Information Processing and Management 56 (2019) 1166–1184

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification

Ovidiu Şerban*,[1],[a], Nicholas Thapen*,[1],[a], Brendan Maginnis[a], Chris Hankin[a], Virginia Foot[b]

[a] Institute for Security Science and Technology, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[b] The Defence Science and Technology Laboratory (DSTL), Porton Down, Salisbury SP4 0JQ, UK

ARTICLE INFO

Keywords:
Real-time processing
Classification
Clustering
Event detection

MSC:
68T50
68T05
62M20

ABSTRACT

Interest in real-time syndromic surveillance based on social media data has greatly increased in recent years. The ability to detect disease outbreaks earlier than traditional methods would be highly useful for public health officials. This paper describes a software system which is built upon recent developments in machine learning and data processing to achieve this goal. The system is built from reusable modules integrated into data processing pipelines that are easily deployable and configurable. It applies deep learning to the problem of classifying health-related tweets and is able to do so with high accuracy. It has the capability to detect illness outbreaks from Twitter data and then to build up and display information about these outbreaks, including relevant news articles, to provide situational awareness. It also provides nowcasting functionality of current disease levels from previous clinical data combined with Twitter data.

The preliminary results are promising, with the system being able to detect outbreaks of influenza-like illness symptoms which could then be confirmed by existing official sources. The Nowcasting module shows that using social media data can improve prediction for multiple diseases over simply using traditional data sources.

1. Introduction

Interest in syndromic surveillance based on social media data has greatly increased in recent years (Charles-Smith et al., 2015; Paul et al., 2016). Many more such data sources, such as Twitter, have become available due to the massive growth in social media usage (Greenwood, Perrin, & Duggan, 2016). In addition the development of distributed and parallel technologies and modern Machine Learning frameworks have provided a good foundation for real-time data processing (Wu, Zhu, Wu, & Ding, 2014).

This paper is built upon a flurry of recent work in the field of syndromic surveillance through social media. The publication of Google Flu Trends (GFT) in 2009 (Ginsberg et al., 2009) was a landmark in digital disease detection. GFT demonstrated that data not collected for public health purposes could be a useful addition to traditional health data analysis. In recent years social media, especially Twitter data, has been used to positive effect for: disease tracking (Collier, Son, & Nguyen, 2011; Culotta, 2010; Lamb, Paul, & Dredze, 2013; Lampos & Cristianini, 2010; Lampos, De Bie, & Cristianini, 2010), outbreak detection (Aramaki, Maskawa, &

* Corresponding author.
E-mail addresses: o.serban@imperial.ac.uk (O. Şerban), nicholas.thapen@imperial.ac.uk (N. Thapen), b.maginnis@imperial.ac.uk (B. Maginnis), c.hankin@imperial.ac.uk (C. Hankin), vefoot@dstl.gov.uk (V. Foot).
[1] O. Şerban and N. Thapen contributed equally to this work.

https://doi.org/10.1016/j.ipm.2018.04.011
Received 31 May 2017; Received in revised form 31 January 2018; Accepted 28 April 2018
Available online 01 June 2018

Morita, 2011; Bodnar & Salathé, 2013; Diaz-Aviles, Stewart, Velasco, Denecke, & Nejdl, 2012; Li & Cardie, 2013) and predicting the likelihood of individuals becoming ill (Sadilek, Kautz, & Silenzio, 2012). News media has also been used to give early warning of increased disease activity before official sources have reported (Brownstein, Freifeld, Reis, & Mandl, 2008).

This project presents a software system, SENTINEL, which extends the previous version of the system (DEFENDER (Thapen, Simmie, Hankin, & Gillard, 2016)) by focusing on real-time processing and improved health classification and denoising using deep neural networks.

The system ingests social, news and clinical sources and internally performs data extraction, transformation, aggregation and statistical analysis. We provide three main disease surveillance applications:

- Early warning detection (EWD): To provide advance warning of potential health events.
- Situational awareness: To provide contextual information about potential health-related events that may have occurred.
- Nowcasting: To provide predictions of disease levels which incorporate data from current social media activity.

Early warning is provided by running a bio-surveillance outbreak detection algorithm over the time-series of symptomatic tweets for each location. This detects possible outbreak events. Tweet ranking and retrieval of relevant news articles provide situational awareness for each detected event. Nowcasting is a type of forecasting where one attempts to predict the current but still unknown level of a time series (Lampos & Cristianini, 2012). For example, CDC notifiable disease reports[2] are published with a one to two week lag time, so providing an estimate of the current disease level before a report is released can be valuable (Eysenbach, 2009). We employ a model that combines previous CDC disease data and Twitter data to improve nowcasting performance over a model solely incorporating the CDC data.

Overall, the current system achieves a better performance than the previous version, while processing all the data in real-time. It processes an average of 1.8 million tweets per day in normal usage on a single machine, and has the ability to process 90 million per day if more data is available.

2. Related work

When looking at event detection using Twitter various approaches have been attempted. These have included searching for spatial clusters in tweets (Nagar et al., 2014; Walther & Kaisser, 2013), leveraging the social network structure (Aggarwal & Subbian, 2012), analysing the patterns of communication activity (Chierichetti, Kleinberg, Kumar, Mahdian, & Pandey, 2014) and identifying significant keywords by their spatial signature(Abdelhaq, Sengstock, & Gertz, 2013). More recently interesting approaches have been described for multi-scale event detection of spatio-temporal events using a Wavelet transform (Dong, Mavroeidis, Calabrese, & Frossard, 2015) and for fusing data from multiple social networks in order to increase confidence in event detection (Peña-Araya, Quezada, Poblete, & Parra, 2017). Eyewitness (Krumm & Horvitz, 2015) is another event detection system which detects anomalies in time-series of tweets from localised areas at differing temporal and spatial resolutions.

The idea of real-time Twitter data processing has been exploited in the past for Earthquake Reporting (Sakaki, Okazaki, & Matsuo, 2013). This system treated users mentioning earthquakes as sensors, using a particle filter to determine the earthquake epicentre. It was tested against notifications delivered by the Japan Meteorological Agency (JMA), and managed to warn users faster than the JMA's reporting systems. Jasmine (Watanabe, Ochi, Okabe, & Onai, 2011) is another system that focuses on local event detection based on geolocated information propagated on microblogging platforms. Their approach focuses on real-time and location disambiguation for tweets without any location information. Recently the Indiana University Network Science Institute has developed OSoMe (Davis et al., 2016), an open analytics platform designed to facilitate computational social science. This is a distributed real-time processing system built on Apache Hadoop and HBase that leverages a collection of over 70 billion tweets. It provides apps for displaying temporal and geographical diffusion of information across the social network, along with visualisations of the network.

Several software systems which detect events from Twitter and provide visualisation and situational awareness capabilities have been created in recent years. TwitInfo (Marcus et al., 2011) identifies events by finding spikes in the number of tweets mentioning keywords and provides timelines and maps for visualisation. LeadLine (Dou, Wang, Skau, Ribarsky, & Zhou, 2012) provides similar visualisation capabilities while incorporating topic modelling and named entity recognition. Twitris (Sheth et al., 2014) is a comprehensive platform with real-time processing built on Apache Storm, designed to enable spatio-temporal analysis of events on Twitter, including sentiment analysis, incorporation of associated news and Wikipedia content, friend-follower network information and sentiment analysis. Systems focused on disease include Lee, Agrawal, and Choudhary (2013), and Ji, Chun, and Geller (2012), both of which use simple keyword based techniques to identify health-related tweets from Twitter's streaming API and display geo-temporal trends visually. The HealthTweet (Dredze, Cheng, Paul, & Broniatowski, 2014) system extends these by using a statistical classifier to identify those tweets which are truly health-related.

In contrast to these systems SENTINEL examines multiple symptoms and diseases, and uses a more sophisticated classifier using deep neural networks to identify those tweets which are truly health-related. It is built in a modular way and linked together using Apache Kafka (Kleppmann & Kreps, 2015), a publish-subscribe scalable messaging service which allows for an extremely high throughput and low latency.

When looking at nowcasting of disease data using social media various approaches have been employed. Paul, Dredze, and

[2] Please refer to Section 4.3 for a full description of this CDC data.

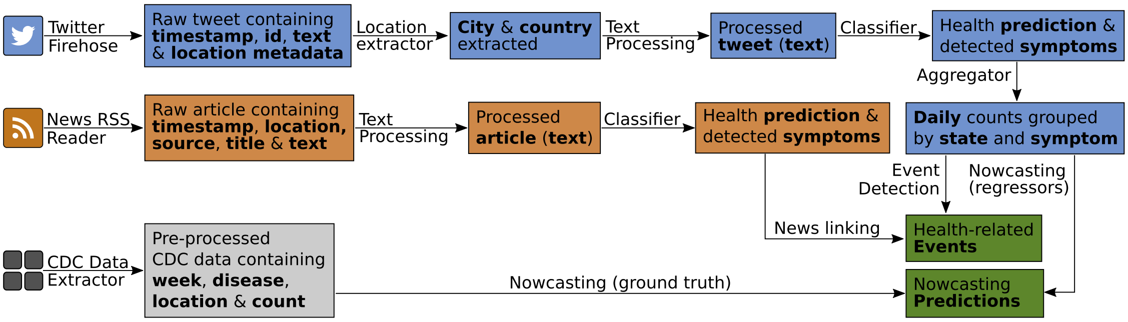

Twitter Firehose
Raw tweet containing **timestamp**, **id**, **text** & **location metadata**
Location extractor
City & country extracted
Text Processing
Processed **tweet (text)**
Classifier
Health **prediction** & detected **symptoms**
Aggregator
News RSS Reader
Raw article containing **timestamp**, **location**, **source**, **title** & **text**
Text Processing
Processed **article (text)**
Classifier
Health **prediction** & detected **symptoms**
Daily counts grouped by **state** and **symptom**
Event Detection
Nowcasting (regressors)
News linking
Health-related **Events**
CDC Data Extractor
Pre-processed CDC data containing **week**, **disease**, **location** & **count**
Nowcasting (ground truth)
Nowcasting **Predictions**

Fig. 1. A data integration diagram, showing the transformation process happening within SENTINEL.

Broniatowski (2014) have shown that including Twitter data improves nowcasting performance to a greater degree than Google Flu Trends data. They focus on influenza-like illness (ILI) data from the CDC, using a linear autoregressive model and incorporating a weekly estimate of the influenza rate derived from Twitter data using the software developed by Lamb et al. (2013). Other studies applying Twitter data to influenza forecasting or nowcasting include Culotta (2010), Li and Cardie (2013), Sadilek, Kautz, and Silenzio (2012), and Santos and Matos (2014), and indeed a literature review on this topic has identified influenza as by far the most popular disease analysed using social media (Charles-Smith et al., 2015). Our approach applies similar statistical techniques to these, but we extend from influenza to a range of other illnesses reported by the CDC. Another study combining multiple sources of data for influenza nowcasting is Santillana et al. (2015), who employ a variety of statistical machine learning techniques in order to achieve excellent results. Other conditions that have been studied using Twitter data include allergies (Lee, Agrawal, & Choudhary, 2015) and gastro-intestinal disorders (Sadilek, Brennan, Kautz, & Silenzio, 2013).

3. System overview

SENTINEL ingests data from multiple sources in order to provide its disease surveillance applications. Fig. 1 shows a simplified work-flow of the transformation and fusion of these different data sources, and this section gives an overview of the system's operation. All of the components performing these data transformations are fully detailed in Section 6.

SENTINEL's Event Detection functionality monitors the Twitter stream, classifying those tweets which are self-reports of illness and storing them. It then creates a daily count of tweets mentioning each monitored symptom for each US state. The CDC's Early Aberration Reporting System (EARS) (Hutwagner, Thompson, Seeman, & Treadwell, 2003) algorithm is used to detect unusual spikes in these daily symptom time-series. Each such spike leads to the generation of an event in the system, which can then be enriched with more data to provide better situational awareness. The news feed processes articles, saving them if they are health-related and tagging them with mentioned symptoms. Those articles which match the location, symptom and date of an event are associated with it.

The Nowcasting functionality ingests disease data from the CDC, which is provided weekly for each US state, but with a 1–2 week lag. For each disease and state it then creates a weekly nowcast using the previous CDC data combined with the additional regressors provided by the Twitter symptom data.

The outputs of the Event Detection and Nowcasting functionality, namely the generated events and the nowcasting predictions, are then shown to the user in the Front-end UI, which runs as a Javascript app.

4. Data acquisition & management

The most important data source used by SENTINEL is Twitter data. It is used as the basis of the Event Detection system as well as being an input to the Nowcasting algorithm. The Nowcasting algorithm combines the previous weeks' CDC data with the up to date daily aggregated counts of Twitter data to forecast the current level of disease, while the News data is linked to the events and displayed in the Front-end to increase confidence in the accuracy of the information.

4.1. Twitter data

The primary data source for SENTINEL is social media, specifically US Twitter data, because of its desired characteristics:

- Timely: tweets are received within seconds of their creation.
- High coverage: 24% of online adults in the US use Twitter, equating to 21% of the US population Greenwood et al. (2016).
- Publicly available: tweets for geographical areas or filtered by keyword sets are available without requiring explicit permission from the post author unless the author has flagged their account as private. Around 5% of users do so Liu, Kliman-Silver, and Mislove (2014), meaning that the vast majority of tweets are available for research.
- Localised: tweets can provide a fine-grained location estimate for an individual if they have opted into that service. Leetaru, Wang,

Table 1
Demographic breakdown of Twitter users in the United States as of April 2016 (Greenwood et al., 2016). Figures shown are the percentage of online adults in each category who use Twitter.

Category	Subcategory	Users (%)	Category	Subcategory	Users (%)
All online adults		24	Gender	Men	24
				Women	25
Age	18–29	36	Income	< $30K/year	23
	30–49	23		$30K–$49,999	18
	50–64	21		$50K–$74,999	28
	65+	10		$75K+	30
Education	High school or less	20	Living area	Urban	26
	Some college	25		Suburban	24
	College+	29		Rural	24

Cao, Padmanabhan, and Shook (2013) found that 1.6 percent of users opt in to geo-locating their Twitter posts. Sloan and Morgan (2015) found similar results for the UK.

Although the benefits of using social media data for this purpose are substantial there are several disadvantages to using this non-curated source:

- Noise: tweets referring to potential illness terms may have nothing to do with health. For example high levels of fever activity may be caused by posts containing the term “Bieber Fever”.
- Low confidence: health related Twitter data is of varying quality. For example, a user may report that they have the flu when actually they have a common cold or people may be discussing a disease such as scarlet fever due to increased media hype.
- Demographic bias: A recent demographic breakdown of American Twitter users is provided in Table 1. The strongest bias is that Twitter is used more commonly by younger people, with 36% of online 18–29 year olds using the platform as opposed to only 10% of those in the 65+ bracket. It also shows that the college educated and those with higher earnings are somewhat more likely to be Twitter users.

All of these disadvantages are addressable. To eliminate the noise problem a health related tweet classifier has been implemented into the system pipeline, only storing count data for tweets related to health. This classifier also partially addresses the second problem, since we attempt to single out only those tweets which are self-reports of a symptom of illness. The classifier is discussed in more detail in Section 6.1.4. The principal way in which we address the low confidence problem is by including additional data from news sources. News articles about a symptomatic event may confer more confidence that a Twitter event is a real health concern, or the temporal dynamics of the event may suggest that the story broke first in the media and is now being propagated through social media as a result. The bias disadvantage is partially resolved by the fusion of multiple data sources. Concerns about demographic bias are important, but these do also apply to many other currently used methods of syndromic surveillance. Participatory studies such as Influenzanet (Guerrisi et al., 2016) only capture a self-selected sample of those who sign up. People who do not visit doctors will not appear in clinical reports such as CDC data, and Google Flu Trends (Ginsberg et al., 2009) only observes those who use this search engine. A diversity of methods is required to capture all segments of the population. As long as the demographic bias of the Twitter data towards younger, richer, college educated individuals is understood from studies such as the Pew report cited above, information derived from it can be useful in a clinical context.

Tweets for the system are collected via Twitter’s live streaming API, using a geographical bounding box encompassing the contiguous 48 US states. We use Twitter’s *hosebird* HTTP streaming library to connect to the API.

4.2. News data

News data is used by SENTINEL for a different purpose than the social media data. Unlike systems such as HealthMap (Brownstein et al., 2008) news reports are not mined independently for health outbreaks. The articles are instead used as a secondary source to add or remove confidence from social media events. Our methodology for linking social media and news data together is detailed in Section 6.2.3.

News data is collected on three different levels:

- World health related news sources, such as ProMed and the World Health Organization News letter. These sources are not localized to the US, but most of their articles and alerts provide the location of the article as part of the RSS metadata information.
- US National news sources, such as CNN, NY Times, USA Today, Chicago Tribune, Reuters, Wall Street Journal. Most of these websites provide a separate health related category on their RSS Feed. In total, 19 national news sources are covered, providing 51 RSS feeds.
- US Regional and State level news sources. These RSS feeds were automatically crawled from various community-based platforms and grouped by the states they cover. In total, 16,803 RSS Feeds are crawled.

4.3. CDC data

The CDC data is used by SENTINEL as an input to the Nowcasting functionality and also as a source of ground truth for evaluation. The main source of official data comes from the Morbidity and Mortality Weekly Reports (MMWR) provided by the National Notifiable Diseases Surveillance System (NNDSS) as part of the Centers for Disease Control and Prevention (CDC) System.[3] The data is published on a weekly basis, with 1 week delay. The reports are available through the Open Data initiative in the US Government, via the newly created Socrata Open Data API (SODA). These reports provide weekly state wide counts of individuals presenting to clinicians with one of the notifiable diseases.

Even though the reports are freely available in various formats (JSON, CSV, etc.), the data have not been normalized or cleaned up, making them difficult to use with the Nowcasting algorithms. Therefore a few simple techniques to clean, remove duplicates and normalize the data have been employed.

The CDC Influenza (Flu) reports are not published by the SODA API. These can be downloaded manually from the FluView app, consisting of laboratory confirmed influenza hospitalizations, available from the Emerging Infections Program (EIP) in 10 US states and Influenza Hospitalization Surveillance Project (IHSP) covering 8 US states. Unfortunately, only California, Colorado, Connecticut, Georgia, Maryland, Minnesota, New Mexico, Oregon and Tennessee full influenza reports were available from EIP during the 2016–2017 season, and Michigan, Ohio and Utah from IHSP. The other 36 contiguous states do not have these reports in a standardized (easy to process) format.

During the detailed event evaluation, some of the events were validated manually using reports available on the state level, in different formats, without having access to the raw data. These are recovered from the Weekly US Influenza Surveillance Reports (https://www.cdc.gov/flu/weekly/).

The weekly counts of individuals affected by a notifiable disease obtained from official data sources are henceforth referred to as 'CDC counts'.

4.4. Data characteristics

As of the time this paper was produced the following amount of data has been collected (between 20 June 2016 and 02 March 2017):

- 466,896,997 tweets, approximately 1.8 million per day on average, with peak days seeing 2.2 million tweets received.
- 2,669,235 news articles, around 18,000 daily on a regular month, 40,000 during the period of the US presidential election
- 49 CDC reports[4] were collected on a weekly basis (52 weeks) for 54 US locations.[5]

The CDC Data is published on a weekly basis and for the scope of our work the collection started in the beginning of 2016. We carried out all evaluations on the above June-March time period where all of our data overlapped. Table 2 shows the average and maximum number of confirmed cases for specific diseases retrieved for all US contiguous states. The CDC published a list of probable cases for some diseases, plus subsets for various age ranges and variants. The full list of diseases used by our Nowcasting model is available in Table A.6.

The only exception to the CDC data collection protocol is Influenza due to its seasonality: it starts in week 40 and ends in week 17 of the next year. Moreover, the data is published only for 12 US states: California, Colorado, Connecticut, Georgia, Maryland, Minnesota, New Mexico, Oregon, Tennessee, Michigan, Ohio and Utah. All the missing reports for Influenza, mainly outside the data collection periods, were assumed to be equal to zero.

5. System architecture

The system comprises a back-end architecture that ingests tweets and other data sources and processes them, and a web front-end to display the results. The system architecture uses a data centred approach, built around lock-free pipelines with reusable components that are combined to transform the input into a desirable format. This design allows each component to be simple and efficient. Due to the large amounts of data and multiple processing steps involved in the system, many tasks are run in the background rather than at user request.

The back-end engine is composed of various data processing pipelines interfacing with the communication and integration library (Apache Kafka) and various storage engines. Inputs to the back-end engine are from the Twitter API, RSS feeds of news sites and the CDC SODA API. The back-end then outputs into a front-end data store (PostgreSQL). The front-end runs as an HTML5 and React.js JavaScript app. Fig. 2 shows the interaction between various components of the system, also focusing on the processing schedule for each pipeline. There are three processing types for the SENTINEL data processing pipeline: real-time, daily and weekly. Real-time data is collected and processed within seconds or less of its conception. The scheduled processing runs daily or weekly, depending on the publishing patterns of the data sources.

[3] https://www.cdc.gov/mmwr/index.html.

[4] The 49 CDC reports include 32 distinct diseases and 17 reports of variants of these diseases.

[5] The CDC publishes reports for the 53 US states and territories including Guam, Puerto Rico and US Virgin Islands. An additional report for the whole country.

Table 2
Statistics of CDC confirmed cases for specific diseases.

Disease	Confirmed cases		Disease	Confirmed cases	
	Avg.	Max.		Avg.	Max.
Babesiosis	0.36	72	Campylobacteriosis	17.13	770
Chlamydia Trachomatis Infection	513.14	18,367	Coccidioidomycosis	5.21	359
Cryptosporidiosis	3.44	391	Dengue	0.00	2
Giardiasis	4.10	246	Gonorrhea	154.71	6,088
Haemophilus Influenzae Invasive	1.40	107	Hepatitis	0.53	52
Invasive Pneumococcal Disease	5.53	305	Legionellosis	1.72	117
Lyme Disease	6.54	635	Malaria	0.46	43
Meningococcal Disease	0.10	10	Mumps	1.50	216
Pertussis	4.25	205	Rabies Animal	1.15	75
Rubella	0.00	2	Salmonellosis	13.16	731
Shiga Toxin	1.46	113	Shigellosis	4.83	294
Spotted Fever	0.56	110	Syphilis	3.41	151
Tetanus	0.01	2	Varicella	2.62	123
Vibriosis	0.35	33	West Nile	0.03	13
*Influenza	219.69	6,580			

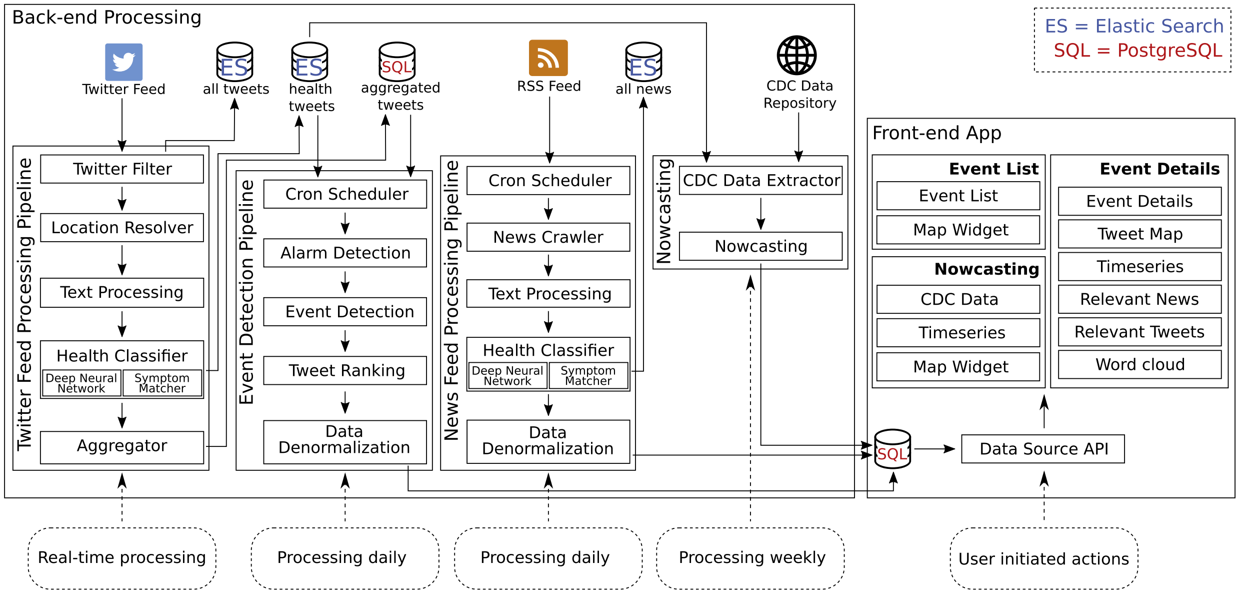


Fig. 2. System architecture.

The processing pipelines split a complex problem into smaller, more manageable parts. These increase the re-usability of these components, some with different parameters or models. One such example is the Text Processing component, used both in the Twitter and News pipelines. The Health Classifier component runs the same code in both pipelines, but the underlying models and parameters are different.

The data processing architecture is stateless, allowing an elastic configuration of resources, by adding or removing components based on demand. This is one of the key roles of Apache Kafka which provides a watermark feature for all message queues, given by each topic and subscriber group combination. Components within the same group will have the message id load balancing feature active.

Dealing with heterogeneous data sources in an optimal way is always difficult (Halevy, Rajaraman, & Ordille, 2006). Storing the data without a major impact on performance, both on update and retrieval, is sometimes very complex. For this system, we use multiple storage engines and strategies, each tailored to specific requirements. Data in numeric tabular formats are stored in PostgreSQL and text documents are inserted into an Elastic Search Index. This offers major benefits on the query strategies available, such as retrieving similar documents. The processing queues are stored by Kafka to allow better load balancing and reply strategies.

In terms of performance, the system regularly processes 1.8 million tweets and 18,000 news articles per day. At its peak, when re-processing all data from scratch, the system achieved a top performance of approximately 90 million tweets per day on a single machine.

6. Back-end: components and algorithms

Splitting the pipelines into smaller components ensures that each one is manageable and has a single responsibility. The processing is split into the Twitter and News Processing Pipelines, the Event Detection Pipeline and the Nowcasting pipeline. This section details each of the components in the system.

6.1. Twitter and news processing pipelines

The Twitter and News Processing Pipelines share most of their components, with the exception that the News pipeline does not require the Location Resolver or Aggregator components. The other difference between the pipelines is in the underlying models and parameters, which are adapted to each domain. Both pipelines ingest textual data, pre-process it and then tag it with metadata such as location or detected symptoms. They then determine if the text is health-related, and if so store it. In addition the Twitter pipeline aggregates the data by symptom, date and location to provide time-series data which can be fed into the Event Detection and Nowcasting pipelines.

6.1.1. Location resolver

The Location Resolver attempts to resolve each tweet's location metadata to a uniform format. In this version of the system, an assumption has been made that the location metadata is correct and validated by Twitter.[6] Nevertheless, depending on the user's privacy settings, the location can be very precise as the exact address (e.g. Mitchell Street, Milwaukee, WI, USA), state or a Point of Interest (POI) (e.g. Manhattan or Statue of Liberty). In all these cases the bounding box of the location is provided by the API. The Location Resolver attempts to translate the string provided in Twitter's metadata into a city and state code. When not possible, (i.e. for a POI), the bounding box of the objective is checked against the state (or city) border using well known city bounding box databases.[7]

6.1.2. Text processing

This component converts raw text into easily processable word tokens suitable for use by our machine learning algorithms. We pre-process the text to:

1. Remove links, email addresses and mentions.
2. Translate html entities (e.g. &nbsp; becomes the space character).
3. Translate emojis and emoticons into their name, according to a dictionary of well-known web emoji (emoticon, n.d.) and ASCII emoticons (gemoji, n.d.).
4. Quoted words are unquoted and prefixed with *quote_* (For example *cough* and "cough" are replaced with quote_cough). This was implemented because words quoted in this way often denote sarcasm.
5. Hashtags are split into the individual words, by applying two different strategies:
   (a) For hashtags written in Camel-Case scripting notation, the words are split according to the case rules.
   (b) Otherwise a prefix-based space prediction algorithm (Aho-Corasick (Aho & Corasick, 1975)) is used to split the hashtag into the minimum possible number of words.
6. All punctuation and excess spaces are removed. Finally, text is converted to lower case.

The semantic hashtag splitting task is the most complex text processing step and has been a research focus in itself (Bansal, Bansal, & Varma, 2015). In our work the problem is simplified since the hashtags are not used directly for the event tracking, but are split into their constituent words and added to the processed text. The Aho–Corasick (Aho & Corasick, 1975) word splitting algorithm was chosen due to its speed in real-time systems (Tumeo, Villa, & Chavarria-Miranda, 2012). This algorithm is biased towards prefixes that form valid longer words when parsing is ambiguous (e.g. *superbowl* will be parsed as a single valid word instead of *superb owl*). In practice we found this to be an advantage since these longer words better captured the intended semantics of the hashtags.

6.1.3. Assigning symptoms

In order to determine which tweets and articles show symptoms of illness we initially employ a keyword matching technique. The process of building up the keyword set is described in a previous work by the authors (Thapen, Simmie, & Hankin, 2016). It is based on a combination of the Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008) ontology's /medicine/symptom tag (Freebase has subsequently been shut down so this is no longer generally available), symptom terms found on the CDC website[8] and manual revision which results in a group of keywords describing each symptom (including common aliases and synonyms for these words). In this work we further enriched these synonym lists by using the word embeddings trained on the Twitter data (described below). For each symptom keyword we generated a list of the 10 closest words in the embedding space by cosine similarity, and manually added

[6] When using a location based filter, the Twitter Firehose API streams only the tweets containing location information.

[7] The state bounding boxes were extracted from http://www.mapdevelopers.com/ and https://www.openstreetmap.org/.

[8] www.cdc.gov and in particular the specific symptom pages such as: Influenza (cdc.gov/flu/consumer/symptoms.html), Campylobacter (cdc.gov/campylobacter/symptoms.html), E. Coli (cdc.gov/ecoli/ecoli-symptoms.html), etc.

those that were appropriate synonyms. The final symptom list contains 38 different symptoms, each with an average of 27 synonyms. Each tweet and article is tagged with any symptoms that match any word in the text. The machine learning classifier is then responsible for determining whether text tagged with symptoms is actually health-related or is using these words in non health-related contexts.

6.1.4. *Machine learning classification*

We employ Machine Learning classifiers to identify those tweets and news articles which are genuinely health-related. The keyword matching technique employed using our symptom keywords throws up many tweets and articles which use other senses of the symptom words. For example the term 'headache' can easily crop up in many non health-related contexts. Only health-related tweets and news articles are stored in our databases.

Previous studies in this area have used methods such as Multinomial Naïve Bayes (Lee et al., 2015; Santos & Matos, 2014) and SVMs (Dredze et al., 2014) to identify health-related tweets. In recent years Deep Neural Networks (DNNs) have set new benchmarks in text classification (Kim, 2014) due to their ability to learn complex representations from the textual data. To test their effectiveness in the health classification task on Twitter we implemented two DNN models, a Convolutional Neural Network (CNN) as in Kim (2014) and a Long Short Term Memory Network (LSTM) (Hochreiter & Schmidhuber, 1997), which is a type of Recurrent Neural Network (RNN). These DNNs were implemented using TensorFlow (Abadi et al., 2016), a software library widely used and well supported within the machine learning community. We also implemented an SVM model using the LibShortText toolkit (Yu, Ho, Juan, & Lin, 2013) and a Multinomial Naïve Bayes model using Scikit-learn (Pedregosa et al., 2011) to serve as baselines, both using TF-IDF (Sparck Jones, 1972) feature vectors.

An advantage of DNNs is their ability to leverage unlabelled data as well as labelled data to aid in classification (Mikolov, Chen, Corrado, & Dean, 2013). It is time-consuming to manually annotate more than a few thousand tweets, but easy to collect many millions of unlabelled tweets. Word embeddings such as GloVe (Pennington, Socher, & Manning, 2014) learn vector representations of words from large corpora of text utilising the distributional hypothesis that similar words will appear in similar contexts. In these vector representations similar words should have similar vectors. Using these word embeddings as inputs to neural network models instead of simpler one-hot representations of words has been shown to increase performance on a variety of natural language processing tasks (Turian, Ratinov, & Bengio, 2010). In particular they allow machine learning models to generalise more effectively beyond their limited number of training examples, since similar words not seen in these examples should produce similar classifier outcomes. We experimented with the Glove (Pennington et al., 2014) and FastText (Bojanowski, Grave, Joulin, & Mikolov, 2016) techniques for generating word embeddings, and present our results in Section 8.1.

In order to train our machine learning classifiers we selected 9353 tweets for manual annotation using a stratified sampling method, attempting to select 10 tweets for each of our 1026 symptom keywords (or as many as available if 10 were not present in our dataset.). These were then annotated as being health-related if they were an instance of a user self-reporting an illness, and non-health related otherwise. Hence a tweet merely discussing illness, such as referring to a flu vaccination campaign, was treated as non health-related for our purposes. 29.3% of this training set were found to be health-related. To account for this imbalance, we ensured that each mini-batch during training was sampled to contain equal numbers of health-related and non health-related tweets. For the news classifier we employed a different approach to annotation, using a distant supervision method. We took a sample of 5761 articles equally split between those in a health-related RSS feed and those from general feeds. Those articles taken from health feeds were labelled health-related for training purposes, and those from general feeds non-health related.

Our CNN uses 128 filters that act on 3 words, 128 filters that act on 4 words and 128 filters that act on 5 words (parameters chosen for their success in Kim, 2014). This produces a total of 384 features which are fed into a final logistic regression function which produces the final classification result. Our RNN uses two LSTM layers, the first of 128 neurons and the second of 256. For regularisation we use dropout on both models with a probability of 0.5, and for the CNN an additional L2 regularisation term with a factor of 0.01. When training word vectors on our Twitter corpus we trained 300 dimensional models on 269,544,449 tweets. Our models were trained on a server running Ubuntu Linux 16.10, with a 48 core CPU, 256GB of RAM and 2 NVidia 1060 GPU cards.

The details of the training and test regimen used are presented in the evaluation in Section 8.1.

6.1.5. *Aggregator*

The Aggregator is a real-time batch processing component. Its input is a stream of the health-related tweets identified by the Machine Learning Classifier and is made up of the original tweet text, the publication date, location and detected symptoms. The algorithm counts the tweets matching each symptom in a given time window, for the specific location. The date, symptom and location will produce a unique aggregation key, used by our event tracking components.

In our current experiments, the time window is one day, but this could be easily adjusted to other frequencies, such as hourly or every 5 minutes, if needed. Each tweet is counted for every detected symptom, when multiple symptoms are considered and at city, state and country level. The Aggregator publishes a database update for the new counts and a notification event when an update is available, which all the updated aggregation keys. The aggregated output is used by the event detection algorithm.

6.2. *Event detection pipeline*

The Event Detection Pipeline ingests the aggregated tweet counts and executes the EARS algorithm to detect relevant events. Each event is then processed to determine the most relevant tweets related to it. Relevant news articles are also linked to each event.

6.2.1. Event detection

The Event Detection module uses time-series of symptom count data generated by the Aggregator to create possible outbreak events. It leverages considerable existing syndromic surveillance research by utilising an algorithm designed and developed by the CDC. The primary surveillance algorithm used is EARS (Hutwagner et al., 2003), specifically the C2 and C3 variants of this technique. Details of our adaptation of EARS can be found in an earlier work by the authors (Thapen et al., 2016). One change that we made for the SENTINEL system was to implement our own version of the algorithm in Java to streamline our software stack by reducing the number of language dependencies.

6.2.2. Situational awareness

Once an event has been detected, the event data needs to be enriched with more details useful for a Situational Awareness tool. Firstly, the set of tweets that make up the event are processed. Stop words are removed, and then TF-IDF vectors are generated for all terms remaining. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Sparck Jones, 1972). The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. A document such as a tweet can be represented by a TF-IDF vector, which contains the TF-IDF value for every word in the document.

We generate a word cloud to provide an overview of the tweets in the event. Words in the word cloud are sized according to their raw term frequency in the tweet corpus. Another overview is provided by the selection of the most relevant tweets, using a ranking method described in Thapen et al. (2016) which is similar to that of Zubiaga, Spina, Amigó, and Gonzalo (2012). It involves ranking the tweets by their cosine similarity to the mean vector of the event corpus (using TF-IDF vector representations). The top five tweets ranked by this measure are returned for presentation to the user.

6.2.3. News linking

News articles are linked in to the event if they are health-related and share a symptom, location and were published within two days of the event. The date of the article is taken to be the date on which it was published, and the article text is scanned using our symptom keywords to obtain a keyword match and assign mentioned symptoms. Articles from local newspapers are assigned a location of the US state in which the newspaper operates, whereas national newspaper articles are taken to match any state location. The linked news articles are then displayed to the user in a list in the front-end UI.

6.3. Nowcasting

The Nowcasting pipeline uses LASSO (Least Absolute Shrinkage and Selection Operator) to make its predictions. LASSO is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It was introduced by Robert Tibshirani in 1996 based on Leo Breimans Nonnegative Garrote (Tibshirani, 1996).

In the previous work (DEFENDER (Thapen et al., 2016)) the Nowcasting module used a Mean Absolute Error (MAE) on a cross-validation window of 4 time periods (over 28 days for training).

The previous work dealt with a limited number of symptoms and diseases, whereas for SENTINEL there are a larger number of variables to consider. For this scenario we found that LASSO offered comparable accuracy with much improved performance, training in several minutes as opposed to several hours for the prior method.

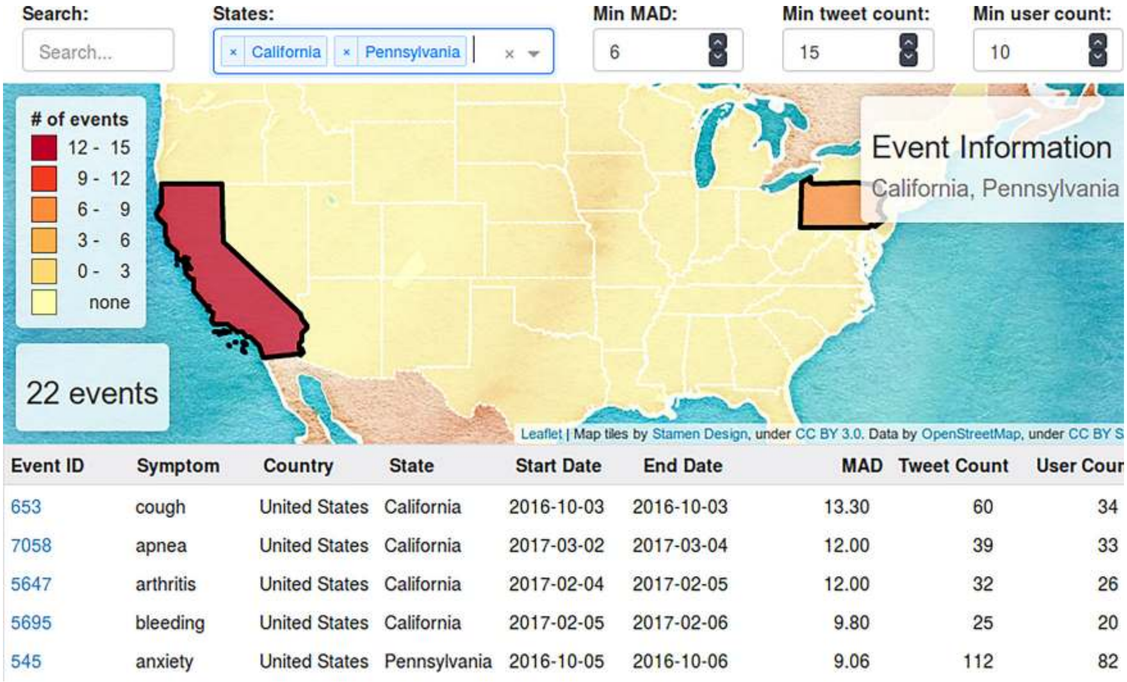
First the weeks where we have both CDC data and Twitter data are selected, and the Twitter data is aggregated into weekly counts for each symptom (from the daily). We take 8 weeks as the minimum training set, so that nowcasts can start to be produced on the 9th week from the start of the coincident data. We define the first week where we have coincident data as t_0. For each CDC disease the predictions for week t_n are made as follows:

- Take the US-wide time series of all Twitter symptoms from t_0... t_{n-1} as regressors.
- Take the CDC counts from t_0... t_{n-1} as the ground truth.
- Feed these into the LASSO model for training.
- This will output the Twitter symptom coefficients y_1... y_n that best fit the model, shrinking most of them to 0.
- Next for each state, take the CDC counts t_0... t_{n-1} for that state, and the Twitter symptom time series y_1... y_n for the same times for that state. Train a LASSO state-specific model using this data.
- Now use this model to predict the CDC count for t_n using the Twitter counts for t_n as regressors.

Selecting the coefficients is done on the US-wide data as this provides the model with the largest volume of data to work with, which should select the best model and be less prone to over-fitting.

7. Front-end: SENTINEL app

The front-end UI is designed as an Early Warning Detection (EWD) and Situational Awareness system, where the user can interact with the data and filter the targeted events. The whole system and UI is not meant to work independently of the user, but aid them in the decision making process and provide a support for data-driven decisions. Fig. 3 shows the list of available events, along some



Search: Search...

States: × California × Pennsylvania

Min MAD: 6

Min tweet count: 15

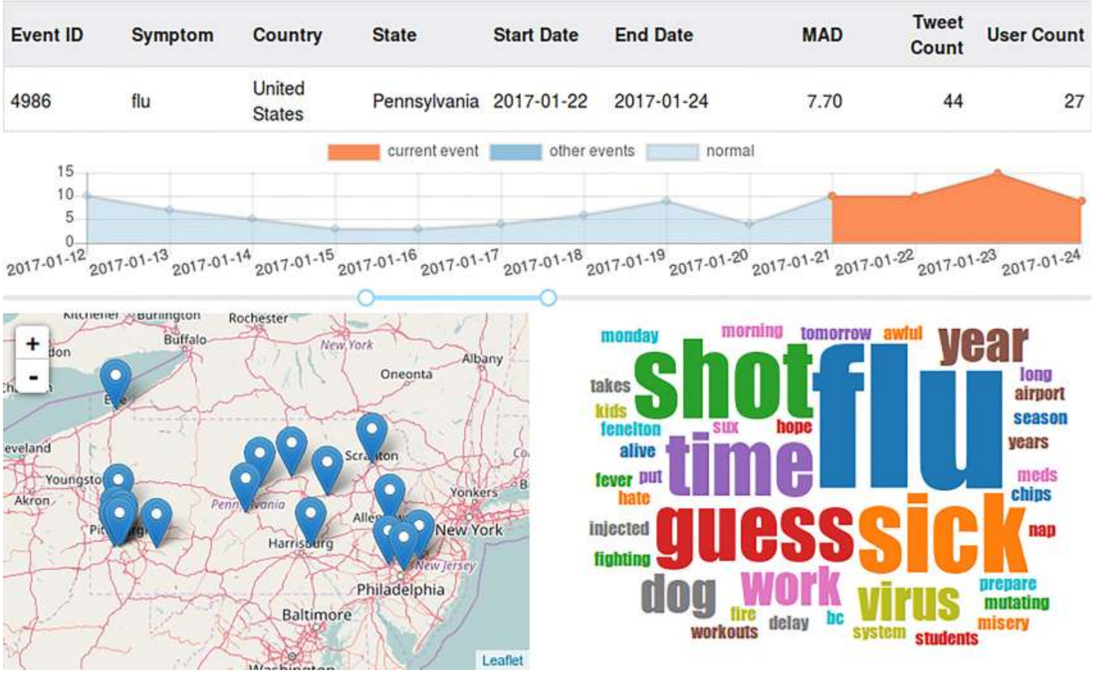
Min user count: 10

Event ID	Symptom	Country	State	Start Date	End Date	MAD	Tweet Count	User Cour
653	cough	United States	California	2016-10-03	2016-10-03	13.30	60	34
7058	apnea	United States	California	2017-03-02	2017-03-04	12.00	39	33
5647	arthritis	United States	California	2017-02-04	2017-02-05	12.00	32	26
5695	bleeding	United States	California	2017-02-05	2017-02-06	9.80	25	20
545	anxiety	United States	Pennsylvania	2016-10-05	2016-10-06	9.06	112	82

Fig. 3. The event list shown in the system.

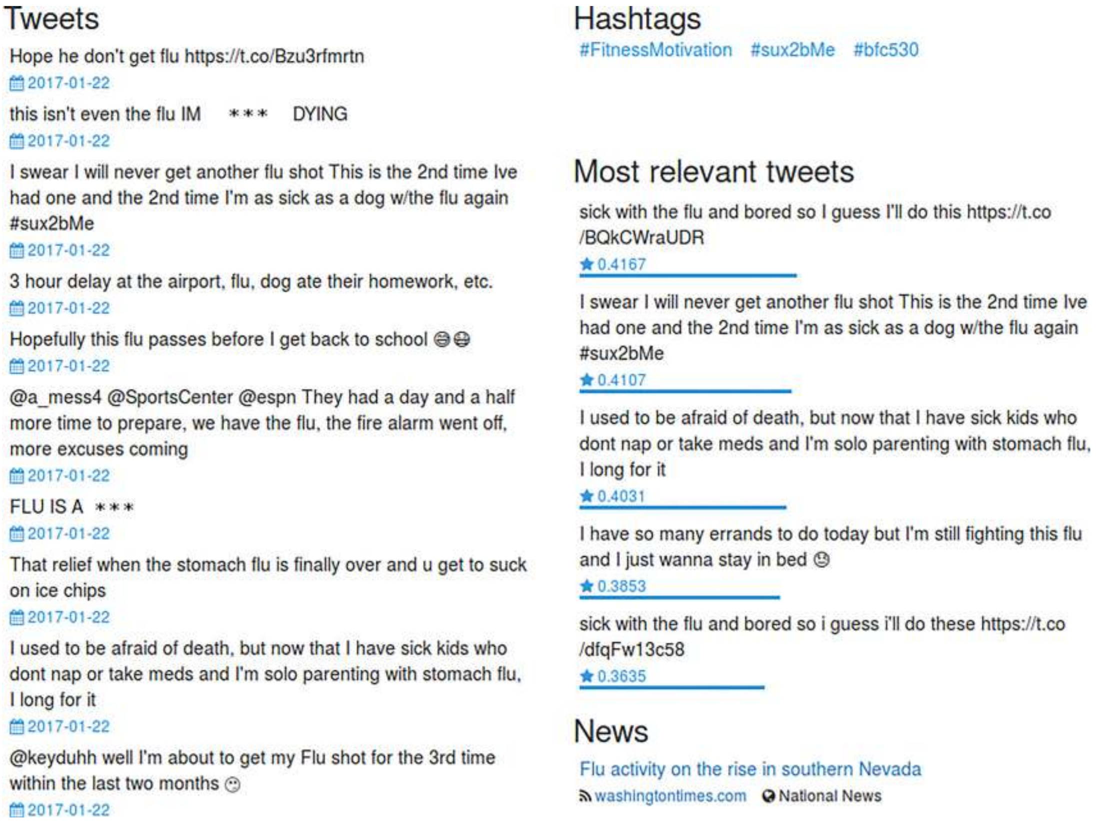
basic data, that can be used for filtering. The number of users, tweets and MAD (Median Absolute Deviation) provide a measure of confidence in the detected events. MAD was chosen as a robust statistic for determining the strength of relative spikes in count-based time series. It can be interpreted similarly to the standard deviation, but is more robust to outliers and non-normal data distributions. The reasoning behind this choice is more fully explored in a previous work by the authors (Thapen et al., 2016).

Figs. 4 and 5 show the Situational Awareness screen and put the event into context, by compiling a list of important details related to the event data: such as the hashtags used in event tweets, the list of all tweets, the list of tweets found to be most relevant and the linked news.



Event ID	Symptom	Country	State	Start Date	End Date	MAD	Tweet Count	User Count
4986	flu	United States	Pennsylvania	2017-01-22	2017-01-24	7.70	44	27

Fig. 4. The Situational Awareness page in the system - top half.



Tweets

Hope he don't get flu https://t.co/Bzu3rfmrtn
📅2017-01-22

this isn't even the flu IM *** DYING
📅2017-01-22

I swear I will never get another flu shot This is the 2nd time Ive had one and the 2nd time I'm as sick as a dog w/the flu again #sux2bMe
📅2017-01-22

3 hour delay at the airport, flu, dog ate their homework, etc.
📅2017-01-22

Hopefully this flu passes before I get back to school
📅2017-01-22

@a_mess4 @SportsCenter @espn They had a day and a half more time to prepare, we have the flu, the fire alarm went off, more excuses coming
📅2017-01-22

FLU IS A ***
📅2017-01-22

That relief when the stomach flu is finally over and u get to suck on ice chips
📅2017-01-22

I used to be afraid of death, but now that I have sick kids who dont nap or take meds and I'm solo parenting with stomach flu, I long for it
📅2017-01-22

@keyduhh well I'm about to get my Flu shot for the 3rd time within the last two months
📅2017-01-22

Hashtags

#FitnessMotivation #sux2bMe #bfc530

Most relevant tweets

sick with the flu and bored so I guess I'll do this https://t.co/BQkCWraUDR
★0.4167

I swear I will never get another flu shot This is the 2nd time Ive had one and the 2nd time I'm as sick as a dog w/the flu again #sux2bMe
★0.4107

I used to be afraid of death, but now that I have sick kids who dont nap or take meds and I'm solo parenting with stomach flu, I long for it
★0.4031

I have so many errands to do today but I'm still fighting this flu and I just wanna stay in bed
★0.3853

sick with the flu and bored so i guess i'll do these https://t.co/dfqFw13c58
★0.3635

News

Flu activity on the rise in southern Nevada
washingtontimes.com National News

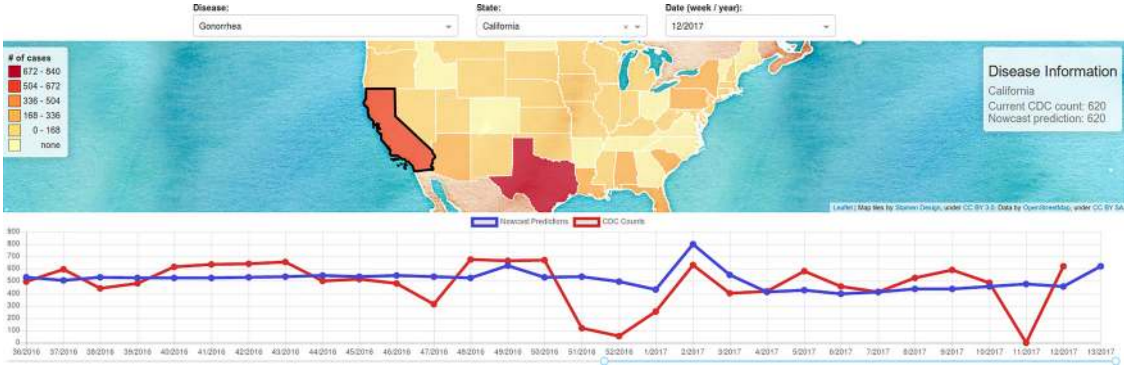
Fig. 5. The Situational Awareness page in the system - bottom half.

The Nowcasting screen, shown in Fig. 6, presents the predictions made for a specific disease, based on existing CDC data and detected symptom-based counts. The data can be navigated by date, location and disease.

8. Evaluation

SENTINEL outputs a range of results, so multiple evaluation protocols had to be employed:

1. Classifier evaluation: testing the efficacy of our Machine Learning models by evaluating their performance on human annotated data;
2. Evaluation of the EWD: performed to assess the accuracy of the outbreak detection algorithm against existing data sources;
3. Nowcasting model evaluation: tests the accuracy of the prediction against the CDC Data.



Disease: Gonorrhea

State: California

Date (week / year): 12/2017

Disease Information
California
Current CDC count: 620
Nowcast prediction: 620

Fig. 6. The Nowcasting screen showing predictions for Gonorrhoea, in California on week 12 (2017).

Table 3
The accuracies and F1 scores for the Twitter classification models.

Corpora	Classification model	Feature model	Accuracy	F1
Twitter	Naïve Bayes	TF-IDF	0.780	0.735
Twitter	SVM	TF-IDF	0.823	0.670
Twitter	RNN	GloVe	0.828	0.831
Twitter	RNN	FastText	0.850	0.850
Twitter	CNN	GloVe	0.836	0.833
Twitter	CNN	FastText	0.854	0.852

Table 4
The accuracies and F1 scores for the News classification models.

Corpora	Classification model	Feature model	Accuracy	F1
News	Naïve Bayes	TF-IDF	0.832	0.814
News	SVM	TF-IDF	0.901	0.828
News	CNN	GloVe	0.934	0.934
News	CNN	FastText	0.939	0.939
News	RNN	FastText	0.940	0.940

8.1. Classifier evaluation

For the classification evaluation, accuracy and F1 were employed as standard measures since they are widely used in Machine Learning applications (Sokolova & Lapalme, 2009). F1 is weighted according to the class weights.

The size of our annotated corpora was 9353 samples for the Twitter task and 5761 articles for the News task. We performed the evaluation using 10-fold cross validation using a split of 80% of data for training and 20% held out for testing.

Tables 3 and 4 show the results for the two tasks. It can be seen that the neural networks outperformed the baseline methods in all cases. For both tasks the models trained using FastText vectors outperformed those using GloVe vectors. The best performing model for the Twitter health classification task was the CNN using FastText, while the RNN using FastText performed best for the news.

The system evaluation, described in the next section, was performed using the CNN trained on FastText vectors, which was the model selected for use in the app. We opted to use a single model to simplify the architecture, and this model performed best on the more important Twitter classification task.

8.2. Early warning detection (EWD) evaluation

Evaluation of outbreak detection can be performed using time-to-detection or examination of successful/erroneous alarms. In general, researchers have evaluated their event detection systems by examining a specific outbreak after they know it has occurred, and back-testing to check whether their system would have detected it. Examples of such research include using a seasonal flu outbreak in the US (Li & Cardie, 2013) or a 2011 *E. coli* outbreak in Germany (Diaz-Aviles et al., 2012).

In the case of SENTINEL there was no prior known event or outbreak that occurred during the data collection period which the evaluation can be assessed against. Instead the system must be evaluated by determining whether the events detected during this period are genuine alarms. In order to do this a source of ground truth is required in order to compare our data with actual real-world events. For this purpose various state and federal level reports have been employed, including CDC data and state-level influenza monitoring (as described in Section 4.3).

8.2.1. Detected events

Between July 2016 and March 2017 the Event detection algorithm generated 1329 events containing more than 15 tweets, from more than 10 users. After an initial manual evaluation we took these values as a cutoff to ensure that a minimum number of users were involved, as events generated with fewer users were almost always spurious. In order to initially further evaluate these events we used the MAD metric to split them into 4 intervals, generated between: $[min, Q_1]$, $(Q_1, Q_2]$, $(Q_2, Q_3]$ and $(Q_3, max]$, where Q_1, Q_2 and Q_3 are the MAD quartiles. For the current data: $min \leftarrow 0.07$,[9] $Q_1 \leftarrow 2.75$, $Q_2 \leftarrow 3.50$, $Q_3 \leftarrow 4.67$ and $max \leftarrow 54.25$.

Figs. 7–9 give an overview of the event data. Fig. 7 shows that headache, nausea and anxiety generated the greatest number of events. Figs. 8 and 9 show that more populous states tended to produce more events, with California and Texas having the greatest number. We also analysed the events by MAD quartile, but this did not reveal any significant patterns in whether certain symptoms or states were more likely to produce higher or lower confidence events.

[9] For convenience on generating the reports, the *min* value is set to 0.00 and the first interval becomes $(0.00, Q_1]$.

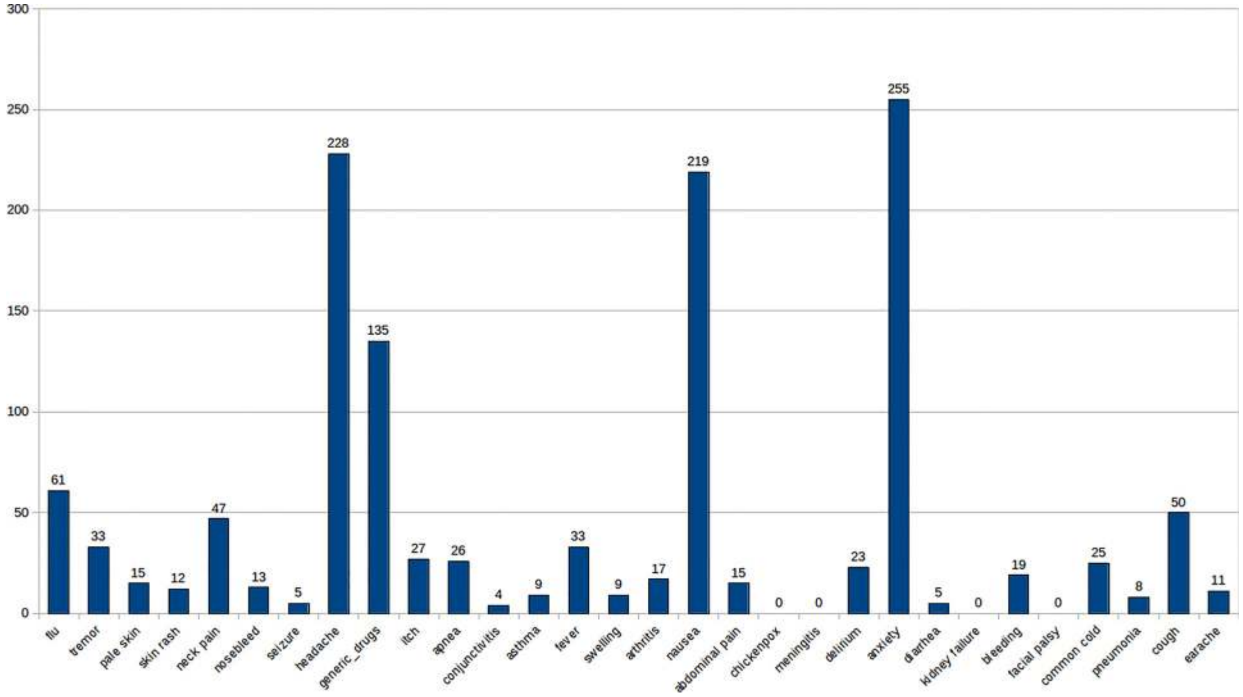


Fig. 7. Event counts for all symptoms tracked.

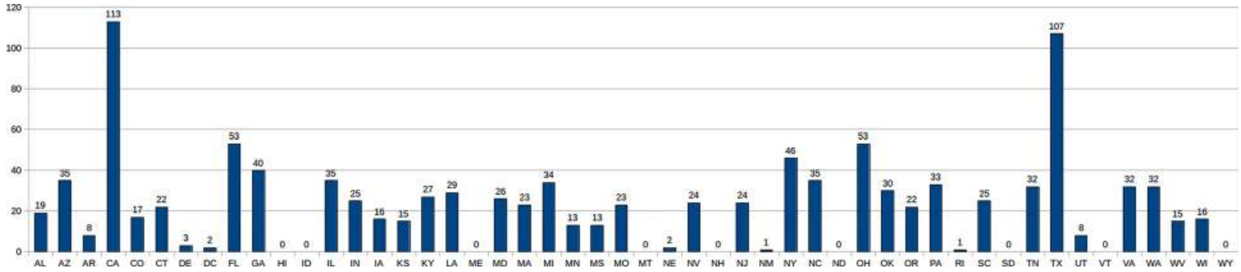


Fig. 8. Raw event counts by US State.

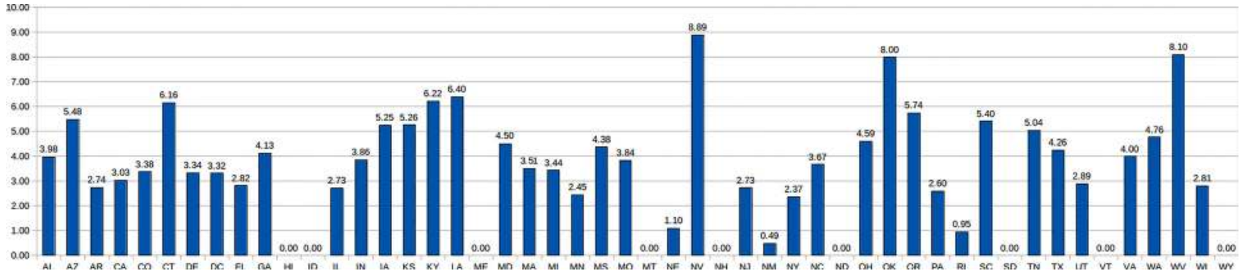


Fig. 9. Event counts normalized by State population.

8.2.2. Evaluation methodology

For the evaluation a sample of these events was manually analysed using the Event Details page contained in the SENTINEL App. For each event the following questions were examined:

- After reading the tweets contained in the event, what is a good summary of their content?
- Were the hashtags used in the tweets useful when creating this summary?
- Did the relevant tweets selected by our algorithm provide a precis of the overall tweet content?
- How many of the news articles were relevant to the tweet summary?

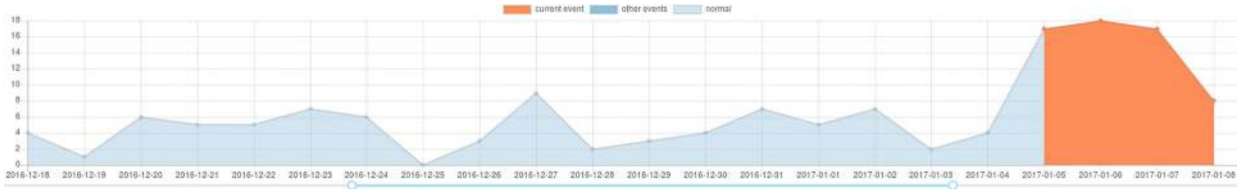


Fig. 10. SENTINEL front-end screenshot showing a flu event in January 2017 detected based on Twitter data. The time series shows the number of tweets referring to cold and flu symptoms in the state of Washington. The area coloured orange is the period where an alarm was triggered by the system.

Table 5
Top 5 most important tweets talking about the flu event detected by the EWD system.

Tweet	Score
I'm so mad I've already lost hella weight and now I'm gonna lose even more because I have the stomach flu	0.3743
Even tho I hate hospitals, i have to feel very lucky that this flu didnt kill me which it would have if I hadn't ta ...	0.3631
@McNarnia I almost feel like I have the flu even though I got my flu shot. That would be my luck.	0.3554
I've had the flu &amp; I've been feeling awful for five days but I woke up feeling so much better today	0.3354
Okay I'm *** dying ! Idk if I have food poisoning the flu idk wtf my whole body is aching and I'm shivering UNDER THE BLANKET	0.3312

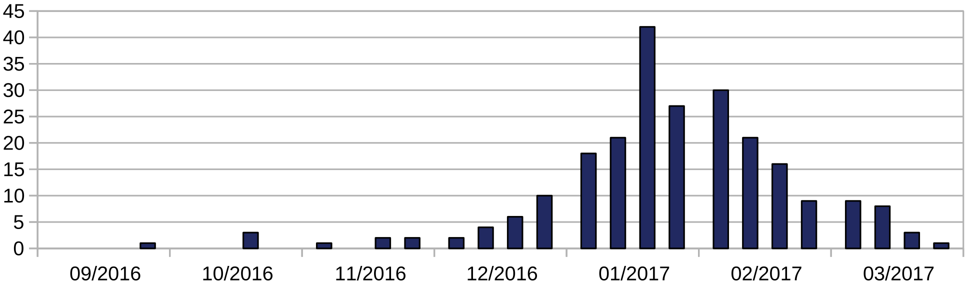


Fig. 11. Reported Lab confirmed ILI hospitalizations from Spokane county, WA.

- Are the bulk of tweets referring to a health event?
- Is there evidence in ground truth data of a health event occurring in this location and time period?

An example of this evaluation is given here for event 133 (flu in Washington state between the 6th and 8th of January 2017). The time series for this event as displayed in the SENTINEL front-end is shown in Fig. 10, and the most relevant tweets selected by our algorithm are shown in Table 5 (along with their score which is their cosine similarity to the mean tweet vector as described in Section 6.2.2). The tweets are mainly complaints about different cold and flu symptoms. There are four hashtags, with one of them being the highly relevant #flu, and the relevant tweets selected by the algorithm are indeed useful. No news articles were found for this event. A manual inspection of the event tweets shows them to be genuine illness reports, showing that the health classifier has worked correctly in this instance.

In order to see if this is a real health event a source of ground truth is required. In this case we consulted the Washington State Influenza Update for Week 1 of 2017.[10] This reveals that there is indeed a spike in influenza activity during this time period, as evidenced by Fig. 11. This is therefore evidence that SENTINEL has detected a genuine health event.

8.2.3. *Qualitative evaluation*

Initially 10 events were randomly selected from each MAD quartile for evaluation, with a constraint being that each event within a quartile should be for a different symptom. From these 40 events the hashtags people used in their tweets were useful in 7 cases, while the list of relevant tweets were found to be useful in 37 cases. A relevant news article was found in 8 of the sample events. 33 of the events were determined to be health-related. However, ground truth evidence was found for only one event in this analysis, an outbreak of flu in AL. Table A.5 presents a summary of the results.

We then examined the events with the highest MAD, to determine if these could be correlated with outbreaks of illness with a higher confidence. Events with a minimum MAD of 8 and a minimum user count of 25 were examined, excluding anxiety since this category was found not to produce high quality events (the tweets in these were found to be generalised expressions of stress with no theme linking them together). 16 events were detected that fulfilled these conditions. 5 of these were found to coincide with

[10] http://www.doh.wa.gov/Portals/1/Documents/5100/420-100-FluUpdate.pdf

outbreaks of influenza-like illness (ILI) in their states, all of them being for symptoms of ILI. A further 3 of them coincided with slight increases in the ILI figures. Others could be potential health events, but were generated for symptoms such as nausea. No ground truth data could be found for diseases such as gastro-enteritis and therefore these could not be evaluated. These results show that with these parameters events were much more likely to be significant and backed up by ground truth data. Further work is required to find methods of mapping non ILI related symptoms to ground truth for evaluation.

8.3. Evaluation of nowcasting model accuracy

In order to evaluate the accuracy of our nowcasting model its predictions must be compared with the actual outcome, i.e. the true level of the CDC case counts one week later. This evaluation has been conducted and the Mean Absolute Error (MAE) computed for each disease. The MAE is defined as follows:

$$MAE = \frac{1}{n}\sum |x_i - y_i| \tag{1}$$

Here x_i is the predicted value and y_i is the actual CDC value for the ith day, with the days numbered from i = 1 up to n.

Our LASSO model incorporating the Twitter data has been evaluated against a baseline ARIMA autoregression model solely based on the CDC data. The results for each disease are displayed in Table A.6, along with the percentage improvement from the baseline to our model. The average percentage improvement is 13.469. Incorporating the Twitter data therefore does provide a real improvement over the baseline model.

9. Conclusion and future work

The system currently collects around 1.8 million tweets per day, processes and stores them. On a daily basis it generates events and associated situational awareness reports. On a weekly basis it downloads CDC data and performs Nowcasting. All areas of the system are built with reliable open-source technologies that embody the current state of the art in software development.

Evaluation of our results show that the health classifiers are robust and accurate, with the chosen classifier giving an F1 of 0.852 for the Twitter classification task and 0.939 for news classification. These classifiers outperformed the baselines, demonstrating that deep learning is useful in this sphere of text classification. Our news crawler is retrieving large numbers of health-related articles, which are being linked to a significant fraction of detected events, although this linkage shows room for improvement. The event detection evaluation shows that the tools made available on the Event Details page of the App are useful in event evaluation, and that given suitable filter parameters around 1/3 of detected events were significant enough to be validated by the ground truth data currently available. Finally the Nowcasting evaluation showed that including our Twitter data provided a 13% boost to Nowcasting accuracy compared to the baseline.

Future work:

- **Incorporating epidemiological models:** Extension of the nowcasting to use the information provided by the specific disease module and epidemiological models. This would allow forecasting disease levels much further into the future.
- **Improved News Linkage:** Topic modelling such as Latent Dirichlet Allocation (LDA) could be used to identify the main topic of each news article. This could then be used to facilitate improved linking of news articles and events, ensuring that only those articles topically referring to the symptom or disease in question are linked.

Acknowledgements

This research was carried out in cooperation with the UK Defence Science and Technology Laboratory, through contract DSTL/AGR/00728/01. It was funded by the U.S. Department of Defense's Defense Threat Reduction Agency (DTRA).

Appendix A. Evaluation results

Table A.5
Qualitative evaluation of events, sampled from various MAD intervals.

Event ID	Symptom	Location	MAD / Tweet Count / User Count	Human summary	Hashtags useful?	Relevant tweets useful?	News useful?	Health related?
2408	fever	CA	2.80/33/20	People complaining about high temperatures - cold/flu symptoms.	No	Yes	No	Yes
1915	itch	CA	3.23/29/16	Complaints about general itching - no specific theme	No	Yes	Yes	Yes
4449	cough	CA	2.77/54/31	People complaining about coughing - cold/flu symptoms	No	Yes	7/13	Yes
2221	swelling	US	3.27/33/21	Mixture of different types of inflammation - no clear theme	No	No	4/22	Yes
4184	neck pain	NY	2.80/22/26	Mixture of headache and caffeine references	Yes	Yes	n/a	No
1142	delirium	US	2.80/145/74	Complaints about insomnia	Yes	Yes	1/4	Yes
2410	pneumonia	US	2.95/29/21	Complaints about flu	No	Yes	No	Yes
2652	headache	NV	2.88/36/20	Complaints about headache / migraine	No	Yes	n/a	Yes
3654	apnea	TX	2.88/18/14	Chatter about caffeine	No	Yes	No	No
1183	generic drugs	PA	3.29/33/21	Chatter about different drugs, including alcohol. Could be related to 4th July celebrations.	No	Yes	n/a	No
4693	tremor	TX	4.85/43/19	Various unrelated uses of the word shake	No	Yes	n/a	No
5424	cough	CA	4.87/53/32	Complaints about coughing	No	Yes	n/a	Yes
3635	anxiety	AL	4.79/51/45	Lots of people saying they're confused	No	Yes	No	No
2310	common cold	AZ	7.00/18/16	Complains about cold/flu symptoms - strep throat	No	Yes	No	Yes
2769	nausea	MO	5.00/17/15	Complaints about general unwellness	No	Yes	n/a	Yes
620	flu	AL	19.5/34/30	Complaints about having the flu	No	Yes	1/5	Yes
5620	asthma	US	7.85/51/38	Complaints about asthma	Yes	Yes	n/a	Yes
4047	conjunctivitis	CA	5.83/35/29	Complains about red sore eyes, some red eye flights	No	Yes	n/a	Yes
2546	abdominal pain	TX	5.09/20/23	Complaints about stomach aches	No	Yes	n/a	Yes
764	swelling	US	5.83/44/25	Complaints about swelling and inflammation	No	Yes	n/a	Yes
285	abdominal pain	CA	2.10/16/19	A combination of people complaining about abdominal pain, headaches and flu symptoms	n/a	Yes	n/a	Yes
727	nausea	AZ	3.50/36/21	Users complaining about being ill	n/a	No	Yes	Yes
1213	arthritis	US	3.61/90/57	Tweets about chest pain, arthritis and body aches	1/4	3/5	n/a	Yes
1213	arthritis	US	3.61/90/57	Tweets about chest pain, arthritis and body aches	1/4	3/5	n/a	Yes
1229	pale skin	CA	3.50/23/24	A combination of people tweeting about alcohol poisoning and skin rashes	n/a	3/5	n/a	Yes
2318	delirium	NY	3.50/19/13	Tweets about insomnia	Yes	Yes	n/a	Yes
2447	tremor	CA	3.50/27/19	Tweets about people shivering for various reasons	No	Yes	n/a	No
2707	flu	FL	1.50/23/12	People complaining about the flu	No	Yes	Yes	Yes
2768	common cold	CA	3.50/16/12	People tweeting their cold symptoms	n/a	Yes	n/a	Yes
2965	generic drugs	GA	3.50/20/11	Users tweeting about allergies and various drugs	n/a	Yes	n/a	Yes
3090	bleeding	US	4.00/114/64	Some of the tweets are sarcastic, but most of them are about users tweeting about bleeding symptoms	No	Yes	No	Yes
3415	headache	AL	0.97/27/12	Users tweeting about their migraines	n/a	Yes	2/15	Yes
3507	common cold	US	0.52/121/55	People tweeting their cold symptoms	No	Yes	No	Yes
3660	cough	PA	3.82/20/15	People tweeting about them coughing	n/a	Yes	n/a	Yes
3768	fever	CA	0.72/35/26	People tweeting having a fever	n/a	Yes	1/20	Yes
3997	diarrhea	US	2.27/44/27	Users swearing and complaining about diarrhea	No	No	No	No
4469	bleeding	CA	1.33/26/20	People tweeting about bleeding	n/a	Yes	No	Yes
4747	arthritis	CA	1.75/16/11	Tweets about chest pain, arthritis and body aches	n/a	Yes	No	Yes
4808	tremor	TX	1.00/24/12	Tweets about people shivering for various reasons	Yes	Yes	No	Yes
5364	delirium	CA	0.91/16/11	Tweets about insomnia	n/a	Yes	n/a	Yes
5460	flu	TX	3.76/6/39	People complaining about their flu symptoms	Yes	Yes	n/a	Yes

Table A.6
Nowcasting mean absolute error per disease.

Disease	LASSO	ARIMA	Percentage difference
Babesiosis	0.25	0.325	23.235
Campylobacteriosis	5.287	6.165	14.250
Chlamydia trachomatis infection	122.018	126.673	3.675
Coccidioidomycosis	1.039	1.069	2.772
Cryptosporidiosis	3.556	3.702	3.942
Dengue	0.006	0.012	46.667
Ehrlichiosis anaplasmosis anaplasma phagocytophilum	0.562	0.628	10.522
Ehrlichiosis anaplasmosis ehrlichia chaffeensis	0.292	0.322	9.113
Ehrlichiosis anaplasmosis undetermined	0.022	0.036	36.957
Giardiasis	2.125	2.275	6.578
Gonorrhea	38.01	39.944	4.844
Haemophilus influenzae invasive all ages all serotypes	0.583	0.641	9.025
Hepatitis viral acute type A	0.436	0.486	10.366
Hepatitis viral acute type B	0.363	0.419	13.444
Hepatitis viral acute type C	0.231	0.305	24.316
Invasive pneumococcal disease age < 5	0.208	0.228	8.537
Invasive pneumococcal disease all ages	1.753	1.76	0.401
Legionellosis	0.996	1.085	8.191
Lyme disease	3.176	4.23	24.911
Malaria	0.358	0.383	6.452
Meningococcal disease invasive all serogroups	0.073	0.093	21.667
Mumps	1.787	1.81	1.268
Pertussis	1.841	1.924	4.274
Rabies animal	0.746	0.838	10.964
Rubella	0.003	0.006	50.000
Salmonellosis	6.077	6.996	13.130
Shiga toxin producing E.coli stec	1.046	1.155	9.429
Shigellosis	2.277	2.472	7.893
Spotted fever rickettsiosis confirmed	0	0	0.000
Spotted fever rickettsiosis including RMSF confirmed	0.023	0.031	25.000
Spotted fever rickettsiosis including RMSF probable	0.599	0.765	21.695
Spotted fever rickettsiosis probable	0.008	0.012	33.333
Syphilis primary and secondary	1.196	1.296	7.771
Tetanus	0.026	0.021	-25.926
Varicella chickenpox	1.039	1.076	3.510
Vibriosis	0.373	0.426	12.422
West nile virus disease neuroinvasive	0.034	0.035	4.348
West nile virus disease non-neuroinvasive	0.003	0.005	42.857

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2018.04.011

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. In *Osdi* (pp. 265–283). (vol. 16).

Abdelhaq, H., Sengstock, C., & Gertz, M. (2013). Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment, 6*(12), 1326–1329.

Aggarwal, C. C., & Subbian, K. (2012). *Event detection in social streams. Sdm12. Sdm* SIAM624–635.

Aho, A. V., & Corasick, M. J. (1975). Efficient string matching: An aid to bibliographic search. *Communications of the ACM, 18*(6), 333–340.

Aramaki, E., Maskawa, S., & Morita, M. (2011). *Twitter catches the flu: Detecting influenza epidemics using twitter. Proceedings of the conference on empirical methods in natural language processing.* Stroudsburg, PA, USA: Association for Computational Linguistics1568–1576 EMNLP '11

Bansal, P., Bansal, R., & Varma, V. (2015). *Towards deep semantic analysis of hashtags. European conference on information retrieval.* Springer453–464.

Bodnar, T., & Salathé, M. (2013). *Validating models for disease detection using twitter. Proceedings of the 22nd international conference on world wide web companion.* Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee699–702 WWW '13 Companion.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv:1607.04606.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). *Freebase: A collaboratively created graph database for structuring human knowledge. Proceedings of the 2008 acm sigmod international conference on management of data.* AcM1247–1250.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Med, 5*, e151.

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen, J. M., et al. (2015). Using social media for actionable disease surveillance and outbreak management: A systematic literature review. PloS one, 10(10).

Chierichetti, F., Kleinberg, J., Kumar, R., Mahdian, M., & Pandey, S. (2014). *Event detection via communication pattern analysis. Eighth international AAAI conference on weblogs and social media (icwsm)*51–60.

Collier, N., Son, N. T., & Nguyen, N. M. (2011). Omg u got flu? Analysis of shared health messages for bio-surveillance. *Journal Biomedical Semantics, 2*(S-5), S9.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing twitter messages. arXiv:1007.4748.

Davis, C. A., Ciampaglia, G. L., Aiello, L. M., Chung, K., Conover, M. D., Ferrara, E., et al. (2016). Osome: The iuni observatory on social media. *PeerJ Computer Science*, 2, e87.

Diaz-Aviles, E., Stewart, A., Velasco, E., Denecke, K., & Nejdl, W. (2012). Epidemic intelligence for the crowd, by the crowd. *ICWSM, 12*, 439–442.

Dong, X., Mavroeidis, D., Calabrese, F., & Frossard, P. (2015). Multiscale event detection in social media. *Data Mining and Knowledge Discovery, 29*(5), 1374–1405.

Dou, W., Wang, X., Skau, D., Ribarsky, W., & Zhou, M. X. (2012). *Leadline: Interactive visual analysis of text data through event identification and exploration. Visual analytics science and technology (vast), 2012 ieee conference on*. IEEE93–102.

Dredze, M., Cheng, R., Paul, M. J., & Broniatowski, D. (2014). *Healthtweets. org: A platform for public health surveillance using twitter. AAAI conference on artificial intelligence*. Citeseer593–596.

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of Medical Internet Research, 11*(1).

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012–1014 doi:10.1038/nature07634.

Greenwood, S., Perrin, A., & Duggan, M. (2016). Social media update 2016. *Pew Research Center, 11*.

Guerrisi, C., Turbelin, C., Blanchon, T., Hanslik, T., Bonmarin, I., Levy-Bruhl, D., et al. (2016). Participatory syndromic surveillance of influenza in europe. The Journal of Infectious Diseases, 214(suppl_4), S386–S392.

Halevy, A., Rajaraman, A., & Ordille, J. (2006). *Data integration: The teenage years. Proceedings of the 32nd international conference on very large data bases*. VLDB Endowment9–16.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hutwagner, M. L., Thompson, M. W., Seeman, G. M., & Treadwell, T. (2003). The bioterrorism preparedness and response early aberration reporting system (ears). *Journal of Urban Health, 80*(1), i89–i96.

Ji, X., Chun, S. A., & Geller, J. (2012). *Epidemic outbreak and spread detection system based on twitter data. Health information science*. Springer152–163.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv:1408.5882.

Kleppmann, M., & Kreps, J. (2015). Kafka, samza and the unix philosophy of distributed data. *IEEE Data Engineering Bulletin, 38*(4), 4–14.

Krumm, J., & Horvitz, E. (2015). *Eyewitness: Identifying local events via space-time signals in twitter feeds. Proceedings of the 23rd sigspatial international conference on advances in geographic information systems*. ACM20.

Lamb, A., Paul, M. J., & Dredze, M. (2013). *Separating fact from fear: Tracking flu infections on twitter. Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. Atlanta, Georgia: Association for Computational Linguistics789–795.

Lampos, V., & Cristianini, N. (2010). *Tracking the flu pandemic by monitoring the social web. Cognitive information processing (cip), 2010 2nd international workshop on*. IEEE411–416.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST), 3*(4), 72.

Lampos, V., De Bie, T., & Cristianini, N. (2010). *Flu detector-tracking epidemics on twitter. Machine learning and knowledge discovery in databases*. Springer599–602.

Lee, K., Agrawal, A., & Choudhary, A. (2013). *Real-time disease surveillance using twitter data: demonstration on flu and cancer. Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining*. ACM1474–1477.

Lee, K., Agrawal, A., & Choudhary, A. (2015). *Mining social media streams to improve public health allergy surveillance. Advances in social networks analysis and mining (asonam), 2015 ieee/acm international conference on*. IEEE815–822.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global twitter heartbeat: The geography of twitter. *First Monday, 18*(5).

Li, J., & Cardie, C. (2013). Early stage influenza detection from twitter. arXiv:1309.7340.

Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). *The tweets they are a-changin: Evolution of twitter users and behavior. Icwsm30. Icwsm* 5–314.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). *Twitinfo: Aggregating and visualizing microblogs for event exploration. Proceedings of the sigchi conference on human factors in computing systems*. ACM227–236.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., et al. (2014). A case study of the New York city 2012–2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *Journal of Medical Internet Research, 16*(10).

Paul, M. J., Dredze, M., & Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLOS Currents Outbreaks*.

Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., et al. (2016). *Social media mining for public health monitoring and surveillance. Biocomputing 2016: Proceedings of the pacific symposium*. World Scientific468–479.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12, (Oct),2825–2830.

Peña-Araya, V., Quezada, M., Poblete, B., & Parra, D. (2017). Gaining historical and international relations insights from social media: Spatio-temporal real-world news analysis using twitter. *EPJ Data Science, 6*(1), 25.

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation. Empirical methods in natural language processing (emnlp)*1532–1543.

Sadilek, A., Brennan, S., Kautz, H., & Silenzio, V. (2013). *nemesis: Which restaurants should you avoid today? First AAAI conference on human computation and crowdsourcing*.

Sadilek, A., Kautz, H. A., & Silenzio, V. (2012). *Modelling spread of disease from social interactions. In sixth AAAI international conference on weblogs and social media (icwsm)*322–329.

Sadilek, A., Kautz, H. A., & Silenzio, V. (2012). *Predicting disease transmission from geo-tagged micro-blog data. AAAI*136–142.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 919–931.

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., & Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Computational Biology, 11*(10), e1004513.

Santos, J. C., & Matos, S. (2014). Analysing twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling, 11*(1), S6.

Sheth, A., Jadhav, A., Kapanipathi, P., Lu, C., Purohit, H., Smith, G. A., et al. (2014). *Twitris: A system for collective social intelligence. Encyclopedia of social network analysis and mining*. Springer2240–2253.

Sloan, L., & Morgan, J. (2015). Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PLoS ONE, 10*(11), e0142209.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Thapen, N., Simmie, D., & Hankin, C. (2016). The early bird catches the term: Combining twitter and news data for event detection and situational awareness. *Journal of Biomedical Semantics, 7*(1), 61.

Thapen, N., Simmie, D., Hankin, C., & Gillard, J. (2016). Defender: Detecting and forecasting epidemics using novel data-analytics for enhanced response. *PloS one, 11*(5), e0155417.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 267–288.

Tumeo, A., Villa, O., & Chavarria-Miranda, D. G. (2012). Aho-corasick string matching on shared and distributed-memory parallel architectures. *IEEE Transactions on Parallel and Distributed Systems, 23*(3), 436–443.

Turian, J., Ratinov, L., & Bengio, Y. (2010). *Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics384–394.

Walther, M., & Kaisser, M. (2013). *Geo-spatial event detection in the twitter stream. Advances in information retrieval*. Springer356–367.

Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011). *Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs.*

Proceedings of the 20th acm international conference on information and knowledge management. New York, NY, USA: ACM2541–2544. http://dx.doi.org/10.1145/2063576.2064014 CIKM '11

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97–107.

Yu, H., Ho, C., Juan, Y., & Lin, C. (2013). Libshorttext: A library for short-text classification and analysis. *Rapport interne, Department of Computer Science* Software available at http://www.csie.ntu.edu.tw/∼cjlin/libshorttext

Zubiaga, A., Spina, D., Amigó, E., & Gonzalo, J. (2012). *Towards real-time summarization of scheduled events from twitter streams. Proceedings of the 23rd acm conference on hypertext and social media*. ACM319–320.