameters is as follows: LoRa rank 8; LoRA applied to all layers; learning rate *1e-4*. The number of epochs is 5. The optimizer is AdamW. The warm-up ratio is 0.1. The data precision is `bf16`. The maximum input length is set to 2048 tokens.

Relevance Check: We manually annotated a *test set* of 300 articles to evaluate the relevance classifier. The test data contain 162 positive and 138 negative instances. We report accuracy, precision, recall, and F1-score of the positive class in Table 2. Our fine-tuned model is `Llama-ft`; `Qwen-zs` and `Llama-zs` are `Qwen/QwQ-32B` and `Llama-3.1-8B-Instruct` performing zero-shot relevance classification, respectively. As can be seen from the Table, zero-shot models have high recall but substantially lower precision, which means that in a real-world setting—where most documents are irrelevant—they would trigger many false positives. Fine-tuning allows us to overcome this problem, and the fine-tuned model yields the highest accuracy and F1-score.

4.3. Stage III: Event-based clustering

In the next stage news articles reporting the same outbreak event are clustered together. The variability in reporting styles and terminology across news sources presents a challenge for this stage. Following previous work on zero-shot IE (Agrawal et al., 2022; Wei et al., 2023), we first extract event attributes from each relevant article, and then create clusters based on the extracted information. We use `Qwen/QwQ-32B` to extract disease and location terms from relevant articles. Clustering is done in real-time, which enables timely grouping of documents published within close temporal proximity and reporting similar diseases and locations.

Attribute	Example Value
Disease	Cholera
Location(s)	`loc1`, `loc2`, ...
Summaries	{ `loc1`: Brief summary 1, `loc2`: Brief summary 2, ... }

Table 3: JSON-style structured summary with disease name, multiple locations, and corresponding location-specific summaries.

The extracted disease and location are normalized using our predefined disease and location ontologies. The ontology maps various terms—e.g., *avian influenza*, *bird flu*, *H5N1*, etc.—to a single canonical name, which is needed for accurate clustering.

Articles are clustered together if they share the same disease entity. The extracted fields for each article are shown in Table 1. We also extract sentences with supporting evidence corresponding to the event attributes.

4.4. Stage IV: Multi-document summarization

In the final stage, we generate a concise outbreak summary for each cluster. To that end, we use *evidence* text from the group articles as an input for Qwen32b and promt it to output a JSON-like summary with 3 attributes as shown in Table 3.

It is possible to generate different summaries by changing the granularity of location—e.g., city, country, continent. Here, for the convenience of evaluation, we use country as the unit of location.

5. Evaluation

To evaluate the end-to-end pipeline we perform three check procedures, as shown in the orange blocks in Figure 1. The relevance check was presented in Section 4.2. To avoid the influence

of prior knowledge present inside the LLMs, we use only recent news articles and ensure that the model's release date precedes the news content.

For each week, if a CDTR sub-report published in that week contains the same disease and location as a summary generated within this or preceding two weeks, we consider the sub-report to be successfully covered by the summary.

To ensure factual accuracy of our generated summaries, we perform manual validation. This involves verifying that the content of the sub-reports does not contradict the generated summaries, particularly on critical details such as numbers of cases and deaths, and the dates.

5.1. Coverage Check: Semantic Evaluation

We evaluate the quality of generated summaries by measuring how well they help retrieve the original ground-truth documents. We use three complementary coverage recall metrics at different granularities.

Setup. We use EpiGator results from the last week of 2015 (December 25 to December 31) as our test set. This set contains multiple disease groups, e.g., H5N1, MERS. Each group has a set of source documents—the original news articles used for generating summaries—and multiple generated summaries divided by different locations. We use each summary as a query to retrieve documents from our database, then check how many source documents are retrieved.

Metric 1: Summary-Level Coverage. *How well each individual summary covers the source documents?*

For each summary, we calculate what percentage of its group's source documents it retrieves. Then we average this percentage across all summaries in the test set. This metric tells us whether individual summaries are effective at retrieving relevant documents.

Metric 2: Group-Level Combined Coverage. *For each disease group, what fraction of source documents are covered, when all summaries for this disease are combined?*

For each disease group the metric combines as follows:

Step 1 – Union of all retrievals: Take the union of all documents retrieved by all summaries in this group. Each document is counted only once even if multiple summaries retrieved it.

Step 2 – Calculate coverage: Check what percentage of the group's source documents appear in this union.

Step 3: Finally, the metric is averaged across all disease groups.

For example, if the H5N1 group has 3 summaries that retrieve:

- Summary 1: {doc A, doc B, doc C}
- Summary 2: {doc B, doc D}
- Summary 3: {doc C, doc E}

The union is {A, B, C, D, E} (5 unique documents), and we check how many of H5N1's source documents are in this set.

Metric 3: Group-Level High-Confidence Coverage. *For each disease group, after filtering low-confidence retrievals and removing duplicates, what fraction of source documents are covered?*

This metric evaluates quality at the group level with three steps:

Step 1 – Filter low-confidence retrievals: For each summary, we only keep documents with retrieval scores above 0.4, discarding uncertain matches.

Step 2 – Merge within each group: For each disease group, we combine all high-confidence retrievals from all its summaries and remove duplicates. This gives us the unique set of documents retrieved by the group.

Step 3 – Calculate group coverage: For each group, we calculate what percentage of source documents appear in this merged set. Finally, we average across all groups.

Results and Discussion. As shown in Table 4, our summaries achieve satisfactory coverage across all three metrics. The progression from 64.19% (individual summaries) to 68.97% (pooled results) to 71.21% (filtered and deduplicated groups) demonstrates that: (1) individual summaries are reasonably effective, (2) combining multiple summaries improves coverage, and (3) focusing on high-confidence retrievals yields the best results. These findings confirm that our generated summaries effectively capture the key information from the original documents.

Model	Summary Coverage	Overall Coverage	Group Coverage
EPIGATOR	64.19%	68.97%	71.21%

Table 4: Semantic coverage metrics evaluating how well generated summaries retrieve ground-truth documents.

5.2. Fact Checking

To evaluate the early detection capability of EpiGator, we employ authoritative reports from the European Centre for Disease Prevention and Control

ECDC Sub-report	EPIGATOR Summary
On **6 January 2025**, the US CDC and the Louisiana Department of Health reported that the patient that was hospitalised with severe ***avian influenza H5N1*** in the state has passed away. This is the first **death** from ***H5N1*** reported by the **United States**. The patient was over 65 years-old and according to the reports had underlying conditions. The patient had been exposed to non-commercial backyard flock and wild birds.	The first human **death** due to ***H5N1 bird flu*** in the **United States** was reported in Louisiana on **January 6, 2025**. The individual, older than 65 with underlying medical conditions, was infected after exposure to a backyard flock and wild birds. The **death** marks the first case of its kind in the **U.S.**, with no evidence of human-to-human transmission and no additional cases identified in Louisiana as of **January 2025**.
Since the previous **update** on 10 December 2025, and as of 5 January 2026, five new ***MERS*** cases (including one fatality) have been reported in **Saudi Arabia** with date of onset between September and December 2025. The patients reside in Makkah (2), Riyad (2) and **Najran** (1) regions in Saudi Arabia. All patients are adults,	In **Saudi Arabia**, 17 cases of ***MERS*** have been reported as of December 21, 2025, occurring across various regions including Riyadh, Taif, **Najran**, Hail, and Hafr Al-Batin. The outbreak in Saudi Arabia is **ongoing**, with no specific start date mentioned. This represents a significant portion of the global cases reported during the period.

Table 5: Comparison between ECDC sub-reports and the corresponding generated summaries, illustrating matched coverage.

(ECDC) as ground truth references. Each ECDC report documents outbreak events with the following information:

- Disease name
- Geographic location of outbreak
- Time period of outbreak (which may span extended durations, e.g., 2010–2025)

EpiGator generates a series of summary reports as its final output, which are continuously updated over time. Each summary report follows a structured format where disease information is organized by location as shown in Table 3

Historical versions of these summary reports are preserved to enable temporal analysis and validation. For each disease outbreak record in an ECDC report, we perform a temporal retrospective search through EpiGator's historical summary reports to assess early detection performance.

Two examples of EPIGATOR summary and the corresponding ECDC report are shown in Table 5.

We checked 68 ECDC sub-reports in total. Among them, 51 were successfully matched by EPIGATOR summaries, while 15 did not match, which can be interpreted as 75% recall.

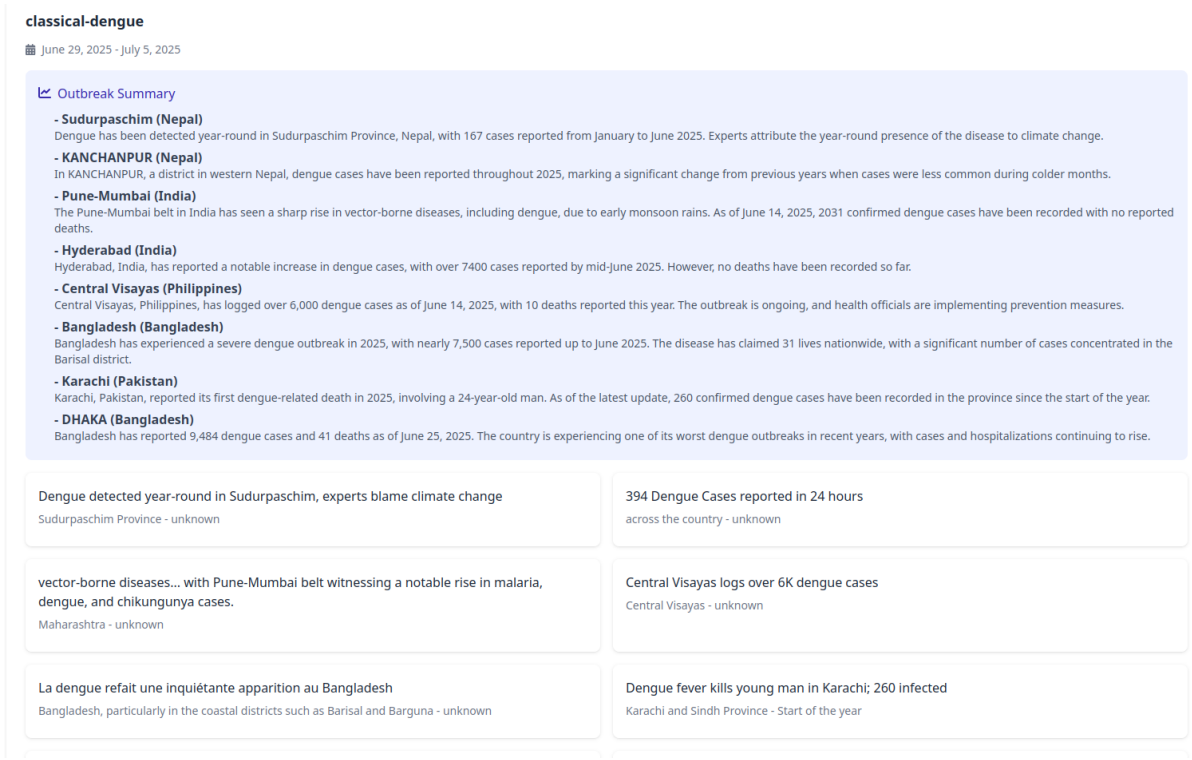
As explained above, ECDC reports are not suitable as such ground truth in our setting, since the reports do not explicitly reference the resources on which they are built, and due to certain mismatches between the ECDC reports and the wider news stream. In particular, ECDC reports have a bias toward events that affect the EU area in particular, while the news is collected from global sources.

6. Conclusion

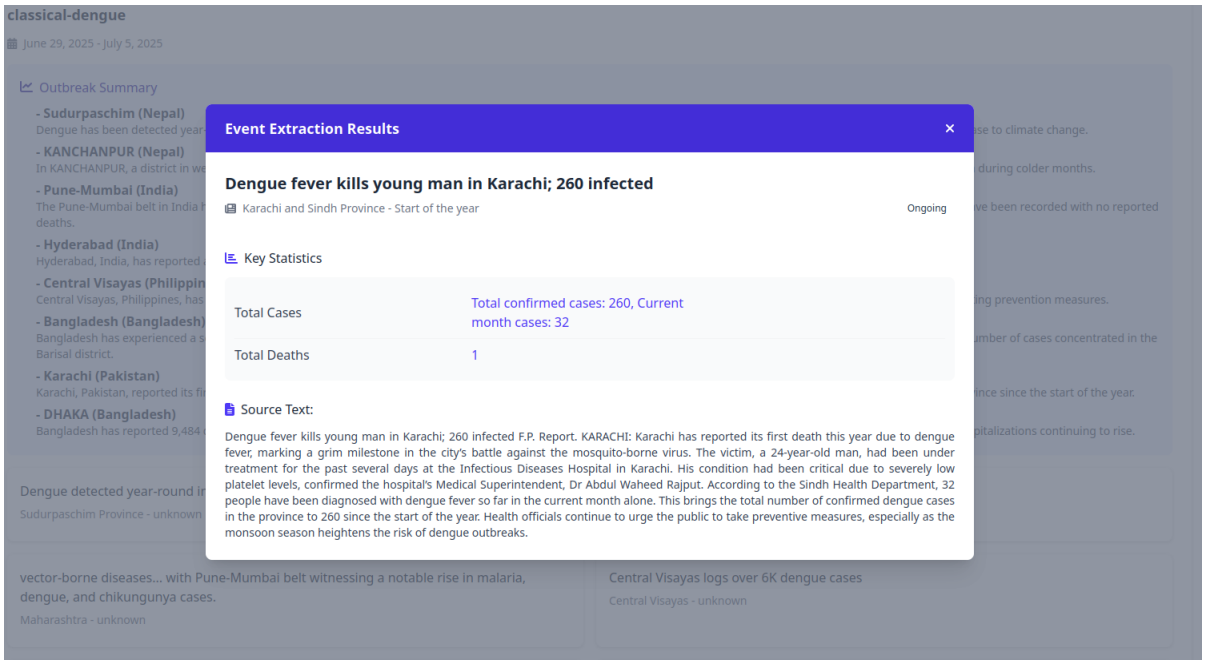
In this work, we present EPIGATOR, an end-to-end epidemic surveillance system that integrates multiple stages of data processing, event extraction, and summarization. EPIGATOR automates the identification and extraction of outbreak-related information from diverse data sources and provides a user-friendly visualization interface, enabling real-time monitoring and intuitive analysis of epidemic events. Through a comprehensive evaluation against authoritative datasets, we demonstrate the effectiveness and reliability of our approach. This framework offers valuable support for public health surveillance and decision-making.

While EPIGATOR shows good performance on detecting and summarizing outbreak events, many challenges remain open, and need to be addressed in future work. A major issue is the mismatch in the timing and the scope between the news articles and official reports, which complicates our formal evaluation. Some outbreaks that receive a great deal of attention in the news are not covered in ECDC reports at all (potentially due to a lack of relevance in the EU context), while some CDTRs may refer to past, historical events, or contain no new information.

In the future, we aim to improve the evaluation by using question-answering-based (QA) or IE-based methods to check the accuracy of the generated summaries (Min et al., 2023; Honovich et al., 2022; Fabbri et al., 2022). We also plan to extend the system by adding summaries at different levels of geographic granularity, such as city or continent, depending on the user requirements.



(a) Description of epidemic of classical dengue.



(b) Detailed view for each news article

Figure 2: User Interface of EPIGATOR

Acknowledgements

This work was supported in part by the Research Council of Finland, Project *"Know-AI"* (Grant 1359285), and by European Regional Development Fund (ERDF) Project *"Generative AI and Knowledge Management"* (Grant 4740347).

References

Auss Abbood, Alexander Ullrich, Rüdiger Busche, and Stéphane Ghozzi. 2020. EventEpi—A natural language processing framework for event-

based surveillance. *PLoS computational biology*, 16(11):e1008277.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Martin Atkinson, Jakub Piskorski, Erik van der Goot, and Roman Yangarber. 2011. Multilingual real-time event extraction for border security intelligence gathering. In U. Kock Wiil, editor, *Counterterrorism and Open Source Intelligence*, pages 355–390. Springer Lecture Notes in Social Networks, Vol. 2.

Philippe Barboza, Laetitia Vaillant, Abla Mawudeku, Noele P Nelson, David M Hartley, Lawrence C Madoff, Jens P Linge, Nigel Collier, John S Brownstein, and Roman Yangarber. 2013. Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events. *PLoS One*, 8(3):e57252.

Catarina G. Belem, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2025. From single to multi: How LLMs hallucinate in multi-document summarization.

Sangeeta Bhatia, Britta Lassmann, Emily Cohn, Angel N Desai, Malwina Carrion, Moritz UG Kraemer, Mark Herringer, John Brownstein, Larry Madoff, and Anne Cori. 2021. Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread. *NPJ digital medicine*, 4(1):73.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Sergio Consoli, Peter Markov, Nikolaos I Stilianakis, Lorenzo Bertolini, Antonio Puertas Gallardo, and Mario Ceresa. 2024. Epidemic information extraction for event-based surveillance using large language models. In *International Congress on Information and Communication Technology*, pages 241–252. Springer Nature Singapore.

Adam W Crawley, Kyeng Mercy, Sabrina Shivji, Hannah Lofgren, Daniella Trowbridge, Christine Manthey, Yenew Kebede Tebeje, Alexey Wil Clara, Kimberly Landry, and Stephanie J Salyer. 2024. An indicator framework for the monitoring and evaluation of event-based surveillance systems. *The Lancet Global Health*, 12(4):e707–e711.

Llorenç Escoter, Lidia Pivovarova, Mian Du, Anisia Katinskaia, and Roman Yangarber. 2017. Grouping business news stories based on salience of named entities. In *15th Conference of the European Chapter of the Association for Computational Linguistics Proceedings of Conference, Volume 1: Long Papers*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David Hartley, Noele Nelson, Ronald Walters, Ray Arthur, Roman Yangarber, Larry Madoff, Jens Linge, Abla Mawudeku, Nigel Collier, and John Brownstein. 2010. The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 3(1):7096.

David M Hartley, Noele P Nelson, RR Arthur, Philippe Barboza, Nigel Collier, Nigel Lightfoot, JP Linge, Erik van der Goot, Abla Mawudeku, and LC Madoff. 2013. An overview of internet biosurveillance. *Clinical Microbiology and Infection*, 19(11):1006–1013.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Silja Huttunen, Roman Yangarber, and Ralph Grishman. 2002. Diversity of scenarios in information extraction. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jens P Linge, Ralf Steinberger, Flavio Fuart, Stefano Bucci, Jenya Belyaeva, Monica Gemo, Delilah Al-Khudhairy, Roman Yangarber, and Erik van der Goot. 2010. MedISys: medical information system. In *Advanced ICTs for disaster management and threat detection: Collaborative and distributed frameworks*, pages 131–142. IGI Global.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore.

Jakub Piskorski and Roman Yangarber. 2013. Information extraction: past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 23–49. Springer.

Lidia Pivovarova, Mian Du, and Roman Yangarber. 2013. Adapting the PULS event extraction framework to analyze Russian text. In *Proceedings of the 4th International Workshop on Balto-Slavic Natural Language Processing*, pages 100–109, Sofia, Bulgaria. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.

Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter, and Roman Yangarber. 2008. Text mining from the web for medical intelligence. In *Mining massive data sets for security*, pages 295–310. IOS Press.

Ralf Steinberger, Bruno Pouliquen, and Erik Van der Goot. 2013. An introduction to the Europe Media Monitor family of applications. *arXiv preprint arXiv:1309.5290*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, and Meishan Zhang. 2023. Zero-shot information extraction via chatting with ChatGPT. *arXiv e-prints*, pages arXiv–2302.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

Roman Yangarber. 2006. Verification of facts across document boundaries. In *Proceedings of the International Workshop on Intelligent Information Access (IIIA-2006)*, Helsinki, Finland.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Zixuan Zhang, Heba Elfardy, Markus Dreyer, Kevin Small, Heng Ji, and Mohit Bansal. 2023. Enhancing multi-document summarization with cross-document graph-based information extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1696–1707.

A. Evaluation: comparison with human-curated reports

This week's topics

1. Overview of respiratory virus epidemiology in the EU/EEA
2. Marburg virus disease (MVD) - Ethiopia - 2025/26
3. Cholera – Multi-country (World) – Monitoring global outbreaks – Monthly update
4. Middle East respiratory syndrome coronavirus (MERS-CoV) – Multi-country – Monthly update
5. Avian influenza A(H9N2) – Multi-country (World) – Monitoring human cases
6. Human cases of swine influenza A(H1N1) virus variant - Multi-country - 2024

Executive summary

Overview of respiratory virus epidemiology in the EU/EEA, week 01, 2026

The number of patients presenting to primary care with symptoms of respiratory illness is elevated in most reporting countries. This indicates that there is currently significant respiratory virus circulation in the European Union/European Economic Area (EU/EEA).

Influenza virus circulation continues to increase in most countries in the EU/EEA, although for some countries, the peak appears to have passed. Influenza A(H3N2) remains the dominant subtype at the EU/EEA level, but A(H1N1)pdm09 is also being detected. All age groups are affected. Increases in hospitalisation are observed in many countries, primarily in adults aged 65 years and above.

Early estimates of seasonal influenza vaccine effectiveness in the EU/EEA for the season 2025-2026 were published by ECDC on 19 December 2025, and match those published for A(H3N2) viruses by the United-Kingdom and Canada.

Respiratory syncytial virus (RSV) circulation continues to slowly increase in several countries. Hospital data show rising RSV-related admissions in a few countries, primarily among children under five years of age. At EU/EEA level, RSV-related hospitalisations are at levels below those observed at this time in the past four seasons.

SARS-CoV-2 circulation continues to decrease in all age groups, and the impact on hospitalisations is currently limited.

All data are provisional and may be affected by reporting delays, incomplete country data, or low testing volumes. A few countries with high testing rates can disproportionately influence pooled data. Further information is available under 'Country notes' and 'Additional resources'.

Marburg virus disease (MVD) - Ethiopia - 2025/26

- Since the CDTR update on 19 December 2025, there has been no additional confirmed cases and no additional deaths of Marburg Virus Disease (MVD) reported in Ethiopia.
- Since the start of the outbreak, and as of 5 January 2026, 17 cases (14 confirmed and three probable) of MVD have been reported, including 12 deaths (nine confirmed and three probable (case fatality rate (CFR): 64.3%)).
- Two areas have been affected across two regions; Jinka town, South Ethiopia Regional State and Hawassa City, Sidama Region.
- The total number of contacts who were monitored is 886, according to a press release from the Ethiopian Ministry of Health on 5 January 2026.
- As of 5 January 2026, there has been no new cases of MVD and no contacts have been monitored for 21 days. The outbreak will be declared over 42 days after the last Marburg patient tests negative and is discharged.
- This is the first MVD outbreak ever reported in Ethiopia.
- The likelihood of exposure to MVD for EU/EEA citizens visiting or living in Ethiopia is assessed as low, with uncertainties connected to the limited epidemiological information available. The impact at population level is assessed to be low as it is not expected that there will be significant numbers of MVD cases among EU/EEA citizens in Ethiopia. Therefore, the overall risk for EU/EEA citizens visiting or living in Ethiopia is low.
- In the event of MVD cases being imported into the EU/EEA, we consider the likelihood of further transmission to be very low, and the associated impact low. Therefore, the overall risk for the EU/EEA is assessed as low.

Figure 3: Example of CDTR: weekly report from ECDC

B. Error Analysis

This section demonstrates representative error cases encountered in our system, including:

1. Hallucinations, where the model generates information not present in the source text.
2. Errors in event information extraction, leading to incomplete or incorrect event representations.

We show 2 error cases in Table 6. These error patterns highlight important limitations that warrant further attention. Future work should focus on developing more robust mechanisms to mitigate these issues and enhance system performance.

Table 6: Examples of Error Cases in the System

Case	Original Text (Excerpt)	Generated Summary (Excerpt)	Error Type
1	*"At least 87 people have died in Argentina after being administered bacteria-infected fentanyl in hospitals..."* Location: Buenos Aires, Argentina Event: Deaths from contaminated fentanyl	**Summary identifies fentanyl as a disease outbreak** "Outbreak Summary: Buenos Aires - As of August 14, 2025, at least 87 deaths have been reported...following the administration of fentanyl contaminated with Klebsiella pneumoniae..."	**Hallucination** / **Misclassification** Fentanyl is a pharmaceutical drug, not disease. System incorrectly categorized this as disease outbreak.
2	*"The study also found that the overall disease severity was relatively low in South Kivu. They saw two deaths out of 403 hospitalized patients..."* Context: Research study examining 13 years of historical data (2010-2023)	**Summary presents historical data as recent outbreak** "Democratic Republic of Congo (DRC): ...it reports two deaths out of 403 hospitalized patients, resulting in a fatality rate of less than one percent. The outbreak is ongoing and rapidly expanding..."	**Event Information Extraction Error** Death statistics are from historical research study, not recent outbreak. System incorrectly presents past data as current event.