How to Exploit Twitter for Public Health Monitoring?

K. Denecke1,5; M. Krieck2; L. Otrusina3; P. Smrz4; P. Dolog6; W. Nejdl6; E. Velasco6

1Innovation Center Computer Assisted Surgery, Leipzig, Germany; 2Niedersächsisches Landesgesundheitsamt, Hannover, Germany; 3Brno University of Technology, Brno, Czech Republic; 4Aalborg University, Aalborg, Denmark; 5Forschungszentrum L3S, Hannover, Germany; 6Robert Koch Institut, Berlin, Germany

Keywords
Textmining, Web science, public health, population surveillance, epidemic intelligence, Medicine 2.0

Summary
Objectives: Detecting hints to public health threats as early as possible is crucial to prevent harm from the population. However, many disease surveillance strategies rely upon data whose collection requires explicit reporting (data transmitted from hospitals, laboratories or physicians). Collecting reports takes time so that the reaction time grows. Moreover, context information on individual cases is often lost in the collection process. This paper describes a system that tries to address these limitations by processing social media for identifying information on public health threats. The primary objective is to study the usefulness of the approach for supporting the monitoring of a population’s health status.

Methods: The developed system works in three main steps: Data from Twitter, blogs, and forums as well as from TV and radio channels are continuously collected and filtered by means of keyword lists. Sentences of relevant texts are classified relevant or irrelevant using a binary classifier based on support vector machines. By means of statistical methods known from biosurveillance, the relevant sentences are further analyzed and signals are generated automatically when unexpected behavior is detected. From the generated signals a subset is selected for presentation to a user by matching with user queries or profiles. In a set of evaluation experiments, public health experts assessed the generated signals with respect to correctness and relevancy. In particular, it was assessed how many relevant and irrelevant signals are generated during a specific time period.

Results: The experiments show that the system provides information on health events identified in social media. Signals are mainly generated from Twitter messages posted by news agencies. Personal tweets, i.e. tweets from persons observing some symptoms, only play a minor role for signal generation given a limited volume of relevant messages. Relevant signals referring to real world outbreaks were generated by the system and monitored by epidemiologists for example during the European football championship. But, the number of relevant signals among generated signals is still very small: The different experiments yielded a proportion between 5 and 20% of signals regarded as “relevant” by the users. Vaccination or education campaigns communicated via Twitter as well as use of medical terms in other contexts than for outbreak reporting led to the generation of irrelevant signals.

Conclusions: The aggregation of information into signals results in a reduction of monitoring effort compared to other existing systems. Against expectations, only few messages are of personal nature, reporting on personal symptoms. Instead, media reports are distributed over social media channels. Despite the high percentage of irrelevant signals generated by the system, the users reported that the effort in monitoring aggregated information in form of signals is less demanding than monitoring huge social-media data streams manually. It remains for the future to develop strategies for reducing false alarms.

1. Introduction

In the last couple of years, public health authorities and epidemiologists became aware that it is no longer sufficient to consider indicator-based data from patient records or collected through active reporting for disease surveillance since the time delay is often immense. Research started to develop methods and technologies that allow monitoring additional data sources for the purpose of detecting public health threats as early as possible. These developments were summarized by the term “Epidemic Intelligence” [1].

Epidemic Intelligence comprises an early identification, assessment and verification of potential public health hazards and a timely dissemination of alerts. It relies heavily on established indicator sources such as number of reported infections, drug prescriptions, etc. With the increased development of the web and in...
particular the increased use of social media tools for information and communication, such sources became also relevant for monitoring purposes; event-based surveillance systems were developed to process these sources. Social media are Internet-based applications that further enable people to share their own information via the Internet. This form of communication is more common than ever before and has gained unprecedented popularity around the world through social networking websites like Facebook or microblogging websites like Twitter. The trend is also recognizable in the health care field, where people are accessing websites for medical advice, joining patient communities and posting information about their own health status [2]. In this context, the research area of health web science came up, among others with research on online health communities and health-related YouTube videos [3–5].

For public security issues (e.g., monitoring terroristic activities), social media is already considered intensively. This matter is acknowledged also by the public health community [6]. Health authorities started to use additional information for the purpose of early warning and disease outbreak prevention. For example, epidemiologists at the European Centre of Disease Prevention and Control (ECDC) use the MedISys system [7] to check local news for disease outbreak information. Scientific assessments of the utility of web mining for the domain of public health showed that it could help in overcoming time delays in reacting to health threats, among others by providing additional information about outbreaks [8].

Accordingly, a couple of systems was developed that monitor mainly online news for disease surveillance purposes (e.g., MedISys, BioCaster, see Section 2). They mainly present texts that are determined as relevant by the system and resist on aggregating information referring to the same health event into signals. Beyond online news, another frequently assessed source are search logs from search engines such as Google Flu Trends. Evaluations of those approaches concentrated so far on a comparison of the results with data from health statistics (indicator-based data). For example, several papers correlated the output of Google Flu Trends with influenza outbreak statistics [9–11]. In this work, we present an approach that considers information from multiple sources, aggregates relevant information into signals that are presented with various presentation options to a user. In contrast to the existing assessments, our experiments do not correlate automatically the results with official health statistics, but involved epidemiologists from health organizations to study the relevance of online media for disease surveillance in daily practice. To the best of our knowledge something similar has not yet been reported for any other system.

The objectives of this paper are three-fold: 1) presenting a system for monitoring medical social media for disease surveillance, 2) assessing the capabilities of the developed algorithms, their possibilities and limitations for exploiting social-media data for disease surveillance purposes, and 3) reporting on the user experiences on the usefulness of social media monitoring.

The paper starts with a summary of the latest developments in the Epidemic Intelligence (Section 2). The assessments, presented in this paper, exploit the M-Eco system that we developed as a project consortium to enable social media monitoring for disease surveillance purposes. The M-Eco system is described in Section 3. Several experiments were performed by epidemiologists with the system. Results and observations when working with the M-Eco system are reported in Section 4. Finally, we draw conclusions regarding possibilities and limitations of social media for disease surveillance and provide recommendations for future developments and research in Section 5.

2. Background

Public availability of user-generated information from social media has been recognized as a potential source of information for surveillance epidemiologists [6]. While the data from indicator-based surveillance rather refers to the health status of individual persons (laboratory results, onset of diseases), social-media data rather describes the context and the perception of affected persons or their surroundings of a public health event. Beyond, in social media, the perception of symptoms are described rather than diagnosed diseases. The complementing content of medical social-media can help to further optimize decision making on specific actions implied by signals from both – indicator-based surveillance systems and event-based surveillance systems. These are at least the assumptions behind social media monitoring for disease surveillance. This paper tries to find out to what extent this assumption is true.

Several groups started in the last years with considering web information to support public health and public policy. G. Eysenbach introduces Infovigil, a proof-of-concept infoveillance system [12]. It can identify, archive, and analyze health-related information from Twitter and other information streams from the Internet and social media, and allows to monitor public opinions and public health-related behavior [12]. Initiatives including ProMED-mail [13], GPHIN [14], BioCaster [15], and HealthMap [16], have demonstrated new mechanisms for acquiring data for disease surveillance purposes. Most of the existing systems for event-based surveillance consider textual resources in multiple languages. They mainly monitor information presented in online news articles. For detecting public health events, handcrafted patterns are exploited by those systems that help in extracting information on disease and location. The extraction is normally considered as a classification task. Explicit and well-structured units of information that are easily interpretable are identified by text mining and information extraction techniques. Results are presented as geolocations in a map, in alerts, or by means of automatically created summaries.

BioCaster [17] is an ontology-based system which uses a domain-specific event ontology to perform named entity recognition on outbreak reports. The system analyzes texts from over 1,700 RSS feeds, classifies them for topical relevance and plots them onto a Google map using geo-code information. PULS [18] employs information extraction techniques to identify diseases and locations of reported health events. PULS is integrated into MedISys [7], which
automatically collects articles concerning public health in various languages from news, and aggregates the extracted facts according to pre-defined categories, in a multi-lingual manner. Other systems such as Proteus-BIO [19] or EpiSpider [20] automatically extract infectious disease outbreak information from several sources including ProMed-mail, the news archive of the World Health Organization (WHO) and medical news web sites. Such event-driven surveillance tools are already widely used by health organizations [21].

Although it is thought that information exchange on the Internet is changing the landscape for detecting potential public health threats, it is not yet clear whether or not it can enhance established indicator-based surveillance of known infectious diseases. In particular, studies are missing that analyze the usefulness of social media to support disease surveillance. Several studies focused on correlating Web search behavior [22] or data streams from Twitter [23, 24] with official health statistics to assess their use for monitoring influenza-like illnesses. Chan et. al [25] applied search-stream surveillance techniques to the monitoring of dengue. They found out that search queries closely track dengue activity as measured by traditional systems. Our assessments are broader in a sense that the M-Eco system allows exploiting social media for monitoring important infectious diseases, not only influenza-like illnesses.

3. Methods

This section introduces the processing pipeline of the M-Eco system. After a general description of the overall system, the single components are described in more detail. Knowing the system’s underlying procedures for content collection and signal generation is crucial to understand the experiments and results presented in Section 4. Although a prototype implementation exists, the system is not yet intended to be publicly available.

3.1 General Overview

M-Eco exploits data from social media and TV/radio for public health monitoring purposes. The system:

- monitors social media, TV, radio and online news,
- aggregates texts into signals,
- visualizes the signals using geographic maps, time series and tag clouds,
- allows searching and filtering signals along various criteria (location, time, medical condition).

In a nutshell, the system works as follows: Data from social media and TV/radio is continuously collected. Social media sources comprise among others texts from Twitter, blogs, and fora. TV and radio transmissions are recorded via satellite and transcribed automatically to text. The transcripts are used as input to the system. In the following, the term „text” is used to refer to some piece of text which can be for example a tweet, a blog posting, or even a transcript of a TV or radio transmission. A text is annotated with linguistic information; disease names, persons and locations are identified. These features provide an input to machine learning algorithms that detect patterns in the data. The patterns are analyzed and signals are generated automatically when unexpected behavior is determined. A signal is a hint to some anomalous event. Since the amount of generated signals can overwhelm a user, recommendation techniques are exploited to filter out those signals that are of potential interest for a particular user. The information related to a signal is shown in charts and through personalized tag clouds to allow users to easily assess signals.

To realize these steps, the M-Eco system consists of a set of web services that cover four areas depicted in Figure 1: 1) Content collection, 2) Signal generation, 3) User modeling and recommendation as well as 4) Visualization in a user interface. The services work in a pipeline fashion and are triggered automatically four times a day. The processing within the four components is described in more detail in the following paragraphs.

3.2 Content Collection and Document Analysis Component

The information database of the system is filled continuously by the Content Collector and Document Analysis Component. It collects data from various sources by means of

![Figure 1 Components and processing pipeline of the M-Eco system](image)
web crawling and streaming APIs (see below), and makes them accessible to other components. The collection focuses on broadcast news from TV and radio, news data from MedISys [7], and social media content from blogs, forums and Twitter. The TV and radio data is collected via satellite and transcribed to written text by SAIL’s Media Mining Indexing System [26].

Blogs and forums are collected from their corresponding RSS feeds. A streaming API from Twitter allows downloading a subset of Twitter data containing keywords from a predefined list. Although the Twitter data is publicly available, it is not possible to download all data from Twitter in real-time due to restrictions imposed by Twitter. Hence, a special treatment is needed. We got access to a special account from Twitter that allows defining lists of up to 10,000 keywords. About 1,300 names of symptoms and diseases are currently used as the keywords, the rest is reserved for specific conditions and locations. These keywords were manually collected by epidemiologists as part of the M-Eco system development process, starting from a list of 18 diseases or group of diseases (Table 1). The epidemiologists were asked to collect synonyms for these diseases, in particular terms used by non-health professionals to describe these diseases and related symptoms. In total, 386 terms referring to disease names and 872 terms referring to symptoms in English and German were collected for the initial version of the system. The list is continuously extended. Further, existing language resources such as WordNet, GermaNet, or the OpenOffice thesaurus were additionally used to extend the keyword lists.

The data collection process is generally language-dependent. However, for the experiments reported in this paper, English and German texts were used.

Within the document analysis, texts are annotated with linguistic and semantic information. All collected relevant texts are processed by a set of natural language processing tools. The data is tokenized and part-of-speech-tagged by the Tree Tagger and parsed by the Stanford Parser. All texts are also semantically annotated. Four different tools are applied: BURGeoN is based on Geonames.org data and is used to annotate locations and to disambiguate geographical features. The identification of terms referring to symptoms and diseases is based on a pre-defined list of known entities (the same list that is used for collecting the texts). The recognition of those entities is realized as a minimal finite state automaton that was built for both, diseases and symptoms. HeidelTime [27] is used to identify temporal expressions and the Stanford NER [28] is used to detect information on affected organisms.

3.3 Signal Generation Component

The Event Detection and Signal Generation Component exploits the annotated texts provided by the Content Collection and Document Analysis component to generate signals. It produces signals with associated information on the disease or symptom the signal is referring to and a location that has been extracted for that signal. In a first step, relevant sentences are separated from irrelevant ones using a transfer-based machine learning classifier based on support vector machines and considering the annotated information as features (sentence position, part-of-speech tag tree, named entities including disease names, location, person) [29]. The classifier is based on the open source implementation of the SVM-TK by Moschitti [30]. The algorithm automatically classifies outbreak reports (from ProMed Mail [13] and WHO) to train a supervised learner. The knowledge acquired from the learning process is then transferred to the task of classifying social media texts. Our experiments showed that with the automatic classification of training data and the transfer approach, an overall precision of 92% and an accuracy of 78.20% is achieved (tested on blogs from the Avian Flu Diary, http://afludiary.blogspot.de/).

Sentences are classified as relevant or irrelevant by the method presented before. For all relevant sentences, entity pairs (location, disease) are exploited to produce time series. A time series is produced for each entity pair occurring in sentences of texts published within one week. The sentences determined as relevant are also made available in the visualization for the user when assessing a signal. This helps to concentrate on the relevant sentences during manual information analysis.

The time series provide then the input for statistical methods for signal generation, CUSUM and Farrington. These two statistical methods have originally been developed for indicator-based surveillance [31]. Cumulative sum or CUSUM methods originated in quality control [32]. They focus on several consecutive periods, and sum up the aberrations in one particular direction. Farrington et al. proposed an approach based on generalized linear models [33], which is by now broadly applied in European countries for indicator based surveillance. The Farrington approach fits a regression model to the data over several years, allowing for a secular trend. Outbreaks in the past are automatically identified and removed, and the statistical distribution fits either to rare counts or to frequent counts. Since there are so far no assessments available on which of these two statistical methods is better suited for generating signals from social-media data, the involved epidemiologists wanted to see signals generated by both algorithms. The user interface allows them to select signals from the one or the other algorithm. Between zero and fifty signals are generated by this procedure every night. The exact

Table 1  Names of diseases that provided the starting point for manual keyword collection

<table>
<thead>
<tr>
<th>Disease Name</th>
<th>Disease Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borrelisosis</td>
<td>MRSA (Staphylococcus aureus)</td>
</tr>
<tr>
<td>campylobacter</td>
<td>norovirus</td>
</tr>
<tr>
<td>EHEC (pathogenetic E. coli)</td>
<td>salmonellosis</td>
</tr>
<tr>
<td>early summer meningocephalitis</td>
<td>SARS</td>
</tr>
<tr>
<td>hepatitis</td>
<td>encephalitis</td>
</tr>
<tr>
<td>meningitis</td>
<td>haemorrhagic fever</td>
</tr>
<tr>
<td>influenza</td>
<td>respiratory diseases</td>
</tr>
<tr>
<td>measles</td>
<td>childhood illnesses</td>
</tr>
<tr>
<td>hospital acquired infection</td>
<td>gastroenteritis</td>
</tr>
</tbody>
</table>
number depends on several variables or factors that influence the generation of signals such as the type of considered data (e.g., Twitter’s update frequency is much higher than of a blog).

The parameters of the two signal generation algorithms were set as follows: Farrington was run with a window size of 5, and an alarm threshold of 0.5 while the CUSUM, specifically C1, was run with a training window of 4 days, a buffer of 1 day and an upper control limit of 1 and an alarm threshold of 0.3. The parameters were chosen after initial testing by the epidemiologists: They were confronted with results produced by the algorithms using different parameters. At the end they decided that it would be better to have more signals, even irrelevant ones, instead of missing something at that period of time when the system is starting to work. Clearly, a more in depth analysis of the behavior of the algorithms and associated parameter settings for processing social media data is open to future work and requires also long-term assessment of the system.

Further, the following properties of the signal generation algorithm need to be considered:
1. In order to contribute to a signal, at least one entity referring to a disease or symptom, e.g. “measles”, has to occur in a sentence (or the text it was taken from).
2. A signal is generated when the threshold for the number of associated sentences/texts with the same disease entity is exceeded. The threshold is based on an empirical value, but at least two sentences/texts with the same keyword need to exist.
3. The sentences/texts that contribute to a signal must have to fall into the same time frame which is set to one day.

3.4 Recommendation Component

The Recommendation Component gets as input the generated signals and either selects those that are of interest for a user according to his profile or ranks the signals appropriately. The component also supports users with personalized presentation options (e.g., tag clouds, list of recommendations) that are visualized in the user interface (see section 3.5). In this way, information or alerts are filtered before being presented to a user which in turn reduces information overload. The recommendation component requires a user profile that consists of information on user behavior from interactions with the system (e.g., ratings, tags, search terms).

The personalization and recommendation of signals mainly relies upon the tagging behavior of a user. Tags are potential indicators of user preference. For instance, a medical expert that has exhaustively assigned the tag “swine flu” to the texts he evaluates, seems to be interested in that disease. Therefore, this knowledge can be utilized to filter out irrelevant recommendations unrelated to “swine flu”. For recommending items to the user, we compare his tags, i.e. tags assigned by him to his texts of interest against the tags assigned to candidate and unknown texts. The comparison is realized by the cosine similarity of two tag vectors, one corresponding the user's tag vector and the other corresponding to the text's tag vector. The text with highest similarity to the user's tag profile is then selected to be recommended. More details on the Recommendation Component and the underlying algorithms can be found in [34, 35].

In order to help users navigating through vast collection of texts and finding new items, we provide a visual representation of texts through a tag cloud component. Besides indexing texts in the corpus, each tag helps users to find new related information of interest. Tag clouds were chosen as the core retrieval interface for exploring specific texts. A subset of texts is retrieved by a clicking on the term available in the tag cloud. Terms called tags displayed in the tag cloud are defined by users or automatic annotation tools. Therefore, various combinations of tags result in different subsets of retrieved texts and all emerging trends and relationships in the available set of texts can be explored.

Tag clouds are a common visualization method in the Web 2.0 community. Studies showed that they enhance the perception of (web) documents [36] and they support an explorative search when it is difficult to specify a concrete query. These benefits perfectly address the challenges of disease surveillance from web documents. In the context of surveillance, using tag clouds for visualization is still unexplored.

In addition to the standard tag clouds we extended the tag cloud interface with semantic and syntactical grouping to avoid tag redundancies in the tag cloud. Semantically similar tags were depicted with the same color and positioned nearby each other. For clustering tags into semantic groups, the K-means algorithm was exploited that considers each tag from a tag space as feature vector [27]. By this clustering of tags, it is easier for a user to notice and discover relations and connections between depicted tags.

Specifically in the case of tags, we proposed tag-based recommender based on tensor factorizations. Tensor based recommenders build 3-dimensional matrix (tensor) by reflecting relationships between all users, items and tags. Afterwards, a factorization technique is performed on the constructed tensor. The tensor approximation usually reveals latent relations between the involved objects. The algorithm is described in detail by Leginus et al. [27].

3.5 User Interface and Visualization

The User Interface allows a user to search for disease names or symptoms and to assess the related signal information by means of a geographic map, a tag cloud or a timeline. These three types of visualizations were selected based on a requirement analysis performed in advance. Geographic map and timeline are visualizations the users know from other surveillance systems. So they desired to see information in a similar way. The geographic map plots the signals to a map. It enables the user to select specifically signals related to locations that are interesting for him. The timeline shows the text volume referring to a specific disease or symptom (or the corresponding signal, respectively) over time. This allows users to learn about the progress of a disease outbreak as reflected in social media and also about seasonal differences. The tag cloud provides a quick overview on the content of the texts associated with a signal. They enable the user to
quickly decide about the relevance of a signal. Access to the original sources that contributed to the signal generation is provided as well as filtering capabilities (e.g. selecting a time span). Beyond, user feedback options were included into the user interface. With “Thumbs up-thumbs down” and a rating scale for signals users can judge the relevancy of the presented signal. This information is fed back to the recommendation process and considered for ranking and filtering. Through those services described before, M-Eco offers 1) additional information through social media monitoring, 2) perception of recommendation and users behavior and 3) visualization and support for risk assessment. A screenshot of the system is shown in Figure 2 and Figure 3.

4. Experiment

In this section, we describe our assessments with the M-Eco system performed to 1) learn more about the behavior of the algorithms when applying on social media, and 2) assess the relevance of social media monitoring for disease surveillance purposes. To test the whole processing pipeline, we performed a test with scenario-driven generated tweets (Section 4.1). To learn more about the usefulness of social media for public health, the system was used by epidemiologists to assess real-world outbreaks (Sections 4.2 and 4.3).

4.1 Scenario-driven Twitter Experiment

4.1.1 Experimental Setup

The main purpose of the scenario-driven Twitter experiment was to learn about the behavior of the M-Eco system and its algorithms within a controlled scenario. The aim was to answer the question, how the algorithms behave when processing social-media data. For such assessment, we need to know which texts should have contributed to a signal and what are the outbreaks they are referring to. In order to study the aforementioned questions, we considered the following three scenarios:

1. Pupils staying home on suspicion of measles (Measles),

Figure 2  Tag cloud view of texts from a selected signal
2. Soccer fans showing symptoms after attending a live transmission of a soccer game during European Championship (Salmonellosis),
3. Travelers coming back from Egypt and experiencing symptoms (Hepatitis A).

For these three scenarios, we asked 13 epidemiologists within a workshop to put themselves into the position of a concerned person, i.e. they had to imagine that they are either sick by themselves or are relatives or friends of a sick person. The participating epidemiologists work for national and global health organizations such as WHO, ECDC, Robert Koch-Institut and Institut de Veille Sanitaire. We decided to ask epidemiologists to join this experiment for two reasons: 1) The data used for the experiment should be “controlled”, i.e. we wanted to know which texts were entered into the system and 2) we wanted to get also the feedback on the generated signals of the epidemiologists (i.e. to learn whether their expectations were in-line with the results).

The experiment was conducted during a two-day workshop with the epidemiologists. They were asked to provide texts in German with a length no longer than 140 characters (which is the maximum length of a message at Twitter) for the three scenarios together with a time and location according to their choice. Example texts are given in Table 2.

All texts were generated within one hour. The 195 generated texts were fed into the processing pipeline, mixed up with texts that were collected from Twitter on that particular day when the experiment was conducted and the M-Eco algorithms generated signals. The exact number of tweets that were collected from Twitter on that day is difficult to estimate, since the content collection process is not counting the texts that are filtered out by the relevance filter. In total 636 texts entered the signal generation process. We measured how many conceived texts contributed to the generation of a signal and assessed characteristics of those texts.

We expected that 75–80% of the conceived texts contribute to signals, since we let epidemiologists conceive the texts who know the relevant terminology and we have concrete description of the scenarios.

4.1.2 Experimental Results
After passing the processing pipeline, signals were generated to which the testing texts contributed. One third of the testing texts were filtered out within the content collection process due to missing matching keywords referring to a disease or symp-
tom. For the measles scenario, almost all of the texts were kept by the content collection process. A reason for this is that the users more often used the concrete disease name "Masern" (measles) in their texts, while in the other scenarios (i.e. Hepatitis A or Salmonellosis) symptoms were preferred over the disease name in the text.

About one third of the texts that contained disease/symptom-referring keywords became part of a signal (124 out of 195 texts, Table 3). 80% of the texts on measles contained relevant keywords while only around 50% of the texts referring to the other two scenarios matched relevant keywords. As a reminder: only texts matching relevant keywords are considered in the signal generation process. Forty-two of the texts matching some keyword became part of a signal in the signal generation process. In total 636 texts including these 42 texts produced by the epidemiologists during the experiment contributed to 21 generated signals. Some texts occurred in more than one signal since they contained more than one keyword or contributed to a signal at a specific location and of another signal without specific location. Out of 42 texts that showed up as a part of a signal, 21 texts were related to measles, 10 were related to Salmonellosis and 11 dealt with Hepatitis.

### 4.1.3 Discussion

From a technical point of view, the experiment has shown that the pipeline worked as intended, i.e., the texts were passed through the components and a subset of texts led to signals that are related to the initial scenarios. However, the proportion of texts captured by the content collection and signal generation procedures was smaller than expected (75–80% were expected; 21% of the texts contributed in total). The first reason for the small proportion of captured texts is the incomplete keyword list underlying the content collection process. More than one third of the conceived texts did not contain keywords from the keyword list and were therefore filtered out. This result reflects a major challenge of the M-Eco system or similar systems: In social media, medical terms are not necessarily used to talk about medical conditions. The terminology is often very hard to predict (e.g. previously undefined slang words or phrases). For a human reader able to detect the slang words or phrases, it may be unquestionably clear that a text contains statements about a specific indication, but automated keyword searches may not identify the same text as referring to a specific indication of a disease.

Another problem was that the experiment participants were allowed to select any location of their choice. Unfortunately, texts referred to so many different locations that they did not produce a cluster in space. This refers to another major challenge of the event-based surveillance systems: Should signals be generated when the texts only build a cluster in time instead of time and space? From Twitter data analysis, we learned that the proportion of real tweets that contain location information is very small, even if we consider user profile information. Thus, more signals would be generated when the location information is not considered for signal generation and only a cluster in time is built. However, for epidemiological risk assessment it is crucial to know where something is happening, i.e. to have a cluster in space. Solutions to this problem remain open for future research.

Letting epidemiologist conceive tweets based on given scenarios as it was done in this scenario-based experiment clearly reflects only partly the reality and could lead to biased results. The texts were probably not always written in a way as it would have been done by a concerned person. However, we decided to run the experiment as presented to learn more about the possibilities and limitations, in particular, on how the system would work under good conditions with respect to text quality, data processing and analysis. “Real” tweets are considered and studied in the experiments described in the following sections.

### 4.2 Monitoring during Mass Gathering

Beyond studying the system behavior in a controlled experiment, we wanted to analyze how the system can be used in real world monitoring and whether signals are generated for real disease outbreaks. For this reason, epidemiologists exploited the M-Eco system for disease monitoring purposes for a certain time period and in particular during a mass gathering event.

| Number of texts and their occurrences in signals in the tweet experiment. The first column shows the disease name of the corresponding scenario for which texts were generated. “Total number of documents” refers to the number of texts that were produced by the involved epidemiologist. The column “With keyword” indicates the percentage of texts out of the total number of texts that contained at least one keyword from the list underlying the content collection process. Column “part of a signal” indicates the percentage of texts with matching keywords that contributed to a signal. |
|---------------------------------|-----------------|-----------------|
|                                | Total number of documents | With keywords | Part of a signal |
| **Hepatitis**                  | 65               | 37 (57%)        | 11 (17%)         |
| **Measles**                    | 65               | 52 (80%)        | 21 (33%)         |
| **Salmonellosis**              | 65               | 35 (54%)        | 10 (15%)         |
| **Total**                      | 195              | 124 (64%)       | 42 (21%)         |

**Examples for the texts provided by the epidemiologists as input for the Twitter experiment. The translation is given in brackets. The translated text was not used in the experiment.**

- Krankes Kind ist nicht toll. Fieber, Husten und Pocken überall (Sick children are so bad. Fever, coughing, smallpox everywhere)
- Masern in der Schule! 3 Kinder betroffen (measles at school! 3 children infected.)
- Die Waldorfer im Kiez haben schon wieder Masern … (Again measles in the area “Waldorfer Kiez”)

**Table 2 Examples for the texts provided by the epidemiologists as input for the Twitter experiment.**

| **Table 2** Examples for the texts provided by the epidemiologists as input for the Twitter experiment. The translation is given in brackets. The translated text was not used in the experiment. |
|---------------------------------|-----------------|-----------------|
|                                | Total number of documents | With keywords | Part of a signal |
| **Hepatitis**                  | 65               | 37 (57%)        | 11 (17%)         |
| **Measles**                    | 65               | 52 (80%)        | 21 (33%)         |
| **Salmonellosis**              | 65               | 35 (54%)        | 10 (15%)         |
| **Total**                      | 195              | 124 (64%)       | 42 (21%)         |
4.2.1 Experimental Setup

For the experiment described in this section, we asked three full-time epidemiologists from the Surveillance Unit of the Department for Infectious Disease Epidemiology at the Robert Koch Institute in Berlin and one full-time epidemiologist at the State Health Agency of Lower Saxony (NLGA) (see Table 4 for their profiles) to determine the relevance of generated signals presented in the M-Eco user interface and to rate and tag them. Relevance was defined as a signal truly referring to a medical condition. A relevant signal might provide new or helpful information and therefore might be interesting for an epidemiologist in charge of monitoring this particular disease. This does not mean, that the content of this signal is reflecting a beforehand completely unknown relevant public health event, as this would almost be impossible to figure out in each case.

Evaluators were also asked to look out for events deemed relevant to their current work, and where possible they were asked to compare any daily findings with other information sources already in use for infectious disease epidemiology. Finally, evaluators were asked to generally comment on their overall experience with the system. The monitoring was done during a mass gathering which was the UEFA European Football Championship “Euro 2012” taking place in Poland/Ukraine from June 8 to July 1, 2012. During that period many soccer fans stayed in Poland and Ukraine. The assumption underlying our assessments was that in particular during such events, people are exchanging information through Twitter and social media, also on health issues. Beyond the regular data collection performed as described before in Section 3.2, additional tweets were collected using hashtags such as #Euro, #Soccer, #Football, #Poland, #Ukraine, #UEFA, #em2012, #UEFAEuro2012. The main objective of the experiment was to measure how many signals are generated by M-Eco during a mass gathering event, how many of them are irrelevant or refer to a real world outbreak. As a reference, the evaluators accessed the information sources that they are normally using in their work when monitoring a mass gathering event, i.e. reports from the World Health Organization, from the European Center of Disease Prevention and Control and online news.

Each signal was judged by at least one epidemiologist. One user rated every signal and the other users rated those signals which appeared to be interesting according to the medical condition and the location presented by the system. The main objective of the evaluation was to find out the relevance of the generated signals to infectious disease epidemiologists. The evaluation encompassed daily monitoring of infectious diseases and related symptoms relevant for Germany using the search interface made available during 12 days of the games.

### Table 4  Evaluator profiles

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Age</th>
<th>Experience (Years)</th>
<th>Academic degree</th>
<th>Job title</th>
<th>Professional focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>30–35</td>
<td>4</td>
<td>PhD</td>
<td>Epidemiologist</td>
<td>Infectious disease epidemiology, social epidemiology, surveillance</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
<td>30–35</td>
<td>2</td>
<td>MSc</td>
<td>Epidemiologist</td>
<td>Infectious disease epidemiology, surveillance</td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>35–40</td>
<td>5</td>
<td>PhD</td>
<td>Epidemiologist</td>
<td>Infectious disease epidemiology, psychology</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>40–45</td>
<td>15</td>
<td>PhD, MSc</td>
<td>Department Head</td>
<td>Infectious disease epidemiology, surveillance, statistical methodology</td>
</tr>
</tbody>
</table>

### Table 5  Signals produced with reference to the EURO 2012 and rated positive (relevant) by at least one out of four users. The other 229 signals were not rated positive by any of the users.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Medical Condition</th>
<th>Location</th>
<th>Positive / Negative ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Measles</td>
<td>Wroclaw</td>
<td>3 / 0</td>
</tr>
<tr>
<td>2</td>
<td>Diarrhea</td>
<td>France</td>
<td>2 / 0</td>
</tr>
<tr>
<td>3</td>
<td>HIV</td>
<td>Ukraine</td>
<td>2 / 0</td>
</tr>
<tr>
<td>4</td>
<td>Measles</td>
<td>European Union</td>
<td>2 / 0</td>
</tr>
<tr>
<td>5</td>
<td>Measles</td>
<td>Poland</td>
<td>2 / 0</td>
</tr>
<tr>
<td>6</td>
<td>Measles</td>
<td>Ukraine</td>
<td>2 / 0</td>
</tr>
<tr>
<td>7</td>
<td>Measles</td>
<td>Ukraine</td>
<td>2 / 0</td>
</tr>
<tr>
<td>8</td>
<td>Measles</td>
<td>Ukraine</td>
<td>1 / 1</td>
</tr>
<tr>
<td>9</td>
<td>Fatigue</td>
<td>Germany</td>
<td>1 / 0</td>
</tr>
<tr>
<td>10</td>
<td>HIV</td>
<td>South Africa</td>
<td>1 / 0</td>
</tr>
<tr>
<td>11</td>
<td>Meningitis</td>
<td>Oslo</td>
<td>1 / 0</td>
</tr>
<tr>
<td>12</td>
<td>Sweating</td>
<td>London</td>
<td>1 / 0</td>
</tr>
<tr>
<td>13</td>
<td>Tiredness</td>
<td>United Kingdom</td>
<td>1 / 0</td>
</tr>
</tbody>
</table>
4.2.3 Discussion

This assessment showed that the average number of signals (around 20 per day) is an easily manageable amount of signals for one person. This means that the monitoring effort for public health authorities can be significantly reduced through the M-Eco signal generation: They don’t have to go through millions of online texts. The epidemiologists reported that the time for deciding whether a M-Eco signal is relevant or not is very short given the tag cloud visualization shown for each signal.

Irrelevant signals can usually be identified quite fast, already by checking the tag cloud. If the tag cloud assessment is insufficient for making such decision, the single texts linked to a signal have to be read. This takes on average less than one minute for the irrelevant signals. Potentially relevant signals of course might require more time for assessment. However, they are considered being worth the time as they usually contain interesting pieces of information. Altogether, the time effort for an epidemiologist to check this average number of signals would be less than one hour per day.

From a technical perspective, the number of false positive signals is very high. Future work thus needs to concentrate on processing ambiguity and filtering out irrelevant signals to decrease the number of false positives. Since the people are very creative in inventing new word compounds, one possible solution is to retrain or update algorithms regularly. Despite the large number of false positive signals assessed during the experiments, the users were impressed by the possibility of monitoring using social media. Additionally, most of the signals classified as relevant in this experiment did not refer to unknown public health threats, but to situations which were known beforehand and communicated during the Euro 2012 as part of prevention strategies, like the high HIV prevalence in the Ukraine and the ongoing measles outbreak in the Ukraine. We conclude that the presented system can support health officials in monitoring population’s health during mass gatherings. Early warnings for disease outbreaks could not yet been determined in the assessment.

4.3 Epidemiological Assessment of M-Eco Signals

In the following, we describe another epidemiological assessment of signals that reflects the daily work of epidemiologists.
4.3.1 Experimental Setup

The evaluation encompassed daily monitoring of infectious diseases and related symptoms relevant for Germany using the M-Eco interface made available during the time period Oct 9 – Oct 27, 2012. The evaluation was completed by the same four epidemiologists that performed the assessment reported in section 4.2 (Table 4). The main aim of the evaluation was to study the relevance of the signals to infectious disease epidemiologists.

Evaluators were also asked to look out for events deemed relevant to their current work, and they were asked – where possible – to compare signals with other information sources already in use for infectious disease epidemiological surveillance. Finally, evaluators were asked to generally comment on their overall experience with the system.

The monitoring concentrated on specific diseases and symptoms (Table 6). In an expert consultation and discussion diseases (and up to five symptoms) were chosen for monitoring that were deemed to be more prevalent in Germany. Additionally, priority was given to those diseases and symptoms likely to be discussed in the general population via social media (due to popularity and general ubiquity or due to previous outbreak history or historical media coverage) or those less likely to induce social stigma. We excluded diseases that were deemed seasonally irrelevant for the time period, such as Tick-borne encephalitis (TBE) that primarily occurs in the summer months. Other diseases were excluded because they occur so rarely that experts found a high likelihood for them to remain unmentioned in social media: e.g. Q-Fever; or due to an uncommon prevalence, or a faster/more severe onset of disease (and therefore higher likelihood to be detected by other sources) in Germany: e.g. Hemorrhagic (West Nile) Fever, Tuberculosis.

The assessment procedure was as follows: Each search term was allocated to one of four evaluators (Table 4) and evaluators monitored their terms daily with regards to the following attributes by entering their terms into the search field of the M-Eco user interface and going through the results:

- Get the number of signals returned to their term.
- Check whether there is indication of a larger event, i.e. an outbreak.
- Check whether signals are relevant, i.e. whether the signal truly referred to a medical condition.
- Check whether the search results were found in other epidemiological surveillance sources, i.e. the Robert Koch Institute's weekly epidemiological surveillance conference call and report (reference standard).

4.3.2 Experimental Results

Figure 4 shows the number of returned signals over time during the monitoring phase. Overall, the assessed signals show no specific upward or downward trend over the entire time period. The peaks on 19.10.2012 and 26.10.2012 show the most...
overall number of signals. When looking more closely at the diseases, we see that the overwhelming majority of signals can be attributed to the terms “Influenza” (N = 483) and “Grippe” (English: “Flu”; N = 324). These trends are met by signals from the term “MRSA” (N = 104) and “Hantavirus” (N = 76) and by the common symptom-terms “Kopfschmerzen” (English: “headache”; N = 235), “Fieber” (English: “fever”; N = 212), “Husten” (English: “cough”; N = 151). To a much lesser extent, the trend was met by “Durchfall” (English: “diarrhea”; N = 33). The terms “Norovirus” (N = 10), “Salmonellen” (English: “Salmonella”; N = 0), “Anthrax” (N = 0), and “Masern” (English: “measles”; N = 3), yielded very few, if any, results.

### 4.3.3 Discussion

A high number of texts referred to “Influenza” and “Grippe”, even though technically the evaluation period was outside the flu-season. This can be explained by the overall discussion in the general public about the availability of flu vaccinations in Germany, where a shortage was identified and publicly discussed in this time period. The comparison of the M-Eco signals and the notified outbreaks for Germany did not produce a definitive match. Nonetheless, it cannot be excluded that there are outbreaks underlying the signals, which were not detected by the indicator-based systems because the symptoms “headache” and “fever”, which were mentioned often, are likely related to many diseases that do not fall under a mandatory reporting clause, e.g. common colds.

In addition, our evaluation during this period uncovered signals referring to true outbreaks in other countries. For example, a signal on 18.10.2012 provided information regarding a gastrointestinal outbreak in Belgium. This outbreak was already known in Belgium and it has been reported about it by international health agencies. Thus, the M-Eco system provided no new information to the Robert Koch-Institute. However, one can point out that it might have been helpful information for a German user in a smaller health agency, and especially to those neighboring Belgium, since the regional health agencies in Germany do not have direct access to the international information channels otherwise common at national health authorities.

Concerning the presented signals, their number was smaller than expected initially by the evaluators. This could by part be explained by the restriction of the experiment to Germany: The amount of social media texts and thus the number of signals in German language is much smaller than in English language. A second reason could be the perceived social stigma associated with certain terms associated with diseases or symptoms that yielded fewer results. There is no social stigma talking about headache, fever or flu. However, gastrointestinal diseases, although sometimes mentioned by the media during large outbreaks, are not necessarily those illnesses most frequently discussed. One is more likely to publicly talk about a “headache” than “bloody diarrhea”. This is in contrast to the official notification system, where gastrointestinal diseases play an important role compared to flu like illnesses.

### 5. Discussion and Conclusions

In this paper, the M-Eco system was described and evaluated. The system allows monitoring of social-media data for generating epidemic intelligence. Through the evaluations, we were able to provide examples where M-Eco generates early signals on infectious disease outbreaks.

#### 5.1 Discussion of the Experiments

This work provides first insights into the possibilities, social media is providing for public health monitoring. It became clear that it is a continuing task to improve the algorithms to better match a constantly changing media landscape, and the language and socio-cultural handling of social media by the populations whose health we wish to monitor. More work on adaptive algorithms is required in particular for dealing with the terminology change that occurs quite often in social media (e.g. “football fever” happened during the European Soccer Championship).

The presented assessments and user feedback showed that the system is able to reduce the overwhelming amount of information to a manageable amount of signals. The different experiments yielded a proportion between 5 and 20% of signals regarded as “relevant” by the users. It is difficult to compare this value over systems, but in indicator-based surveillance epidemiologists from Robert Koch-Institute estimate one “relevant” outbreak per 100 notified cases, and in MedISys they estimate one “relevant” event (leading to a report) per 500 assessed texts.
quality changes when extending the keyword list. However, the epidemiologists involved in our evaluations were not interested in such retrospective analysis, but wanted to know and assess what the system contributes to their daily work. For this reason, the evaluations presented in this work were performed as described.

The main source that contributed to the signal generation was Twitter. The system also exploits additional sources. We did not studied in the experiments to what extent the single sources (TV, radio, blogs, online news, Twitter) contribute to the signal generation. Also the relevance or reliability of signals could differ depending on the source (e.g. information from online news is more reliable than information from Twitter). Information on reliability of signals could also be useful for users. This clearly shows that there is still room for improving algorithms and running additional evaluations to study the system and its results in more depth.

5.2 Comparison to Other Systems

Existing systems such as MedISys rely on a predefined set of media channels and thus they can only detect events reported through these channels. In contrast, M-Eco can detect each media item when it is fed into social media like Twitter. The system also allows to “learn” new relevant sources that can be integrated into the processing pipeline. For national or international public health authorities, this early knowledge is helpful, even if there is already a local awareness. This is because not every source of information might be forwarded quickly enough within the official notification system, and also there are infectious disease events that are not in the focus of the notification system (either the underlying disease is not notifiable, or the event occurs abroad).

To get an impression of the effectiveness of M-Eco to monitor health events, we can compare the proportion of 1 out of 20 information items being regarded as relevant (see evaluation during the EURO 2012 reported in Section 4.2) or, if we want to be more specific, the value of one outbreak event (Wroclaw) filtered out of 242 signals. We want to compare these numbers to results produced by existing systems: a) MedISys and b) the traditional infectious disease surveillance according to the infectious disease law.

a) For MedISys, if we take one week as example, the number of retrieved items between November 4 and November 10 was 858,175. A total of 924 items were reported to the Early alerting and reporting portal. After automatic deduplication, 694 items were displayed to the users. Four hundred and thirty were assessed to be relevant, but only for one an “incident under review” was created. This means that from 694 items presented to the users only one led to a concrete action.

b) In the German infectious disease notification system, i.e. the indicator-based system based on reliable medical notification data, each year between 300,000 and 400,000 case notifications are reported to the Robert Koch Institute. If we focus on public health relevant outbreak events, which is often defined as outbreaks with more than four cases, i.e. outbreaks larger than the typical household size, this number varies between 3000 and 4000 per year. Thus, the ratio between such outbreaks and number of notified cases is about 1 to 100. Here it has to be considered that already a lot of workload has been invested to identify and verify these cases.

Of course the comparison between the different surveillance systems suffers from the different definitions of what is regarded as relevant. However, if we compare M-Eco with MedISys and even if we only take the Wroclaw measles outbreak as a single relevant event, then the order of magnitude of relevant events in M-Eco is similar if not better than in MedISys. If the traditional notification system is considered as reference, it has to be taken in mind, that this system requires a high amount of work for collecting and verifying the case information. But even then the ratio of 1 outbreak per 100 notified cases makes it obvious that from the perspective of outbreak detection the effectiveness of M-Eco is in the same order of magnitude.

5.3 Lessons Learnt

Contrasting the initial expectation, the signals were not generated from clustered reports on personally reported symptoms, but on news reports that were fed into social media, and replicated or forwarded by interested users. Therefore, M-Eco was not the first instance to detect the public health event, because there were local actors who had already detected and reported about the event. But M-Eco brought such reports quickly to a broader attention.

Concerning one of the initial objectives of the project, we were not yet able to present an example where M-Eco was the first to detect an outbreak by a clustering of social media contributions with similar symptoms in space and time, and where the outbreak was afterwards confirmed by the traditional notification system. In part, this might be due to the fact that people are more likely to talk via social media about having a headache than about having diarrhea, but more than 95% of the outbreaks detected by the German notification system refer to diarrheal diseases. Thus, the M-Eco system should be regarded rather as a complement to the existing system than as a competitor.

In the assessments, we made observations on the characteristics of social media that are relevant for disease surveillance. First, we recognized that the texts that contributed to signals rated as relevant by the epidemiologist often linked to media reports or so-called secondary reports. This experience lets us conclude that there might be a trend in social media whereby users tend to write less often about their personal specific symptoms, but most often forward information from reliable sources such as news sites, or as our experiment showed here, HIV prevention efforts from authorities.

Second, most signals were generated from Twitter data. The volume of relevant Twitter data that is processed by the system is much higher than from any other source considered as input. We did not yet assess to what extent the other sources (blogs, forums, TV, radio) contributed to signal generation.

Beyond, the coverage of social media is still an open issue. It is unclear who is providing relevant health information via social media, which age groups, personal background of persons might play a role, geographic coverage etc. Another challenge
is the quality of content collected from social media and the difficulty to automatically decide whether it is a real outbreak or not. Many of the social media texts present vague reports of illnesses. This means that technology is in principle ready for monitoring social-media data for disease surveillance purposes. However, there are still some open questions to address as they were mentioned before. Further, we did not yet considered the legal and ethical issues related to the use of data from TV/radio and social-media for public health surveillance as well as the reliability of the collected information. Before implementing a tool for monitoring social media and TV / radio such as M-Eco into disease surveillance workflows, those questions need to be addressed.

Acknowledgment

This work was done within M-Eco project partly funded by the European Commission under 247829. We acknowledge all technical partners developing the M-Eco system that performs the basis for our evaluation.

References