ORIGINAL ARTICLE

Twitter sentiment classification for measuring public health concerns

Xiang Ji[1] · Soon Ae Chun[2] · Zhi Wei[1] · James Geller[1]

Received: 2 November 2014 / Revised: 18 April 2015 / Accepted: 26 April 2015 / Published online: 12 May 2015

Abstract An important task of public health officials is to keep track of health issues, such as spreading epidemics. In this paper, we are addressing the issue of spreading *public concern about* epidemics. Public concern about a communicable disease can be seen as a problem of its own. Keeping track of trends in concern about public health and identifying peaks of public concern are therefore crucial tasks. However, monitoring public health concerns is not only expensive with traditional surveillance systems, but also suffers from limited coverage and significant delays. To address these problems, we are using Twitter messages, which are available free of cost, are generated world-wide, and are posted in real time. We are measuring public concern using a two-step sentiment classification approach. In the first step, we distinguish Personal tweets from News (i.e., Non-Personal) tweets. In the second step, we further separate Personal Negative from Personal Non-Negative tweets. Both these steps consist themselves of two sub-steps. In the first sub-step (of both steps), our programs automatically generate training data using an emotion-oriented, clue-based method. In the second sub-step, we are training and testing three different Machine Learning (ML) models with the training data from the first sub-step; this allows us to determine the best ML model for different datasets. Furthermore, we are testing the already trained ML models with a human annotated, disjoint dataset. Based on the number of tweets classified as Personal Negative, we compute a Measure of Concern (MOC) and a timeline of the MOC. We attempt to correlate peaks of the MOC timeline to peaks of the News (Non-Personal) timeline. Our best accuracy results are achieved using the two-step method with a Naïve Bayes classifier for the Epidemic domain (six datasets) and the Mental Health domain (three datasets).

Keywords Twitter mining · Sentiment analysis · Public health · Measure of concern · Automatic sentiment labeling · Sentiment classification · Social analytics

1 Introduction

Public health surveillance is critical to monitoring the spread of infectious diseases and deploying rapid responses when there is an indication of an epidemic emerging. Different surveillance strategies have been developed to meet different needs. These strategies include sentinel surveillance systems, household surveys, laboratory-based surveillance, and most recently Integrated Disease Surveillance and Response (IDSR) (DCP 2014). Besides monitoring the spread of a disease itself, monitoring *emotional changes* of the general public, brought about by epidemics, is becoming increasingly important for public health specialists.

The importance of monitoring the public's concerns about an epidemic is illustrated by the recent Ebola scare in the United States. Since the end of September 2014, Ebola concerns have spread in the United States after a Liberian visitor to Dallas became the first person to be diagnosed in the USA. The immigration examination and the medical system's ability to deal with Ebola were widely questioned

✉ Xiang Ji
xj25@njit.edu

[1] Department of Computer Science, New Jersey Institute of Technology, Martin Luther King Blvd, Newark, NJ 07102, USA

[2] City University of New York, College of Staten Island, 2800 Victory Blvd, Staten Island, NY 10314, USA

by the general public (Reuters 2014) due to a series of missteps when the Liberian was issued a visitor visa and was not diagnosed by a Dallas hospital. For example, a tweet on October 15th of 2014, stated that, "*Co-worker LEGITIMATELY thinks #Ebola was caused by one of two things:* (1) *Gov't attempts at population control.* (2) *ISIS THIS IS NOT A JOKE*". As the public opinion will potentially affect the government's public health decisions, President Obama attempted to calm the public by stating that "This is a serious disease, but we can't give into hysteria or fear" (Reuters 2014).

Zhu et al. (2008) studied the changes in mental state of the Chinese public during the outbreak of SARS (2003). They found that, during the outbreak, most of the people surveyed (96.4 %) reported emotional changes and negative emotions such as panic (54.8 %), nervousness (34.0 %), and fear (7.6 %). Psychological changes might lead to unpredictable behavior. Of all the people surveyed, 23.3 % admitted to "irrational" behaviors such as going on a shopping spree, or to actions such as seeking shelter, preparing provisions, etc.

Another example is the public's reaction to Japan's nuclear emergency in March 2011 (Guardian 2011). Text messages about nuclear plumes spread throughout Asia. In China, the rumors that iodized salt could help ward off radiation poisoning amid Japan's nuclear emergency triggered panic buying all over the country. In Vietnam, students were kept indoors by schools, some companies allowed staff to leave early to avoid rainfall after the rumor spread that rain would burn the skin and cause cancer. A university in Manila canceled classes due to a similar scare.

As the above examples illustrate, monitoring public panic about health issues is critical not only to public health specialists but also to government decision makers. However, for traditional public health surveillance systems, it is hard to detect and monitor health-related concerns and changes in public attitudes to health-related issues. Due to their expenses, the existing surveillance methods, such as questionnaires and clinical tests, can only cover a limited number of people and results often appear with significant delays. To supplement the current surveillance systems, a novel tool must be developed. This tool must be able to track real-time statistics of emotions related to different health matters, such as epidemics, to provide early warning, and to help the government decision makers prevent or respond to potential social crises that might be the impact of these health-related emergencies.

The Web has created unprecedented resources for tracking threats to public health. Ginsberg et al. (2009) relied exclusively on search engine logs, in which users submitted queries in reference to issues that they were concerned about, to approach this problem. Their thread of research led to the realization that an aggregation of large numbers of queries might show patterns that are useful for the early detection of epidemics. Twitter, a micro-blog service provider, shows several advantages over search engines for disease surveillance. It is up-to-date and it has more than 500 million users in total. There are more than 340 million tweets posted by Twitter users per day (Twitter 2014a). Most tweets are public and the Twitter API (Twitter 2014b) enables researchers to retrieve the tweets as well as related information, such as geographical location and hyperlinks included.

We explore the potential of mining social network data, such as tweets, to provide a tool for public health specialists and government decision makers to gauge the measure of concern (MOC) expressed by Twitter users under the impact of diseases. To derive the MOC from Twitter, we developed a two-step classification approach to analyze sentiments in disease-related tweets. We first distinguish Personal from News (Non-Personal) tweets. Many news articles released by online media organizations are used for 're-tweets' by Twitter users. We consider these News tweets as Non-Personal, as opposed to Personal tweets posted by individual Twitter users. We refer to the former as *News tweets* and the latter as *Personal tweets*. In the second stage, the sentiment analysis is applied only to Personal tweets to distinguish Negative from Non-Negative tweets.

Although News tweets may also express concerns about a certain disease, they tend not to reflect the direct emotional impact of that disease on people. A person re-tweeting a News message about a disease, which is comparable to forwarding an email message, is most likely not directly affected by it, while a user sending out a Personal tweet with emotional expressions might be directly affected. Note that the two-step sentiment classification problem we present is different from the traditional Twitter sentiment classification, which is categorizing tweets into positive/negative or positive/neutral/negative tweets (Zhang et al. 2011; Liu and Zhang 2012; Mohammad et al. 2013; Saif et al. 2013; Aramaki et al. 2011) without distinguishing Personal from Non-Personal tweets first. Our sentiment classification method is able to identify Personal tweets (including Personal Negative and Personal Non-Negative) and News (Non-Personal) tweets. In addition, we subsequently use the results of the classification to compute the correlation between sentiment-carrying tweets and News tweets, as the classification results provide all the necessary data for this computation.

We need to differentiate between the spread of concern about a disease and the spread of the disease itself. For example, the tweet: "Wiz looks like he got the measles and Ross just dark as hell. I can't tell if they're tattoos or wrinkles http://twitpic.com/4geuc2" is annotated as a Non-Negative tweet, because it shows no concern. However, it is a strong clue to track the spread of measles. We focus on studying the Twitter users' concerns about diseases instead of the outbreak of the disease itself, which has been extensively studied (Brownstein et al. 2008; Collier and Doan 2012; Signorini et al. 2011; Aramaki et al. 2011; Lampos and Cristianini 2010).

Using the sentiment classification results, we quantify the MOC based on the number of Personal Negative tweets per day. The MOC increases with the relative growth of Personal Negative tweets and with the absolute growth of Personal Negative tweets. Previous research (Sha et al. 2014; Ji et al. 2013) found that sentiment surges co-occurred with health events on a timeline. Different from the previous work, we calculated the correlation between MOC timeline (i.e., change over time) and News timeline and the correlation between Non-Negative timeline and News timeline using the Jaccard Coefficient (Liben Nowell and Kleinberg 2007). Using the MOC to track public health concerns can help government officials to make timely decisions to refute rumors, and thus prevent potential social crises such as the past case of Chinese panic buying of salt. Monitoring of the public concern using social network data can provide public health specialists with a surveillance capability for large segments of the population, in real-time, and with low expenses.

We summarize our contributions as follows:

1. We developed a two-step sentiment classification method by combining clue-based search and Machine Learning (ML) methods by first automatically labeling the training datasets, and then building classifiers for Personal tweets and classifiers for tweet sentiments. As previously discussed, the traditional Twitter sentiment classification methods classify the tweets into positive/negative or positive/neutral/negative. Our two-step algorithm is different from the traditional methods because it filters out the Personal tweets and News tweets in the first step, and then the Personal Negative tweets are used for defining a MOC. To the best of our knowledge, while previous research has extracted objective tweets, it has not explicitly extracted the News tweets, and has not utilized Personal Negative, Personal Non-Negative, and News tweets to define a MOC to quantify the sentiment trends on the timeline. Thus, using this novel method, one can combine the sentiment classification results into a MOC to reveal the sentiment timeline trends.
2. We quantified the MOC using the results of sentiment classification, and used it to reveal the timeline trends of sentiments of tweets. We both quantitatively and qualitatively correlated the sentiment timeline trends and the News timeline trends, and calculated the correlation between MOC timeline and News timeline and the correlation between Non-Negative timeline and News timeline using the Jaccard Coefficient. We performed the correlation analysis among different tweet sentiment classes. The experimental results show that the peaks of the MOC and the peaks of NN (Non-Negative) tweets are weakly correlated with the peaks on the News timeline without any appreciable time delay/lead.
3. We applied our sentiment classification method and the MOC to other topical domains, such as mental health monitoring and crisis management. The experimental results support the hypothesis that our approach is generalizable to other domains.

The rest of the paper is organized as follows. In Sect. 2, related work and open problems are discussed. In Sect. 3, we give formal definitions of the concepts used in this paper. In Sect. 4, sentiment classification methods and results are introduced in detail. In Sect. 5, the sentiment timeline trend analysis results are illustrated, interpreted, and discussed. Section 6 contains conclusions and suggestions for future research.

2 Related work

2.1 Disease and emergency monitoring with Twitter

Since the year 2008, concepts and systems have been developed to monitor disease outbreaks and emergencies with Twitter. Artman et al. (2011) introduced the concept of dialogical emergency management, which emphasizes the screening of vast and quickly spread information on the Internet, to help the emergency management staff gain a better strategic awareness of the public. The Alert4All Screening of New Media (SNM) tool (Johansson et al. 2012) was developed based on this concept to analyze emotion recognition/affect in social media, e.g., Twitter and Facebook, regarding crisis management. Brownstein et al. (2008) used online News to perform surveillance of epidemics. Their system, Healthmap, collects reports from online News aggregators, such as Google News. By categorizing the News into epidemics-related and epidemics-unrelated reports, and filtering the epidemics-related documents into "breaking News", "warnings", and "old News", the system is able to trigger alerts based on "breaking News". To detect disease outbreaks and monitor their progression over time and location, we have previously implemented the Epidemics Outbreak and Spread Detection System (EOSDS) (Ji et al. 2012) by monitoring social media data (specifically, Twitter data). EOSDS provides the functionality to perform geographic visual analysis of tweets with three different kinds of maps. The static map shows each individual tweet's location. The distribution (intensity) map displays absolute and relative frequencies of tweets from every USA state. The filter map provides users with a functionality to monitor the spread of a particular epidemic.

The other thread of research focused on building models, primarily supervised learning models, to detect disease and emergency events from Twitter. Collier and Doan (2012) developed a model to automatically classify Twitter messages into six fixed classes of syndromes, such as Respiratory and Gastrointestinal. Aramaki et al. (2011) applied a Support Vector Machine (SVM) to distinguish influenza-related tweets from tweets that are irrelevant. Signorini et al. (2011) also used an SVM-based estimator to analyze H1N1-related tweets, and estimated the Influenza-like Illness (ILI) rate, which is usually regarded as the ground truth, preceding the official announcement of an H1N1 outbreak by 1–2 weeks. Similarly, (Culotta 2010b) experimented with a number of regression models to correlate Twitter messages with statistics from the Center for Disease Control and Prevention (CDC) and provided a relatively simple method to track the ILI rate using a large number of Twitter messages (Culotta 2010a). Lampos and Cristianini (2010) used an approach to automatically learn a set of markers to help compute flu scores, and achieved a high correlation with the HPA flu score, which is the equivalent of the CDC score in the UK. All of the above research projects studied how to use Twitter to detect the outbreak of diseases instead of the sentiment trend caused by epidemics, which is the focus of this paper.

2.2 Sentiment analysis

Sentiment Analysis has been an active research area since the 2000s. With an increasing number of datasets from various data sources, such as blogs, review sites, News articles, and micro-blogs available, researchers have become interested in mining high-level sentiments from them. Sentiments are also closely related to information spread. Their relationship was shown in different contexts, such as social transmission (Berger 2011), News broadcasts (Heath 1996), and online social media, such as Twitter (Stieglitz and Dang-Xuan 2013). By analyzing the sentiments of opinion leaders, the public health officials will be able to monitor the viral effects in social media communication, and take early actions to prevent unnecessary panic.

A survey of sentiment analysis was done by Pang and Lee (2008). According to the target of analysis, the research on sentiment analysis can be categorized into the following levels: document-level (Pang et al. 2002), blog-level (Mishne 2005), sentence-level (Wilson et al. 2005), tweet-level (Johansson et al. 2012; Brynielsson et al. 2014; Saif et al. 2014) with the sub-category non-English tweet-level (Refaee and Rieser 2014), and tweet-entity-level (Saif et al. 2012). Due to the large number of available tweets and their real-time nature, tweets are ideal for sentiment classification and quantification for disease monitoring, and more broadly, for crisis monitoring.

2.3 Twitter sentiment classification

Extensive research has been done in the sub-area of Twitter sentiment classification since 2009 (Barbosa and Feng 2010; Bifet and Frank 2010; Pak and Paroubek 2010; Jiang et al. 2011; Mohammad et al. 2013; Zhou et al. 2014; Brynielsson et al. 2014). Most of this thread of research used Machine Learning-based approaches such as Naïve Bayes, Multinomial Naïve Bayes, and Support Vector Machine. The Naïve Bayes classifier is a derivative of the Bayes decision rule (Fukunaga 1990), and it assumes that all features are independent from each other. Good performance of Naïve Bayes (NB) was reported in several sentiment analysis papers (Barbosa and Feng 2010; Brynielsson et al. 2014; Zhou et al. 2014). Multinomial Naive Bayes (MNB) is a model that works well on sentiment classification (Bifet and Frank 2010; Pak and Paroubek 2010; Zhou et al. 2014). MNB takes into account the number of occurrences and relative frequency of each word. Support Vector Machine (Cortes and Vapnik 1995) is also a popular ML-based classification method that works well on tweets (Jiang et al. 2011; Brynielsson et al. 2014). In Natural Language Processing, SVM with a polynomial kernel is more popular (Chang and Lin 2011).

Mohammad et al. (2013) explored an extensive list of features such as clusters, negation, and *n*-grams, and used a Support Vector Machine (SVM) to classify Twitter messages into positive, negative, and neutral. Barbosa and Feng (2010) focused on automation of the training data generation process. Their work combined sentiment-labeled tweets coming from three sources: Twendz, Twitter Sentiment, and Tweet Feel. A moderate Cohen's Kappa Coefficient served as evidence that the combination of several sources reduced the bias of the individual sources. In this way, the combination improved the polarity classification.

The above sentiment classification studies have two drawbacks:

1. They classified Twitter messages into either positive/negative or positive/negative/neutral with the assumption that all Twitter messages express ones' opinion. However, this assumption does not hold in many situations, especially when the tweets are about epidemics or more broadly, about crises. In these situations, as we found when we randomly sampled

100 tweets, many tweets (up to 30 %) of the samples are repetitions of the News without any personal opinion. Since they are not explicitly labeled with retweet symbols, it is not easy for a stopword-based preprocessing filter to detect them. We attempt to solve a different problem, which is how to classify tweets into three categories: Personal Negative tweets, Personal Non-Negative tweets, and News tweets (tweets that are non-Personal tweets). We are not singling out positive tweets, as few people would post positive tweets about a spreading epidemic.

Instead of identifying News tweets, Brynielsson et al. (2014) used manual labeling to classify tweets into "angry", "fear", "positive", or "other" (irrelevant). Salathe and Khandelwal (2011) also identified irrelevant tweets together with sentiment classifications. Without considering irrelevant tweets, they calculated the H1N1 vaccine sentiment score from the relative difference of positive and negative messages. As we will show later, by the two-step classification method, we can automatically extract News tweets and perform the sentiment analysis, and the results of sentiment classification are the input for computing the correlation between sentiments and News trends. In this way, the goals of sentiment classification and measuring the public concern can be achieved in an integrated framework.

2. Secondly, although the above research approaches have developed sophisticated models to improve the precision and recall of sentiment classification, they did not quantify the results of the sentiment classification to measure timeline trends, and correlate them with real-world incidents, to provide insights for public health specialists and government decision makers. We developed the MOC to quantify the sentiments, and we correlate sentiment trends and News trends to provide better knowledge of Twitter users' reactions toward crises, such as epidemics, mental health problems, clinical science problems, etc.

2.4 Quantifying Twitter sentiment on timeline

The objective of sentiment quantification is to convert natural language text to a numerical value or a timeline of numerical values to gain insights into the sentiment trends. Zhuang et al. (2006) generated a quantification of sentiments about movie elements, such as special effects, plot, dialog, etc. Their quantification contains a positive score and a negative score toward a specific movie element.

For tweet-level sentiment quantification on a timeline, Chew and Eysenbach (2010) used a statistical approach to computing the relative proportion of all tweets expressing concerns about H1N1 and visualized the temporal trend of positive/negative sentiments based on their proportion. Similar research was done by O'Connor et al. (2010). In their thread of research, they quantified the sentiments as a timeline by deriving a day-to-day (positive and negative) sentiment score simply by counting how many positive and negative words of one tweet appear in the subjectivity lexicon of OpinionFinder (Wilson et al. 2005), which is a list containing words marked as positive or negative. By analyzing Chinese micro-blogs, Sha et al. (2014) found that the sentiment fluctuations on a timeline were associated with the announcements of new regulations or government actions.

There are two drawbacks of the existing Twitter sentiment quantification research: (1) the clue-based sentiment extraction models used by the above researchers are often too limited. As pointed out by Wiebe and Riloff (2005), identifying positive or negative tweets by counting words in a dictionary usually has high precision but low recall. In the case of Twitter sentiment analysis, the performance will be even worse, since many words in tweets are not recorded in a dictionary. For example, LMAO is a positive "word" in Twitter, but it does not match any word in MPQA (Riloff and Wiebe 2003), which is a popular sentiment dictionary. (2) The correlation between sentiments and News events are only studied visually by observing their co-occurrence on a timeline (Sha et al. 2014; O'Connor et al. 2010), but to the best of our knowledge, there is no prior work that both quantitatively and qualitatively studies these correlations between Twitter sentiment and the News in Twitter to identify concerns caused by diseases and crisis.

As we summarized the Twitter sentiment classification and Twitter sentiment quantification research, there is a research gap between them. More specifically, the existing sentiment classification research does not quantify sentiment timeline trends from the classification results to provide insights into the sentiments. On the other hand, the existing sentiment quantification research often used a clue-based model, which has a low recall in terms of identifying sentiment tweets. In addition, the existing sentiment quantification work has only qualitatively correlated the sentiment timeline with real-world events, but has not provided a comprehensive, quantitative correlation between the sentiment timeline trend and the News timeline trend. This work is our attempt to fill this gap.

There are two objectives to achieve. The first objective is to automatically label datasets for training a Twitter sentiment classifier together with identifying News (Non-Personal) tweets. The purpose of identifying News tweets is that it can help filtering them out in the first step and then the Negative vs. Non-Negative classifier can be applied only to the Personal tweets. The second objective is to quantify the sentiment trends and News timeline trends from sentiment classification results, and compute a quantitative measure of correlation between them, to better understand the sentiment timeline trends relative to events in the real world.

3 Definitions

Definition 1 (*Personal Tweet*) A *Personal Tweet* is defined to be one that expresses its author's private states (Wilson and Wiebe 2003; Wilson et al. 2005). A private state can be a sentiment, opinion, speculation, emotion, or evaluation, and it cannot be verified by objective observation. In addition, if a tweet talks about a fact observed by the Twitter user, it is also defined as a Personal Tweet.

Example (*Personal Tweet*) "The boyfriend is STILL sick from the @fatburger he ate last Thursday. The doctor suspects listeria. :("

The purpose of this definition is to distinguish the tweets written word-by-word by the Twitter users from the News tweets redistributed in the Twitter environment, as mentioned above.

Definition 2 (*News Tweet*) A *News Tweet* (denoted with NT) is a tweet that is not a Personal Tweet. A News Tweet states an objective fact.

Example (*News Tweet*) "Measles outbreak reported in Honiara, Solomon Islands | Outbreak News Today http://fb.me/1hMxpNmrh"

Definition 3 (*Personal Negative Tweet and Personal Non-Negative Tweet*) If a tweet is a Personal Tweet, and it expresses negative emotions or attitude, it is a *Personal Negative Tweet* (denoted as PN). Otherwise, it is a *Personal Non-Negative Tweet* (denoted as PNN). Personal Non-Negative Tweets include personal neutral or personal positive tweets. A Personal Tweet is either a PN or a PNN.

Definition 4 (*Raw Tweet*) A *Raw Tweet tw* is defined as a tuple

$$tw = \ < tid, te, ty, h, t > \tag{1}$$

where *tid* is the unique tweet identifier; *te* is the tweet text; *ty* is the tweet type, which is a disease or crisis as the topic of this tweet (e.g., "Swine Flu" or "Bipolar Disorder", etc.); *h* is the tweet holder ("tweeter"), which is the person who posted the tweet *tw*; and *t* is the time when *tw* was posted.

Example (*Raw Tweet*) *tw* = <*TFare1*,"Everytime someone writes #tb i think of tuberculosis", *tuberculosis, StealYoKidney, 10/8/2014>.*

It means that the user "StealYoKidney" posted a tweet of "tuberculosis" on 10/8/2014, and the tweet's text is "Everytime someone writes #tb i think of tuberculosis".

Definition 5 (*Label*) Given a Raw Tweet *tw*, *O(tw)* is defined as the *Label* of the tweet, where

$$O(tw)\{PN, PNN, NT\} \tag{2}$$

such that *PN* is a Personal Negative Tweet, *PNN* is a Personal Non-Negative Tweet, and *NT* is a News Tweet.

Definition 6 (*Tweet Label*) Given a Raw Tweet *tw*, a *Tweet Label ts* is defined as a tuple

$$ts = \ < ty(tw), O(tw), h, t > \tag{3}$$

where *ty*(*tw*) is the type of tweet *tw* (e.g., listeria); *O*(*tw*) is the Label as defined in Definition 5; *h* is the Label holder, who is the person who posted the tweet *tw;* and *t* is the time when *tw* was posted.

Example (*Tweet Label*) Tweet Label for the tweet "Then I gotta go get the damn tuberculosis test?" is <tuberculosis, *PN*, user1, 4/5/2014>;

Tweet Label for the tweet "Signed off again with viral meningitis! Nice one" is <meningitis, *PNN*, user2, 5/7/2014>;

Tweet Label for the tweet "Listeria risk prompts recall of cheese from Unicer Foods in Canada" is <listeria, *NT*, user3, 6/8/2014>.

Definition 7 (*Tweet Label Dataset*) A Tweet Label Dataset is defined as a set of Tweet Labels of the same type collected at a specific time *t*.

$$TS_t = \{ts_1, ts_2, \ldots ts_n\} \tag{4}$$

Example (*Tweet Label Dataset*) two Raw Tweets were collected

<1, "Back to the wonderful world of listeria! WOO HOO! Here I come!", workin_with_S, 6/23/2014>

<2,"#Listeria outbreak causes Roos Foods to shut plant #FoodSafety http://t.co/mUOJMZqbUB #LegalUpdates", Nlegal_IMC, 6/23/2014>

The Tweet Label Dataset $TS_{6/23/2014}$ = {<listeria, *PNN*, workin_with_S, 6/23/2014>, <listeria, *NT*, Nlegal_IMC, 6/23/2014>}, where there are two Tweet Labels of "listeria" type on 6/23/2014.

Definition 8a (*Measure of Concern*) Given is TS_i as a Tweet Label Dataset (as defined in Definition 7) of a particular type (e.g., listeria) at a time *i*. The MOC M_i is defined as follows:

$$M_i = \left(\sum\nolimits_1^n \sigma(ts_j)\right)^2 \Big/ |TS_i| \tag{5}$$

where $\sigma(ts_j) = 1$ if $O(ts_j) = PN$ and $ts_j \in TS_i$; $\sigma(ts_j) = 0$ otherwise. Intuitively, M_i is the square of the total number of Personal Negative tweets that are posted at time i, divided by the total number of Raw Tweets of a particular type at the same time i. The MOC increases with the relative growth of Personal Negative tweets and with the absolute growth of Personal Negative tweets.

Definition 8b (*Non-Negative Sentiment*) Similarly, the *Non-Negative Sentiment* NN_i is defined as follows:

$$NN_i = \left(\sum\nolimits_1^n \alpha(ts_j)\right)^2 \Big/ |TS_i| \qquad (6)$$

where $\alpha(ts_j) = 1$ if $O(ts_j) = PNN$ and $ts_j \in TS_i$; $\alpha(ts_j) = 0$ otherwise. Intuitively, NN_i is the square of the total number of Personal Non-Negative tweets that are posted on time i, divided by the total number of Raw Tweets of a particular type at the same time i.

Definition 8c (*News Count*) Finally, the *News Count* NE_i is defined as follows:

$$NE_i = \sum\nolimits_1^n \beta(ts_j) \qquad (7)$$

where $\beta(ts_j) = 1$ if $O(ts_j) = NT$ and $ts_j \in TS_i$; $\beta(ts_j) = 0$ otherwise. NE_i is the total number of News Tweets at the time i. Note that the News Count is not normalized by the total number of Raw Tweets. The reason is that we are interested in studying the relationship between sentiment trends and News popularity trends (see Sect. 5). An absolute News Count is able to better represent the popularity of News.

Definition 9 (*Measure of Concern Timeline*) Given a series of time points $T = (1, 2, \ldots, j, k, \ldots n)$, where $j < k$, the *Measure of Concern Timeline* is defined as the time series

$$MOC[1:n] = (M_1, M_2, M_i, \ldots M_n) \qquad (8)$$

where M_i is the MOC at time i. Similarly, $NN[1{:}n]$ and $NE[1{:}n]$ are defined as the *Non-Negative Sentiment Timeline*, and the *News Count Timeline*, respectively. Figure 5 in Sect. 5 will show the visualization of an MOC Timeline.

Definition 10 (*Peak*) Given a timeline, a value X_i on the timeline is defined as a peak if and only if X_i is the largest value in a given time interval $[i - b,\ i + a]$. The time intervals $a > 0,\ b > 0$ can be chosen according to each specific case to limit the number of peaks. Peaks are defined for MOC timelines, Non-Negative timelines, and News Count timelines. Figure 5 in Sect. 5 will show the peaks as red or black dots on an MOC Timeline.

4 Two-step sentiment classification

In this section, we present the two-step sentiment classification and quantification method. As discussed earlier, the goal in sentiment classification is different from the one of classic sentiment classification of Tweets. Many News tweets are re-tweeted in Twitter. Classifying the tweets into Personal and News tweets in the first step can help consider only Personal tweets in a sentiment analysis in the next step (Negative vs. Non-Negative classification). Since we are interested in studying the correlation between the timeline trend of sentiments and of News, the detection of News tweets needs to be seamlessly integrated. Thus, our approach of classifying a tweet into one of the three classes—Personal Negative, Personal Non-Negative (including neutral and positive), and News—allows not only the classification but also correlation studies. An overview of our method is shown in Fig. 1.

Only English tweets, which were automatically detected during the data collection phase (see Table 5 for the data sets), are considered. As shown in Fig. 1, the sentiment classification problem is approached in two steps. First, for all English tweets, we separated Personal from News (Non-Personal) tweets. Second, after the Personal tweets were extracted by the most successful of the Personal/News Machine Learning classifier, these Personal tweets were used as input to another Machine Learning classifier, to identify Negative tweets. After News tweets, Personal Negative tweets, and Personal Non-Negative tweets were extracted, these tweets were used to compute the correlation between the sentiment trend and the News trend. The details of each "box" in Fig. 1 will be introduced in the rest of this section.

4.1 Pre-processing of features

In cases of disease surveillance on Twitter, the classical division of sentiments into positive and negative is inappropriate, because diseases are generally classified as negative. Positive emotions could arise as a result of relief about an epidemic subsiding, but we ignore this possibility. Thus, a two-point "Likert scale" with the points positive and negative would not cover this spectrum well. Rather, we started with an asymmetric four-point Likert scale of "strongly negative", "negative", "neutral", and "positive". We then combined "strongly negative" and "negative" into one category, and "neutral" and "positive" into

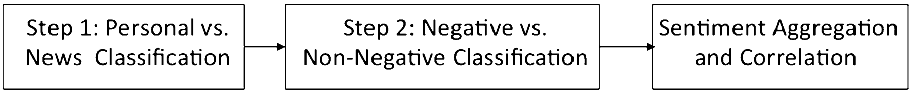

Step 1: Personal vs. News Classification
Step 2: Negative vs. Non-Negative Classification
Sentiment Aggregation and Correlation

Fig. 1 Overview of the two-step sentiment classification and quantification method

another. We use “Negative” as the name of the first category and “Non-Negative” for the second one. Thus, the problem reduces to a two-class classification problem, and a Personal tweet can either be a Negative tweet or a Non-Negative tweet.

Some features need to be removed or replaced. We first deleted the tweets starting with “RT”, which indicates that they are re-tweets without comments to avoid duplications. For the remaining tweets, the special characters were removed. The URLs in Twitter were replaced by the string “url”. Twitter’s special character “@” was replaced by “tag”. For punctuations, “!” and “?” were substituted by “excl” and “ques”, respectively, and any of “.,:;−|+=/” were replaced by “symb”. Twitter messages were transformed into vectors of words, such that every word was used as one feature, and only unigrams were utilized for simplicity.

4.2 Tweet sentiment classification

In the following, we present Personal vs. News classifiers and Negative vs. Non-Negative classifier:

4.2.1 Clue-based tweet Labeling

The clue-based classifier parses each tweet into a set of tokens and matches them with a corpus of Personal clues. There is no available corpus of clues for Personal versus News classification, so we used a subjective corpus MPQA (Riloff and Wiebe 2003) instead, on the assumption that if the number of strongly subjective clues and weakly subjective clues in the tweet is beyond a certain threshold (e.g., two strongly subjective clues and one weakly subjective clue), it can be regarded as Personal tweet, otherwise it is a News tweet. The MPQA corpus contains a total of 8221 words, including 3250 adjectives, 329 adverbs, 1146 any-position words, 2167 nouns, and 1322 verbs. As for the sentiment polarity, among all 8221 words, 4912 are negatives, 570 are neutrals, 2718 are positives, and 21 can be both negative and positive. In terms of strength of subjectivity, among all words, 5569 are strongly subjective words, and the other 2652 are weakly subjective words.

Twitter users tend to express their personal opinions in a more casual way compared with other documents, such as News, online reviews, and article comments. It is expected that the existence of any profanity might lead to the conclusion that the tweet is a Personal tweet. We added a set of 247 selected profanity words (Ji 2014a) to the corpus described in the previous paragraph. USA law, enforced by the Federal Communication Commission, prohibits the use of a short list of profanity words in TV and radio broadcasts (FederalCommunicationsCommittee 2014). Thus, any word from this list in a tweet clearly indicates that the tweet is not a News item.

We counted the number of strongly subjective terms and the number of weakly subjective terms, checked for the presence of profanity words in each tweet and experimented with different thresholds. A tweet is labeled as Personal if its count of subjective words surpasses the chosen threshold; otherwise it is labeled as a News tweet.

In clue-based classification, if the threshold is set too low, the precision might not be good enough. On the other hand, if the threshold is set too high, the recall will be decreased. The advantage of a clue-based classifier is that it is able to automatically extract Personal tweets with more precision when the threshold is set to a higher value.

Because only the tweets fulfilling the threshold criteria are selected for training the “Personal vs. News” classifier, we would like to make sure that the selected tweets are indeed Personal with high precision. Thus, the threshold that leads to the highest precision in terms of selecting Personal tweets is the best threshold for this purpose.

The performance of the clue-based approach with different thresholds on human-annotated test datasets is shown in Table 1. More detailed information about the human-annotated dataset is shown in Sect. 4.3.2.2. Among all the thresholds, s3w3 (3 strong, 3 weak) achieves the highest precision on all three human annotated datasets. In other words, when the threshold is set so that the minimum number of strongly subjective terms is 3 and the minimum number of weakly subjective terms is 3, the clue-based classifier is able to classify Personal tweets with the highest precision of 100 % but with a low recall (15 % for epidemic, 7 % for mental health, 1 % for clinical science).

Table 1 Results of Personal tweets classification with different thresholds (Precision/Recall)

Threshold	Dataset		
	Epidemic	Mental health	Clinical science
s1w0	0.61/0.69	0.55/0.74	0.48/0.58
s1w1	0.64/0.48	0.53/0.63	0.51/0.52
s1w2	0.70/0.24	0.53/0.38	0.61/0.40
s1w3	0.75/0.18	0.50/0.20	0.58/0.22
s2w0	0.86/0.37	0.53/0.40	0.75/0.42
s2w1	0.86/0.28	0.53/0.38	0.73/0.38
s2w2	0.91/0.15	0.51/0.24	0.76/0.26
s2w3	0.91/0.15	0.37/0.10	0.80/0.16
s3w0	1.00/0.21	0.79/0.21	0.89/0.16
s3w1	1.00/0.21	0.79/0.21	0.88/0.14
s3w2	1.00/0.15	0.84/0.15	0.86/0.12
s3w3	1.00/0.15	1.00/0.07	1.00/0.01

4.2.2 *Machine learning classifiers for personal tweet classification*

To overcome the drawback of low recall in the clue-based approach, we combined the high precision of clue-based classification with Machine Learning-based classification in the Personal vs. News classification, as shown in Fig. 2. Suppose that the collection of Raw Tweets of a unique type (e.g., tuberculosis) is T. After the pre-processing step, which filters out non-English tweets, re-tweets, and near-duplicate tweets, the resulting tweet dataset is $T' = \{tw_1, tw_2, tw_3, \ldots, tw_n\}$, which is a subset of T, and is used as the input for the clue-based method for automatically labeling datasets for training a Personal vs. News classifier as shown in Fig. 2.

In the clue-based step for labeling training datasets, each tw_i of T' is compared with the MPQA dictionary (Riloff and Wiebe 2003). If tw_i contains at least three strongly subjective clues and at least three weakly subjective clues, tw_i is labeled as a Personal tweet. Similarly, tw_i is compared with a News stopword list (Ji 2014b) and a profanity list (Ji 2014a). The News stopword list contains 20+ names of highly influential public health News sources and the profanity list has 340 commonly used profanity words. If tw_i contains at least one word from the News stopword list and does not contain any profanity word, tw_i is labeled as a News tweet. For example, the tweet "Atlanta confronts tuberculosis outbreak in homeless shelters: By David Beasley ATLANTA (Reuters)—Th… http://yhoo.it/1r88Lnc #Atlanta" is labeled as a News tweet, because it contains at least one word from the News stopword list and does not contain any profanity word. We mark the set of labeled Personal tweets as $T_p{'}$, and the set of labeled News tweets as $T_n{'}$, note that $(T_p{'} \cup T_n{'}) \subseteq T'$.

The next step is the Machine Learning-based method. The two classes of data T'_p and T'_n from the clue-based labeling are used as training datasets to train the Machine Learning models. We used three popular models: Naïve Bayes, Multinomial Naïve Bayes, and polynomial-kernel Support Vector Machine. After the Personal vs. News classifier is trained, the classifier is used to make predictions on each tw_i in T', which is the preprocessed tweets dataset. The goal of Personal vs. News classification is to obtain the Label for each tw_i in the tweet database T', where the Label $O(ts_i)$ is either *Personal* or *NT* (News Tweet). Label was introduced in Definition 5, whereby Personal could be PN or PNN.

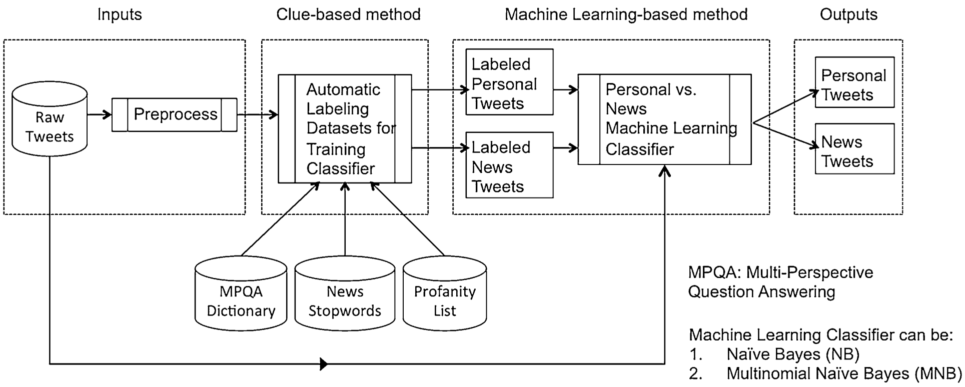
4.2.3 *Negative sentiment classifier*

As shown in Fig. 1, after a classifier for Personal tweets in step 1 is built, the second step in the sentiment classification is to classify the set of *Personal* tweets $T'' = \{tw_i : O(tw_i) = \text{Personal}, tw_i \in T'\}$ into Personal Negative (PN) or Personal Non-Negative (PNN) tweets. Figure 3 shows the process of classification in this second step. In the rest of this section, Negative is used to refer to the Personal Negative and Non-Negative is used to refer to the Personal Non-Negative.

In terms of training the classifier for Negative vs. Non-Negative classification, the ideal training dataset must be large and contain little noise. Manual annotation of a training dataset is possible, but this process usually requires different annotators to independently label each tweet and to calculate their degree of agreement. This limits the fast generation of large-sized training datasets. Pang and Lee (2008) listed a few annotated corpuses used in previous work in the field of sentiment analysis. These corpuses cover topics such as customer reviews of products and restaurants. However, to the best of our knowledge, there is no disease-related annotated corpus that can be used as a training dataset to distinguish Negative tweets from Non-Negative tweets.

In order to build the training datasets for Negative versus Non-Negative classification (TR-NN), we formed a whitelist and blacklist of stopwords using predefined

Fig. 2 Personal vs. News (Non-Personal) Classification



MPQA: Multi-Perspective Question Answering

Machine Learning Classifier can be:
1. Naïve Bayes (NB)
2. Multinomial Naïve Bayes (MNB)
3. Support Vector Machine (SVM)

Fig. 3 Negative vs. Non-Negative Classification

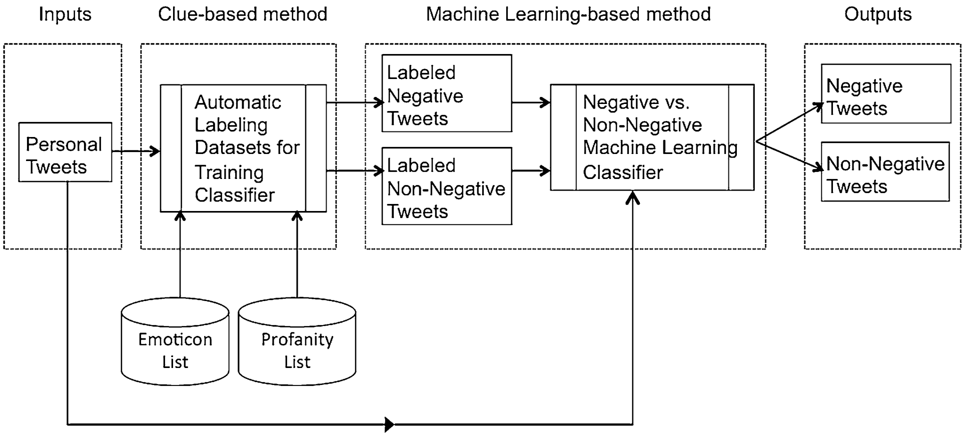

Inputs
Clue-based method
Machine Learning-based method
Outputs
Personal Tweets
Automatic Labeling Datasets for Training Classifier
Labeled Negative Tweets
Labeled Non-Negative Tweets
Negative vs. Non-Negative Machine Learning Classifier
Negative Tweets
Non-Negative Tweets
Emoticon List
Profanity List

Table 2 Partial list of the emoticons used

Negative	Non-Negative
-.-	:o)
:C	:]
:c	:]
;c	:3
;C	:c)

emoticons. An emoticon is a combination of characters that form a pictorial expression of one's emotions. Emoticons have been used as important indicators of sentiments in previous research. We combined the emoticon lists used by Go et al. (2009), Pak and Paroubek (2010), and Agarwal et al. (2011). A partial list of emoticons is in Table 2.

The whitelist and blacklist of stopwords for building TR-NN are described in Table 3. The whitelist is used for extracting while the blacklist is used for eliminating information. A tweet is extracted as a Negative tweet if and only if this tweet contains at least one stopword (or emoticon) from the Negative whitelist, and does not contain any stopword (or emoticon) from the Negative blacklist. A tweet is extracted as Non-Negative using similar lists, a Non-Negative whitelist, and a corresponding blacklist. For example, the tweet "They are going to take fluid from around the spinal cord to see if she has meningitis… :(" is extracted as a Negative tweet, because it contains at least one stopword from the Negative whitelist and no words from the Negative blacklist.

As shown in Fig. 3, the emoticons contained in the tweets are used to generate the training dataset TR-NN. Tweets were labeled as *PN* or *PNN* based on the emoticons they contained. More specifically, if a tweet contains at least one negative emoticon or at least one word from the profanity list that has 247 selected profanity words (Ji 2014a), it is labeled as *PN*. If a tweet contains at least one non-negative emoticon or at least one positive emoticon, it is labeled as a *PNN*. These two categories (*PN* and *PNN*) of labeled tweets were combined into the training dataset TR-NN for Negative vs. Non-Negative classification. Table 4 shows examples of tweets in TR-NN. The set of labeled *PN* tweets is marked as T''_{ne}, and the set of labeled *PNN* tweets is marked as T''_{nn}, and $(T''_{ne} \cup T''_{nn}) \subseteq T'$. Similarly, T''_{ne} and T''_{nn} are used to train the Negative vs. Non-Negative classifier, and the classifier is used to make predictions on each tw_i in T'', which is the set of Personal tweets. The goal of Negative vs. Non-Negative classification is to obtain the Label for each tw_i in the tweet database T'', where the Label $O(tw_i)$ is either *PN* or *PNN*. (There are no News tweets at this stage).

After step 1 (Personal tweets classification) and step 2 (sentiment classification), for a unique type of tweets (e.g., tuberculosis), the Raw Tweet dataset T is transformed into a series of Tweet Label datasets TS_i. Recall from the definition section that TS_i is the Tweet Label dataset for time i, and $TS_i = \{ts_1, ts_2, ts_3,\ldots, ts_n\}$, where $O(ts_i)$ is either *PN*, or *PNN*, or *NT*.

4.3 Experimental results of the classification approach

4.3.1 Data collection and description

We implemented a data collector using the Twitter API version 1.1 and Twitter4J library (Twitter4J 2014) to collect real-time tweets containing certain specified health-

Table 3 Whitelist and blacklist of stop words for building TR-NN

	Negative	Non-Negative
Whitelist	Negative emoticons and profanities	Neutral and positive emoticons
Blacklist	News keywords, retweet	News keywords, retweet

Table 4 Examples of Personal Negative and Personal Non-Negative tweets in training dataset TR-NN

Personal Negative	I hate TuBerculosis. they get on my damn nerves. They the reason Chrissy don't lotion his ankles or elbows
	Uh ohhhh!:(CDC: 1 dead, 7 others sickened by listeria traced to cheese
Personal Non-Negative	Car's so fresh and so clean. Time to lay out in the sun with some ruby beer and work on my melanoma:)
	Preventing swine flu, one ham at a time.:)

Table 5 The statistics of the collected dataset

Dataset Id	Tweet type	Total number of tweets	Number of non-english tweets	Number of tweets after preprocessing
1	Listeria	13,572	1979	4544
2	Influenza	1,509,609	716,901	527,489
3	Swine Flu	73,974	35,970	20,430
4	Measles	166,555	8808	60,016
5	Meningitis	159,393	52,824	42,229
6	Tuberculosis	215,083	147,350	33,030
7	Major Depression	2,269,885	121,649	884,304
8	Generalized Anxiety Disorder	380,094	271,758	71,978
9	Obsessive–compulsive Disorder	434,571	168,061	171,211
10	Bipolar Disorder	51,520	7416	20,915
11	Air Disaster	15,871	681	5765
12	Melanoma Experimental Drug	86,757	9858	40,261

related keywords (e.g., listeria), along with associated user profile information for subsequent analysis. The overall data collection process can be described as "ETL" (Extract-Transform-Load) approach, as it is widely used in Data Warehousing. The data was collected in JSON format from the Twitter Streaming API. (This is the "Extract" step). Then the raw JSON data was parsed into relational data, such as tweets, tweet_mentions, tweet_place, tweet_tags, tweet_urls, and users (Transform step). Finally, the relational data were stored into our MySQL relational database (Load step).

The current prototype system has collected a total of 15+ million tweets in 12 datasets. These datasets include six infectious diseases: Listeria, influenza, swine flu, measles, meningitis, and tuberculosis; four mental health problems: Major depression, generalized anxiety disorder, obsessive–compulsive disorder, and bipolar disorder; one crisis: Air disaster; and one clinical science issue: Melanoma experimental drug. The core component uses the Twitter Streaming API for collecting epidemics-related real-time tweets. The tweets were collected from March 13 2014 to June 29 2014. The statistics of the collected datasets are shown in Table 5.

For each tweet type, the tweets were collected according to the keywords of the dataset. These keywords are shown in the "Appendix" Section. The language of tweets is automatically identified by Twitter4J library during the data collection phase. For example, if the value of the tweet attribute "lang" is "en", that means this tweet is an English tweet. If the value of tweet attribute is "fr", it means that this tweet is a French tweet. Only English tweets are used in our experiments. As shown in Table 5, some datasets have a larger portion of non-English tweets, for example, influenza, swine flu, and tuberculosis compared with other datasets.

The pre-processing step filters out re-tweets and near-duplicate tweets. Two tweets are considered near-duplicates of each other, if they contain the same tokens (words) in the same order; however, they may contain different capitalization of words, different URLs and different special characters such as @, # etc. For example, the two tweets (1) "SEVEN TONS OF #HUMMUS RECALLED OVER LISTERIA FEARS… http://t.co/IUU5SiJgjG" and (2) "seven tons of hummus recalled over @listeria fears—http://t.co/dBgAk1heo4." are near-duplicates, thus only one tweet (randomly chosen) is kept in the database.

4.3.2 Evaluation

To the best of our knowledge, there are no evaluation datasets for the performance of sentiment classification of health-related tweets. To compare the three previously discussed classifiers, Naïve Bayes, Two-Step Multinomial Naïve Bayes, and Two-Step Polynomial-Kernel Support Vector Machine, we created one group of test datasets using the clue-based method and a second group of test datasets using human annotation, in order to evaluate the usability of our approach. Weka's implementations (Hall et al. 2009) of Naïve Bayes, Multinomial Naïve Bayes, and polynomial-kernel SVM with default parameter configurations were used for the experiments.

4.3.2.1 Clue-based annotation for test dataset The clue-based annotation of the test dataset was done as follows. We first automatically extracted the Personal tweets and News tweets by the clue-based approach described in Sect. 4.2.1 and labeled them as Personal or News. Then we randomly divided the labeled dataset into three partitions and used two partitions for training the three different classifiers. Finally, we compared the different classifiers' accuracies on the third partition of labeled data. For example, for Dataset 3 in Table 5, in the classification step, 2899 Personal tweets and 508 News tweets were automatically extracted using the MPQA corpus (Riloff and Wiebe 2003). We randomly divided these tweets into training and test datasets, resulting in 1933 Personal and 339 News tweets as training dataset, and the remaining 966 Personal tweets and 169 News tweets as test dataset. A similar emoticon-based approach was used to automatically generate a training dataset and a test dataset for Negative vs. Non-Negative classification.

4.3.2.2 Human annotation for test dataset Because the clue-based annotation method is automatic, it is relatively easy to generate large samples. However, the drawback is that the training and testing datasets are extracted by the same clue-based annotation rule, thus the results might carry a certain bias. In order to more fairly evaluate the usability of our approach, we created a second test dataset by human annotation, which is described as follows.

We extracted three test data subsets by random sampling from all tweets from the three domains epidemic, clinical science, and mental health, collected in the year 2015. Each of these subsets contains 200 tweets. Note that the test tweets are independent from the training tweets that were collected in the year 2014. One professor and five graduate students annotated the tweets, with each tweet annotated by three people. The instructions for annotators are shown in the "Appendix". Annotators were asked to assign a value of 1 if they considered a tweet to be Personal, and a value of 0 if they considered it to be News, according to the instructions they were given. If a tweet was labeled as a Personal tweet by an annotator, s/he was asked to further label it as Personal Negative or Personal Non-Negative tweet. We utilized Fleiss' Kappa (Fleiss 1971) to measure the inter-rater agreement between the three annotators of each tweet. Table 6 presents the agreement between human annotators. For each tweet, if at least two out of three annotators agreed on a Label (Personal Negative, Personal Non-Negative, or News), we labeled the tweet with this sentiment. Table 7 shows the numbers of tweets with different labels. For example, the fraction 25/200 for Negative tweets in "epidemic" means that out of the 200 human-annotated epidemic tweets, 25 tweets were labeled as Personal Negative tweets. The total number of tweets in each dataset does not add up to 200, because in some cases each of the three annotators classified a tweet differently. Tweets for which no majority existed were omitted from the analysis.

4.3.3 Classification results

The results of the two-step classification approach are shown in this section. The performance was tested separately with the clue-based annotated test dataset and the human annotated test dataset.

4.3.3.1 Results with clue-based annotated test dataset We compared the previously discussed classifiers: Two-Step Naïve Bayes, Two-Step Multinomial Naïve Bayes, and Two-Step Polynomial-Kernel Support Vector Machine. As previously discussed, the labeled dataset was

Table 6 Agreement between human annotators

Domains	Epidemic	Clinical science	Mental health
Total number of tweets	200	200	200
At least two annotators agree	192/200	194/200	188/200
Fleiss' Kappa Coefficient	0.4	0.54	0.33

Table 7 Statistics regarding human annotated dataset

Domains	Epidemic	Clinical science	Mental health
Total number of tweets	200	200	200
Personal Negative tweets	25/200	10/200	34/200
Personal Non-Negative tweets	34/200	34/200	58/200
News tweets	133/200	150/200	96/200

Table 8 Size of experimental training and test datasets for two-step classification (PN is Personal Negative and PNN is Personal Non-Negative)

Classifier	Step 1		Step 2					
	MNB/NB/SVM		MNB		NB		SVM	
Dataset Id	Training (Personal/News)	Testing (Personal/News)	Training (PN/PNN)	Testing (PN/PNN)	Training (PN/PNN)	Testing (PN/PNN)	Training (PN/PNN)	Testing (PN/PNN)
1	206/238	102/119	18/8	8/4	19/8	9/4	20/8	9/4
2	83,032/7206	41,515/3602	32,689/5346	16,344/2672	32,420/5244	16,209/2621	32,700/5359	16,350/2679
3	1933/339	966/169	634/226	316/113	629/228	314/113	636/226	317/113
4	5808/3770	2904/1885	630/112	314/55	618/112	309/56	647/114	323/56
5	3501/1094	1750/546	658/306	329/152	650/306	325/152	662/307	330/153
6	2863/756	1431/378	412/144	205/72	402/132	201/65	414/147	207/73
7	262,991/5163	131,495/2581	29,153/4320	14,576/2160	29,178/4314	14,589/2157	29,189/4326	14,594/2163
8	8159/1301	4079/650	2446/725	1222/362	2428/720	1213/360	2454/732	1226/365
9	27,972/673	13,985/336	5714/2046	2856/1023	5680/2030	2839/1014	5714/2060	2857/1029
10	5160/303	2580/151	548/92	273/46	546/90	272/45	548/95	274/47
11	313/314	156/156	28/8	13/3	28/7	14/3	30/10	14/5
12	7180/1154	3590/576	648/160	324/79	640/158	320/78	648/160	323/79

randomly divided into three partitions and we used two partitions for training the three different classifiers. The detailed training and test dataset sizes are shown in Table 8. Note that the test datasets for each classifier in step 2 can be different. The reason is that different classifiers extract different numbers of Personal tweets in the first step, thus the test data in the second step, which is extracted from the previously extracted Personal tweets, can also be different for the three classifiers. The two-step sentiment classification accuracy on individual datasets (1–12) is shown in Table 9 and confusion matrices of the best classifiers in terms of accuracy are shown in Table 10; similarly, the classification accuracy and confusion matrices of the best classifiers for the three domains (epidemic, mental health, clinical science) are shown in Tables 11 and 12, respectively.

On individual datasets, all three two-step methods show good performance. SVM is slightly better than the other two classifiers for most of the datasets. For the domain datasets, which combine individual datasets according to their domains, all three two-step methods also exhibit good performance. SVM again slightly outperforms the other two classifiers in all three domains.

4.3.3.2 Results with human annotated test dataset In order to evaluate the usability of two-step classification, Personal vs. News classification and Negative vs. Non-Negative classification were also evaluated with human annotated datasets.

Table 9 Results of S1A/S2A (S1A = step one accuracy and S2A = step two accuracy) on individual dataset (rounded to 2 decimal places)

Dataset Id	2S-MNB	2S-NB	2S-SVM
1	0.91/0.92	0.90/0.77	0.99/1.00
2	0.97/0.95	0.96/0.92	1.00/0.97
3	0.97/0.90	0.95/0.94	1.00/0.97
4	0.94/0.89	0.90/0.97	1.00/0.97
5	0.95/0.91	0.93/0.97	1.00/0.98
6	0.96/0.86	0.92/0.97	1.00/0.99
7	0.98/0.97	0.98/0.98	1.00/0.99
8	0.96/0.90	0.95/0.96	1.00/0.96
9	0.98/0.96	0.96/0.98	1.00/0.98
10	0.96/0.90	0.95/0.98	1.00/1.00
11	0.89/0.81	0.88/0.82	0.96/0.95
12	0.92/0.87	0.89/0.98	1.00/0.98

- *Personal vs. News Classification* We compared our Personal vs. News classification method with three baseline methods. (1) A naïve algorithm that randomly picks a class. (2) The clue-based classification method described in Sect. 4.2.1. Recall that in the clue-based method, if a tweet contains more than a certain number of strongly subjective terms and a certain number of weakly subjective terms, it is regarded as a Personal tweet, otherwise as a News tweet. (3) A URL-based method. In URL-based method, if a tweet contains an URL, it is classified as a News tweet; otherwise the

Table 10 Confusion matrices of the best classifier on each dataset (*Step 1* positive class is Personal and Negative class is News, *Step 2* positive class is Personal Negative and Negative class is Personal Non-Negative)

Dataset Id	Step 1					Step 2			
	Best classifier	True pos.	False neg.	False pos.	True neg.	True pos.	False neg.	False pos.	True neg.
1	2S-SVM	101	1	2	117	9	0	0	4
2	2S-SVM	41,513	2	12	3590	16,121	229	372	2307
3	2S-SVM	966	0	3	166	311	6	5	108
4	2S-SVM	2904	0	3	1882	320	3	9	47
5	2S-SVM	1749	1	1	545	323	7	5	148
6	2S-SVM	1431	0	5	373	207	0	4	69
7	2S-SVM	131,494	1	8	2573	14,494	100	135	2028
8	2S-SVM	4079	0	2	648	1205	21	40	325
9	2S-SVM	13,984	1	1	335	2819	38	51	978
10	2S-SVM	2580	0	5	146	274	0	0	47
11	2S-SVM	156	0	11	145	14	0	1	4
12	2S-SVM	3571	19	1	575	318	5	3	76

Table 11 Results of S1A/S2A (*S1A* step one accuracy and *S2A* step two accuracy) on individual domain

Dataset Id	2S-MNB	2S-NB	2S-SVM
Epidemic (1, 2, 3, 4, 5, 6)	0.95/0.95	0.94/0.93	0.99/0.97
Mental health (8, 9, 10)	0.97/0.96	0.96/0.97	1.00/0.97
Clinical science (12)	0.92/0.87	0.89/0.98	1.00/0.98

tweet is classified as a Personal tweet. The classification accuracies of different methods and confusion matrices of the best classifiers are presented in Tables 13 and 14, respectively. The results show that 2S-MNB and 2S-NB outperform all three baselines in most of the cases. Surprisingly, 2S-SVM does not perform as well as on the clue-based annotated test dataset. It is possible that SVM overfitted to the clue-based annotated dataset, since SVM is a relatively complex model and it infers too much from the training datasets. Overall, all methods exhibit a better performance on the epidemic dataset than on the other two datasets. In addition, as we compare the ML-based approaches (2S-MNB, 2S-NB, 2S-SVM), the ML-based approaches outperform the clue-based approaches in most of the cases. This means that although the ML-based approaches utilize the simple clue-based rules to automatically label the training data, they also learn some emotional patterns that cannot be distinguished by MPQA corpus. Some unigrams are learned by the ML-based methods and are shown to be useful for the classification, which will be discussed later.

- *Negative vs. Non-Negative Classification* The second step in the two-step classification algorithm is to separate Negative tweets from Non-Negative tweets. As discussed in Sect. 4.2, the training datasets are automatically labeled with emoticons and words from a profanity list, and then the classifier is trained by one of the three models, Multinomial Naïve Bayes (MNB), Naïve Bayes (NB), and Support Vector Machine (SVM). The accuracies of Negative vs. Non-Negative classification and confusion matrices of the best classifiers for human annotated datasets are shown in Tables 15 and 16, respectively. 2S-MNB outperforms the other two algorithms on the epidemic dataset, and 2S-NB outperforms the other two algorithms on the mental health and clinical science datasets. All three classifiers perform better than the random-select baseline, which generates an average of 50 % accuracy. We

Table 12 Confusion matrices of the best classifier on individual domain (*Step 1* positive class is Personal and Negative class is News, *Step 2* positive class is Personal Negative and Negative class is Personal Non-Negative)

Dataset Id	Best classifier	Step 1				Step 2			
		True pos.	False neg.	False pos.	True neg.	True pos.	False neg.	False pos.	True neg.
Epidemic	2S-SVM	47,916	9	18	6695	17,046	245	398	2652
Mental health	2S-SVM	20,602	6	3	1137	4290	69	88	1353
Clinical science	2S-SVM	3571	19	1	575	318	5	3	76

Table 13 Accuracy of Personal vs. News classification on human annotated datasets

Dataset	Random	Clue-based	URL-based	2S-MNB	2S-NB	2S-SVM
Epidemic	0.52	0.77	0.82	0.86	0.87	0.71
Mental health	0.48	0.56	0.68	0.72	0.78	0.59
Clinical science	0.49	0.82	0.72	0.74	0.71	0.36

Table 14 Confusion matrices of the best Personal vs. News classifier on human annotated datasets (positive class is Personal and Negative class is News)

Dataset Id	Best classifier	True positive	False negative	False positive	True negative
Epidemic	2S-NB	52	15	11	122
Mental health	2S-NB	81	23	21	75
Clinical science	Clue-based	21	29	7	143

Table 15 Negative vs. Non-Negative classification results on human annotated datasets

Dataset Id	2S-MNB	2S-NB	2S-SVM
Epidemic	0.73	0.59	0.59
Mental health	0.63	0.65	0.57
Clinical science	0.64	0.73	0.68

can see that although the classifier is trained with tweets containing profanity and tweets containing emoticons, the classifier is still able to perform with an average accuracy of 70+% on human annotated test datasets. Overall, 2S-NB and 2S-MNB both achieved good Negative vs. Non-Negative classification accuracy in terms of accuracy and simplicity, followed by 2S-SVM.

4.3.4 Error analysis of sentiment classification output

We analyzed the output of sentiment classification. As discussed in Sect. 4.3.2, we manually annotated 600 tweets as Personal Negative, Personal Non-Negative, and News. We used 2S-MNB, which achieved the best accuracy in our experiments described in Sect. 4.3.3, to classify each of the 600 manually annotated tweets as Personal Negative, Personal Non-Negative, or News. Then we analyzed the tweets that were assigned different labels by 2S-MNB and by the human annotators.

For the Personal vs. News classification, we found two major types of errors.

1. The tweet is in fact a Personal tweet, but is classified as a News tweet. By manually checking the content, we found that these tweets are often users' comments on News items (Pointing by URL) or users are citing the News. There are 27 out of all 140 errors belonging to this type. One possible solution to reduce this type of error is that we can calculate what percentage of the tweet text appears in the web page pointed to by the URL. If this percentage is low, it is probably a Personal tweet since most of the tweet text is the user's comment or discussion, etc. Otherwise, if the percentage is near 100 %, it is more likely a News tweet since the title of a news article is often pasted into the tweet text.
2. The tweet is in fact a News item, but is classified as a Personal tweet. Those misclassified tweets are News items that have "personal" titles, and mostly have a question as title. There are 48 out of all 140 errors belonging to this type. One possible solution is to check the similarity between the tweet text and the title of the web page content pointed to by the URL. If both are highly similar to each other, the tweet is more likely a News item. Those two types of errors together cover 54 % (75/140) of the errors in Personal vs. News classification.

For Negative vs. Non-Negative classification, in 50 % (30/60) of all errors, the tweet is in fact Negative, but is classified as Non-Negative. One possible improvement is to incorporate "Negative phrase identification" to

Table 16 Confusion matrices of the best Personal Negative vs. Personal Non-Negative classifier on human annotated datasets (Positive class is Personal Negative and Negative class is Personal Non-Negative)

Dataset Id	Best classifier	True positive	False negative	False positive	True negative
Epidemic	2S-MNB	17	8	8	26
Mental health	2S-NB	18	16	16	42
Clinical science	2S-NB	4	6	6	28

complement the current ML paradigm. The appearance of negative phrases such as “I feel bad”, “poor XX”, and “no more XX” are possible indicators of Negative tweets. Examples of misclassified tweets are as follows:

“This is the scariest chart I’ve made in awhile http://t.co/3MH5exZjSh http://t.co/oc9lyEO0XY” (Personal tweet classified as News tweet).

“My OCD has been solved! Get our newsletter here: http://t.co/fAxsHjaIn4 http://t.co/1Jhkbta2Px” (Personal tweet classified as News tweet).

“What is Generalized Anxiety Disorder? (GAD #1) http://t.co/y32GmkYhkh #Celebrity #Charity http://t.co/EYDupOLxY8” (News tweet classified as Personal tweet).

“Basal Cell Carcinoma is the most common form of skin cancer. Do you know what to look for? http://t.co/hmofWTApG9” (News tweet classified as Personal tweet).

“@Jonathan_harrod I know there is some research going on, but… Measles kills and us easily spread. @mercola” (Negative tweet classified as Non-Negative tweet).

“Having a boyfriend with diagnosed OCD is not easy task, let me tell ya” (Negative tweet classified as Non-Negative tweet).

4.3.5 *Contribution of unigrams*

In order to illustrate which unigrams are most useful for the classifiers’ predictions, ablation experiments were performed on Personal vs. News classification and Negative vs. Non-Negative classification on the three human annotated test datasets. The classifier 2S-MNB was used since it took less time to train and has one of the best average accuracies on human-annotated test dataset. 2S-MNB was trained with the automatically generated data from the Epidemic, Mental Health, and Clinical Science domains collected in the year 2014. Then the trained classifiers were used to classify the sentiments of human annotated datasets collected in the year 2015, where unigrams were removed from the test dataset one at a time, in order to study each removed unigram’s effect on accuracy. The change of classification accuracy was recorded each time, and the unigram that leads to the largest decrease in accuracy (when removed) is the most useful one for predictions. Table 17 shows the ablation experiments for Personal vs. News classification. For example, the unigrams “i”, “plz”, “lol” are not in MPQA corpus but are learned by the ML classifier 2S-MNB as the most important unigrams contributing to classification. Some words that are closely related to sentiment polarity are also shown in the list. For example, “bitch”, “love”, and “risk” are strong indicators for Personal vs. News classification. We did not find any useful unigram in Negative vs. Non-Negative classification by this ablation experiment.

4.3.6 *Bias of Twitter data*

Twitter may give a biased view, since people who are tweeting are not necessarily a very representative sample of the population. As pointed out by Bruns and Stieglitz (2014), there are two questions to be addressed in terms of generalizing collected Twitter data. (1) Does Twitter data represent Twitter? (2) Does Twitter represent society? To answer the first question, according to the documentation (Twitter 2014b), the Twitter Streaming API returns at most 1 % of all the tweets produced on Twitter at any given time. Once the number of tweets matching given parameters (keywords, geographical boundary, user ID) is beyond the 1 % of all the tweets, Twitter will begin to sample the data that it returns to the user. To mitigate this, we utilized highly specific keywords (e.g., h1n1, h5n1) for each tweet type (e.g., flu) to increase the coverage of collected data (Morstatter et al. 2013). These keywords are shown in “Appendix” Section. As for the second question, Mislove et al. (2011) has found that the Twitter users significantly over-represent the densely populated regions of the USA, are predominantly male, and represent a highly non-random sample of the race/ethnicity distribution. To reduce the bias of collected Twitter data, we defined the MOC in relative terms in Sect. 3. It depends on the fraction of all tweets obtained during the day that have been classified as “Personal Negative” tweets. The MOC analysis will be discussed in more detail in Sect. 5.

Table 17 Most important unigrams in Personal vs. News classification

Dataset	Unigrams with most importance
Epidemic	url, i, case, but, o, plz, night, etc., same, children, you, really, he, what, would, thing, vaccine, am, of, me, don’t, bitch, actually, love, this, know, kind, 92, flu, with, the
Mental health	url, disorder, often, bipolar, psychology, i, had, much, brown, his, into, lol, everything, 2014, forever, ocd, says, depression, my, isn’t, if, is, im, overdrive, to, so, have, the
Clinical science	Melanoma, health, http, co, risk, prevention, app, video, shown, shows, i, new, s, bladder, your, daily, now, skincancer, cases, skin, police, increased, number, breast, be, get, and, former, am, an, of, study, beauty, is, tri, yervoy, raises, via, for, problems, researchers

5 Concern sentiment trend analysis in public health

We are interested in making the sentiment classification results available for public health monitoring, especially the results of computing the MOC, to monitor public sentiments toward different types of diseases. Unlike the previous research on *qualitatively* comparing the co-occurrence of sentiment trends with News broadcasts, this paper approaches the problem of *quantitatively* studying the correlation between Twitter sentiment trends and News trends caused by various epidemics. The correlation process is shown in Fig. 4. There are three inputs for the correlation process. The News tweets are the outputs in the first step, as shown in Fig. 2; the Personal Negative tweets and the Personal Non-Negative tweets are the outputs in the second step, as shown in Fig. 3.

Given a tweet type, after the two-step sentiment classification method has been applied to the raw tweets, Tweet Label datasets D, introduced in Definition 7, are generated. By Definitions 8a, b, c and 9, we can produce three timelines: MOC[1:n], NN[1:n], NE[1:n], which are timelines for MOC, Non-Negative sentiment, and News, respectively.

Next, three sets of peaks P_1, P_2, and P_3 are generated from NE[1:n], MOC[1:n], and NN[1:n], respectively. Peaks were introduced in Definition 10 of Sect. 3. The time interval tx is set to 7 days. We are interested in the correlation between P_1 and P_2 (peaks of News and peaks of MOC), and the correlation between P_1 and P_3 (peaks of News and peaks of Non-Negative sentiments). The Pearson Correlation Coefficient (PCC) appears to be a natural way to measure the correlation between two time series, since the PCC is good at measuring the similarity of two linearly dependent variables. However, for the problem addressed here, as we are interested in the News about outbreaks of epidemics, it makes more sense to measure the similarity between the peaks. We utilized the Jaccard Coefficient for this purpose and define the correlations as follows:

$$JC(\mathrm{MOC}, \mathrm{NEWS}, t) = \frac{|P_{2,c+t} \cap P_{1,c}|}{|P_{2,c+t} \cup P_{1,c}|} \tag{9}$$

$$JC(NN, \mathrm{NEWS}, t) = \frac{|P_{3,c+t} \cap P_{1,c}|}{|P_{3,c+t} \cup P_{1,c}|} \tag{10}$$

$P_{2,c+t}$ is meant to assign a time lag or time lead of t days (depending on the sign of t) to the collection of MOC peaks, thus in (9), the News peak at date c will be compared with the MOC peak at date $c + t$. Similarly, $P_{3,c+t}$ is meant to assign a time lag or time lead of t days to the collection of Non-Negative peaks, thus the News peak at date c will be compared with the Non-Negative peak at date $c + t$. The Jaccard Coefficient will have a value between 0 and 1, and the higher the value, the better the two time series correlate with each other.

Figure 5 presents an example of using the Jaccard Coefficient (JC) to measure the correlation between peaks of MOC (in green) and peaks of News (in purple). As Fig. 5 shows, the MOC timeline has seven peaks and the News timeline has six peaks. Three peaks of MOC and another three peaks of News (they are marked by red disks) are pair-wise matched. The remaining four peaks of MOC and the remaining three peaks of News (marked by black disks) are not matched. The JC between the peaks of MOC and the peaks of News is calculated by the size of the intersection divided by the size of the union. In this example, the JC is 3/(7 + 6 − 3) = 0.3.

5.1 Quantitative correlation of peaks

Table 18 summarizes the number of peaks in each of the three time series: MOC (Negative sentiment), NN sentiment, and News. The best Jaccard Coefficient between MOC peaks and News peaks for a given dataset was computed as follows: Firstly, we directly computed the JC between MOC peaks and News peaks without any time delay or lead, and we recorded the result. Secondly, we added 1, 2, or 3 days of lead to the original MOC, computed the correlation between the revised MOC peaks and

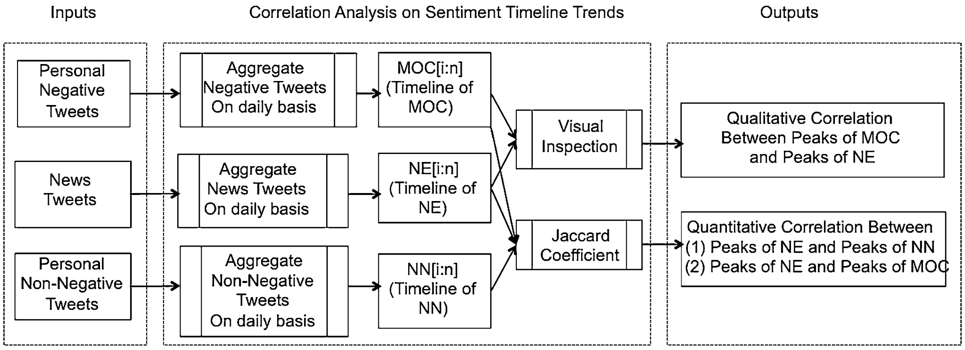

Inputs
Correlation Analysis on Sentiment Timeline Trends
Outputs
Personal Negative Tweets
News Tweets
Personal Non-Negative Tweets
Aggregate Negative Tweets On daily basis
Aggregate News Tweets On daily basis
Aggregate Non-Negative Tweets On daily basis
MOC[i:n] (Timeline of MOC)
NE[i:n] (Timeline of NE)
NN[i:n] (Timeline of NN)
Visual Inspection
Jaccard Coefficient
Qualitative Correlation Between Peaks of MOC and Peaks of NE
Quantitative Correlation Between (1) Peaks of NE and Peaks of NN (2) Peaks of NE and Peaks of MOC

Fig. 4 Correlation between sentiment trends and News trends

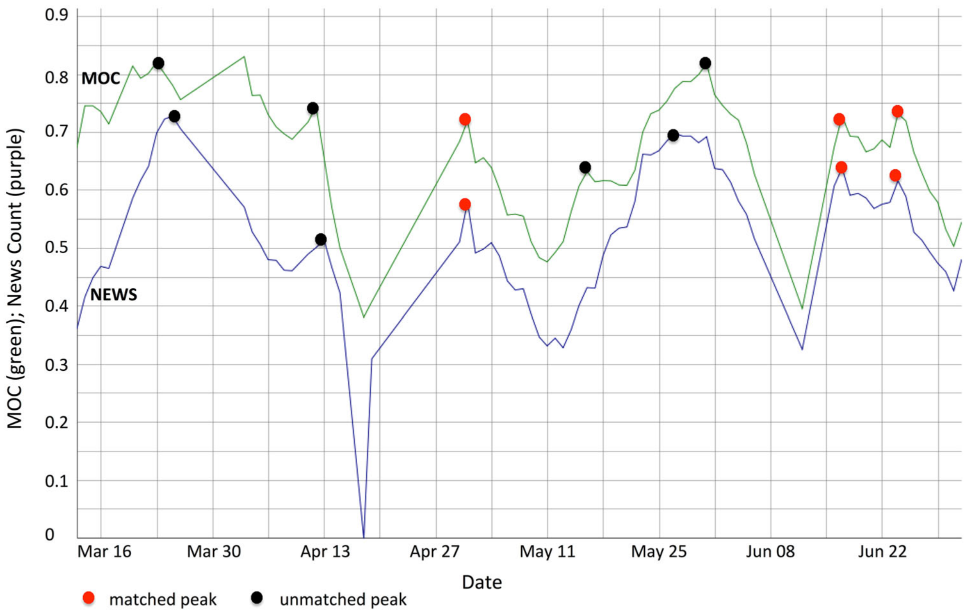


Fig. 5 An example of calculating the Jaccard Coefficient between peaks of MOC and peaks of News

Table 18 The correlation results of Measure of Concern vs. News and Non-Negative vs. News

Dataset Id	# of peaks in MOC	# of Peaks in NN	# of Peaks in News	Best JC (MOC vs. News)	MOC vs. News time delay/lead	Best JC (NN vs. News)	NN vs. News time delay/lead
Epidemic	7	8	8	0.25	0	0.231	0
Mental health	7	6	7	0.273	0	0.3	0
Clinical science	2	2	3	0	0	0.25	−1

the original News peaks, respectively, and recorded these three results. Thirdly, we added 1, 2, or 3 days of delay to the original MOC, and we recorded three more results. Finally, we chose the highest measure from the above seven results as the best correlation between MOC and News. The best correlation between NN sentiment and News was computed similarly.

The best Jaccard Coefficients between MOC peaks vs. News peaks and between NN peaks vs. News peaks for the three domain datasets are shown in Table 18. The $+t$ time means that we delay all MOC peaks or NN peaks to t days later, and the -t time means that we move all MOC or NN peaks to t days earlier. Note that two peaks overlap with each other if and only if the two peaks happen on exactly the same day.

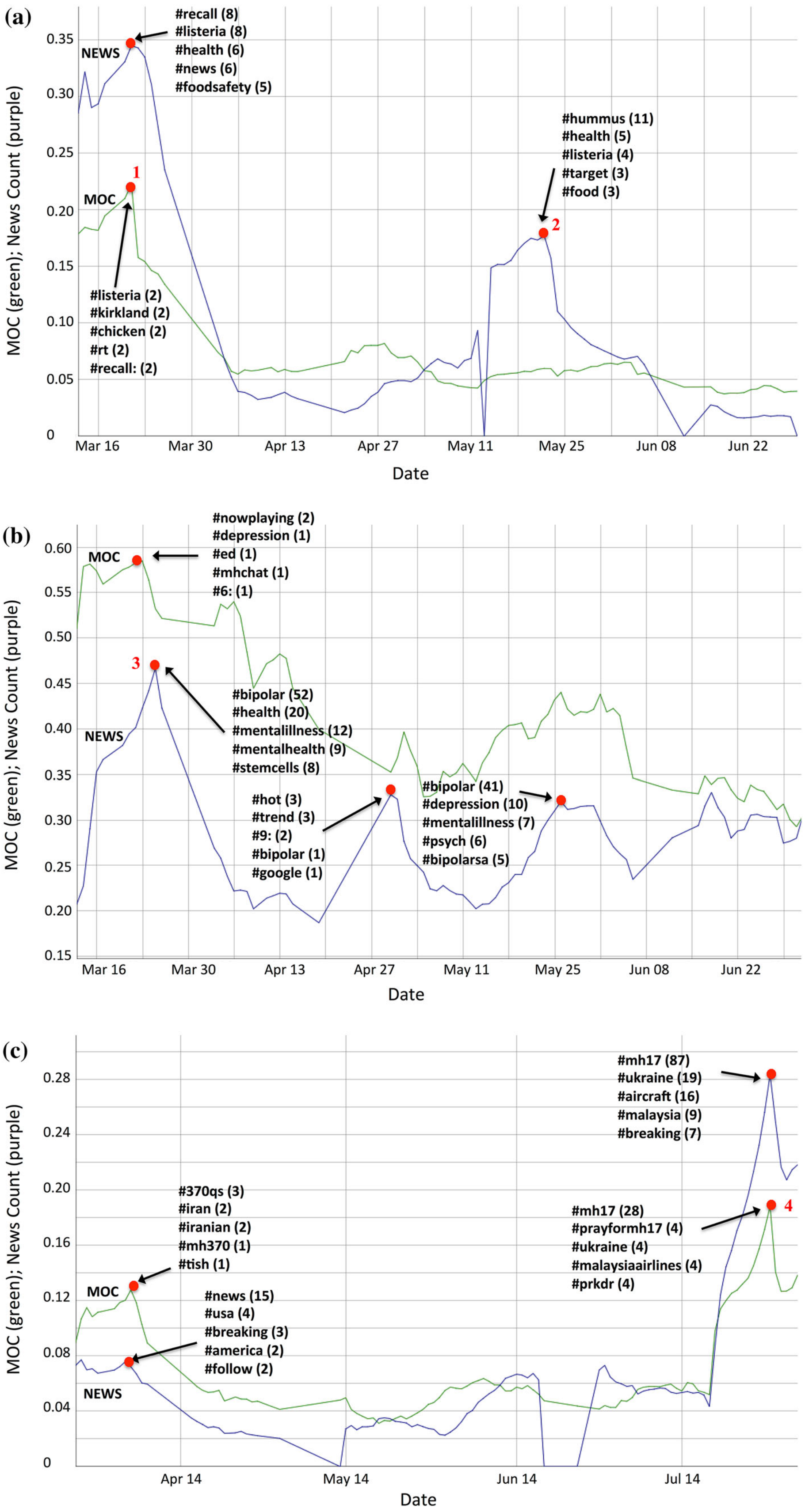
From Table 18, we can see that without any time delay/lead, the peaks of MOC and the peaks of NN correlated with the peaks of News in all datasets with a Jaccard Coefficient of 0.2–0.3. One exception is in the clinical science dataset, where the peaks of MOC do not correlate with the peaks of News. One possible reason is that there are only two peaks for MOC and three peaks for News.

5.2 Qualitative correlation of peaks

We also qualitatively studied the surges in News and MOC, and how those surges co-occurred with the surges of TV and Internet broadcasts and newspaper articles about real-world events. The timeline trends of (1) listeria, (2) bipolar disorder, and (3) air disaster are shown in Fig. 6, where the MOC, NN, and NE are min–max normalized, and a 10-day moving average is used to reduce the spikes in values. The top 5 most frequently mentioned topic terms (hash tags) for the tweets on each peak date are also shown in Fig. 6. For listeria in Fig. 6a, the News Peak 1 occurred because on that same day, several food items produced by Parkers Farm were recalled due to a listeria contamination (FoxNews 2014a). We observe that there was a surge in MOC as well. News Peak 2 was caused by the News broadcast that a company is voluntarily recalling more than 14,000 pounds of hummus and dips due to listeria concerns (FoxNews 2014b).

For bipolar disorder in Fig. 6b, the News Peak 3 was recorded on 03/25/2014. On that day, researchers reported creating stem cells from the skin of people with bipolar disorder to directly measure cellular differences between

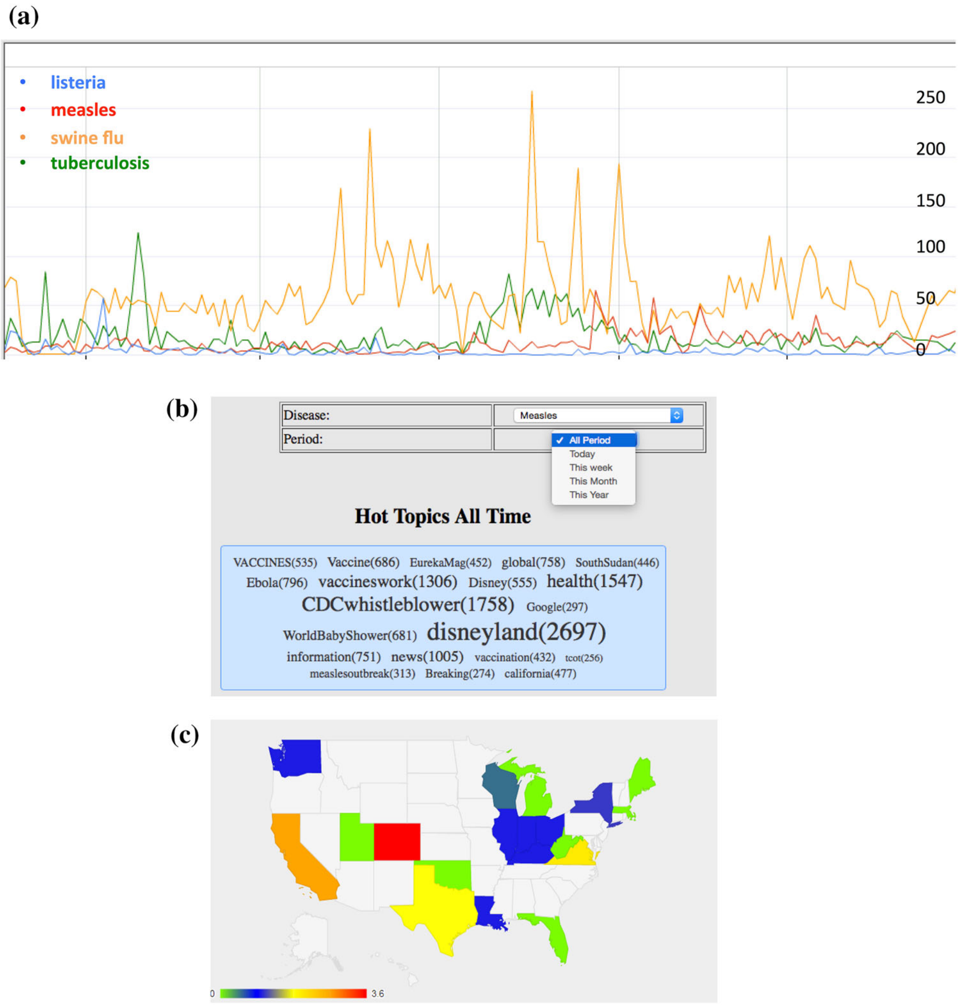
Fig. 6 Measure of Concern timeline trend (*green*) vs. News Timeline Trend (*purple*): in **a** listeria, **b** bipolar disorder and **c** air disaster with most frequent topic terms in different peaks



people with bipolar disorder and people without (Discover 2014). It is surprising to find that the MOC peaks correlated well with this News peak. For air disasters in Fig. 6c, the News Peak 4 was recorded on 07/17/2014. On that day, Malaysian Airlines flight MH17 crashed in the Ukraine (Independent 2014). There are surges of MOC on the same day as well. This qualitative correlation reveals that in most of the cases, the surges of News generated by our method indeed correlated well with the surges of TV, Internet, and newspaper reports of real-world events. Surprisingly, the surges of MOC also correlate with the surges of News, which shows that the general public tends to express negative emotions according to News peaks during these circumstances.

5.3 Prototype system

To monitor the timeline and geographic distribution of public concern, we expanded the Epidemics Outbreak and Spread Detection System visual analytics tools with (1) a concern timeline chart to track the public concern trends on the timeline; (2) a tag cloud for discovering the popular topics within a certain time period; and (3) a concern map that shows the geographic distribution of concern. The public health specialists can utilize the concern timeline chart, as shown in Fig. 7a, to monitor (e.g., identify concern peaks) and compare public concern timeline trends for various diseases. Then the specialists might be interested in what topics people are discussing on social media during the "unusual situations" discovered with the help of the concern timeline chart. To answer this question, they can use the tag cloud, as shown in Fig. 7b to browse the top topics within a certain time period for different diseases. In addition, the concern map, as shown in Fig. 7c, can help public health specialists and government officials to identify parts of the country with different MOCs toward a particular disease or crisis; thus appropriate preventive actions can be taken in high-concern regions.



Disease: Measles
Period: ✓ All Period
Today
This week
This Month
This Year

Hot Topics All Time

VACCINES(535) Vaccine(686) EurekaMag(452) global(758) SouthSudan(446) Ebola(796) vaccineswork(1306) Disney(555) health(1547) CDCwhistleblower(1758) Google(297) WorldBabyShower(681) disneyland(2697) information(751) news(1005) vaccination(432) tcot(256) measlesoutbreak(313) Breaking(274) california(477)

Fig. 7 EOSDS visual analytics tools for public concern monitoring **a** sentiment timeline chart, **b** topics cloud, **c** concern map

6 Conclusions and future work

We discussed the difficulties of measuring and monitoring public health concerns by traditional public health surveillance systems, due to high expenses, limited coverage, and significant time delays. To address these problems, we used the MOC, derived from the social network site Twitter, to monitor the public's concern about common health and disaster issues.

To derive the MOC and understand its relationship with the News Count timeline on Twitter, we developed a two-step sentiment classification approach: In the first step, we classify health tweets into Personal tweets versus News tweets. This step separates the tweets that carry the personal opinions of tweeters from those that are third-party factual reports such as News articles. It uses a subjective clue-based lexicon and News stopwords to automatically extract training datasets: labeled Personal tweets and labeled News tweets. These auto-generated training datasets are then used to train Machine Learning models to classify whether a tweet is Personal or News. After filtering out News tweets, in the second step, we utilized an emotion-oriented clue-based method to automatically extract training datasets and generate another classifier to predict whether a Personal tweet is Negative or Non-Negative.

We used the MOC to quantify the health concerns of the tweeting public, and defined a method to both qualitatively categorize and quantitatively measure the correlation between MOC timeline and News Count timeline.

In order to more fairly evaluate the two-step classification method, we created a test dataset by human annotation for three domains: epidemic, clinical science, and mental health. The Fleiss's Kappa values between annotators were 0.40, 0.54, and 0.33 for epidemic, clinical science, and mental health, respectively. According to the criteria presented by Landis and Koch (1977), the annotators reached a moderate agreement on the epidemic and clinical science datasets, and a fair agreement on the mental health dataset. This result illustrates the complexity of the sentiment classification task, since even humans exhibit relatively low agreement on the labels of tweets.

Experimental results show that (1) in sentiment classification, by combining a clue-based method with a Machine Learning method, our two-step algorithm is able to classify a tweet as Personal Negative, Personal Non-Negative, or News tweet with good accuracy. This overcomes the drawbacks of the clue-based method and the Machine Learning methods when used separately. (2) Quantitatively, the peaks of MOC and the peaks of NN (Non-Negative) correlated with the peaks of News with Jaccard Coefficients of 0.2–0.3. Note that this range of Jaccard Coefficient is still too low to make useful predictions. (3) Qualitatively, as we expected, the peaks of News correlated well with the surges of TV, Internet, and newspaper reports about real-world events. Surprisingly, the surges of MOC also correlated with the surges of News in some cases. This suggests that the general public tends to express negative emotions according to News peaks during these circumstances. (4) As shown in the experiments, our method to derive the MOC is generic and can be applied to topics in other domains, such as mental health monitoring, and clinical science.

Future work includes the following topics.

1. Negation and irony in tweets. In logic, negation is an operation that transforms a proposition p into another proposition "not p". Wiegand et al. (2010) presented a survey on the role of negation in sentiment analysis and investigated several negation models. In our sentiment classification approach, we utilized profanity words to automatically annotate Negative tweets with the assumption that profanity words indicate a negative emotion. However, this assumption may not hold in a negative context. Kiritchenko et al. (2014) showed that in a negative context, both positive and negative terms tend to convey a negative sentiment (e.g., "I know what it feels like to make a thousand dollars in one day, thanks to my tax returns. I'm still not satisfied"). We also observe that profanity words sometimes reverse their polarity when used to modify a positive term (e.g., "damn" is a positive term in "Marriage feels pretty damn great!"). To further improve our two-step sentiment classification, we plan to utilize the negated and affirmative context lexicons (Kiritchenko et al. 2014) to give different polarities and sentiment scores to a word depending on whether it appears in a negated context or in an affirmative context.
2. Irony is another difficult problem, since an ironic statement is used to express the opposite of what is being said (Quintilien and Butler 1953). Utsumi (1996) proposed one of the first computational theories to formalize an ironic environment but the model is too abstract to represent a non-hearer–listener interaction (Reyes et al. 2013). In addition, irony detection requires knowledge of cultural and social stereotypes and tends to be subjective and personal. For the sentiment classification task, the appearance of irony often indicates the opposite of the literal meaning of a statement. Recently, a few techniques were proposed to detect irony in tweets, by investigating irony features. Reyes et al. (2013) proposed four types of conceptual features, which include signatures, unexpectedness, style, and emotional scenarios. More recently, Barbieri and Saggion (2014) designed

another set of features that take into account the sentiments. We plan to investigate these features to better reveal the real sentiments underneath the literal ones to further improve the accuracy of sentiment classification.

3. Measure of Concern is currently based on the number of Personal Negative tweets and total number of tweets on the same day. The MOC was used to define the fraction of tweets that are Personal Negative tweets. We plan to fine grain this definition to quantify the number of tweets expressing real concern. Previous research has been done to classify fine-grained emotions. Brynielsson et al. (2014) manually labeled angry, fear, and positive tweets and trained classifiers with the labeled data to predict which emotion category each tweet belongs to. In the future, we will extend the current work using a more specific sentiment lexicon, such as LIWC (Pennebaker and Francis 1999), to automatically label the tweets that express concerns. In this way, the classifier will be better able to directly identify the tweets with concerns.
4. To improve the performance of classification, we plan to extend the current feature set to include more features specific to micro-blogs, such as slang terms and intensifiers to capture the unique language in micro-blogs. Slang replacement (Piskorski et al. 2013) is able to reveal the semantics and sentiment by translating the slang terms into their original meaning. For example, “ugh” is translated into “disgusted”. As reported by Kouloumpis et al. (2011), the presence of intensifiers, such as all caps and character repetitions in micro-blogs, is also a useful feature for sentiment classification. In Personal vs. News classification, we chose to work in the Machine Learning-based paradigm. However, we note that some lightweight knowledge-based approaches could possibly produce competitive results. For example, if the tweet is of the form “TEXT URL” and the TEXT appears on the web page that the URL points to, the tweet is a News Tweet. The intuition behind this approach is that the title of a news article is often pasted into the tweet body followed by the URL to that news article. We would like to perform a comparison of these knowledge-based approaches with our ML approach in the future.
5. Although it is difficult to find the ground truth for sentiment trends, we would like to conduct a systematic experiment on comparing the sentiments derived by our methods with the epidemic cases reported by other available tools, and with authoritative data sources, such as Health Map and CDC reports. The sentiment trends for *topics* will also be studied by combining the sentiment analysis algorithms with *topic modeling algorithms*.

Acknowledgments This work was partially funded by a CUNY PSC research grant and partially sponsored by Intelligent Automation Inc. for the summer internship training for X. Ji.

Appendix

Data collection keywords

See Table 19.

Table 19 Keywords for collecting tweets in each dataset

Dataset	Keywords
Epidemic	Listeria, Listeriosis, flu, influenza, h1n1, h5n1, ah1n1, adenovirus, h3n2, h3n8, h7n3, Swine Flu, Swine influenza, pig influenza, hog flu, pig flu, Swine influenza virus, swine-origin influenza virus, measles, measle, rubeola, coryza, morbilli, koplik spots, meningitis, encephalitis, meningococcal, brain infection, meningoencephalitis, meningococcus, neisseria meningitidis, mollarets, tuberculosis, tuberculose, tuberculous, mantoux test, mdr tb, bcg vaccine, phthisis, tdr tb
Mental health	Generalized anxiety disorder, Obsessive–compulsive disorder, Obsessive–compulsive neurosis, OCD, Bipolar disorder, Manic depression, Bipolar affective disorder
Clinical science	Skin cancer, melanoma, nivolumab, IMCgp100, PV-10, lambrolizumab, T-Vec, TVEC, imatinib, methotrexate, MPDL3280A

Instructions for human annotation

1. Task

Task 1 Label each tweet as Personal or Non-Personal. If the tweet is a Personal tweet, fill 1 into the PERSONAL cell. Otherwise, the tweet is a Non-Personal tweet, fill 1 into the NEWS (NON-PERSONAL) cell.

Task 2 If the tweet is labeled as PERSONAL tweet in task 1, judge whether the tweet is a PERSONAL NEGATIVE or PERSONAL NON-NEGATIVE. Fill 1 into the corresponding cell.

2. Definitions of PERSONAL and NON-PERSONAL

A Personal tweet is defined to be one that expresses its author’s private states. A private state can be a sentiment, opinion, speculation, emotion, or evaluation, and it cannot be verified by objective observation. In addition, if a tweet talks about a fact observed by the Twitter user, such as

“The boyfriend is STILL sick from the @fatburger he ate last Thursday. The doctor suspects listeria.:(”, this tweet is also defined as Personal. All tweets that are not Personal are defined as Non-Personal tweets.

3. Definitions of PERSONAL NEGATIVE and PERSONAL NON-NEGATIVE

If a Personal tweet expresses negative emotions or attitude, it is a Personal Negative tweet. Otherwise, it is a Personal Non-Negative tweet. Neutral or positive tweets are both Personal Non-Negative tweets.

4. Examples of PERSONAL NON-NEGATIVE:

   (1) RT @sunetrac: Narendra Modi has swine flu- i don’t know why but this news is really exciting me
   (2) #RememberWhen everyone had the swine flu in 7th grade “
   (3) I watched that movie when I had swine flu” - guess who

5. Examples of PERSONAL NEGATIVE

   (1) No more potential skin cancer! huzzah
   (2) Depression is the worst.
   (3) How can you rape a 14 year old tuberculosis patient? What kinda Konji is that?
   (4) @creightonkauss @professor_gram3 meningitis is a bitch

6. Examples of NEWS (NON-PERSONAL)

   (1) Metformin shows promise as anti-tuberculosis drug #Pharmacy http://t.co/lUXLx5NA7R
   (2) Disneyland says unvaccinated kids not welcome amid measles outbreak http://t.co/eSztH9mIy0
   (3) 67 Confirmed cases of measles in California-centered outbreak—LA Times http://t.co/mzokIrJdyk #SmartNews.

References

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of Twitter data. In: Proceedings of the Workshop on Languages in Social Media. Portland, Oregon, pp 30–38

Aramaki E, Maskawa S, Morita M (2011) Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, United Kingdom, pp 1568–1576

Artman H, Brynielsson J, Johansson BJE, Trnka J (2011) Dialogical Emergency Management and Strategic Awareness in Emergency Communication. In: Proceedings of the 8th International ISCRAM Conference

Barbieri F, Saggion H (2014) Modelling Irony in Twitter. In: Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp 56–64

Barbosa L, Feng J (2010) Robust sentiment detection on Twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, pp 36–44

Berger J (2011) Arousal increases social transmission of information. Psychol Sci 22(7):891–893

Bifet A, Frank E (2010) Sentiment knowledge discovery in twitter streaming data. In: Discovery Science, 2010. Springer, pp 1–15

Brownstein JS, Freifeld CC, Reis BY, Mandl KD (2008) Surveillance Sans Frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med 5(7):e151

Bruns A, Stieglitz S (2014) Twitter data: What do they represent? It Inf Technol 56(5):240–245

Brynielsson J, Johansson F, Jonsson C, Westling A (2014) Emotion classification of social media posts for estimating people’s reactions to communicated alert messages during crises. Security Inf 3(1):1–11

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27

Chew C, Eysenbach G (2010) Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 5(11):e14118

Collier N, Doan S (2012) Syndromic Classification of Twitter Messages. In: Kostkova P, Szomszor M, Fowler D (eds) Electronic Healthcare. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 91. pp 186–195

Cortes C, Vapnik V (1995) Support-Vector Networks. Mach Learn 20(3):273–297

Culotta A (2010a) Detecting influenza outbreaks by analyzing Twitter messages. arXiv:10074748

Culotta A (2010) Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the First Workshop on Social Media Analytics. Washington D.C., District of Columbia, pp 115–122

DCP (2014) Disease Control Priorities Project. http://www.dcp-3.org/dcp2

Discover (2014) Stem Cells Shed Light on Treatments for Bipolar Disorder. http://blogs.discovermagazine.com/d-brief/2014/03/26/stem-cells-shed-light-on-treatments-for-bipolar-disorder/-U-wKD4BdXN8

FederalCommunicationsCommittee (2014) Obscenity, Indecency and Profanity Guide. http://www.fcc.gov/guides/obscenity-indecency-and-profanity

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378

FoxNews (2014a) Company Recalls Several Food Products Due to Listeria. http://fox8.com/2014/03/23/several-nationally-distributed-food-products-recalled-due-to-listeria/

FoxNews (2014b) Food Fear: Hummus, Dips from Target, Giant Eagle, Trader Joe’s Recalled. http://fox8.com/2014/05/20/food-fear-hummus-dips-from-target-giant-eagle-trader-joes-recalled/

Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014

Go A, Bhayani R, Huang L (2009) Twitter Sentiment Classification using Distant Supervision. Technical Report

Guardian (2011) Chinese panic-buy salt over Japan nuclear threat. http://www.guardian.co.uk/world/2011/mar/17/chinese-panic-buy-salt-japan

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor Newsl 11(1):10–18

Heath C (1996) Do people prefer to pass along good or bad news? Valence and relevance of news as predictors of transmission propensity. Organ Behav Hum Decis Process 68(2):79–94

Independent (2014) Malaysia Airlines flight MH17 crash. http://www.independent.co.uk/news/world/europe/malaysia-airlines-plane-crash-boeing-jet-carrying-295-people-crashes-in-ukraine-9612882.html

Ji X (2014a) Profanity Filter Word List. http://web.njit.edu/~xj25/eosds_beta/files/profanity_list.txt

Ji X (2014b) Stopwords. http://web.njit.edu/~xj25/eosds_beta/files/news_stopwords.txt

Ji X, Chun SA, Geller J (2012) Epidemic outbreak and spread detection system based on twitter data. In: Proceedings of the First international conference on Health Information Science, Beijing, China. pp 152–163

Ji X, Chun SA, Geller J (2013) Monitoring Public Health Concerns Using Twitter Sentiment Classifications. In: Proceedings of IEEE International Conference on Healthcare Informatics. Philadelphia

Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1. pp 151–160

Johansson F, Brynielsson J, Quijano MN (2012)Estimating citizen alertness in crises using social media monitoring and analysis. In: Intelligence and Security Informatics Conference (EISIC). pp 189–196

Kiritchenko S, Zhu X, Mohammad SM (2014) Sentiment analysis of short informal texts. J Artif Intell Res:723–762

Kouloumpis E, Wilson T, Moore J (2011) Twitter sentiment analysis: the good the bad and the omg! Icwsm 11:538–541

Lampos V, Cristianini N (2010) Tracking the flu pandemic by monitoring the Social Web. In: Proceedings of IEEE International Conference on Digital Ecosystems and Technologies. pp 411–416

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics:159–174

Liben Nowell D, Kleinberg J (2007) The link prediction problem for social networks. J Am Soc Inform Sci Technol 58(7):1019–1031

Liu B, Zhang L (2012) A survey of opinion mining and sentiment analysis. In: Mining Text Data. pp 415–463

Mishne G (2005) Experiments with mood classification in blog posts. In: Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access

Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist JN (2011) Understanding the Demographics of Twitter Users. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11). pp 554–557

Mohammad SM, Kiritchenko S, Zhu X (2013) NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. arXiv:13086242

Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose. arXiv:13065204

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of International Conference on Weblogs and Social Media. pp 122–129

Pak A, Paroubek P (2010) Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10, Valletta, Malta

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-vol. 10. pp 79–86

Pennebaker J, Francis M (1999) Linguistic Inquiry and Word Count. Lawrence Erlbaum (citeulike-article-id:2863249)

Piskorski J, Tanev H, Balahur A (2013) Exploiting Twitter for Border Security-Related Intelligence Gathering. In: European Intelligence and Security Informatics Conference (EISIC). pp 239–246

Quintilien, Butler HE (1953) The Institutio Oratoria of Quintilian. With an English Translation by HE Butler

Refaee E, Rieser V (2014) An Arabic twitter corpus for subjectivity and sentiment analysis. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland

Reuters (2014) Americans 'can't give into hysteria or fear' over Ebola: Obama. http://www.reuters.com/article/2014/10/18/us-health-ebola-usa-idUSKCN0I61BO20141018

Reyes A, Rosso P, Veale T (2013) A multidimensional approach for detecting irony in twitter. Lang Resour Eval 47(1):239–268

Riloff E, Wiebe J (2003) Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 conference on Empirical methods in natural language processing. pp 105–112

Saif H, He Y, Alani H (2012) Semantic sentiment analysis of twitter. In: The Semantic Web–ISWC 2012. pp 508–524

Saif H, Fernandez M, He Y, Alani H (2013) Evaluation datasets for twitter sentiment analysis. In: Proceedings of 1st Workshop on Emotion and Sentiment in Social and Expressive Media (ESSEM) in Conjunction with AI* IA Conference, Turin, Italy

Saif H, Fernández M, Alani H (2014) Automatic stopword generation using contextual semantics for sentiment analysis of Twitter. In: Proceedings of CEUR Workshop

Salathe M, Khandelwal S (2011) Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 7(10)

Sha Y, Yan J, Cai G (2014) Detecting Public Sentiment Over PM2.5 Pollution Hazards through analysis of Chinese Microblog. In: ISCRAM: The 11th International Conference on Information Systems for Crisis Response and Management

Signorini A, Segre AM, Polgreen PM (2011) The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 6(5):e19467

Stieglitz S, Dang-Xuan L (2013) Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. J Manag Inf Syst 29(4):217–248

Twitter (2014a) Twitter. http://en.wikipedia.org/wiki/Twitter

Twitter (2014b) Twitter Developers Documentation. https://dev.twitter.com/docs

Twitter4J (2014) Twitter4J. http://twitter4j.org/en/

Utsumi A (1996) A unified theory of irony and its computational formalization. In: Proceedings of the 16th conference on Computational linguistics, vol. 2. Association for Computational Linguistics, pp 962–967

Wiebe J, Riloff E (2005) Creating subjective and objective sentence classifiers from unannotated texts. In: Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing. Mexico City, Mexico, pp 486–497

Wiegand M, Balahur A, Roth B, Klakow D, Montoyo A (2010) A survey on the role of negation in sentiment analysis. In: Proceedings of the workshop on negation and speculation in natural language processing. pp 60–68

Wilson T, Wiebe J (2003) Annotating opinions in the World Press. In: Proceedings of SIGdial-03. pp 13–22

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)

Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B (2011) Combining lexicon-based and learning-based methods for Twitter sentiment analysis

Zhou Z, Zhang X, Sanderson M (2014) Sentiment Analysis on Twitter through Topic-Based Lexicon Expansion. In: Databases Theory and Applications. pp 98–109

Zhu X, Wu S, Miao D, Li Y (2008) Changes in emotion of the Chinese public in regard to the SARS period. Social Behav Personal 36(4):447

Zhuang L, Jing F, Zhu XY (2006) Movie review mining and summarization. In: Proceedings of the 15th ACM international conference on Information and knowledge management. pp 43–50