International Journal on Digital Libraries
https://doi.org/10.1007/s00799-021-00308-9

Unified approach to retrospective event detection for event- based epidemic intelligence

Marco Fisichella[1]

Received: 9 April 2021 / Revised: 1 September 2021 / Accepted: 4 September 2021

Abstract
Inferring the magnitude and occurrence of real-world events from natural language text is a crucial task in various domains. Particularly in the domain of public health, the state-of-the-art document and token centric event detection approaches have not kept the pace with the growing need for more robust event detection in public health. In this paper, we propose UPHED, a unified approach, which combines both the document and token centric event detection techniques in an unsupervised manner such that events which are: rare (aperiodic); reoccurring (periodic) can be detected using a generative model for the domain of public health. We evaluate the efficiency of our approach as well as its effectiveness for two real-world case studies with respect to the quality of document clusters. Our results show that we are able to achieve a precision of 60% and a recall of 71% analyzed using manually annotated real-world data. Finally, we also make a comparative analysis of our work with the well-established rule-based system of MedISys and find that UPHED can be used in a cooperative way with MedISys to not only detect similar anomalies, but can also deliver more information about the specific outbreak of reported diseases.

Keywords Retrospective public health event detection · Clustering · Event-based epidemic intelligence

1 Introduction

A public health event is defined as a specific infectious disease that is affecting a population at a specific time and place. An important strategy, used by public health officials to mitigate the impact of potential threats, is to find ways to detect the signs of a public health event as early as possible. Unstructured and informal Web documents are used as data sources to detect facts about current infectious disease activity within a population [19]. The body of work devoted to this effort is known as event-based Epidemic Intelligence (e-EI) [35].

Existing event-based e-EI systems rely upon the enumeration of possible types of medical reporting patterns, or rules (e.g., MediSys [47]). This presents a huge limitation, since given the variety of natural language, many rules may be required, and the recall for identifying relevant events can be low. Systems that only rely upon rule-based event detection are limited, since the only threat indicators (e.g., keywords) they can detect are those that are explicitly under surveillance. One way to overcome the aforementioned limitations, is to cast a new light on the task of public health event detection—so that it is done in an unsupervised manner.

1.1 Limitations of existing systems

There are two major approaches to unsupervised event detection and they can be distinguished by the type of observation variable used to model an event.

In document-centric detection [30,46], the observation variable is the coupling (co-occurrence) between documents and words. Generative models are often used to capture this coupling by introducing an unobserved (hidden) variable to infer an event. In token-centric approaches an event is modeled by the temporal correlation of bursty tokens (features) within a document collection [18,20].

The drawback of the document-centric approaches is that temporal aspects are not explicitly incorporated into the event model, and no prior burst analysis is done on these representations. In contrast, approaches based on the correlation of bursty features can filter out a vast number of potentially irrelevant features; yet a model of the temporal co-occurrence of words within a document is lost. Moreover, many types of features that are relevant for public health event detec-

✉ Marco Fisichella
mfisichella@L3S.de

[1] L3S Research Center of Leibniz University of Hannover, Hannover, Germany

Published online: 09 October 2021

tion (i.e., symptoms, victims, or medical conditions) are not modeled.

1.2 Proposed solution

In our approach, UPHED—a Unified Approach to Public Health Event Detection, we seek to overcome the limitations of using each type of event detection approach individually; and propose a unified approach to public health event detection for e-EI. Further, we propose that by applying an unsupervised algorithm to public health event detection, we help to overcome the limitations of existing e-EI systems.

To justify our hypothesis, we adapt an unsupervised approach to our problem domain so that events which are: rare (aperiodic); reoccurring (periodic); and domain or task-specific, can be detected. In more detail, we combine burst function analysis with the entity-centric feature representation in a generative model for probabilistic event detection. Going beyond a random initialization of the probabilities in this generative process, we instead exploit a known distribution of the features that are obtained directly from the burst function. Additionally, in our burst analysis, we refine the approach to feature representation by incorporating a Cauchy–Lorentz distribution to more closely model the true behavior of periodic, non-burst (trough) activity.

1.3 Our contributions

The contributions of this work are:

- Use of an approach to unsupervised event detection and adaption of its feature set to the domain of public health event.
- Presentation of a general model which, in contrast to previous approaches, incorporates two main techniques: the burst function spectral analysis and the entity-centric feature representation of documents in a generative model. Compared with existing solutions, our UPHED approach results in a more efficient and accurate method to predict public health events.
- Refining the model for representing periodic, non-burst features with the Cauchy–Lorentz distribution. The better sampling achieved by such a distribution, is shown to be more efficient with respect to the previous representations that use Gaussian distributions [20].
- Exploration of the cooperative nature of the proposed approach UPHED with the well-established rule-based system MedISys. Through a comparative analysis, the suggested strategy not only detects similar anomalies, but can also deliver more information about the specific outbreak of reported diseases.

The remainder of this journal is organized as follows: we discuss related work in Sect. 2. In Sect. 3, we present details of our approach. Then, in Sect. 4 we present the experimental results. Finally, we provide our conclusions in Sect. 5.

2 Related work

2.1 Event detection in public health

Three main approaches exist for the detection in public health. They are rule-based, supervised, and hybrid.

2.1.1 Rule-based systems

A common approach to detect public health events in e-EI, is using rule-based approaches [31,39] where regular expressions are used to detect events from unstructured text. One of the drawbacks of rule-based approaches is in building (and maintaining) the pattern base. As reported in the survey [48], the early method of event extraction was mainly based on rule-based methods, and later developed into a method based on pattern matching. These methods are essentially the same, that is, they need to build rules or templates. The event extraction method based on pattern matching refers to a method of matching the event sentence to be extracted with the corresponding template. The method based on pattern matching is better applied in a specific field, but this method has poor portability and flexibility. As an example of such system, the French Animal Health Epidemic Intelligence System has been monitoring signals of the emergence of new and exotic animal infectious diseases worldwide. The core component is a combined information extraction method founded on rule-based systems and data mining techniques. The information extraction approach allows extraction of key information on diseases, locations, dates, hosts and the number of cases mentioned in the news [2].

We seek to go beyond these limitations by considering an unsupervised approach to public health event detection and compare our work to the well-established rule-based system of MediSys.

2.1.2 Supervised detection

Recently, the work done in [11] proposes GRITS that uses the binary relevance method[1] to predict the disease referred to by a body of text. This uses an ensemble of logistic regression classifiers, one for each disease label (approximately 120). Each classifier estimates the probability that a text passage is associated with a single disease, given the vector of features extracted by GRITS' NLP algorithms. The HealthMap data

[1] sklearn.multiclass.OneVsRestClassifier.

used to train the GRITS classifiers is sufficiently large, but each article is only labeled with one disease, even when a text may mention multiple diseases. This means that disease traits extracted from an article may not map specifically to the disease that article is labeled with, negatively impacting classifier training.

Many machine learning methods are based on trigger words for event recognition. Current event detection research lacks comprehensive consideration of the context of the trigger words. In [45], the contextual information of word is divided into sentence-level and document-level in the method. The contextual information is captured based on BiLSTM model. At the same time, a word representation method suitable for trigger word classification tasks is proposed in this paper. The word representation incorporates semantic information, grammar information, and document-level context information of word. The word vectors in the sentence are sequentially input into BiLSTM model to obtain output vectors containing sentence-level contextual information. However, the method based on trigger words introduces a large number of counterexamples in training, resulting in imbalances between positive and negative examples.

Numerous supervised classifiers exist for detecting public health events within unstructured text. In all cases, the authors incorporate the use of some type of semantics, such as: roles [15], hedges [12], or ngrams [26,50] in order to capture relevant entity co-occurrences within a document. A limitation however is that they all also use manually labeled data to build their models. Although automatic labeling is exploited in the work of Stewart et al. [41], this approach has some limitations since the full sentence parsing techniques is not only expensive, but results in semantic ambiguity, given the parse tree representations used.

In conclusion, the current dominant role in event extraction research is method based on machine learning [48], but this method requires large-scale labeled training corpus. If the training corpus is not enough or the category is single, it will seriously affect the extraction effect of the event, and the corpus construction becomes an important task. However, the construction of the corpus takes a lot of manpower and time. In order to alleviate this problem, the scholars further explored the method of deep learning. In our work, we seek to go beyond the human effort associated with building a supervised classifier by taking an unsupervised approach to event detection. We compare our approach to a well-established system: MediSys.

In the rest of the article, we rely on the event formulation depicted in Fig. 1. It presents a graphical representation of this model, where F is the term space size of all kinds of entities (e.g., in figure the count of all kinds of entities is f). Furthermore, within the figure the concept/node Event E is on top of the other nodes, since in the unsupervised event detection an Event is a latent variable, whose value is defined

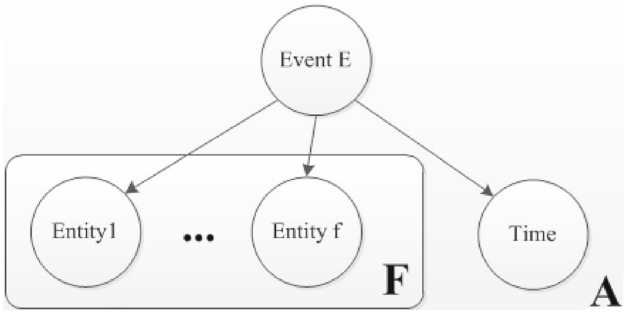

Fig. 1 Graphical model representation

with respect to the observed content of articles, i.e., *Entities* and *Time*, by a generative process. A indicates the article set.

2.1.3 Hybrid models

Researches which bridge the gap between a fully supervised and unsupervised approach are the works done by Paul et al. [36] and more recently by Burchard et al. in [5]. Using labeled messages for relevance to health, they grouped together symptoms and treatments into health-related topics in an unsupervised manner. They apply the Ailment Topic Aspect Model (ATAM) to over one and a half million health-related tweets and discover mentions of over a dozen ailments, including allergies, obesity and insomnia. Like probabilistic topic models, such as latent Dirichlet allocation (LDA), associate word tokens with latent topics. Documents are distributions over topics, and topics are distributions over words, often forming a semantically coherent word set. ATAM, which models how users express their illnesses and ailments in tweets, builds on the notion of topics. It assumes that for each health-related tweet reflects a latent ailment such as flu, allergies, or cancer. Similar to a topic, an ailment indexes a distribution over words. Furthermore, a recent work conducted a data-driven exploratory study of COVID-19 information using machine learning and representation learning methods on Twitter users' data [10].

With these works, authors showed how Twitter has broad applicability for public health research. The disadvantage of these methods is that they require large-scale labeled training corpus. Furthermore, much debate and polarization exist about the impact of social media on the health of patients that might be considered as a non-ideal source of data [38]. The limitation of social media also from a demographic point of view are also described by [37,43], in which Twitter users tend to be younger and healthier than the average of the entire population, biasing all systems and models built on top.

2.1.4 Multilingual event extraction models

Building effective epidemiological surveillance systems is of high importance these days. Detection of news reports on disease outbreaks is a crucial requirement of such systems. In the paper [33], authors study in detail the performance of different methods on the task of epidemiological news report detection. The evidence presented in their work suggests that the models based on fine-tuned language models and/or graph convolutional networks (training data are really relevant here) achieve very good performance ($>90\%$) on the classification of multilingual epidemiological texts, not only for high-resource languages but also for low-resource languages.

Authors in [29] present *Daniel* as a text genre-based information extraction (IE) system devoted to news. It is efficient at distinguishing irrelevant documents in epidemic surveillance and at filtering streams of documents with low-resourced languages. When no classical IE system is available or training data is scarce, Daniel can fill the gap efficiently. The method described increases coverage in number of languages at low cost, rather than optimizing results with a particular language. Wikipedia is used to screen some common disease names to be matched with repeated character strings. The language variations, such as declensions, are handled by processing text at the character level, rather than at the word level. This additionally allows Daniel to handle various writing systems in a similar fashion. With an average F1-measure of 0.85, Daniel scores are below state-of-the-art systems (Puls used by MediSys or Biocaster), as authors confirmed with their comparative evaluation [28].

Finally, we describe MediSys, the system with which we compare our approach, that allows the selection of articles about any subject via Boolean combinations of search words or lists of search words, organized into classes such as *Countries*, *Communicable Diseases*, *Animal Diseases*, *Organizations*, etc. All the search words are multilingual. Each subject definition is called *alert*, which, according to the nature of the search words, is multilingual [31]. Section 4.5.1 will provide more details. MediSys has proven to be useful and effective for finding documents from a large number of Web sources [39].

In contrast to MediSys, UPHED identifies *events* as clusters of documents associated with labels, i.e., a set of diseases and locations describing clusters. In an operational setting, we propose that after MediSys identifies documents for which alerts are generated, UPHED can deliver more information about the specific outbreak of the diseases reported in those documents, by aggregating documents into larger units than alerts, namely *events*. This can advantage UPHED to use the multilingualism extraction of MediSys. With the help of domain experts, we experiment with such a setting and present the results in the discussion that follows.

2.2 Unsupervised detection

Two main approaches exist for the unsupervised detection of events from raw text. They are document-centric and token-centric.

2.2.1 Document-centric unsupervised detection

Retrospective event detection In document centric approaches, a document is assumed to contain the textual mention of one or more real-world events. An event is inferred by first modeling a document as a weighted set of tokens and then grouping-related documents into clusters based on the similarity between vectors. When no new events are assumed to evolve over time, the problem is one of classical document cluster and referred to as Retrospective Event Detection (RED). State-of-the-art document-centric approaches use generative or probabilistic models. So, instead of directly associating documents to words (as in the non-probabilistic detection), generative models associate each document with some event and each event with some significant words, and is the approach we take in UPHED.

Unlike previous works done in this area [6,17], we also explore how our system can be used in a complementary manner with an existing e-EI system.

Online event detection Conversely from RED, Online (or New) Event Detection [4,46], the total number of events is unknown, and increases over time. The latest document of an incoming text stream is assigned to either an existing cluster based on the similarity of its vector to the existing cluster prototypes. When no existing cluster assignment is possible, a new event is assumed to be detected.

In the study [8,9], authors presented a system to automatize the event validation process by predicting whether a given event has evidence within a set of non-annotated documents, thus simplifying the task of manually searching for event-related information to confirm or deny its verity. The developed system allows to specify an event, retrieves candidate web documents, and assesses what are the documents (if any) where it occurs. The validation method relies on a state of the art model for event validation. The user can review the documents and revise the validation judgments given by the system. Given the possibility for users to provide their own validation judgments, the application can also be used to acquire ground truth data for a given set of input events. Authors chose the Web as a source for documents, due to its easy accessibility and wide event coverage. With our approach we tailor the problem to medical news articles available by the Joint Research Centre, later introduced, and we focus more on event extraction rather than event validation.

Other works focus on collecting and structuring large sets of events, like YAGO2 [7,24], DBpedia,[2] and Wikipedia Current Events portal[3] [16]. Despite the well-structured event-related information, these works do not particularly focus on relations between events and supporting documents. The work in [27] presents methods for populating knowledge bases by automatically extracting and organizing named events from news corpora. The generated corpus is made of 25,000 events and 300,000 news articles, but the ground truth used to evaluate the grouping of documents into events is much smaller: in total, it consists of around 100 named events and 1600 articles in Wikinews and news sources referenced in Wikipedia articles. Moreover, such ground truth is built based on event names and the document categorization of Wikinews and Wikipedia Current Events, without reporting mutual conformation of event participants with temporal constraints. These approaches are more generic and do not focus on health event.

In UPHED, we use a generative model for detecting events and seek to explicitly model the temporal behavior of tokens in a manner similar to online detection, but instead, we address the problem of selecting the salient and significant features to include in the generative model—more similar to token-centric approaches discussed below.

2.2.2 Token-centric unsupervised detection

In contrast to the document-centric approaches of retrospective and online event detection, token-centric approaches infer an event by modeling the temporal behavior of tokens (features) within a collection.

Burst detection In burst detection, bursty tokens exhibits high document frequency over a finite time window. The underlying assumption is that if two tokens co-occur frequently in the same temporal window, then they are assumed to be semantically associated and infer an event [18]. Extensive work has also been remarkably done in on-line detection by He et al. [21,23]. Notably in [22], the authors propose a theoretically elegant, effective, and simple probabilistic model for both offline and online topic detection tasks, leveraging feature selection and temporally discriminative weights. They show how temporal information can be incorporated into more sophisticated generative models like von-Mises Fisher (vMF) [3], but, as stated in their work, no significant improvements in topic detection performance were obtained. Furthermore, they demonstrated that for a generative model, like vMF, due to its generative smoothing process, the utility of discriminative features is attenuated.

In our work, we focus on unsupervised offline topic detection. We succeeded in incorporating temporal information into a sophisticated generative model, and in doing so, we demonstrate the utility of discriminative features.

Feature trajectory The work in [20,44] considers the problem of analyzing features trajectories in both time and frequency domains, with the specific goal of identifying important and less-reported, periodic and aperiodic features. The problem of analyzing feature trajectories for event detection uses a well-known technique in signal processing to identify distribution of all features by spectral analysis. A set of features with identical trends can be grouped together to reconstruct an event in a completely unsupervised manner. In Sect. 3.2, we present the details of such analysis, forming the building blocks for our approach. Finally, spectral analysis techniques have previously been used in [1] to identify periodicities and bursts from query logs. The authors' focus was on detecting multiple periodicities from the power spectrum graph, which were then used to index words for "query-by-burst" search.

In our study, we use spectral analysis to classify word features along two dimensions, namely periodicity and power spectrum. These features with their dimensions are later input to the generative model for detecting events (for details see Sect. 3.2).

3 UPHED: unified approach to public health event detection

An event is defined as a specific episode happening at a specific time and place [14], which may be consecutively reported by many articles in a period. The goal of this work is to introduce an approach to detect events in an unsupervised manner. The model can also be used as a baseline for detecting any anomalies and for building a predictive model for the near future.

3.1 Entity-centric feature representation

As first step, we process raw text to build an entity-centric feature representation of each document. Given a collection of text documents, we define a finite set of articles, $a_i \in \mathcal{A}$, as well as an Event Template denoted as: $\mathcal{T} :=$ (victims, diseases, locations, time).

The template $\mathcal{T}$ represents a set of feature types, which are important for describing events. For a public health event, the template provides information on who was involved; $what$ is the affecting disease, $where$, and $when$. The template was motivated by the work presented in [40].

Figure 2 is a graphical representation of this model for the medical case, where F is the term space size of the three kinds of entities (i.e., Victim, Disease, and Location). A indicates the article set.

[2] http://wiki.dbpedia.org.

[3] https://en.wikipedia.org/wiki/Portal:Current_events.

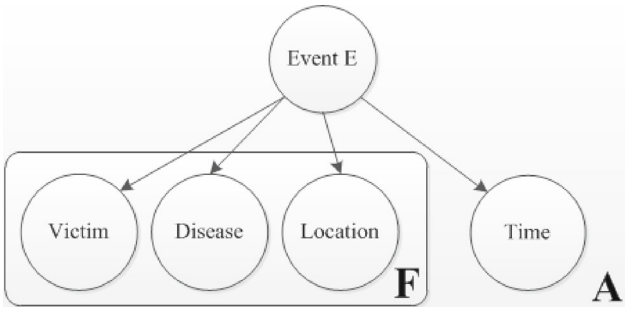

Fig. 2 Graphical model representation for the medical case

Content: The content of each article is represented by a bag-of-entities, whose types are given by $\mathcal{T}$. For each article, a_i, a vector is created for each of the feature types; each entry in the vector corresponds to the frequency with which an entity of a given type, appears in the bag-of-words representation. For sake of clarity, the vector victims (diseases and locations are defined similarly) is considered to be a list

$$< \text{victim}_{i1}, \ldots, \text{victim}_{iN} >$$

and each element is the occurrence count of corresponding entity in a_i.

Time: Each event e_j corresponds to a peak on an article count versus time distribution. In other words, the model is a mixture of many distributions of events. A peak is modeled by a Gaussian function, where the mean is the position of the peak and the variance is the event's duration; i.e., the period between the earliest and latest timestamp (or discrete time value) for the articles of an event. A Gaussian Mixture Model is chosen to model time.

In order to simplify our model, we assume that all feature types of an article, given an event e_j, are conditional independent. The probability of an article a_i to be associated with an event e_j is given by the product of the following individual probabilities:

$$\begin{aligned} p(a_i|e_j) = & p(\text{victims}_i|e_j) * p(\text{diseases}_i|e_j) \\ & * p(\text{locations}_i|e_j) * p(\text{time}_i|e_j) \end{aligned} \quad (1)$$

where the probabilities $p(\text{victims}_i|e_j)$, $p(\text{diseases}_i|e_j)$, and $p(\text{locations}_i|e_j)$ are computed by the Multinomial distributions imposed by our generative model, as presented hereafter in Algorithm 2. Finally, the probability $p(\text{time}_i|e_j)$ follows the Gaussian distribution.

3.2 Feature analysis

In our work, we posit that event detection involves a dual task: the detection of periodic as well as aperiodic events. With respect to a window of 1 year, for example, aperiodic events are also important, since they can represent an event that is annual, e.g., season flu, or quite severe and life threatening, such as a sudden outbreak of EHEC.

Detection of periodic as well as aperiodic events is based on identification of periodic and aperiodic features as described in [20] using a common technique such as the spectral analysis. In this approach, features are classified with respect to their periodicity (P_f) and their dominant power spectrum (S_f).

The periodicity of a feature refers to its frequency of appearances. If the feature is *aperiodic*, then it occurs once within the period P, and its P_f has a value equal to the period itself. If the feature is *periodic*, then it happens regularly with a fixed periodicity, i.e., $P_f \leq \lceil P/2 \rceil$. The periodicity is a function of the dominant power spectrum which is computed via the discrete Fourier transform applied to the feature distributions.

3.2.1 Representative features

Let P be the duration/period (in days) of a collection of articles, and F represents the complete feature space. The representation vector of a feature $f \in F$ is defined as follows:

Definition 1 (*Feature Trajectory*) The trajectory of a feature f can be written as the sequence

$$y_f = [y_f(1), y_f(2), \ldots, y_f(P)]$$

where each element $y_f(t)$ is a measure of feature f at time t, which could be defined using the normalized DF-IDF score

$$y_f(t) = \frac{\text{DF}_f(t)}{N(t)} * \log\left(\frac{N}{\text{DF}_f}\right)$$

where $\text{DF}_f(t)$ is the number of articles containing feature f at day t; DF_f is the total number of articles containing feature f over P; $N(t)$ is the number of articles for day t; and N is the total number of articles over P.

3.2.2 Spectral analysis for dominant period

Given a feature f, we decompose its feature trajectory y_f into the sequence of P complex numbers $[X_1, \ldots, X_P]$ via the discrete Fourier transform DFT:

$$X_k = \sum_{t=1}^{P} y_f(t) * e^{-\frac{2\pi i}{P}(k-1)t}, \; k = 1, 2, \ldots, P$$

DFT can represent the original time series as a linear combination of complex sinusoids, which is illustrated by the

inverse discrete Fourier transform IDFT:

$$y_f(t) = \frac{1}{P}\sum_{k=1}^{P} X_k * e^{-\frac{2\pi i}{P}(k-1)t},\ k = 1, 2, \ldots, P$$

where the Fourier coefficient X_k denotes the amplitude of the sinusoid with frequency k/P.

The original trajectory can be reconstructed with just the dominant frequencies, which can be determined from the power spectrum using the popular periodogram estimator. The *periodogram* is a sequence of the squared magnitude of the Fourier coefficients

$$||X_k||^2, \quad k = 1, 2, \ldots, P/2$$

which indicates the signal power at frequency k/P in the spectrum.

From the power spectrum, the dominant period is chosen as the inverse of the frequency with respect to the highest power spectrum, as follows.

Definition 2 (*Dominant Period*) The dominant period of a given feature f is

$$P_f = \frac{P}{\arg\max_k ||X_k||^2}$$

Accordingly, we have

Definition 3 (*Dominant Power Spectrum*) The dominant power spectrum of a given feature f is

$$S_f = ||X_k||^2, \quad \text{with} \quad ||X_k||^2 \geq ||X_j||^2,\ \forall j \neq k$$

In conclusion, the dominant power spectrum, S_f, of a feature f is a strong indicator of its activeness at the specified frequency; the higher is the S_f, the more likely the feature is to be relevant within the dataset. Thus, S_f can be considered to filter out irrelevant features, i.e., features with a dominant power spectrum less than a pre-fixed threshold chosen according to the domain. After filtering out irrelevant features, the remaining features are meaningful and could potentially be representative for some events [20].

3.2.3 Identifying burst for aperiodic features

Let y_f be the trajectory of feature f over the period P under observation. Then, for each aperiodic feature f_{ap}, we keep only the bursty period which is modeled by a Gaussian distribution whose tails thin down quickly, preserving more the importance for features close to the burst and reducing their significance proportionally to their distance from the burst.

$$f_{\text{ap}}(y_f) = \frac{1}{\sqrt{2\pi\sigma_f^2}} * e^{-\frac{1}{2\sigma_f^2}(y_f(t)-\mu_f)^2} \tag{2}$$

The well-known Expectation Maximization (EM) algorithm is used to compute the Gaussian density parameters μ_f and σ_f [13].

3.2.4 Identifying bursts for periodic features

For periodic features, it is important to preserve their significance from one burst and the next one, specifically in the point of the trough where the tails of their distributions are. To model each periodic feature f_p, we chose a mixture of K Cauchy–Lorentz distributions, where $K = \lfloor P/P_f \rfloor$. The property from Cauchy–Lorentz distribution to maintain its tails thicker, with respect to the Gaussian distribution, reflects better the behavior of the feature far from the bursts. This property, as observed from the computed y_w, reflects better the distribution of periodic features, since, even for t far from the peak of the burst, generally the feature trajectory y_f reports values important to be considered. The mixture is described as follows

$$f_p(y_f) = \sum_{k=1}^{K} \alpha_k * \frac{1}{\pi}\left[\frac{\gamma}{(y_f(t)-\mu_k)+\gamma_k^2}\right] \tag{3}$$

for the mixture proportions α_k of assigning y_f into the kth Cauchy–Lorentz distribution

$$0 \leq \alpha_k \leq 1 \quad \text{where} \quad \sum_{k=1}^{K} \alpha_k = 1, \forall k \in [1, K] \subset \mathbb{N} \tag{4}$$

Furthermore, μ_k is the location parameter, specifying where is the peak of the distribution, and γ_k is the scale parameter which specifies the half-width at half-maximum. μ_k, γ_k and α_k are computed using the EM algorithm [13].

3.2.5 Feature burst distributions algorithm

In this section, we present the algorithm for computing the feature burst distributions. Algorithm 1 wraps together the concepts and the approaches explained so far. The output of this algorithm is all the feature burst distributions which will be used as input for Algorithm 2. We use the notation θ_{type} to identify the burst distributions for all features of a specific type, i.e., θ_v for victim, θ_d for disease, and θ_l for location).

In detail, the algorithm works as follows. For each feature f, we compute the feature trajectory y_f, as presented in Definition 1 of Sect. 3.2.1. Then, we decompose y_f into the

sequence of complex number via the discrete Fourier transform *DFT*, and we compute the Dominant Period P_f and the Dominant Power Spectrum S_f, according to Definitions 2 and 3 of Sect. 3.2.2. With respect to P_f and S_f, we decide to transform the vector y_f in one of the two possible modeled vectors: (i) vector $f_{\text{ap}}(y_f)$, following Sect. 3.2.3, if the feature is aperiodic; (ii) vector $f_p(y_f)$, following Sect. 3.2.4, if the feature is periodic. Finally, considering the type of the feature, we store $f_{\text{ap}}(y_f)$ or $f_p(y_f)$ vector in one of the corresponding matrix θ_{type} (with respect to type) which stores all the feature of one type over the dates t. Successfully using the feature burst distributions for having a more representative model will be shown in Sect. 4.

3.3 Detecting public health events

A core step in the unsupervised detection of events is the clustering of articles and generation of events. Formally, from this stage we get sets of conditional probabilities, already introduced in Sect. 3.1 with Equation 1, and in the following better explained: (i) $p(a_i|e_j)$ is the set of conditional probabilities for an article a_i, given an event e_j; (ii) $p(\text{victims}_i|e_j)$ ($p(\text{diseases}_i|e_j)$ and $p(\text{locations}_i|e_j)$ are defined similarly) is the set of conditional probabilities for occurrences of feature type victims in a_i, given an event e_j; and (iii) $p(e_j)$ is the set of probabilities for an event e_j. We use these probabilities, as a basis for determining that an event has occurred.

3.3.1 Generative model for public health events

Numerous techniques exist for detecting events in an unsupervised way (see Sect. 2). Events in the unsupervised event detection are latent variables, whose value is defined with respect to the observed content of articles by a generative model. In this work, we choose to apply a retrospective event detection algorithm since it is important in e-EI to use data historical collection, in order to build a predictive model of public health events for the near future. The same idea is used in statistical methods for public health to analyze event data from indicator-based systems (e.g., the Farrington Algorithm). Additionally, we have chosen a probabilistic generative model for event detection, because it has been proven to be a more unified framework for handling the multiple modalities (i.e., time and content) of an article [30].

For these reasons, we base our unsupervised event detection algorithm on the Retrospective Event Detection (RED) algorithm presented by Li et al. [30]. It relies on a generative model where the articles are produced using Multinomial distributions over features of multiple types. These articles are used later as starting points for a clustering relying on the iterative EM algorithm. In addition, in their work, the Multinomial distributions are initialized with random probabilities. Thus, the generated articles are randomly picked.

Algorithm 1: Feature Analysis using feature burst distributions

Input: A set of extracted features F; a set of articles A; a fixed threshold τ
Output: all the feature burst distributions θ_{type}, i.e., θ_v, θ_d, and θ_l
begin
  N := count the number of articles within A;
  D := count the number of distinct date t, according to articles' timestamps, in A;
  $P[|F|]$:= array storing the dominant period P_f for each feature $f \in F$;
  $S[|F|]$:= array storing the dominant power spectrum S_f for each feature $f \in F$;
  $FeatureDistributions[|F|][D]$:= matrix storing the vectors of feature trajectories y_f for each feature f over the dates t;
  $FourierFeatureDistributions[|F|][D]$:= matrix storing the decomposition of the vectors of feature trajectories y_f into the sequence of complex vectors via the discrete Fourier transform DFT;
  θ_{type} := matrix storing the modeled vectors of aperiodic feature $f_{ap}(y_f)$ or periodic feature $f_p(y_f)$ of feature trajectories y_f for each feature f of a $type$ over the dates t;
  for *each distinct date t in A* **do**
    $N(t)$:= count the number of articles at date t;
  for *each feature f in F* **do**
    DF_f := count the total number of articles containing entity f;
    for *each distinct date t in A* **do**
      $DF_f(t)$:= count the number of articles containing feature f at date t;
      DF-IDF := $\frac{DF_f(t)}{N(t)} * \log\left(\frac{N}{DF_f}\right)$;
      Store DF-IDF into $FeatureDistributions[f][t]$;
  Compute $FourierFeatureDistributions$ using DFT on $FeatureDistributions$;
  for *each feature f in F* **do**
    $type$:= type of feature f;
    $P[f]$:= compute the dominant period P_f of the corresponding feature;
    $S[f]$:= compute the dominant power spectrum S_f of the corresponding feature;
    if $S[f] \geq \tau$ **then**
      if $P[f] > \left\lceil \frac{P}{2} \right\rceil$ **then**
        Model the feature by a Gaussian distribution (aperiodic feature f_{ap});
        Insert $f_{ap}(y_f)$ in θ_{type};
      else
        Model the feature by a mixture of $K = \lfloor P/P_f \rfloor$ Cauchy–Lorentz distributions (periodic feature f_p);
        Insert $f_p(y_f)$ in θ_{type};

As part of our approach, we refine the RED algorithm by going beyond this random initialization of probabilities—exploiting the feature distributions from our *Feature Analysis* stage (Sect. 3.2). The underlying intuition for our approach is based on proven results [49], which show that an initial starting point estimated in a better-than-random way can, in fact,

be expected to speed up the iterative EM algorithm converging closer to the optimum of the computed log-likelihood of a collection of articles, than an initial point that is picked at random. In our approach, we aggregate the computed feature distributions over the articles, and use this information into the Multinomial distributions of the generative model. Thus, the generated articles, used as starting points by EM algorithm, are not totally randomly picked.

Although it has been proven that retrieved events are not influenced by the starting points [25,42], the EM algorithm needs to be restarted several times with several different random starting points in order to get a good approximation of events. Supported by the analysis in [49], we do not need multiple restarts of the EM algorithm, since an initial starting point estimated in this way, can be expected to be closer to the optimum than a randomly picked initial point.

Our generative model is described in Algorithm 2.

Algorithm 2: Detection of Public Health Events: the generative model

begin
- Choose an event $e_j \sim$ Multinomial(θ_j);
- Generate a medical article $a_i \sim p(a_i|e_j)$;
- Draw a timestamp $time_i \sim N(\mu_j, \sigma_j)$;
- **for** *each feature of* a_i*, according to the type of current feature* **do**
  - Set $victim_{iv} \sim$ Multinomial($\theta_v|time_i$);
  - Set $disease_{id} \sim$ Multinomial($\theta_d|time_i$);
  - Set $location_{il} \sim$ Multinomial($\theta_l|time_i$);

In the algorithm, the vector θ_j represents *event* probabilities initially instantiated randomly (here the definition of *event* is according to the formalization of the Multinomial distribution); μ_j and σ_j are parameters of the conditional Gaussian distribution given an event e_j; θ_v, θ_d, and θ_l are feature burst distributions output by Algorithm 1 aggregating the burst distributions for all features of type *victim*, *disease*, and *location* over the $time_i$ of a given event e_j. Finally, we associate to each feature of type victim—same consideration for features of type disease and location—the probability value extracted by its previously computed burst distribution, over the time$_i$ of a given event e_j.

3.3.2 Learning generative model parameters

After the initialization part presented in Algorithm 2, we need to refine all the model parameters. They can be estimated by Maximum Likelihood method following the approach described in [30]. By introducing latent variable, i.e., events, we can write the log-likelihood of the joint distribution as:

$$l(X;\theta) \propto= \sum_{i=1}^{A} \log \left(\sum_{j=1}^{K} p(e_j)p(a_i|e_j) \right) \tag{5}$$

where X is the corpus of articles; A and K are the number of articles and the number of events, respectively; and θ represents all the model parameters introduced in Algorithm 2, such as θ_j, μ_j, and σ_j for each event e_j. Furthermore, as described in Sect. 3.1 (Eq. 1), given an event e_j all kinds of information of the ith article are conditional independent.

Expectation Maximization (EM) algorithm is applied to maximize log-likelihood. The parameters are estimated by running E-step and M-step alternatively.

In **E-step**, we compute the posteriors, $p(e_j|a_i)$, by:

$$p(e_j|a_i)^{(t+1)} = \frac{p(e_j)^{(t)} p(a_i|e_j)^{(t)}}{p(a_i)^{(t)}} \propto p(e_j)^{(t)} p(a_i|e_j)^{(t)} \tag{6}$$

where the upper script (t) indicates the tth iteration.

In **M-step**, we update the parameters of our model. Since victims, diseases, and locations are modeled similarly, i.e., with independent mixture of unigram models, so their update equations are the same. Then, for sake of clarity we show the update only for the nth feature of type victim. Parameters are updated by:

$$p(\text{victim}_n|e_j)^{(t+1)} = \frac{1 + \sum_{i=1}^{A} p(e_j|a_i)^{(t+1)} * tf(i,n)}{N + \sum_{i=1}^{A} \left(p(e_j|a_i)^{(t+1)} * \sum_{s=1}^{N} tf(i,s) \right)} \tag{7}$$

where tf(i,n) is the count of entity victim$_n$ in a_i and N is the vocabulary size. For each type of entities, N is the size of corresponding term space. Since the co-occurrence matrix is very sparse, we apply Laplace smoothing [34] to prevent zero probabilities for infrequently occurring entities in Eq. 7.

The parameters of the *Gaussian Mixture Model* are updated by:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{A} p(e_j|a_i)^{(t+1)} * \text{time}_i}{\sum_{i=1}^{A} p(e_j|a_i)^{(t+1)}} \tag{8}$$

and

$$\sigma_j^{(t+1)} = \frac{\sum_{i=1}^{A} p(e_j|a_i)^{(t+1)} * (\text{time}_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^{A} p(e_j|a_i)^{(t+1)}} \tag{9}$$

It is important to note that because both the means and variances of the Gaussian functions change consistently with the whole model, the Gaussian functions work like sliding

windows on a time line. In this way, we overcome the shortcomings caused by the fixed windows or the fixed decaying function parameters used in traditional news event detection algorithms [4]. At last, the mixture proportions are updated by:

$$p(e_j)^{(t+1)} = \frac{\sum_{i=1}^{A} p(e_j|a_i)^{(t+1)}}{A} \quad (10)$$

The EM algorithm increases the log-likelihood consistently, while it will stop at a local maximum.

4 Experiments and evaluation

In our experiments, we evaluate the efficiency of our approach against two different event detection strategies: Rdm [30] (baseline) and GaussApp [20] (an approximation of our approach).

4.1 Experimental goals

We examine the effectiveness of our system in detecting a major outbreak of enterohemorrhagic Escherichia coli (EHEC), which occurred in Northern Germany. Finally, going beyond the EHEC case study, we conduct an extensive comparative analysis of our UPHED algorithm with the well-established rule-based system of MediSys by analyzing the *alerts* generated by MediSys versus the *events* detected by our UPHED method.

For sake of transparency, in our previous works [17], we exposed different set of evaluations: experiments on tuning our algorithm, on features pruning, and on the selection of the number of events.

4.2 Feature set

Table 1 presents the main categories of features collected and their counts. The entities have been extracted using two different named entity recognition tools: UMLS MetaMap[4] and OpenCalais.[5]

OpenCalais was used to recognize *diseases* and all variants of locations. MetaMap was used to identify the victim features. MetaMap has originally been developed for indexing biomedical literature and relies upon the Unified Medical Language System (UMLS) Metathesaurus, a very rich biomedical vocabulary designed and maintained by the US National Library of Medicine. Thus, it allows extracting highly domain-specific concepts, but leads when applied to social media or news articles to false positives. For our feature set we are only interested in disease names and symptoms which are more unambiguously detected by OpenCalais. In contrast, the more detailed information on victims provided by MetaMap is very useful for our algorithm. For these reasons, we decided to exploit these two different named entity recognition tools.

Table 1 Overview on the collected features

Feature types	Feature categories	Norm	Unnorm
Victims	Population group	28	4100
	Age group		
	Family group		
	Animal		
Diseases	Diseases	917	2754
	Symptoms		
Locations	City	955	982
	Province or state		
	Country		
	Continent		

Norm is the number of normalized features; Unnorm is the total number of features before the normalization process

Through manual inspection, we further found that noise introduced into the algorithm due to multi-word expressions caused an explosion of the number of features. This is particularly acute for a feature-centric approach such as ours in the medical domain, in which features, consisting of many multi-word expressions, quite commonly exacerbate the problem of producing irrelevant events. We normalized the features using the UMLS MetaMap semantic network. As an example, terms such as *boy, girl, baby, child, kid* were normalized to the single feature, *child*.

4.3 Experiment I: efficiency comparison

The intention of this analysis is to show that the selection of a good starting point can boost the EM algorithm to converge quickly to the optimum and that it is unnecessary to restart the EM algorithm multiple times with different random starting points, as done previously—thereby improving its run time performance.

In this section, we compare three strategies for detecting events:

1. *Rdm:* The baseline of our method, which initializes the EM algorithm with random points, as done in [30], and adapted to the medical domain.
2. *GaussApp:* An approximation of our method identifying bursts for periodic features using a mixture of $K = \lfloor P/P_w \rfloor$ Gaussians, as in [20].
3. *UPHED:* Our revised and proposed method.

[4] http://mmtx.nlm.nih.gov.

[5] http://www.opencalais.com.

Table 2 Efficiency comparison of three different strategies

	Rdm	GaussApp	UPHED
Optimum (log-likelihood)	5807	6140	**6624**
Best starting point (log-likelihood)	4105	4997	**5981**
Average running **time** (s)	401	287	**263**
Average number of iterations for EM	91	45	**38**
Best trial: number of iterations for EM	63	**31**	**30**
Worst trial: number of iterations for EM	121	58	**50**
Average number of **restart** of EM to get the optimum	9–10	**1–2**	**1–2**

Bold values indicate best results

Dataset To build our data set, we collected source documents (articles) from MediSys feed. The data were gathered for a 2 months period, from May 1 to June 30, 2011. In total 13,076 documents were collected.

Convergence time The experimental results are shown in Table 2. Here, we report the log-likelihood both for the best ending points—each one named optimum—and for the best starting points of the three strategies. The log-likelihood indicates how likely the documents are generated by models (where larger log-likelihood values are better). Averaging over several iterations, we show the time taken for the EM algorithm to converge as well as the number of iterations under the best, average and worst case scenarios.

Number of restarts A consideration when using an EM algorithm is that the convergence to a local maximum can prematurely mislead one to use results that are sub-optimal (i.e., the algorithm has not reached the global maximum). Also shown in Table 2 is the number of restarts needed for the EM algorithm to converge to its optimum.

From these results, we can conclude that in all the reported measures, our proposed method UPHED performs more efficiently than Rdm and GaussApp. We can see that UPHED reaches a better optimum compared to Rdm, which starts from a random point. Also we notice that UPHED needs less time and fewer iterations to reach this optimal convergence.

4.4 Experiment II: effectiveness

In this experiment, we study the large outbreak of enterohemorrhagic Escherichia coli (EHEC) which occurred in Northern Germany during May and June 2011. In this section, we illustrate how our approach effectively detected this medical event. Also, our clustering method allowed us to identify other non-EHEC-related medical events that occurred during the same period.

We run our method on 13,076 documents of which 4757 were categorized in six medical events related to the EHEC outbreak, as shown in Table 4. In addition, 4639 were clustered into six additional non-EHEC-related medical events, as reported in Table 5. Both tables show the characterizing cluster terms resulting from our clustering. For each article a_i we evaluated its conditional probability given an event e_j, i.e., $p(a_i|e_j)$, setting a threshold of $\tau_p = 5\%$ for each $p(a_i|e_j)$; for $p(a_i|e_j)$ below this threshold, the article a_i was not associated with the event e_j. Given these settings, the total number of articles associated with the set of events is 9396. In addition, the number of documents assigned to a cluster is shown, as well as a manually created description of the event. The event terms were collected by automatically selecting for each feature type, i.e., diseases and locations, the most probable features having their conditional probability given an event, i.e., $p(\text{diseases}_i|e_j)$ or $p(\text{locations}_i|e_j)$, exceeding the probability threshold τ_f of 40%; for $p(\text{diseases}_i|e_j)$ or $p(\text{locations}_i|e_j)$ below τ_f, the feature was not associated with the event e_j. Figure 3 shows an example of the conditional probability of the top ten terms and entity types being associated event E_1.

Cluster quality As mentioned in [48], a ground truth data set for public health event detection is unavailable. To evaluate the correctness of cluster-document assignment, we performed a manual evaluation. The algorithm was applied to our data set and twelve clusters were created. These clusters were manually assessed by three subjects who had to decide, for each document, whether it was assigned to the correct cluster. In more detail, for each created cluster, the testers were confronted with the two most probable entities for disease, location and victim. These six terms were considered to be descriptive for the cluster, or the documents belonging to this cluster, respectively. The testers had to decide whether the document under consideration described an event taking place in the location specified by the cluster term labels. They had to decide whether it dealt with the disease and the victims mentioned in the cluster description. When all three criteria were fulfilled, the tester was asked to label this document as correctly assigned. The quality of cluster assignment was measured in terms of precision and recall. We are able to achieve a precision of 60% and a recall of 71% with $\tau_p = 5\%$.

For clarity, we reproduce in Fig. 3 the rest of the experiments with different values of τ_p. As expected, decreasing τ_p to three resulted in an increase in recall but a decrease in precision. It is less intuitive when the values of precision and recall for τ_p are equal to two. In this case, precision con-

Table 3 Precision and recall for different values of τ_p

τ_p (%)	Precision (%)	Recall (%)
5	60	71
3	41	86
2	25	18

tinues to decrease, but unexpectedly, recall also decreases. After investigation, we found that this is due to the fact that articles are increasingly associated with multiple clusters at the same time and it is difficult to select the correct one.

4.4.1 Detection of EHEC-related events

All medical events reported in Table 4 are relevant to the EHEC outbreak during May and June 2011 and show how the situation and geographic focus changed over time. In Fig. 4, we can see the temporal distribution of articles for all six EHEC-related medical events.

Similar results for the medical events related to the EHEC outbreak in Europe were also detected and documented in the MediSys Report [32]. Our method complements these results. In contrast, we note that we were able to arrive at a similar outcome as MediSys, without having to perform a manual analysis document by document, as was done in [32]. The advantage of using our unsupervised approach, is that we were able to: 1) achieve a speed up in the process of detecting outbreaks; and 2) better direct the investigators attention to documents containing important information about potential medical outbreak.

Surely, human inspection is always needed in this scenario and cannot be left out, but we can do provide a faster complementary inspection tool.

To better understand the EHEC events in Table 4 and in Fig. 4, we summarize the key temporal developments:

- E_1 presents an outbreak of enterohemorrhagic Escherichia coli (EHEC) occurred in Northern Germany, with a sudden rise on articles first detected at the end of May; later, another rise was observed between June 5 and June 11 according to German authorities announcing that bean sprouts were the source of infection.
- E_2 reports that the contagion was caused by contaminated Spanish vegetables. This statement was alleged at the end of May, as can be seen in the graph with an increasing trend of the curve, while there was another rise on June 1 when Spanish farmers announced that Spanish cucumbers had been tested negative for EHEC.
- E_3 cites when Russia applied trade restriction for European vegetable products with a peak of documents on June 3.
- E_4 refers to many cases of EHEC contagion in Southern France observed between June 3 and June 7.
- E_5 reports an increasing trend with articles on May 27, on the first Swedish tourist group visiting Northern Germany who denounced EHEC infection.
- E_6 clusters together all articles mentioning the European Commission during the discussion on the alleged contaminations of Spanish cucumbers and vegetables, with a peak on sources during June 1 and June 3; furthermore, the European Commission and Parliament were involved on themes about the risk assessment in terms of public health at EU level, with a rise between June 6 and June 9.

4.4.2 Detection of non-EHEC-related events

In Table 5, we see that the medical events presented are related to other public health events that were detected also during the EHEC outbreak. Figure 5 also depicts the temporal development of events.

With the help of domain experts, we summarize the key temporal developments of the non-EHEC events as follows:

- E_7, E_8, and E_9 refer to a big occurrence of measles cases, around ten thousand reported cases in Europe. Of the total, 72% were detected in France, almost with an incidence of ten times more than measles cases occurred in the same period 1 year before (2010). Regarding E_9, several countries were selected as event terms for the feature type locations, which we manually summarized with the location term Europe.
- E_{10} reports that India's scientific community is ready to launch a research programme which will bring to vaccinate millions of people and save another 6.4 million lives over the current decade.
- E_{11} clusters together articles mentioning medical improvements in China during the past 20 years through better infectious disease surveillance systems which have led to decrease the mortality for tuberculosis of 8–6% per year. The success in China has been based on sustained efforts that have progressively achieved coverage of the country's vast population by tuberculosis treatment and surveillance.
- E_{12} cites when the Australian research team antimalarial, which began in 2002 under the direction of Dr. Margaret Phillips, identified a promising inhibitor of a specific enzyme that the malaria parasite requires for survival. For that, her research team won the International Project Of The Year Award.

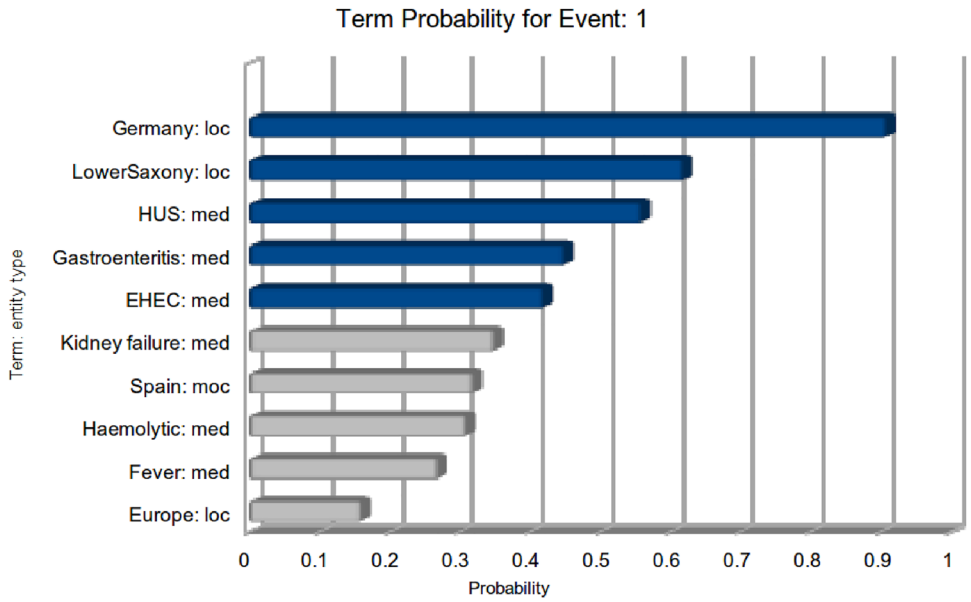

Fig. 3 Conditional probability of the top ten medical condition (med) and location (loc) terms corresponding to EHEC event, E_1, using threshold probability for terms greater than 40%

Table 4 Detected EHEC outbreaks during May and June 2011

Event id	Event terms	fNo. docs	Event description
E_1	Germany, LowerSaxony, EHEC, HUS, gastroenteritis	1,723	Outbreak of enterohemorrhagic *Escherichia coli* (EHEC) occurred in Northern Germany
E_2	Spain, EHEC, hemolytic	957	The contagion was caused by contaminated Spanish vegetables
E_3	Russia, EHEC	450	Russia applied trade restrictions for European vegetable products
E_4	France, EHEC, gastroenteritis, vomiting	387	Many cases of EHEC contagion in France
E_5	Sweden, Germany, EHEC, diarrhea, HUS	801	First Swedish tourist group visiting Northern Germany denounced EHEC infection
E_6	Brussels, Luxembourg, EHEC, HUS	439	Documents treating the economical repercussion on the European market of the entire EHEC outbreak

The columns, respectively, show: the extracted terms, number of documents, and brief description of the real events

4.5 Experiment III: UPHED in comparison with MediSys

In this section, we go beyond the EHEC case study and conduct a more extensive comparison of our UPHED algorithm with MediSys. The comparison was carried out by analyzing the *alerts* generated by MediSys versus the *events* detected by our UPHED method. In Sect. 4.5.1, we first provide a background on MediSys. In Sect. 4.5.2, we describe how MediSys alerts are generated. Then, in Sect. 4.5.3, we present a comparative analysis of UPHED and MediSys.

4.5.1 Overview of MediSys

MediSys is a fully automatic public health surveillance and alerting system run by the Health Threats Unit at Directorate General Health and Consumer Affairs of the European Commission, in collaboration with the Joint Research Centre (JRC) in Ispra, Italy.[6]

MediSys allows one to perform multilingual search over its collection of health-related news articles via Boolean combinations of search terms categorized into *Countries*, *Communicable Diseases*, *Animal Diseases*, *Organizations*, etc. Each subject definition is called *alert*.

[6] https://ec.europa.eu/jrc/en/about/jrc-site/ispra.

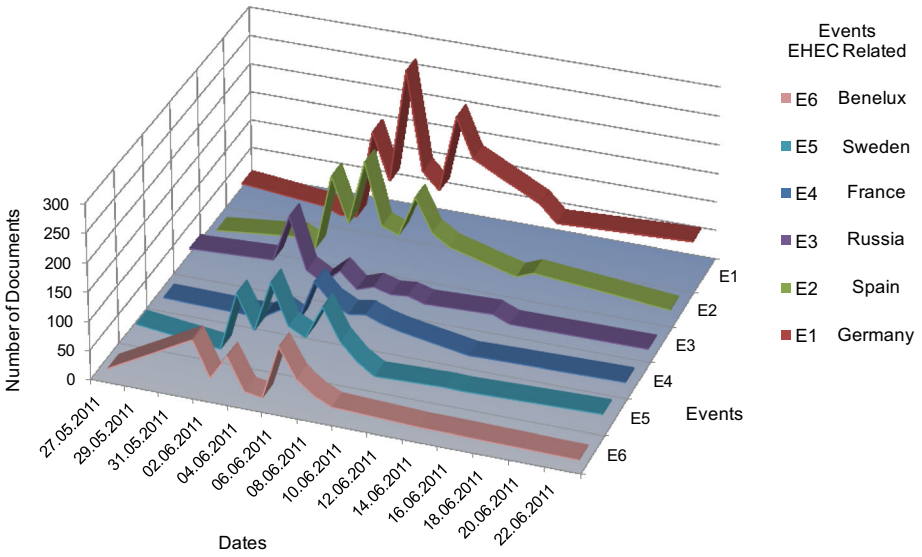

Fig. 4 Documents distributions for each extracted medical event relevant to EHEC outbreak

Table 5 Further detected medical events during May and June 2011

Event id	Event terms	fNo. docs	Event description
E_7	Italy, measles, epidemic	322	Outbreak of measles occurred in Northern Italy
E_8	France, measles, rubella, fever	448	The majority of measles cases in Europe arose in France, around 72% of all measles occurrences
E_9	Europe, measles	2574	The second quarter 2011 registered a peak of measles cases in all Europe
E_{10}	India, New Delhi, malaria	571	India's scientific community is all set to launch a research programme on how to better combat vector-borne diseases, after many cases of malaria in the country
E_{11}	China, Beijing, tuberculosis, malaria	407	China has been sustaining efforts that have progressively achieved coverage of the country's vast population by tuberculosis treatment and surveillance
E_{12}	Australia, Melbourne, malaria	317	"UT Southwestern Research Team's Anti-Malarial Work" wins the "International Project Of The Year Award"

One of the advantages of MediSys is that it has a large news coverage; so, it captures many reports that would go unnoticed by those who only read a few news sources. On the other hand, one of the drawback of having multiple sources is in reporting the same or near duplicate documents to users and subsequently triggering many more alerts than actuality.

To cope with duplicates, MediSys adopts a similarity measure to prune near-duplicate documents. The similarity measure for the news articles is based on cosine similarity of a simple vector-space representation of the first 200 word tokens of each article. This means that not only multiple reports of the same story, but also similar reports about different cases for the same disease may be grouped together and filtered out. This method allows to discard entire groups of non-influent articles at once.

4.5.2 Generating alerts in MediSys

Similar to our notion of an event, a MediSys alert is an indication, that some real-world health-related activity is taking place. An event in MediSys is inferred by using a rule-based approach to extract pattern from unstructured text of news articles. Event Template in MediSys consists of the pair: $\mathcal{T} =<$ Category*, Country $>$; where Category$^* =<$ Disease, AnimalDisease, Organizations $>$. Each instance of the template (textual extraction) is referred to as an alert.

To quantify an alert, MediSys keeps a running count of all alerts for each country. It maintains the average of all documents mentioning a specific category instance and country, over a time window of 2 weeks. An alerting function detects a sudden increase in the number of articles for a given cate-

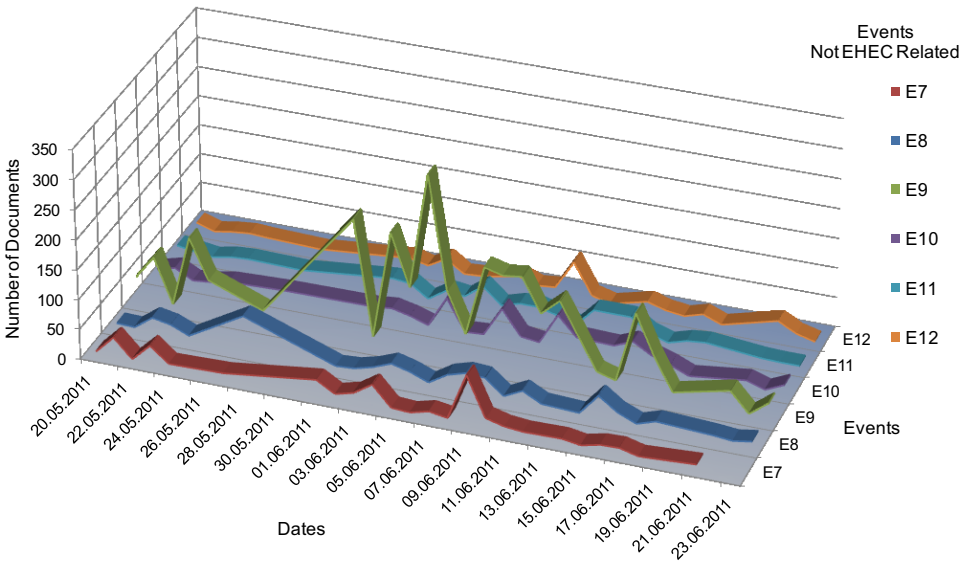

Fig. 5 Documents distributions for each extracted medical event NOT relevant to EHEC outbreak; from E_7 to E_{12}

gory and country, by comparing the statistics for the last day with the 2-week rolling average. The more articles there are for a given *category-country* combination compared to the expected number of articles (i.e., the 2-week average), the higher is the alert level.

The histogram in logarithmic scale in Fig. 6 illustrates seemingly how MediSys presents statistics on the alerts on its web site. On the left side of the figure, alerts with a high level are shown. In particular, red bars identify peaks of documents which triggered high level alerts. On the right side of the figure, alerts with a medium level are reported and represented by yellow bars. Finally, blue bars show the average number of documents of the last 2 weeks for the combination under inspection. As can be noticed, the importance of an alert higher is directly proportional to the deviation of the peak from the 14-day average value.

Alert levels in MediSys are calculated by assuming a normal distribution of articles per category over time. Alert levels are high, if the number of articles found is at least three times the standard deviation over the last 14 days, while alert levels are medium, if the number of articles found is between twice and three times the considered standard deviation. As the total number of articles varies during the week (fewer articles on Sunday and Monday), a correction is applied to the documents' frequencies according to the day of the week [39].

To accomplish our comparison, a RSS tunnel feed has been set up between MediSys and UPHED. At present, UPHED processes only English-language documents. Our method triggers MediSys feeds every minute. MediSys sends through the tunnel documents which it categorizes as relevant to the medical domain. Currently, the documents arrive as plain text and UPHED applies entity extraction (Sect. 4.2). This is done in addition to the normal processing on the MediSys side, where running averages are monitored for all alerts, etc.

4.5.3 Comparative analysis of UPHED and MediSys

MediSys has proven to be useful and effective for finding documents from a large number of Web sources [39]. In contrast to MediSys, UPHED identifies *events* as clusters of documents associated with labels, i.e., a set of diseases and locations describing clusters. In an operational setting, we propose that after MediSys identifies documents for which alerts are generated, UPHED can deliver more information about the specific outbreak of the diseases reported in those documents, by aggregating documents into larger units than alerts, namely *events*. With the help of domain experts, we experiment with such a setting and present the results in the discussion that follows.

Our experiment was based on a collection of 13,076 news articles for a 2 months period (May–June 2011). Each document was selected if it contained specific keywords and their synonyms, e.g., *EHEC, Malaria, Measles, Rubella, Tuberculosis,*, etc. This collection was made available to us by *JRC*.

We computed MediSys alerts based on the procedure outlined in Sect. 4.5.2. In total 255 alerts were found. For each alert, we considered the set of news articles associated and computed the overlaps with each of the 12 events (i.e., clusters) of articles extracted by UPHED and reported in Table 4 and Table 5. Also, a domain expert judged the accuracy of the alert, by analyzing a sample of documents reported together to each alert, assigning it to one of four coded categories. The categories are outlined below. For more details, we report the entire evaluation in "Appendix." To summarize, each category was identified by a color:

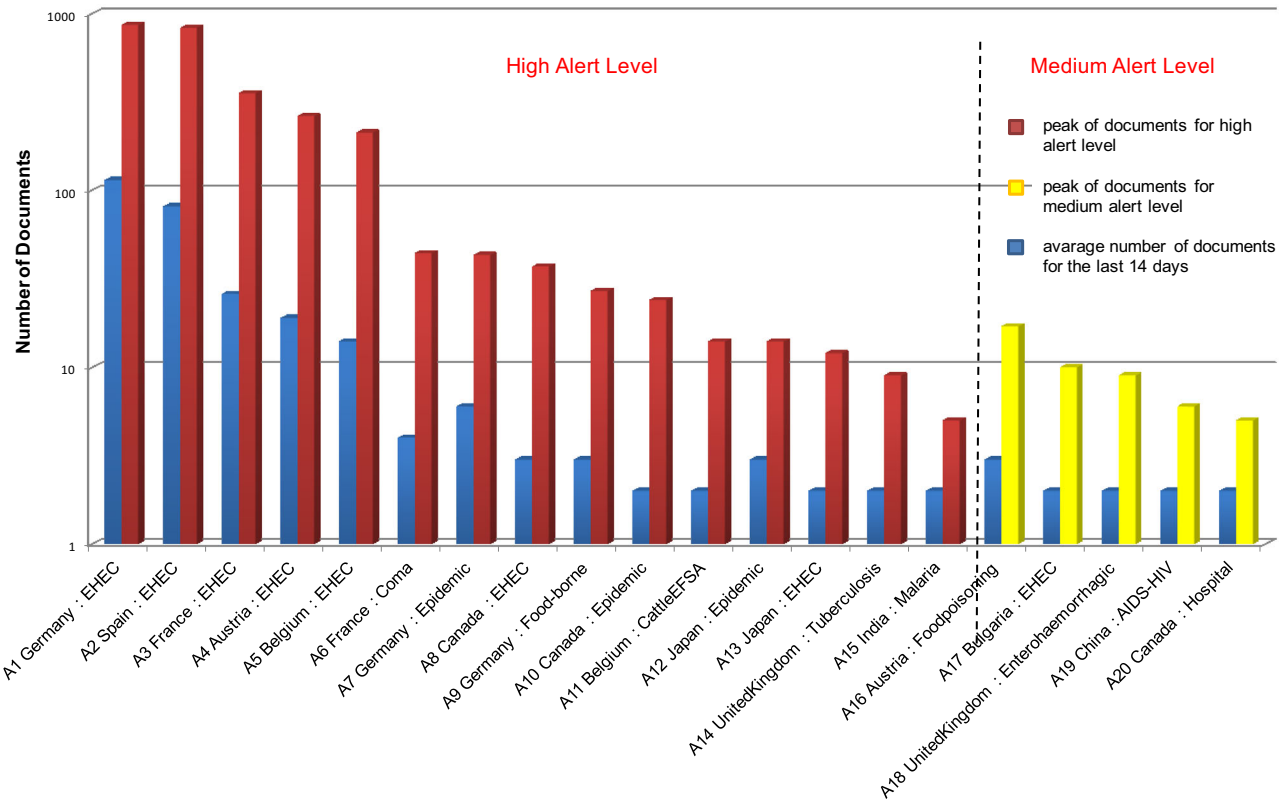

Fig. 6 Selection of 20 alert statistics from MediSys from beginning of May till end of June 2011. Term EFSA identifies the European Food Safety Authority

– Category Green: MediSys alert is related to one cluster (event) detected by UPHED.
– Category Orange: MediSys alert is appropriated and related to no cluster detected by UPHED.
– Category Yellow: MediSys alert reports disease outbreaks happening in locations different from the one shown in the alert or just inappropriate. This category contains cases where the disease name was mentioned but only to inform and to report an outbreak burst somewhere else in the world.
– Category Red: MediSys alert groups together articles not categorized in UPHED events.

The categories are described in detail below and a sample is shown in Table 6.

Category Green: detected by UPHED The documents contained in the specific alert are the same news articles related to one particular event detected by UPHED. Within the alert, articles mentioned one disease outbreak event. Important to be noticed is that it is always the case where one alert, falling in such category, contains only a small subset of articles associated to one UPHED event. The reason is that MediSys computes alerts based on documents which exactly match the keywords pre-specified within the system, and it is not a semantic approach. Furthermore, the alerts categorized here contain documents simply treating just one or few diseases mostly only in the location mentioned by the pre-defined alert; thus, it is reasonable that these documents fall also in the event relevant to that particular location and disease. On the other hand, since UPHED detects events and, in a simplistic way, co-occurrences of feature-instances (i.e., keywords of several type), thus, each event can correspond to multiple alerts. As an example, let the reader observe in Table 6 as multiple alerts in the first category often are related to just one event; this is the case of all alerts considering the country Germany and several diseases related to event E_1. In conclusion, 20 alerts fell into this category and additional 11 alerts showed a big subset of documents related to one cluster.

Category Orange: not detected by UPHED The MediSys alert was appropriate and was related to no events detected by UPHED. The reason why UPHED was not able to detect these events has to be retrieved in the explanation that there were few documents treating the particular outbreak. To let the reader better understand, we can consider the events reported in Tables 4 and 5. It is easy to identify that each event contains hundreds of document associated with, out of a total of less than 10,000 articles of the entire dataset. Thus, it is difficult for the UPHED algorithm to detect events asso-

Table 6 Sample of categorized alerts

Alerts	Description
Category Green: Cluster detected by UPHED	
Belgium:EscherichiacoliInfection	Related to event E_6, about documents treating the economical repercussion of the entire outbreak on the European market
Belgium:CattleEFSA	Related to event E_9, about peak of measles cases in all Europe
France:EscherichiacoliInfection, France:Coma	Related to E_4, about many cases of EHEC contagion in France with some case of coma
Germany:Epidemic, Germany:EscherichiacoliInfection, Germany:Fever, Germany:Food-borne	Related to E_1, about EHEC cases in Germany
India:Malaria, India:WHO	Related to E_{10}, about India's scientific community is all set to launch a research programme on how to combat vector-borne diseases, after many cases of malaria
Spain:EscherichiacoliInfection	Related to E_2, about EHEC contagion which was caused by contaminated Spanish vegetables. Since the entire Europe was talking about Spain's vegetable as the reason of EHEC outbreak, this alert contains also many other documents reporting EHEC cases in Europe, but the biggest overlap is with E_2
Category Orange: Events not detected by UPHED	
China:AIDS-HIV	Cases of AIDS-HIV in China. Too few documents to be detected by UPHED
Japan:CattleEFSA, Japan:Epidemic, Japan:EscherichiacoliInfection	EHEC cases of different nature compared to those occurred in Europe. The burst was caused by infected meat served in a restaurant chain. Few documents to be detected by UPHED
UnitedKingdom:Enterohaemorrhagic, UnitedKingdom:Haemorrhage, UnitedKingdom:Tuberculosis	Cases of tuberculosis in UK
Category Yellow: Alerts reporting disease outbreaks happening in different locations	
Austria:EscherichiacoliInfection, Austria:Foodpoisoning, Austria:WHO	Austria was always mentioned with other European countries for statistics on EHEC infection. In some case were reported Austrians visiting Germany and reporting EHEC
Bulgaria:EscherichiacoliInfection	Bulgarian authorities specially worried about EHEC in Germany and reporting cases in the continent
Canada:Epidemic, Canada:EscherichiacoliInfection, Canada:Hospital	In Canada, many news reporting the European EHEC outbreak. Also, Canada launched food inspection on food coming from EU
Germany:CattleEFSA, Germany:Coma, Germany:Diarrhoea, Germany:Communicabledisease, Germany:Foodpoisoning	Report on EHEC cases all over Europe

For each group of alerts a description is provided with a reference to the UPHED clusters when available. Mentioned clusters are related to events detected in Sect. 4.4. All alerts reported in Fig. 6 are presented

ciated with only tens of documents. Of course, it is always important the proportion with respect to the entire set. On the other hand, MediSys was capable to trigger an alert since it detects a sudden increase in the number of articles for a given *category-country*, by comparing the statistics for the last day with the 2-weeks rolling average. Thus, for MediSys it is sufficient a specific increasing of the standard deviation of the particular combination *category-country* under observation to generate an alert. In conclusion, 11 alerts fell into this category.

Category Yellow: alerts reporting disease outbreaks happening in different locations The third group consists of MediSys alerts reporting disease outbreaks happening in locations different from the one shown in the alert or just inappropriate. Articles falling in this category truly contain the disease name mentioned in the alert, but such disease is not related to the *country* in the alert. Also, many of these documents contain the *country* of the alert only to identify the location where the article was published or where the journalist had her newspaper-headquarter. Most of these articles described the situation in many other countries. As an example, let us consider in Table 6 the alert mentioning Canada and several diseases. In such a case, Canadian journalists described the outbreak of EHEC happening in the European countries. Thus, it is often the case where resources associated with alerts of this category are related to multiple UPHED events. Furthermore, since these documents include many countries (locations), then they constitute elements of noise since one article can contribute in increasing the counter of different alerts. This is because alerts will

match all the countries and diseases reported in them. In summary, this category contains cases where the disease name was mentioned but only to inform and report an outbreak burst somewhere else in the globe. In conclusion, 116 alerts fell into this category.

Category Red: MediSys alert with no UPHED overlap The last group consists of alerts referring to articles not categorized in UPHED clusters. These articles did not exceed the imposed threshold τ_p to be associated with an event (recall Sect. 4.4). Often was the case that many documents cited diseases in contexts unrelated to epidemics and outbreaks. The remaining alerts fell into this category.

5 Conclusions

In this research, we observed and exploited two main characteristics of text documents, i.e., their content and their timestamps, to build an approach for clustering articles in events with an unsupervised learner.

Both the contents and the time information of articles are modeled explicitly and effectively. Particularly, the model of timestamps works like auto-adaptive sliding windows on a time line, which overcomes the inflexible usage of time windows in traditional retrospective event detection algorithms.

Also, our method incorporates two main techniques: the burst function analysis and the entity-centric feature representation. The burst function analysis and entity-centric feature representation were combined in a generative model and forms the basis of our UPHED algorithm. The event model was refined for representing periodic, non-burst features with the Cauchy–Lorentz distribution. Our evaluations showed that better sampling is reached by using such a distribution, resulting in a more efficient algorithm, which is also easy to implement, in practice.

Furthermore, we proved the goodness of our theoretical study adapting our unsupervised learner to the public health domain, extracting a particular instance of events within the context of Epidemic Intelligence. More specifically, the adaptations included the consideration of domain specific features that allow detecting public health-related events.

For the specific domain we considered, no annotated data set were available; thus, we performed our analysis on real-world data sets. We demonstrated the effectiveness of our approach in detecting a recent outbreak of enterohemorrhagic Escherichia coli (EHEC), which occurred in Northern Germany in May of 2011.

Finally, going beyond the EHEC case study, we conducted an extensive comparative analysis of our UPHED algorithm with the well established rule-based system of MediSys by analyzing the *alerts* generated by MediSys versus the *events* detected by our UPHED method. We conclude that the combination of the two initially independent systems, MediSys and UPHED, can lead to a stronger application offering users complementary functionalities.

In conclusion, MediSys computes statistics based on exact matching of keywords in different languages. MediSys is not a semantic approach and is not able to detect alerts with several locations or diseases representing the same medical burst or outbreak. Also, since MediSys does not explicitly select for outbreaks, but for mentions of diseases in any context, it is expectable that many documents might cite diseases in contexts unrelated to epidemics and outbreaks.

On the contrary, UPHED semantically recognizes equal public health events and it clusters together documents treating the same topic. Furthermore, our approach computes medical events based on multiple diseases and locations at the same time, as can be observed in Tables 4 and 5 by the labels extracted to describe each event.

The combination of the two initially independent systems, MediSys and UPHED, can lead to a stronger application offering users complementary functionalities. For disease outbreaks, which are covered by both systems, the combination can lead to additional advantages overtaking the drawbacks of both:

1. UPHED is computationally heavier and might be applied to the document collection pre-filtered by MediSys.
2. The medical event extraction by UPHED might act as a filter for users to identify only disease outbreak reports.
3. UPHED is not a pattern matching-based approach to extract events; thus it is not language dependent. Thus, UPHED might benefit from the wider categorization of news items by MediSys as useful tool for the analysis performed by our method.

These issues are to be tackled in future work. The entities have been extracted using two different named entity recognition tools: UMLS MetaMap and OpenCalais. In future work, we could also explore different tools for entities extraction in combination or in parallel to the two suggested ones: one possibility can be using ScispaCy [7] for processing biomedical, scientific, or clinical documents. Also, other sources of data containing labels to validate clustering results will be considered. Possibly, one source could be the WHO's update headings which are entries that group-related outbreak reports together. According to observations so far illustrated, we think that MediSys can be complemented with our UPHED to provide a better medical information system able to rely on a semantic detection approach.

As a final remark, our approach has been designed to address the limitations of existing public health event detection systems as pointed out by health officials. As a result, our algorithm allows events that are both rare and reoccurring

[7] https://allenai.github.io/scispacy/.

to be detected. The implications for such work is that public health officials can rely upon alternative sources of corroborating information about public health events—an important aspect in event-based Epidemic Intelligence—since diverse information sources can offer an additional means of mitigating the impact of potential health threats.

Acknowledgements The work was partially funded by the European Commission for the eXplainable Artificial Intelligence in healthcare Management (xAIM) project, agreement No INEA/CEF/ICT/A2020/2276680.

Appendix

In this section, we provide details about the evaluation on *Alerts* extracted from Medisys. For each alert we consider all the associated documents; then, for each *Event* detected by UPHED and reported in Tables 4 and 5, we consider all the associated documents and we compute the overlap of documents between alert-document-set and event-document-set. In the following table, on the first column we report alerts. Each alert-row intersects the columns where events are presented. The intersection identifies the number of documents in common (i.e., overlaps) between the corresponding alert and the respective event. For each overlap, a human expert analyzed a sample of articles and judged the accuracy of the alert. Thus, each alert was assigned to one of the four categories presented in Sect. 4.5.3 and hereafter identified by four different colors.

In the following Table are reported the **Level** of each alert, i.e., *High* or *Medium*, the **Date** of the alert's burst, the **Category** it belongs to, and the overlaps between alert's documents with the documents associated with each event. To summarize, each category was identified by a color:

1. Category is identified in green if the Medisys alert is related to one cluster detected by UPHED.
2. Category is identified in orange if the Medisys alert is appropriated and related to NO one cluster detected by UPHED.
3. Category is identified in yellow if Medisys alert reports disease outbreaks happening in locations different from the one shown in the alert or just inappropriate. This category contains cases where the disease name was mentioned but only to inform and to report an outbreak burst somewhere else in the world.
4. Category is identified in light red if Medisys alert groups together articles not written in English-language or without overlap with UPHED events.

Alerts	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	Date	Level
Legenda														
Related to one cluster detected by UPHED														
Related to one cluster NOT detected by UPHED														
Related to several outbreaks in several countries														
No English documents or No overlaps with clusters by UPHED														
Afghanistan:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
AmericanSamoa:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	07.06.2011	Medium
Australia:disease	0	0	0	0	0	0	0	0	2	0	1	8	03.06.2011	High
Austria:Diarrhoea	1	0	0	1	0	0	0	0	0	0	0	0	02.06.2011	High
Austria:ECDC	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	Medium
Austria:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	Medium
Austria:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Austria:EscherichiacoliInfection	20	16	1	10	17	3	0	0	6	0	0	0	31.05.2011	High
Austria:Fever	0	0	0	0	0	0	0	0	0	0	0	0	04.06.2011	High
Austria:Foodpoisoning	13	13	0	5	12	0	0	0	7	0	0	0	02.06.2011	Medium
Austria:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Austria:WHO	88	77	2	34	77	0	9	0	41	0	6	0	31.05.2011	High
Azerbaijan:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	07.06.2011	High
Belarus:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	21.06.2011	High
Belgium:CattleEFSA	0	0	0	0	0	0	0	0	3	0	0	0	16.06.2011	High
Belgium:Communicabledisease	1	1	0	0	0	1	0	0	2	0	0	0	07.06.2011	High
Belgium:Diarrhoea	0	0	0	0	0	0	0	0	0	0	0	0	25.06.2011	Medium
Belgium:ECDC	2	2	0	0	2	2	0	0	2	0	0	0	02.06.2011	Medium
Belgium:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	Medium
Belgium:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Belgium:EscherichiacoliInfection	5	4	0	0	1	22	1	0	2	0	0	0	31.05.2011	High
Belgium:Food-borne	17	16	16	0	1	17	0	0	3	0	0	0	03.06.2011	Medium
Belgium:Foodpoisoning	5	5	0	0	0	5	0	0	2	0	0	0	16.06.2011	High
Belgium:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High
Belgium:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	16.06.2011	High
Belgium:Pharmaceuticals	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	High
Belgium:WHO	9	8	9	0	1	9	0	0	4	0	0	1	02.06.2011	High
Brazil:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	25.06.2011	High
Bulgaria:EscherichiacoliInfection	9	5	3	0	4	0	3	0	0	0	0	0	31.05.2011	Medium
Canada:Epidemic	1	1	1	0	0	0	0	0	1	0	0	0	07.06.2011	High
Canada:EscherichiacoliInfection	15	12	6	0	5	0	3	0	7	0	1	0	02.06.2011	High
Canada:Hospital	10	4	8	2	4	0	0	0	0	0	0	0	16.06.2011	Medium
Chile:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High
China:AIDS-HIV	0	0	0	0	0	0	0	0	0	1	10	0	03.06.2011	Medium
China:EscherichiacoliInfection	3	2	0	0	2	0	0	0	2	0	0	0	02.06.2011	High
China:Hospital	1	0	0	0	0	0	0	0	0	0	2	0	04.06.2011	Medium
China:Pharmaceuticals	4	1	0	1	2	0	0	0	2	0	3	2	03.06.2011	High
China:Tuberculosis	0	0	0	0	0	0	0	0	0	1	8	0	19.05.2011	High
Croatia:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	16.06.2011	Medium
CzechRepublic:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
CzechRepublic:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	28.06.2011	Medium
CzechRepublic:EscherichiacoliInfection	2	1	0	0	1	0	0	0	2	0	0	0	28.06.2011	High
CzechRepublic:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	16.06.2011	High
CzechRepublic:WHO	0	0	0	0	0	0	0	0	0	0	0	0	04.06.2011	High
Denmark:Diarrhoea	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Denmark:ECDC	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	Medium
Denmark:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High
Denmark:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Denmark:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	28.06.2011	High
Denmark:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	Medium

Alerts	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	Date	Level
Denmark:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Denmark:IntensiveCareUnit	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	Medium
Denmark:WHO	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Egypt:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	30.06.2011	High
Estonia:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	07.06.2011	High
Finland:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	10.06.2011	Medium
Finland:EscherichiacoliInfection	3	0	0	0	0	0	0	0	1	0	0	0	28.06.2011	Medium
Finland:WHO	0	0	0	0	0	0	0	0	0	0	0	0	04.06.2011	Medium
France:CattleEFSA	0	0	0	0	1	0	1	6	6	0	0	0	17.06.2011	Medium
France:Coma	2	0	0	8	0	0	2	0	0	0	0	0	18.06.2011	High
France:Diarrhoea	0	0	0	0	0	0	0	0	0	0	0	0	16.06.2011	High
France:ECDC	0	0	0	0	0	0	0	0	0	0	0	0	30.06.2011	High
France:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	16.06.2011	High
France:EscherichiacoliInfection	2	1	1	9	1	1	2	0	0	0	0	0	16.06.2011	High
France:WHO	0	0	0	0	0	0	2	1	2	1	0	0	02.06.2011	High
Georgia:EscherichiacoliInfection	1	1	0	0	0	0	0	0	1	0	0	0	08.06.2011	Medium
Germany:CattleEFSA	23	15	2	1	4	0	0	0	13	0	0	0	16.06.2011	High
Germany:Coma	23	12	0	4	12	1	1	0	7	0	0	0	17.06.2011	High
Germany:Communicabledisease	49	24	13	5	21	1	3	2	37	0	4	4	23.06.2011	High
Germany:Diarrhoea	75	47	14	5	39	9	0	0	22	0	0	0	02.06.2011	High
Germany:ECDC	28	23	1	0	19	2	0	0	7	0	0	0	31.05.2011	High
Germany:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Germany:Epidemic	13	1	5	0	3	0	0	0	3	0	0	0	31.05.2011	High
Germany:EscherichiacoliInfection	314	54	10	3	30	5	3	1	26	0	3	0	31.05.2011	High
Germany:Fever	45	0	10	0	11	0	0	0	9	0	0	0	31.05.2011	High
Germany:Food-borne	73	2	0	0	1	0	0	1	9	0	0	0	02.06.2011	High
Germany:Foodpoisoning	37	37	1	10	30	4	0	0	12	0	0	0	31.05.2011	High
Germany:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	Medium
Germany:Hospital	52	8	7	3	15	0	0	1	21	0	1	0	16.06.2011	High
Germany:IntensiveCareUnit	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Germany:MRSA	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	High
Germany:Pathogens	11	8	0	1	5	0	0	0	6	0	0	0	01.06.2011	Medium
Germany:Pharmaceuticals	7	4	0	1	4	0	0	0	6	0	2	2	03.06.2011	High
Germany:Travel	6	2	0	1	1	0	0	0	4	0	0	0	03.06.2011	Medium
Germany:TravelHealth	16	11	0	2	3	0	0	0	8	0	0	0	02.06.2011	High
Germany:UnknownDisease	0	0	0	0	0	0	0	0	0	0	0	0	04.06.2011	High
Germany:WHO	163	112	34	17	105	9	9	2	55	0	7	1	31.05.2011	High
Ghana:Malaria	0	0	0	0	0	0	0	0	0	0	1	0	21.06.2011	Medium
Greece:Epidemic	0	0	0	0	0	0	0	0	1	0	0	0	07.06.2011	High
Greece:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	High
HongKong:EscherichiacoliInfection	4	1	0	0	0	0	0	0	0	0	0	0	03.06.2011	Medium
Hungary:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High
Hungary:WHO	8	0	8	0	8	0	0	0	1	0	0	0	04.06.2011	Medium
India:EscherichiacoliInfection	1	0	0	0	0	0	0	0	2	1	0	0	25.06.2011	Medium
India:Malaria	0	0	0	0	0	0	0	0	0	6	3	0	09.06.2011	High
India:WHO	0	0	0	0	0	0	0	0	0	3	2	0	02.06.2011	Medium
Ireland:EscherichiacoliInfection	14	9	9	0	3	9	2	0	6	0	2	2	25.06.2011	High
Israel:EscherichiacoliInfection	3	3	2	0	2	0	0	0	0	0	0	0	03.06.2011	Medium
Italy:Diarrhoea	12	2	10	0	12	0	0	0	2	0	0	0	31.05.2011	Medium
Italy:Epidemic	12	2	10	0	12	0	0	0	2	0	0	0	31.05.2011	Medium
Italy:EscherichiacoliInfection	22	12	10	0	20	8	9	0	2	0	0	0	31.05.2011	High
Italy:Fever	2	1	1	0	1	0	15	0	1	0	0	0	04.06.2011	High
Italy:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	16.06.2011	High
Italy:WHO	2	2	1	0	1	0	13	0	2	0	0	0	02.06.2011	High
Japan:CattleEFSA	1	1	0	0	0	0	0	0	0	0	0	0	03.06.2011	High
Japan:Epidemic	1	1	0	1	1	0	0	0	0	0	0	0	07.06.2011	High
Japan:EscherichiacoliInfection	20	15	0	2	13	5	2	0	11	0	0	0	31.05.2011	High

Alerts	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	Date	Level
Kazakhstan:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	10.06.2011	Medium
Kenya:Malaria	0	0	0	0	0	0	0	0	0	0	0	0	17.06.2011	Medium
Latvia:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	Medium
Lebanon:EscherichiacoliInfection	9	0	9	0	9	0	0	0	1	0	0	0	04.06.2011	Medium
Lebanon:WHO	8	0	8	0	8	0	0	0	0	0	0	0	04.06.2011	High
Libya:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	22.06.2011	Medium
Lithuania:EscherichiacoliInfection	7	7	7	7	0	0	0	0	0	0	0	0	28.06.2011	High
Luxemburg:ECDC	4	3	0	0	3	3	0	0	0	0	0	0	07.06.2011	Medium
Luxemburg:Epidemic	0	0	0	0	0	0	0	0	1	0	0	0	07.06.2011	High
Luxemburg:EscherichiacoliInfection	27	20	5	0	14	16	0	4	25	0	0	0	28.06.2011	High
Luxemburg:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	07.06.2011	Medium
Luxemburg:WHO	3	0	1	0	0	3	0	0	4	0	0	0	03.06.2011	Medium
Mexico:EscherichiacoliInfection	5	5	0	0	5	0	0	0	5	0	0	0	01.06.2011	High
Mexico:WHO	5	5	0	0	5	0	0	0	5	2	2	0	04.06.2011	High
Netherlands:Diarrhoea	5	5	0	0	5	0	0	0	0	0	0	0	02.06.2011	High
Netherlands:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Netherlands:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	10.06.2011	Medium
Netherlands:EscherichiacoliInfection	26	26	0	6	24	0	0	0	7	0	0	0	31.05.2011	High
Netherlands:Foodpoisoning	9	8	0	5	7	0	0	0	3	0	0	0	16.06.2011	High
Netherlands:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High
Netherlands:Hospital	2	0	0	0	0	0	0	0	0	0	0	0	16.06.2011	High
Netherlands:WHO	5	5	0	0	5	0	0	0	0	0	0	0	31.05.2011	High
Norway:Diarrhoea	1	1	1	0	1	0	0	0	0	0	0	0	03.06.2011	Medium
Norway:ECDC	5	4	0	2	1	1	0	0	0	0	0	0	03.06.2011	Medium
Norway:Epidemic	2	2	0	2	2	0	0	0	0	0	0	0	03.06.2011	High
Norway:EscherichiacoliInfection	43	37	1	5	36	0	4	0	25	0	4	0	31.05.2011	High
Norway:Foodpoisoning	25	23	0	7	23	0	0	0	16	0	0	0	02.06.2011	High
Norway:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	04.06.2011	High
Norway:Pathogens	15	12	0	8	12	0	0	0	6	0	0	0	04.06.2011	Medium
Norway:WHO	59	49	2	14	49	0	6	0	32	0	6	0	02.06.2011	High
Pakistan:EscherichiacoliInfection	0	0	0	0	0	0	0	0	1	0	0	0	08.06.2011	Medium
Poland:ECDC	3	2	0	2	1	1	0	0	0	0	0	0	07.06.2011	High
Poland:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	Medium
Poland:EscherichiacoliInfection	13	8	4	8	2	2	2	2	3	0	2	0	28.06.2011	Medium
Poland:Foodpoisoning	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	Medium
Portugal:EscherichiacoliInfection	1	1	1	0	1	0	0	0	0	0	0	0	31.05.2011	Medium
Romania:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	07.06.2011	High
Romania:EscherichiacoliInfection	4	2	4	0	0	1	0	0	3	0	0	0	07.06.2011	Medium
Russia:ECDC	0	0	0	0	0	0	0	0	12	0	0	0	02.06.2011	High
Russia:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	Medium
Russia:Epidemic	5	1	5	0	2	0	0	0	20	0	0	0	02.06.2011	High
Russia:EscherichiacoliInfection	48	41	66	6	21	7	4	2	35	0	2	0	31.05.2011	High
Russia:Fever	0	0	10	0	8	0	0	0	3	0	0	0	04.06.2011	High
Russia:Food-borne	13	12	13	0	0	8	3	0	1	0	0	0	03.06.2011	High
Russia:Foodpoisoning	5	2	5	0	0	0	1	0	3	0	0	0	02.06.2011	High
Russia:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Russia:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	Medium
Russia:Pharmaceuticals	0	0	0	0	0	0	0	0	1	0	0	0	03.06.2011	High
Russia:WHO	29	22	29	0	7	9	8	0	15	0	4	1	02.06.2011	High
Samoa:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	07.06.2011	Medium
SaudiArabia:WHO	1	0	1	0	0	0	0	0	0	0	0	0	10.06.2011	Medium
Serbia:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High
Slovakia:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High
Slovakia:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	09.06.2011	High
Slovakia:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Slovakia:WHO	0	0	0	0	0	0	0	0	0	0	0	0	09.06.2011	Medium
Slovenia:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High

Alerts	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	Date	Level
Slovenia:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	08.06.2011	Medium
Slovenia:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	28.06.2011	Medium
SouthKorea:EscherichiacoliInfection	7	7	0	0	5	0	0	0	7	0	1	0	28.06.2011	Medium
Spain:CattleEFSA	1	4	1	0	3	0	0	0	5	0	0	0	02.06.2011	Medium
Spain:Communicabledisease	25	25	2	5	16	1	2	0	19	0	0	0	03.06.2011	High
Spain:Diarrhoea	11	11	0	1	9	0	0	0	4	0	0	0	02.06.2011	High
Spain:ECDC	22	22	0	0	19	2	0	0	5	0	0	0	31.05.2011	High
Spain:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Spain:Epidemic	1	1	1	0	0	0	0	0	0	0	0	0	31.05.2011	High
Spain:EscherichiacoliInfection	187	180	21	21	118	15	2	0	68	0	2	0	31.05.2011	High
Spain:Fever	20	6	10	0	16	0	0	0	7	0	0	0	31.05.2011	High
Spain:Food-borne	23	21	12	0	8	8	3	0	9	0	0	0	02.06.2011	High
Spain:Foodpoisoning	15	15	0	5	12	2	0	0	4	0	0	0	01.06.2011	High
Spain:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Spain:Hospital	2	2	0	0	0	0	0	0	2	0	0	0	31.05.2011	High
Spain:IntensiveCareUnit	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Spain:Pathogens	9	9	0	1	4	1	0	0	6	0	0	0	01.06.2011	Medium
Spain:Pharmaceuticals	6	6	0	1	2	0	0	0	6	0	0	0	02.06.2011	High
Spain:TravelHealth	10	8	0	2	3	0	0	0	3	0	0	0	02.06.2011	High
Spain:WHO	116	107	27	15	80	8	9	0	47	1	5	3	28.06.2011	High
Sweden:Antimicrobialresist	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Sweden:CattleEFSA	1	0	0	1	1	0	0	1	2	0	0	0	16.06.2011	High
Sweden:Communicabledisease	19	13	2	5	19	0	0	0	8	0	0	0	03.06.2011	Medium
Sweden:Diarrhoea	19	14	0	1	16	0	0	0	4	0	0	0	02.06.2011	High
Sweden:ECDC	15	14	0	0	14	2	0	0	3	0	0	0	31.05.2011	High
Sweden:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Sweden:Epidemic	6	3	2	1	6	0	0	0	2	0	0	0	31.05.2011	High
Sweden:EscherichiacoliInfection	170	129	8	21	145	18	2	1	10	0	3	0	31.05.2011	High
Sweden:Fever	30	6	10	0	23	0	0	0	11	0	0	0	03.06.2011	High
Sweden:Food-borne	22	18	9	1	11	9	0	0	8	0	0	0	03.06.2011	High
Sweden:Foodpoisoning	12	12	0	4	11	0	0	0	4	0	0	0	01.06.2011	High
Sweden:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Sweden:Hospital	14	1	4	2	13	0	0	1	10	0	1	0	16.06.2011	Medium
Sweden:Pharmaceuticals	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	High
Sweden:TravelHealth	5	3	0	0	3	0	0	0	3	0	0	0	02.06.2011	Medium
Sweden:WHO	122	84	16	16	106	5	5	2	42	0	6	0	28.06.2011	Medium
Switzerland:Communicabledisease	2	2	0	2	2	0	0	0	0	0	0	0	03.06.2011	High
Switzerland:Diarrhoea	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Switzerland:ECDC	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
Switzerland:Epidemic	2	2	0	2	2	0	0	0	0	0	0	0	02.06.2011	High
Switzerland:EscherichiacoliInfection	60	54	3	24	52	6	2	0	15	0	2	0	31.05.2011	High
Switzerland:Fever	0	0	0	0	0	0	0	0	0	0	0	0	04.06.2011	High
Switzerland:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	01.06.2011	High
Switzerland:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Switzerland:Malaria	0	0	0	0	0	0	0	0	0	1	0	0	31.05.2011	Medium
Switzerland:WHO	88	77	2	34	77	0	6	0	37	4	6	0	02.06.2011	High
Syria:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	04.06.2011	Medium
Thailand:Epidemic	0	0	0	0	0	0	0	0	0	0	0	0	11.06.2011	Medium
Thailand:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	11.06.2011	High
Turkey:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	High
Turkey:WHO	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	High
Ukraine:EscherichiacoliInfection	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
UnitedKingdom:Antimicrobialresist	0	0	0	0	0	0	0	0	2	0	2	2	03.06.2011	Medium
UnitedKingdom:CattleEFSA	5	5	0	0	2	0	0	0	4	0	0	0	22.06.2011	Medium
UnitedKingdom:Communicabledisease	15	9	6	1	5	0	0	1	19	0	4	4	03.06.2011	High
UnitedKingdom:Diarrhoea	17	16	0	0	12	0	0	0	5	0	0	0	02.06.2011	High
UnitedKingdom:ECDC	13	12	0	0	10	0	0	0	3	0	0	0	31.05.2011	High

Alerts	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	Date	Level
UnitedKingdom:Enterohaemorrhagic	0	0	0	0	0	0	0	0	0	0	0	0	31.05.2011	Medium
UnitedKingdom:Epidemic	1	1	1	0	0	0	0	0	0	0	0	0	31.05.2011	High
UnitedKingdom:EscherichiacoliInfection	65	40	2	1	44	2	2	1	12	0	3	0	25.06.2011	High
UnitedKingdom:Fever	4	0	2	0	0	0	0	0	2	0	0	0	03.06.2011	Medium
UnitedKingdom:Food-borne	5	2	0	1	3	1	0	1	2	0	0	0	07.06.2011	Medium
UnitedKingdom:Foodpoisoning	16	16	1	0	11	4	0	0	12	0	0	0	02.06.2011	Medium
UnitedKingdom:Haemorrhage	0	0	0	0	0	0	0	0	0	0	0	0	02.06.2011	High
UnitedKingdom:Hospital	4	2	0	0	2	0	0	0	4	0	0	0	31.05.2011	Medium
UnitedKingdom:Malaria	0	0	0	0	0	0	0	3	8	0	0	1	29.06.2011	Medium
UnitedKingdom:Pathogens	7	7	1	0	6	0	0	0	7	0	0	0	02.06.2011	High
UnitedKingdom:Pharmaceuticals	1	1	0	0	0	0	0	0	3	0	2	2	03.06.2011	High
UnitedKingdom:TravelHealth	4	2	0	0	1	0	0	0	1	0	0	0	03.06.2011	High
UnitedKingdom:tropicalmedicine	22	21	2	0	16	0	0	0	19	0	0	0	02.06.2011	Medium
UnitedKingdom:Tuberculosis	0	0	0	0	0	0	0	0	2	1	0	1	29.06.2011	High
UnitedKingdom:WHO	42	33	2	1	35	0	2	1	22	0	3	0	28.06.2011	High
USA:AIDS-HIV	0	0	0	0	0	0	0	0	0	0	2	0	09.06.2011	High
USA:CattleEFSA	9	10	2	0	0	0	0	0	6	0	0	0	03.06.2011	Medium
USA:Coma	9	5	0	2	5	0	1	0	5	0	0	0	17.06.2011	Medium
USA:Communicabledisease	13	10	2	5	6	1	0	1	7	0	2	0	03.06.2011	High
USA:Diarrhoea	34	12	11	6	19	0	0	1	1	0	1	0	03.06.2011	High
USA:ECDC	0	0	0	0	0	0	0	0	0	0	0	0	03.06.2011	Medium
USA:Epidemic	7	3	4	0	3	0	0	0	1	0	0	0	02.06.2011	High
USA:EscherichiacoliInfection	172	114	31	36	87	16	2	3	2	0	3	1	28.06.2011	Medium
USA:Fever	21	4	10	2	14	0	0	2	2	0	0	0	03.06.2011	High
USA:Food-borne	2	1	0	0	0	0	0	1	7	0	0	0	07.06.2011	High
USA:Foodpoisoning	10	10	1	0	7	2	0	0	8	0	0	0	02.06.2011	Medium
USA:Hospital	0	0	0	0	0	0	0	0	0	0	0	0	25.06.2011	Medium
USA:Malaria	0	0	0	0	0	0	0	0	0	2	2	10	26.05.2011	High
USA:Pathogens	24	19	2	7	15	0	0	0	2	0	1	0	02.06.2011	High
USA:Pharmaceuticals	2	0	0	0	0	0	0	0	3	0	0	0	23.06.2011	High
USA:Salmonellosis	2	0	0	0	0	0	0	0	5	0	0	0	08.06.2011	Medium
USA:Tuberculosis	0	0	0	0	0	0	0	0	12	1	4	0	22.06.2011	Medium
USA:WHO	111	86	26	30	79	0	0	0	55	11	13	1	28.06.2011	Medium
Yemen:EscherichiacoliInfection	1	1	0	0	1	0	0	0	0	0	0	0	04.06.2011	High

References

1. Al Tamime, R., Giordano, R., Hall, W.: Observing burstiness in wikipedia articles during new disease outbreaks. In: Proceedings of the 10th ACM Conference on Web Science, WebSci '18, pp. 117–126. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3201064.3201080
2. Arsevska, E., Valentin, S., Rabatel, J., de Goër de Hervé, J., Falala, S., Lancelot, R., Roche, M.: Web monitoring of emerging animal infectious diseases integrated in the french animal health epidemic intelligence system. PLOS ONE **13**(8), 1–25 (2018). https://doi.org/10.1371/journal.pone.0199960
3. Banerjee, A., Dhillon, I.S., Ghosh, J., Sra, S.: Generative model-based clustering of directional data. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2003)
4. Brants, T., Chen, F., Farahat, A.: A System for new event detection. In: In SIGIR, pp. 330–337. ACM, New York, NY, USA (2003). https://doi.org/10.1145/860435.860495
5. Burchard, L., Schroeder, D.T., Becker, S., Langguth, J.: Resource efficient algorithms for message sampling in online social networks. In: 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–8 (2020). https://doi.org/10.1109/SNAMS52053.2020.9336530
6. Ceroni, A., Fisichella, M.: Towards an entity-based automatic event validation. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C., de Jong, F., Radinsky, K., Hofmann, K. (eds.) Advances in Information Retrieval, pp. 605–611. Springer, Cham (2014)
7. Ceroni, A., Gadiraju, U., Fisichella, M.: Justevents: a crowdsourced corpus for event validation with strict temporal constraints. In: Jose, J.M., Hauff, C., Altıngovde, I.S., Song, D., Albakour, D., Watt, S., Tait, J. (eds.) Advances in Information Retrieval, pp. 484–492. Springer, Cham (2017)
8. Ceroni, A., Gadiraju, U., Matschke, J., Wingert, S., Fisichella, M.: Where the event lies: predicting event occurrence in textual documents. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, p. 1157–1160. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2911451.2911452
9. Ceroni, A., Gadiraju, U.K., Fisichella, M.: Improving event detection by automatically assessing validity of event occurrence in text. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, pp. 1815–1818. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806624
10. Chen, N., Zhong, Z., Pang, J.: An exploratory study of Covid-19 information on twitter in the greater region. Big Data Cogn. Comput. **5**(1), 5 (2021). https://doi.org/10.3390/bdcc5010005
11. Cinti, S., Huff, A.G., Breit, N., Allen, T., Whiting, K., Kiley, C.: Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources. Interdiscip. Perspect. Infect. Dis. **2016**, 5080746 (2016). https://doi.org/10.1155/2016/5080746
12. Conway, M., Collier, N., Doan, S.: Using hedges to enhance a disease outbreak report text mining system. In: BioNLP '09: Proceedings of the Workshop on BioNLP, pp. 142–143. Association for Computational Linguistics, Morristown, NJ, USA (2009)
13. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**(1), 1–38 (1977)
14. Detection, T., project, T.T.: https://www.nist.gov/publications/topic-detection-and-tracking-evaluation-overview
15. Doan, S., Kawazoe, A., Conway, M., Collier, N.: Towards role-based filtering of disease outbreak reports. J. Biomed. Inform. (2008). https://doi.org/10.1016/j.jbi.2008.12.009
16. Fisichella, M., Ceroni, A.: Event detection in Wikipedia edit history improved by documents web based automatic assessment. Big Data Cogn. Comput. **5**(3), 34 (2021). https://doi.org/10.3390/bdcc5030034
17. Fisichella, M., Stewart, A., Cuzzocrea, A., Denecke, K.: Detecting health events on the social web to enable epidemic intelligence. In: SPIRE, pp. 87–103 (2011)
18. Fung, G.P.C., Yu, J.X., Yu, P.S., Lu, H.: Parameter free bursty events detection in text streams. In: VLDB '05: Proceedings of the 31st international conference on Very large data bases, pp. 181–192. VLDB Endowment (2005)
19. Hartley, D., et al.: The landscape of international event-based biosurveillance. Emerg. Health Threats **3**, 7096 (2010)
20. He, Q., Chang, K., Lim, E.P.: Analyzing feature trajectories for event detection. In: SIGIR, pp. 207–214 (2007)
21. He, Q., Chang, K., Lim, E.P.: Using burstiness to improve clustering of topics in news streams. In: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, ICDM '07, pp. 493–498. IEEE Computer Society, Washington, DC, USA (2007). https://doi.org/10.1109/ICDM.2007.17
22. He, Q., Chang, K., Lim, E.P., Banerjee, A.: Keep it simple with time: a reexamination of probabilistic topic detection models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(10), 1795–1808 (2010). https://doi.org/10.1109/TPAMI.2009.203
23. He, Q., Chang, K., Lim, E.P., Zhang, J.: Bursty feature representation for clustering text streams. In: Proceedings of the 2007 SIAM International Conference on Data Mining, pp. 491–496 (2007)
24. Hoffart, J., Suchanek, F., Berberich, K., Weikum, G.: Yago2: a spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. **194**, 28–61 (2012)
25. Hofmann, T.: Probabilistic latent semantic analysis. In: UAI, pp. 289–296 (1999)
26. Keller, M., Blench, M., Tolentino, H., et al.: Use of unstructured event-based reports for global infectious disease surveillance. Emerg. Infect. Dis. **15**(5), 689 (2009)
27. Kuzey, E., Vreeken, J., Weikum, G.: A fresh look on knowledge bases: distilling named events from news. In: CIKM '14 (2014)
28. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Daniel: Language independent character-based news surveillance. In: Isahara, H., Kanzaki, K. (eds.) Advances in Natural Language Processing, pp. 64–75. Springer, Berlin (2012)
29. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Multilingual event extraction for epidemic detection. Artif. Intell. Med. **65**(2), 131–143 (2015). https://doi.org/10.1016/j.artmed.2015.06.005
30. Li, Z., Wang, B., Li, M., Ma, W.Y.: A probabilistic model for retrospective news event detection. In: SIGIR (2005)
31. Linge, J., Steinberger, R., Fuart, F., Bucci, S., Belyaeva, J., Gemo, M.: Medisys: medical information system. In: Asimakopoulou, Eleana, Bessis, Nik (eds.) Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks, pp. 131–142. IGI Global, Hershey (2010)
32. Linge, J.P., Mantero, J., Fuart, F., Belyaeva, J., Atkinson, M., van der Goot, E.: Tracking media reports on the shiga toxin-producing *Escherichia coli*. In: In Proceedings of the Electronic Healthcare International Conference (eHealth). Springer (2011)
33. Mutuvi, S., Boros, E., Doucet, A., Jatowt, A., Lejeune, G., Odeo, M.: Multilingual epidemiological text classification: a comparative study. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6172–6183. International Committee on Computational Linguistics, Barcelona, Spain (2020). https://doi.org/10.18653/v1/2020.coling-main.543
34. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM.

Mach. Learn. **39**, 103–134 (2000). https://doi.org/10.1023/A:1007692713085
35. Paquet, C., Coulombier, D., Kaiser, R., Ciotti, M.: Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveill. **11**(12), 212–214 (2006)
36. Paul, M.J., Dredze, M.: You are what you tweet: analyzing twitter for public health. Artif. Intell. **I**, 265–272 (2011)
37. Rao, D., Paul, M., Fink, C., Yarowsky, D., Oates, T., Coppersmith, G.: Hierarchical Bayesian models for latent attribute detection in social media. In: ICWSM (2011)
38. Smailhodzic, E., Hooijsma, W., Boonstra, A., Langley, D.J.: Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Serv. Res. **16**(1), 442 (2016). https://doi.org/10.1186/s12913-016-1691-0
39. Steinberger, R., Fuart, F., van der Groot, E., Best, C., von Etter, P., Yangarber, R.: Text mining from the web for medical intelligence. Min. Massive Data Sets Secur. **19**, 295–310 (2008)
40. Stewart, A., Fisichella, M., Denecke, K.: Detecting public health indicators from the web for epidemic intelligence. In: eHealth, pp. 10–17 (2010)
41. Stewart, A., Smith, M., Nejdl, W.: A transfer approach to detecting disease reporting events in blog social media. In: Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia, HT '11, pp. 271–280. ACM, New York, NY, USA (2011). https://doi.org/10.1145/1995966.1996001
42. Steyvers, M., Griffiths, T.: Probabilistic Topic Models. Lawrence Erlbaum Associates, Mahwah (2007)
43. Ullah, I., Khan, S., Imran, M., Lee, Y.K.: Rweetminer: automatic identification and categorization of help requests on twitter during disasters. Expert Syst. Appl. **176**, 114787 (2021). https://doi.org/10.1016/j.eswa.2021.114787
44. Vlachos, M.: Identifying similarities, periodicities and bursts for online search queries. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 131–142. ACM Press (2004)
45. Xu, G., Meng, Y., Zhou, X., Yu, Z., Wu, X., Zhang, L.: Chinese event detection based on multi-feature fusion and BiLSTM. IEEE Access **7**, 134992–135004 (2019). https://doi.org/10.1109/ACCESS.2019.2941653
46. Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and online event detection. In: SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 28–36. ACM, New York, NY, USA (1998). https://doi.org/10.1145/290941.290953
47. Yangarber, R.: Verification of facts across document boundaries. In: Proceedings International Workshop on Intelligent Information Access (2006)
48. Zhan, L., Jiang, X.: Survey on event extraction technology in information extraction research area. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 2121–2126 (2019). https://doi.org/10.1109/ITNEC.2019.8729158
49. Zhang, D., Zhai, C., Han, J., Srivastava, A., Oza, N.: Topic modeling for OLAP on multidimensional text databases: topic cube and its applications. Stat. Anal. Data Min. **2**(5–6), 378–395 (2009). https://doi.org/10.1002/sam.v2:5/6
50. Zhang, Y.: Automatic extraction of outbreak information from news. Ph.D. thesis, University of Illinois (2008)