Knowledge-Based Tweet Classification for Disease Sentiment Monitoring

Xiang Ji, Soon Ae Chun and James Geller

Abstract Disease monitoring and tracking is of tremendous value, not only for containing the spread of contagious diseases but also for avoiding unnecessary public concerns and even panic. In this chapter, we present a near real-time sentiment analysis service of public health-related tweets. Traditionally, it is impossible for humans to effectively measure the degree of public health concerns due to limited resources and significant time delays. To solve this problem, we have developed a computational intelligence approach for Epidemic Sentiment Monitoring System (ESMOS) to automatically analyze the disease sentiments and gauge the Measure of Concern (MOC) expressed by Twitter users. More specifically, we present a knowledge-based approach that employs a disease ontology to detect the outbreak of diseases and to analyze the linguistic expressions that convey subjective expressions and sentiment polarity of emotions, feelings, opinions, personal attitudes, etc. with a sentiment classifier. The two-step sentiment classification method utilizes the subjective vocabulary corpus (MPQA), sentiment strength corpus (AFINN), as well as emoticons and profanity words that are often used in social media postings. It first automatically classifies the tweets into personal and non-personal classes, eliminating many tweets such as non-personal "retweets" of news articles from further consideration. In the second stage, the personal tweets are classified into Negative and non-Negative sentiments. In addition, we present a model to quantify the public's Measure of Concern (MOC) about a disease, based on sentiment classification results. The trends of the public MOC are visualized on a timeline. Correlation analyses between MOC timeline and disease-related sentiment category timelines show that the peaks of the MOC are

X. Ji · J. Geller
Department of Computer Science, New Jersey Institute of Technology,
323 Martin Luther King Blvd, Newark, NJ, USA
e-mail: xj25@gmail.com

J. Geller
e-mail: james.geller@njit.edu

S.A. Chun (✉)
Information Systems and Informatics, City University of New York,
College of Staten Island, 2800 Victory Blvd,
Staten Island, NY, USA
e-mail: soon.chun@csi.cuny.edu

W. Pedrycz and S.-M. Chen (eds.), *Sentiment Analysis and Ontology Engineering*,
Studies in Computational Intelligence 639, DOI 10.1007/978-3-319-30319-2_17

weakly correlated with the peaks of the News timeline without any appreciable time delay or lead. Our sentiment analysis method and the MOC trend analyses can be generalized to other topical domains, such as mental health monitoring and crisis management. We present the ESMOS prototype for public health-related disease monitoring, for public concern trending and for mapping analyses.

Keywords Computational intelligence · Sentiment analysis · Public health concern monitoring · Social data analytics

1 Introduction

Disease monitoring and tracking is of tremendous value, not only for containing the spread of contagious diseases but also for avoiding unnecessary public concerns and even panic. In this chapter, we focus on studying the Twitter users' concerns about diseases instead of the outbreak of the disease itself, which has been extensively studied [1–5]. Recent outbreaks of Ebola in Africa, measles on the West Coast of the USA and MERS in South Korea have shown how important it is to monitor and understand the public sentiments in addition to tracking the location, trend and potential trajectory of disease outbreaks. Public health concerns can also develop into dangerous panic states, resulting in irrational behaviors, unjustified fear, discrimination against patients, mistrust in governments' containment efforts as well as a negative overall economic impact. For instance, the Korean MERS incidents scared the public and prompted school and hospital closures, as well as decreased businesses activities and cancelled tourist visits. These, in turn, paralyzed the economy severely within a very short period of time and the central bank had to cut interest rates to prevent a further downward spiral (Wikipedia 2015).

Another example illustrating the consequence of public health concerns is that since the Ebola outbreak in 2014, the immigration examination and the medical system's ability to deal with Ebola have been widely mistrusted by the general public. Even the president of the United States addressed that issue [6], due to a series of mistakes when a traveler was issued a visitor visa and was not diagnosed by the local hospital. For example, a tweet complained that, "I know. Our government is a FAIL. People were more upset about Ebola than they are about thugs killing our cops." Another case is the SARS outbreak in China in 2003. Zhu, Wu, Miao and Li [7] reviewed the mental state changes of the general public during the SARS outbreak in China.

As shown by these examples, it is of great value if public health specialists and government decision makers can actively monitor public health concerns. However, the existing public health concern surveillance methods, such as questionnaires and clinical tests, are not able to cover a large number of respondents due to their expenses and furthermore the survey results are often published with significant time delays. A last example is the public's reaction to Japan's nuclear emergency in March 2011 [8]. Text messages about nuclear plumes spread throughout Asia. In China, Vietnam, and the Philippines rumors spread about possible disastrous consequences.

Ginsberg et al. [9] used search engine logs, in which users submitted queries in reference to issues that they were concerned about, to approach the disease monitoring 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. However, comprehensive access to search engine logs is limited to the search engine providers. Twitter, a popular social network site, has more than 500 million users out of which more than 302 million are active users [10]. This shows Twitter's potential to address the limitations of traditional public health surveillance methods, and of search keyword logs. A percentage of Twitter messages are publicly available and researchers are able to retrieve the tweets as well as related information through the Twitter API [11].

We have developed a method to gauge the Measure of Concern (MOC) expressed by Twitter users for public health specialists and government decision makers [12]. More specifically, we developed a two-step sentiment classification approach. Firstly, personal tweets are distinguished from News tweets. News tweets are considered as Non-Personal, as opposed to Personal tweets posted by individual Twitter users. In the second stage, the Personal tweets are further classified into Personal Negative tweets or Personal Non-Negative tweets. The two-step sentiment classification problem addressed in this chapter is different from the traditional Twitter sentiment classification problem, which classified tweets into positive/negative or positive/neutral/negative tweets [4, 13–16] without distinguishing Personal from Non-Personal tweets first. 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.

The sentiment classification method presented in this chapter is able to identify Personal tweets and News (Non-Personal) tweets in the first place. More importantly, using the sentiment classification results, we quantified the Measure of Concern (MOC) using 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 [17, 18] visually noticed that sentiment surges co-occurred with health events on a timeline. Different from the previous work, which is based on visual observation, we correlated the peaks on the MOC timeline (i.e., change over time) and the peaks of the News timeline and also the peaks of the Non-Negative timeline and the peaks of News timeline using the Jaccard Coefficient [19]. Government officials can use MOC to track public health concerns on the timeline to help make timely decisions, disprove rumors, intervene, and prevent unnecessary social crises at the earliest possible stage. More importantly, public health concern monitoring using social network data is faster and cheaper than the traditional method.

The rest of this chapter is organized as follows. In Sect. 2, we discuss an intelligent system for health sentiment monitoring with background and related. In Sect. 3, sentiment classification methods and results are introduced in detail. In Sect. 4, the sentiment timeline trend analysis results are illustrated, interpreted, and discussed. Section 5 contains the chapter summary.

2 Cognitive Approach for Monitoring Public Health Sentiments

To enable public health sentiment monitoring, an intelligent system called the Epidemic Sentiment Monitoring System (ESMOS) is proposed. It monitors social media data (specifically, Twitter data) and "recognizes" and collects tweets about public-health related diseases. ESMOS then analyzes the data to automatically classify them into sentiment categories, and calculates the public degree of concerns for a particular disease outbreak and its spread. ESMOS also provides visual tools such as the intensity map (heat map) of sentiments to understand the geographic distribution and concentration of public concerns, and a health concern trending map to be able to track the public health concerns over time. The trending map will also allow users to compare collected tweets with the disease-associated news events to qualitatively investigate how the news coverage may influence the public concern. Figure 1 shows the component architecture of ESMOS.

The goal of computational intelligence is to build cognitive systems that can act like and interact with humans, to provide insights and intelligence that are needed in complex problem solving and decision-making situations. Cognitive systems

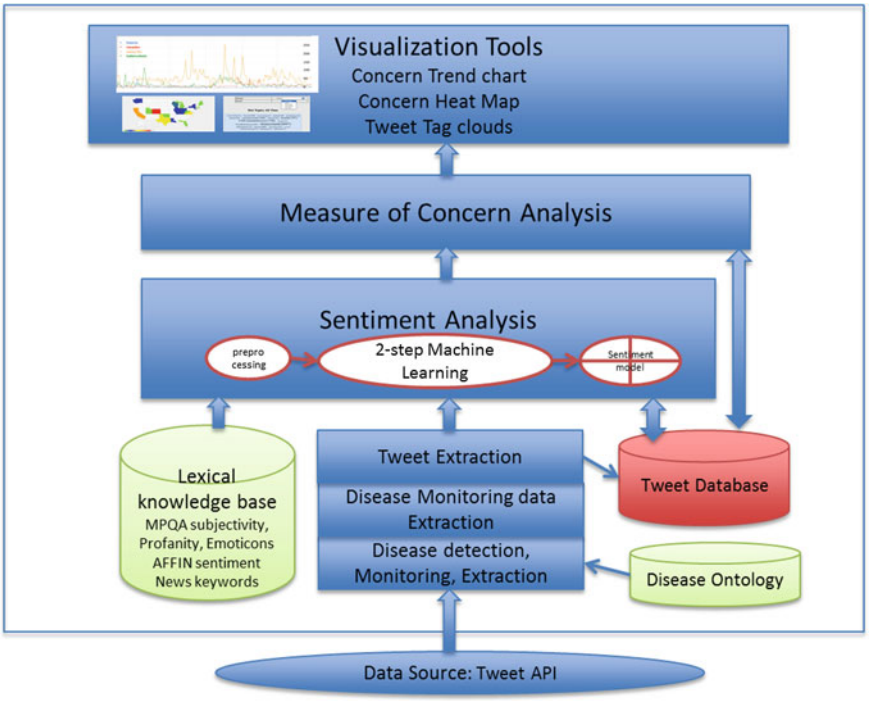

Fig. 1 Epidemic Sentiment Monitoring System (ESMOS) architecture

process [20] structured and unstructured data and learn specific domain knowledge by experience, much in the same way humans do. Cognitive systems use Natural Language Processing and image and speech recognition to understand the world and to interact with humans and other smart systems seamlessly. Unlike many expert systems, cognitive systems not only match and fire pre-defined rules with anticipated possible actions, but also can be trained using AI and Machine Learning methods to sense, integrate, analyze, predict and reason as human brains do for various tasks.

ESMOS utilizes both a linguistic knowledge base and automated machine learning methods to exhibit a degree of computational intelligence. It also utilizes a disease ontology to identify the disease names to collect and monitor tweet data. In addition, the machine learning algorithm used for sentiment classification utilizes automated labeling for the training data set using the linguistic knowledge base, instead of having human-labeled, supervised training data. It distinguishes personal tweets from non-personal tweets such as news articles. In the second step, it automatically labels the positive or negative sentiment tweets to generate a training dataset, avoiding another human labeling step. The training data sets are used to build the sentiment analysis model. The system uses unsupervised learning to learn classifiers using the lexical knowledge.

2.1 Disease Ontology

ESMOS contains 8,043 inherited, developmental and acquired human disease names from the Disease Ontology (DO), an open-source medical ontology, developed by Schriml et al. [21].

A partial structure of the ontology used by ESMOS is shown in Fig. 2. The extracted disease vocabulary is used to supply the keywords to monitor disease tweets for signs of epidemics. The current prototype system monitors a handful of infectious diseases. However, all DO disease terms may be used, given sufficient computational resources. The core component uses the Twitter Streaming API for collecting epidemics-related real-time tweets. The advantage of using the Disease Ontology is that it is linked to medical terminology codes in other ontologies with a set of synonyms. For instance, the Ebola virus is known under different labels. It has the following synonyms and cross-references with three of the most important medical terminology collections: NCI, SNOMED CT and UMLS:

```
id: DOID:4325
name: Hemorrhagic Fever, Ebola
synonym:"Ebola Hemorrhagic Fever" EXACT [NCI2004_11_17:C36171]
synonym: "Ebola virus disease" EXACT [SNOMEDCT_2005_07_31:186746000]
xref: UMLS_CUI:C0282687
```

Fig. 2 Partial structure of the disease ontology used to extract disease keywords

The use of other collections of medical terms (e.g. Wikipedia's list of ICD-9 codes 001–139: infectious and parasitic diseases [22]) may further support the purpose. The major issue with the use of formal ontologies of medical terms is that the laymen's disease terms (e.g. Ebola) may have to be matched with the scientific terms used in medical terminologies (e.g. Ebola Hemorrhagic Fever) [23]. This requires fuzzy matching and similarity matching, which are not addressed in this chapter.

2.2 Lexical Knowledge

Understanding human emotions (also called sentiment analysis) is a difficult task for a machine, for which the computational intelligence approach may provide better results. Often, linguistic expressions as well as paralinguistic features in spoken languages (e.g., pitch, loudness, tempo, etc.) reveal the sentiments or emotional states of individuals. Prior research studies have developed sentiment lexicons using a dictionary approach and a corpus approach [24].

The MPQA (Multi Perspective Question Answering) lexicon [25] was constructed through human annotations of a corpus of 10,657 sentences in 535 documents that contain English-based news from 187 different sources in a variety of countries, dated from June 2001 to May 2002. The annotated lexicon represents opinions and other private states, such as beliefs, emotions, sentiments, speculations, etc. The subjective and objective expressions were also annotated with values of intensity and polarity to indicate the degree of subjectivity, and a negative, positive or neutral sentiment. AFINN is another affective lexicon with a list of English words with sentiment valence ratings between minus five (negative) and plus five (positive). These words have been manually labeled and evaluated in 2009–2011 [26]. Most of the positive words were labeled with +2 and most of the negative words with

−2, and strongly obscene words were rated with either −4 or −5. AFINN includes 2477 words with 1598 negative words (65 %), 878 positive words and one neutral word. Another lexical resource is SentiWordNet [27], which associates each synset of WordNet with three numerical scores Obj(s), Pos(s) and Neg(s). The numerical scores describe the polarity for the terms in the synset, i.e., how Objective, Positive, or Negative the terms contained in the synset are.

In this chapter, we applied the MPQA lexicon to distinguish the Personal from Non-Personal tweets (such as News article tweets) since the subjectivity of the lexicon helps to distinguish Personal expressions from Non-Personal ones. We used the valence ratings in the AFINN lexicon to further distinguish the Personal Negative from Personal Non-Negative tweets. Our goal is to develop an automated sentiment analysis system for the social heath tweets generated by the general public. This system can provide public health officials with the capability to monitor the viral effects in social media communication, so that they can take early actions to prevent unnecessary panic regarding public health related diseases.

2.3 Twitter Sentiment Classification

Since the 2000s, with the tremendous amount of user-generated data from various data sources, such as blogs, review sites, News articles, and micro-blogs becoming available, researchers have become interested in mining high-level sentiments from the user data. Pang and Lee [28] reviewed the sentiment analysis work. This thread of research, depending on the analysis target, can be classified into the one of these levels: document-level [29], blog-level [30], sentence-level [31], tweet-level [32–34] with the sub-category non-English tweet level [35], and tweet-entity-level [36]. Since 2009, extensive research has been carried out on the topic of Twitter sentiment classification. [15, 33, 37–41].

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 [42], and it assumes that all features are independent from each other. Good performance of Naïve Bayes (NB) was reported in several sentiment analysis papers [33, 37, 41]. Multinomial Naive Bayes (MNB) is a model that works well on sentiment classification [38, 39, 41]. MNB takes into account the number of occurrences and relative frequency of each word. Support Vector Machine [43] is also a popular ML-based classification method that works well on tweets [33, 40]. In Natural Language Processing, SVM with a polynomial kernel is more popular [44].

There are two drawbacks of the previous sentiment classification work. Firstly, Twitter messages were classified 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 re-tweet symbols, it is not easy for a stop-word based pre-processing 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).

Recently, some researchers identified irrelevant tweets. Brynielsson, Johansson, Jonsson and Westling [33] used manual labeling to classify tweets into "angry," "fear," "positive," or "other" (irrelevant). Salathe and Khandelwal [45] 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. In our two-step classification method, we can automatically extract News tweets and perform the sentiment analysis. The results of sentiment classification are used 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. Secondly, although sophisticated models were developed by the above research, the results of the sentiment classification were not utilized to provide insights. We provide the Measure of Concern timeline trends as a useful sentiment monitoring tool for public health specialists and government decision makers.

2.4 Quantifying and Visualizing Twitter Sentiment Trends on Timeline

Sentiment quantification is a method to process unstructured text and generate a numerical value or a timeline of numerical values to gain insights into the sentiment trends. Zhuang et al. [46] generated a quantification of sentiments about movie elements, such as special effects, plot, dialogue, etc. Their quantification contains a positive score and a negative score towards a specific movie element. For tweet-level sentiment quantification on a timeline, Chew and Eysenbach [47] 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. [48], in which they calculated a daily (positive and negative) sentiment score by recording the number of positive and negative words of one tweet appearing in the subjectivity lexicon of OpinionFinder [31]. Sha et al. [17] found that the sentiment fluctuations on Sina Weibo were associated with either the new policy announcements or government actions.

The drawbacks of the existing Twitter sentiment quantification research are twofold. Firstly, the lexicon-based sentiment extraction models have limited coverage of words. As pointed out by Wiebe and Riloff [49], identifying positive or negative tweets by counting words in a dictionary or lexicon usually has high precision but low recall. In the case of Twitter sentiment analysis, the performance suffers

more. The lexicon or dictionary does not contain the slang words that are common in social media. For example, LMAO (Laughing My A** Off), is a positive "word" in Twitter, but it does not match any word in MPQA [25], which is a popular sentiment dictionary. In this study, we consider these profanity or slang words as well as the emoticons. Secondly, the existing sentiment quantification work [17, 48] has shown the correlation between sentiments and real-world events (e.g. news) through observing their co-occurrence on a timeline, but has not provided a comprehensive, quantitative correlation between the sentiment timeline trend and the News timeline trend. 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 crises.

In the next two sections, we first describe the machine learning approach to automatically labeling the training datasets to generate the sentiment classifiers, and then the quantitative model for measuring the Measure of Concern and its trend line analysis to see any correlations with news events on the traditional broadcast media.

3 Two-Step Tweet Sentiment Classification

In this section, we present our two-step sentiment classification method. As discussed earlier, our approach to sentiment classification is different from the classic sentiment classification of Tweets. The first step of our method involves training a Twitter sentiment classifier for distinguishing Personal tweets from News (Non-personal) tweets. The second step builds the sentiment classifier using only the personal tweets to identify the Negative versus Non-Negative tweets. The formal definitions of Personal Tweet, News Tweet, Personal Negative Tweet, and Personal Non-Negative Tweet are shown below.

Definition 1 (*Personal Tweet*) A Personal Tweet is a tweet that conveys its author's private states [31, 50]. 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. The goal of this definition is to distinguish the tweets written by the Twitter users from scratch from the News tweets that are re-tweeted in the Twittersphere.

Example (Personal Tweet)
"Since when does a commercial aircraft accident become a matter of National Security Interests? #diegogarcia#mh370"

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):
"#UPDATE Cyanide levels 350x standard limits detected in water close to the site of explosions in China's Tianjin http://u.afp.com/Z5ab"

Definition 3 (*Personal Negative Tweet and Personal Non-Negative Tweet*) A tweet is a Personal Negative Tweet (denoted as PN) if it conveys negative emotions or attitude and it is a Personal Tweet. 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. A Personal Negative Tweet expresses a user's negative sentiment, such as panic, depression, anxiety, etc. Note that this definition is focused on the user's negative emotional state as opposed to expressing the absence of an illness, e.g., getting a negative test result. Two examples are as follows.

Example (Personal Negative Tweet):
"Apparently #Ebola doesn't read our textbooks - it keeps changing the rules as it goes along. Frightening news :("

Example (Personal Non-Negative Tweet):
"To ensure eliminating the #Measles disease in the country, it is important to vaccinate people who are at risk."

As 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 versus Non-Negative classification). An overview of sentiment classification method is shown in Fig. 3. We only consider English tweets, which were automatically detected during the data collection phase in this chapter. As shown in Fig. 3, 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 Personal Negative tweets and Personal Non-Negative tweets. After the Personal Negative tweets, Personal Negative tweets, and News tweets were all identified, they were utilized to compute the Measure of Concern and the quantitative correlation between the sentiment timeline trend and the News timeline trend.

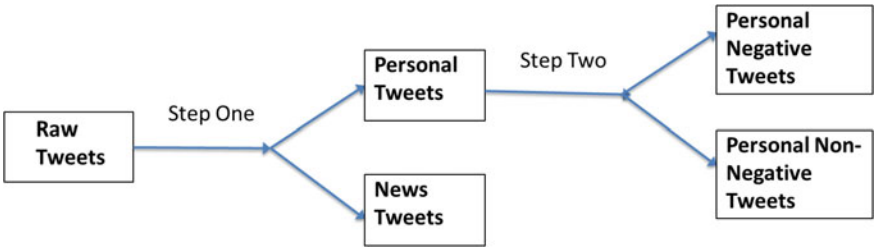

Fig. 3 Overview of the two-step sentiment classification method

3.1 Data Collection

We developed a real-time Twitter data collector with the Twitter API version 1.1 and Twitter4J library [51]. It collects real-time tweets containing the pre-defined public health-related keywords (e.g., listeria). We can describe the overall data collection process as "ETL" (Extract-Transform-Load) pipeline, which is a frequently used term in the area of Data Warehousing. In the first step, the data collector collected tweets in JSON format with Twitter Streaming API. (Extract step). In the second step, the data in JSON format was parsed into relational data, such as tweets, tweet_mentions, tweet_place, tweet_tags, tweet_urls, and users (Transform step). In the last step, the relational data was stored into our MySQL relational database (Load step).

We monitored 12 diseases including 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 preprocessing step filters out re-tweets and converts special characters into corresponding unigrams. More specifically, all tweets starting with "RT," were deleted because "RT" indicates that they are re-tweets without comments to avoid duplications. For the tweets that have a non-starting string "RT," the "RT" was removed.

One member of each pair of tweets that contain the same tokens (words) in the same order was deleted. For example, of the below two tweets only one is kept in the database.

(1) "27 test positive for tuberculosis at high school. http://t.co/Ss4QT1EPP2#CNN"

(2) "27 test positive for tuberculosis at high school. http://t.co/M4D6rgzYaI-@CNN."

3.2 Sentiment Classification of Twitter Data

3.2.1 Automatic Tweet Labeling Based on Subjective Clue Corpus

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 [25] 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 8,221 words, including 3,250 adjectives, 329 adverbs, 1,146 any-position words, 2,167 nouns, and 1,322 verbs. As for the sentiment polarity, among all 8,221 words, 4,912 are negatives, 570 are neutrals, 2,718 are positives, and 21 can be both negative and positive. In terms of

strength of subjectivity, among all words, 5,569 are strongly subjective words, and the other 2,652 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 340 selected profanity words [52] to the corpus described in the previous paragraph. US law, enforced by the Federal Communication Commission prohibits the use of a short list of profanity words in TV and radio broadcasts [53]. 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. 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 versus 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 [12]. 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

Table 1 Results of Personal tweets classification with thresholds (Precision/Recall) [12]

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**

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).

3.2.2 Personal versus News Tweet Classification Based on Machine Learning

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 versus News classification, as shown in Fig. 4. Suppose the collection of Raw Tweets of a unique type (e.g. tuberculosis) is T. After the preprocessing step, which filters out non-English tweets, re-tweets and near-duplicate tweets, the resulting tweet dataset is $\boldsymbol{T'} = \{tw_1, tw_2, tw_3, \ldots, tw_n\}$, 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 versus News classifier as shown in Fig. 4, where the blue part is the automatic labeling of tweets with lexicons (highlighted in green) and the yellow part is the Machine Learning classifiers for classifying Personal tweets from News tweets. We choose three Machine Learning classifiers, Naïve Bayes, Multinomial Naïve Bayes, and Support Vector Machine, as these classifiers achieved good results for similar tasks as discussed in Sect. 2.

In the lexicon-based step for labeling training datasets, each tw_i of $\boldsymbol{T'}$ is compared with the MPQA dictionary [25]. 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 stop word list [54] and a profanity list [52].

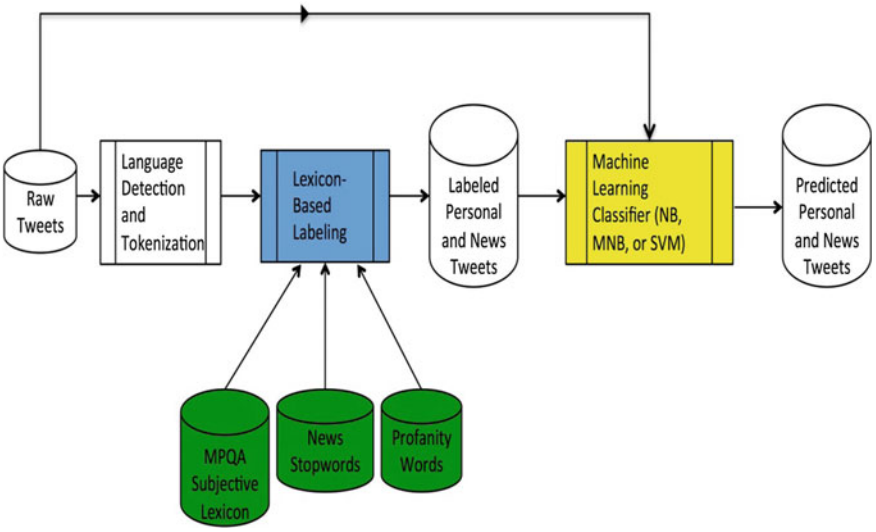

Fig. 4 Personal versus News (Non-Personal) classification

The News stop word 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 stop word list and does not contain any profanity word, tw_i is labeled as a News tweet. For example, the tweet "*UN official: Ebola epidemic could be defeated by end of 2015–41 NBC News* http://bit.ly/1J2rQ1F #world#health" is labeled as a News tweet, because it contains at least one word from the News stop word list and does not contain any profanity word. We mark the set of labeled Personal tweets as $\boldsymbol{T}'_p$, and the set of labeled News tweets as $\boldsymbol{T}'_n$, note that $\boldsymbol{T}'_p \cup \boldsymbol{T}'_n \subseteq \boldsymbol{T}'$.

The next step is the Machine Learning-based method. The two classes of $\boldsymbol{T}'_p$ data and $\boldsymbol{T}'_n$ 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 versus News classifier is trained, the classifier is used to make predictions on each tw_i in $\boldsymbol{T}'$, which is the preprocessed tweets dataset. The goal of Personal versus News classification is to obtain the Label for each tw_i in the tweet database $\boldsymbol{T}'$, where the Label is either *Personal* or *NT* (News Tweet). Personal could be *PN* or *PNN*.

3.2.3 Negative Versus Non-Negative Sentiment Classifier

Ji, Chun, Wei and Geller [12] discussed automatic labeling of Personal Negative and Personal Non-Negative tweets using a sequential approach. In this method, firstly, a profanity list is used to test if a tweet contains any word from the profanity list. If the tweet contains a profanity word, it is labeled as a Personal Negative tweet. Secondly, for the tweets that do not contain a profanity word, negative and non-negative emoticon lists are used to test whether the tweet contains a negative emoticon or a non-negative emoticon. A partial list of emoticons is shown in Table 2. If the tweet contains a negative emoticon, it is labeled as a Personal Negative tweet, and if the tweet contains a non-negative emoticon, it is labeled as a Personal Non-Negative tweet.

This approach has limitations in terms of coverage and sentiment strength. For coverage, this previous method only considered the existence of profanity and emoticons, but it did not take into account the frequency of them. A single use of a profanity word is relatively common for Twitter users to express their emotions, but multiple

Table 2 Partial list of the emoticons used

Negative	Non-Negative
;(	x-D
:'(	:')
:{	:-))
:$	}{'
:'-(	;-)

Table 3 Whitelist of stop words for building TR-NN

	Negative	Non-Negative
Whitelist	Negative emoticons, profanities, AFINN negative words	Neutral and positive emoticons, AFINN non-negative words

uses of profanity words indicate a strong negative sentiment. In addition, the number of profanity words or emoticons is relatively small, since the profanity list contains "only" 340 words and the emoticon list consists of 33 emoticons. It is quite possible to miss potential Personal Negative or Personal Non-Negative tweets with this approach. For the sentiment strength, this previous method only considered the existence of profanity or emoticon, but did not consider the various sentiment strengths of words, which are good indicators for the tweet sentiment detection.

In this chapter, to address these previous limitations, we have developed a new Personal Negative versus Personal Non-Negative labeling method. This new method uses metrics generated from the AFINN lexicon as shown in Table 3, in addition to emoticons and profanities. AFINN [26] is a publicly available list of 2,477 English words and phrases rated for valence with an integer between -5 (negative) and $+5$ (positive). To label a Personal tweet as Personal Negative or Non-Negative, we aggregated the frequencies of profanity words, emoticons, and AFINN words into two metrics: Negative Score and Non-Negative Score. Subsequently we used a threshold to determine the label of the tweet. The Negative Score and Non-Negative Score are defined as follows.

$$NegativeScore = \frac{PR*5+NC_3*3+NC_4*4+NC_5*5+NE*3}{PR*5+NC_3*3+NC_4*4+NC_5*5+NE*3+NNC_3*3+NNC_4*4+NNC_5*5+NNE*3}$$

$$NonNegativeScore = \frac{NNC_3*3+NNC_4*4+NNC_5*5+NNE*3}{PR*5+NC_3*3+NC_4*4+NC_5*5+NE*3+NNC_3*3+NNC_4*4+NNC_5*5+NNE*3}$$

In the above formulas, PR, NC_3, NC_4, NC_5, and NE are the numbers of profanity words, words having -3 valence in AFINN, words having -4 valence, words having -5 valence, and the number of negative emoticons in a tweet. Analogously, NNC_3, NNC_4, NNC_5, NNE are the numbers of words having $+3$ valence in AFINN, $+4$ valence, $+5$ valence, and the number of non-negative emoticons in a tweet. The two metrics express the proportion of negative words and the proportion of non-negative words, respectively.

The threshold is set to 0.95, which means that if the Negative Score of a tweet is greater than or equal to 0.95, it is labeled as a Personal Negative tweet. A Personal Non-Negative tweet is labeled similarly. These two metrics use a larger number of words to label tweets than our previous method (2850 compared with 373). This allows us to generate a larger size of training data and to use sentiment strength to label tweets more accurately. The experimental results that compare the current method and our previous method are shown in Sect. 3.3.2.

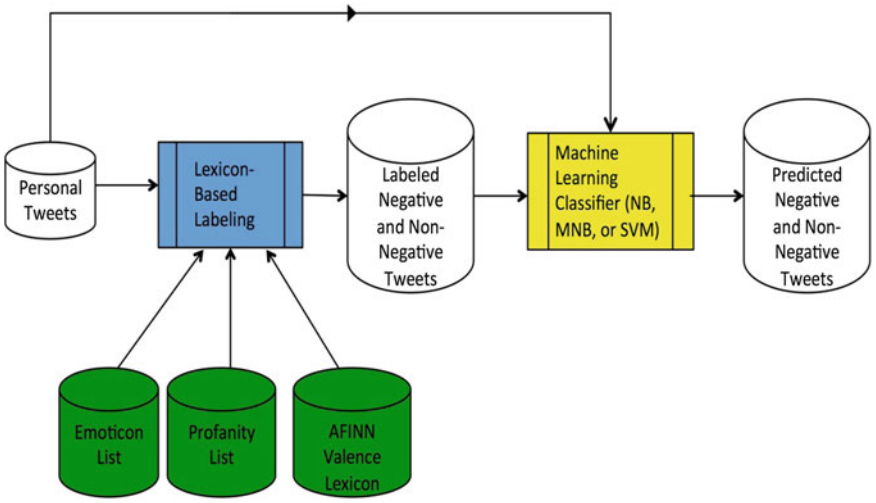

Fig. 5 Negative versus Non-Negative classification

Figure 5 shows the process of Negative versus Non-Negative classification, where the blue part represents the automatic labeling of tweets; the green part represents lexicons; the yellow part represents the Machine Learning classifiers. 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 tweets.

Tweets were labeled as *PN* (Personal Negative) or *PNN* (Personal Non-Negative) using the labeling method described above. These two categories (*PN* and *PNN*) of labeled tweets were combined into the training dataset TR-NN for Negative versus 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''_{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 versus 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 versus 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* (Personal Negative) or *PNN* (Personal Non-Negative). (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. TS_i is the tweet label dataset for time i, and $TS_i = \{ts_1, ts_2, ts_3, \ldots, ts_n\}$, where the label of ts_i is either *PN* (Personal Negative), or *PNN* (Personal Non-Negative), or *NT* (News Tweet).

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

Personal Negative	I remember the Ebola panic days .. Damn. If you were ill during that time, everyone would look at you funny legionnaires disease is killing #newyorkcity one block at a time :(
Personal Non-Negative	So excited for @CancerCenter's #GenomicsChat that starts in ONE HOUR! :) Who's joining me? Learn about new precise treatments for #melanoma. Did it! Woke early today to run 10 miles! First time since the melanoma on my knee 2 years ago :) Praise God!

3.3 Experimental Results of the Two-Step Sentiment Classification

The dataset was collected from March 13 2014 to June 29 2014 and was used by our previous work [12]. The statistics of the collected datasets are shown in Table 5. Only English tweets are used in our experiments. Some datasets have a larger portion of non-English tweets, for example, influenza, swine flu, and tuberculosis compared with other datasets.

Table 5 The statistics of the collected dataset [12]

Dataset id	Tweet type	Total number of tweets	Number of non-English tweets	Number of tweets after preprocessing
1	Listeria	13,572	1,979	4,544
2	Influenza	1,509,609	716,901	527,489
3	Swine flu	73,974	35,970	20,430
4	Measles	166,555	8,808	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	7,416	20,915
11	Air disaster	15,871	681	5,765
12	Melanoma experimental drug	86,757	9,858	40,261

3.3.1 Evaluation Based on Human Annotated Test Dataset

To compare the Naïve Bayes, Two-Step Multinomial Naïve Bayes, and Two-Step Polynomial-Kernel Support Vector Machine classifiers, we created a test dataset using human annotation. Weka's implementations [55] of these classifiers was used. 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.

Each tweet was annotated by three people out of a set of six contributors. In the evaluation, 1 was assigned to Personal and a 0 was assigned to News. If a tweet was labeled as a Personal tweet, the annotator was asked to label it as Personal Negative or Personal Non-Negative tweet. Fleiss' Kappa [56] was used to measure the inter-rater agreement between the three annotators. Table 6 [12] 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.

3.3.2 Sentiment Classification Results

The results of the two-step classification approach are shown in this section. In order to evaluate the usability of two-step classification, Personal versus News classification and Negative versus Non-Negative classification were also evaluated with human annotated datasets. For Personal versus News Classification, we compared our Personal versus News classification method with three baseline methods.

- A random selection method
- The clue-based classification method described above
- A URL-based method, in which a tweet that contains an URL is classified as a News tweet; otherwise a Personal tweet.

The classification accuracies of different methods are presented in Table 7 [12]. The results show that 2S-MNB and 2S-NB outperforms all three baselines in most of the cases. 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 base line clue-based classification approaches in most of the cases.

Table 6 Agreement between human annotators [12]

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 Accuracy of Personal versus News classification on human annotated datasets [12]

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

Some unigrams are learned by the ML-based methods and are shown to be useful for the classification. To better understand this effect, ablation experiments were carried out with the Personal versus News classification on the human annotated datasets. The classifier 2S-MNB was used, since it took much less time to train than the best classifier 2S-NB on the human-annotated test dataset. More precisely, it was trained with the automatically generated data from the Epidemic, Mental Health, and Clinical Science domains collected in 2014.

The trained classifiers were used to classify the sentiments of human annotated datasets from 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 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 8 shows the results of ablation experiments for Personal versus News classification. For example, the unigrams "i", "http", "app", "url" are not in MPQA corpus but are learned by the ML classifier 2S-MNB as the most important unigrams contributing to classification.

The second step in the two-step classification algorithm is to separate Negative tweets from Non-Negative tweets. As discussed in Sect. 3.2.1, the training datasets are automatically labeled, if the tweet has a higher-than-threshold Negative Score or Non-Negative Score. Both scores are calculated by considering the occurrence frequency of words from the profanity list, words from AFINN, and emoticons from the emoticon lists. We compare the performance of Negative versus Non-Negative classification with previous labeling method and with the current labeling method. In this chapter, we call the previous method EPE, as it is an **E**xistence-based method using **P**rofanity list and **E**moticons, and we called the current labeling method FPEA, as it is a **F**requency-based method using **P**rofanity, **E**moticon, and **A**FINN. The classifier is trained by each of the three models, Multinomial Naïve Bayes (MNB), Naïve Bayes (NB), and Support Vector Machine (SVM). The accuracies of Negative versus Non-Negative classification and confusion matrices of the best classifiers

Table 8 Most important unigrams in Personal versus News classification

Dataset	Unigrams with most importance
Epidemic	url, i, case, but
Mental Health	url, disorder, often, bipolar
Clinical Science	melanoma, health, http, co, risk, prevention, app

Table 9 Negative versus Non-Negative Classification Results on Human Annotated Datasets (EPE/FPEA)

Dataset Id	2S-MNB	2S-NB	2S-SVM
Epidemic	0.73/**0.80**	0.59/0.59	0.59/0.66
Mental Health	0.63/0.61	0.65/**0.67**	0.57/0.57
Clinical Science	0.64/0.73	0.73/0.68	0.68/0.68

Table 10 Confusion Matrices of The Best Personal Negative versus 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 (FPEA)	19	6	6	28
Mental Health	2S-NB (FPEA)	19	15	15	43
Clinical Science	2S-MNB (FPEA)	4	6	6	28

for human annotated datasets are shown in Tables 9 and 10, respectively. Overall, the **F**requency-based method with **P**rofanity, **E**moticon, and **A**FINN increases the accuracy of the best classifier by 7 % in the epidemic dataset, and by 2 % in the mental health dataset compared with the previous **E**xistence-based method using **P**rofanity list and **E**moticons. Overall, 2S-MNB (FPEA) achieved the best Negative versus Non-Negative result in terms of accuracy while being faster than 2S-SVM and 2S-NB.

3.3.3 Error Analysis of Sentiment Classification Output

We analyzed the output of sentiment classification. As discussed in Sect. 3.3.1, 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. 3.3.2, 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 versus News classification, we found two major types of errors. The first type of error is that the tweet is 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 (Pointed to by a 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 that appears in the web page pointed to by the URL. If the percentage is low, it is probably a Personal tweet since most of the tweet text is the user's contribution. If the percentage is high, approaching 100 %, it is more likely a News tweet since tweeters often paste the title of a news article into their messages.

The second type of error is that the tweet is in fact a News item, but is classified as a Personal tweet. In total, 48 out of all 140 errors are of this type. A suggested 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 very 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 versus News classification. For Negative versus 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 complement the current Machine Learning paradigm. The appearance of negative phrases such as "I did not like XYZ at all" and "I will not do XYZ any more" are possible indicators of Negative tweets.

3.3.4 Twitter Data Bias

As pointed out in the work of Bruns and Stieglitz [57], there are two questions to be addressed in terms of generalizing collected Twitter data.

- Does Twitter data represent Twitter?
- Does Twitter represent society?

According to the documentation of Twitter [11], the Twitter Streaming API returns at most 1 % of all the tweets at any given time. Once the number of tweets matching the given API parameters (keywords, geographical boundary, user ID) goes beyond the 1 % threshold, Twitter will return a sample of the data to the user. To address this problem, we used domain specific keywords (e.g. h1n1, h5n1) for each tweet type (e.g. listeria) to increase the coverage of collected data [58].

As for the question whether Twitter postings are representative of the society at large, Mislove et al. [59] have found that the Twitter users significantly over-represent the densely populated regions of the US. This might be due to the better availability of high-speed internet in large cities. Twitter users are also overwhelmingly male, and highly biased with respect to race distribution and ethnicity distribution. To reduce the first bias of the collected Twitter data, we defined the Measure of Concern in relative terms. It depends on the fraction of all tweets that have been classified as "Personal Negative" tweets. We assume that as long as the sample of tweets is representative, the Measure of Concern, which is the Personal Negative portion of all tweets, should be similar across different samples sizes, e.g., 1, 10, 100 %, etc.

4 Public Health Concern Trend Analysis

We are interested in making the sentiment classification results available for public health monitoring, especially the results of computing the *Measure of Concern,* to monitor public sentiments towards different types of diseases. The definitions of Measure of Concern, Non-Negative Sentiment, News Count, and Peak are shown below as English text. For a more formal treatment, refer to work by Ji et al. [12].

Definition 4a (*Measure of Concern*) *Measure of Concern (MOC)* 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 Measure of Concern increases with the relative growth of Personal Negative tweets and with the absolute growth of Personal Negative tweets.

The reason for including the relative and the absolute growth of personal tweets into one measure is that, for example, a ratio of 9 : 10 Personal Negative tweets: Personal Tweets appears high, but a lower ratio of 7000 : 10000 should contribute more to the Measure of Concern, because a greater number of the "tweeting public" is involved in this social media discourse.

Definition 4b (*Non-Negative Sentiment*) Similarly, the *Non-Negative Sentiment* NN_i is the square of the total number of Personal Non-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.

Definition 4c (*News Count*) Finally, the *News Count* 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. An absolute News Count is able to better represent the popularity of News.

Definition 5 (*Peak*) Given a timeline of numerical values, 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.

The method for computing the quantitative correlation is shown in Fig. 6. There are three inputs for the correlation process. The News tweets are the outputs of the first step, as shown in Fig. 4; the Personal Negative tweets and the Personal Non-Negative tweets are the outputs in the second step, as shown in Fig. 5. The Jaccard Coefficient is used for computing the correlation.

After the two-step sentiment classification method has been applied to the raw tweets, we can produce three timelines: Measure of Concern timeline, Non-Negative sentiment timeline, and News timeline, respectively. Next, peaks P_1, P_2, and P_3 are generated from these three timelines. The time interval is set to seven days. We are

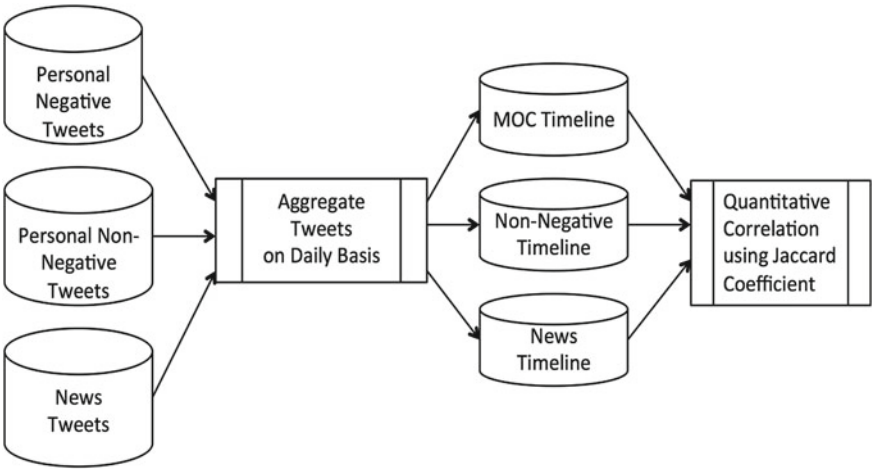

Fig. 6 Correlation between sentiment trends and News trends

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).

We hypothesized that there might be a time delay between the sentiment peaks and the News timeline peaks. For example, an alarming news report might lead to many Twitter users expressing their negative emotions. On the other hand, social media are nowadays often ahead of official news reports, as shown in Broersma and Graham [60] that the news media now often pick up tweets for their news coverage. In the first case, News peaks would precede tweet sentiment peaks. In the second case, tweet sentiment peaks would precede News peaks. We attempted to quantify these alternative choices using the Jaccard Coefficient for this purpose. Thus we defined two correlations as follows:

$$JC(MOC, NEWS, t) = \frac{|P_{2,c+t} \cap P_{1,c}|}{|P_{2,c+t} \cup P_{1,c}|}$$
$$JC(NN, NEWS, t) = \frac{|P_{3,c+t} \cap P_{1,c}|}{|P_{3,c+t} \cup P_{1,c}|}$$

$P_{2,c+t}$ 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 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 $c + t$.

By its definition, the Jaccard Coefficient has a value between 0 and 1. The closer the value is to 1, the better the two time series correlate with each other. To illustrate the calculation of Jaccard Coefficient, we use the following example.

Assume a MOC timeline and a News timeline, where the MOC timeline has nine peaks and the News timeline has eight peaks. Three peaks of MOC and another three peaks of News are pair-wise matched. Note that two peaks match with each other if and only if the two peaks happen on exactly the same day. This is an arbitrary definition, which may be replaced by a finer grain size (hours) or possibly a larger grain size (e.g., weekends). The remaining six peaks of MOC and the remaining five peaks of News are not matched with each other.

Then the Jaccard Coefficient 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. Therefore,

$$\mathrm{JC} = 3/(9 + 8 - 3) = 0.21$$

The best Jaccard Coefficient between MOC peaks and News peaks for a given dataset was computed as follows: First we directly computed the JC between MOC peaks and News peaks without any time delay or lead, and we recorded the result. Then we added one, two, or three days of lead to the original MOC, computed the correlation between the revised MOC peaks and the original News peaks respectively, and recorded these three results. Thirdly, we added one, two, or three 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 peaks of MOC and the peaks of NN (Non-Negative) were correlated with the peaks of News in all datasets with a Jaccard Coefficient of 0.2–0.3.

In order to study how an observable increase in MOC relates to an actual health event (e.g., News count), we quantified the timeline trends of daily MOC and daily News Count for listeria, a potentially lethal foodborne illness, as shown in Fig. 7.

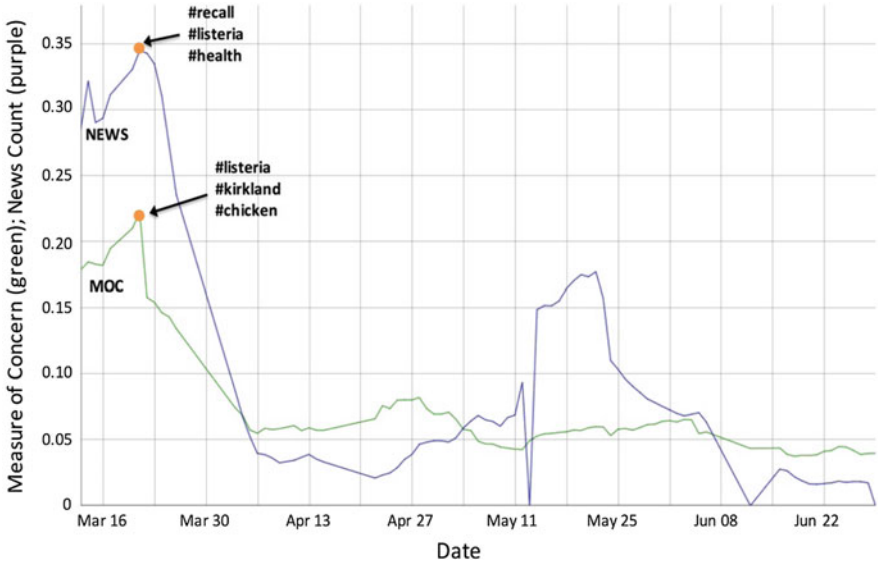

Fig. 7 Comparison between Measure of Concern and News Count timeline trend

The News Count is 0–1 normalized and the top 3 most frequent hash tags for the peak date are shown. The News Count (purple) Peak occurred on March 21st, because on that same day, several food items produced by Parkers Farm were recalled due to a listeria contamination. We note that there was an observable increase of MOC (green) as well, which shows that the general public seemed to express negative emotions according to the news during this circumstance.

Prototype System

We have developed a prototype system of ESMOS (Epidemic Sentiment Monitoring System) to monitor the timeline and topic distribution of public concern, as a part of the Epidemics Outbreak and Spread Detection System [12]. ESMOS displays (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 with a capability to drill down to the individual tweets; and (3) a public health concern map to show the geographic distribution of particular disease concentrations with different granularity (e.g. state, county or individual location level). Figure 8 shows the different visual analytics tools. The public health specialists can utilize the concern timeline chart, as shown in Fig. 8a, to monitor (e.g. identify concern peaks) and compare public concern timeline trends for various diseases. Then the specialists might be interested

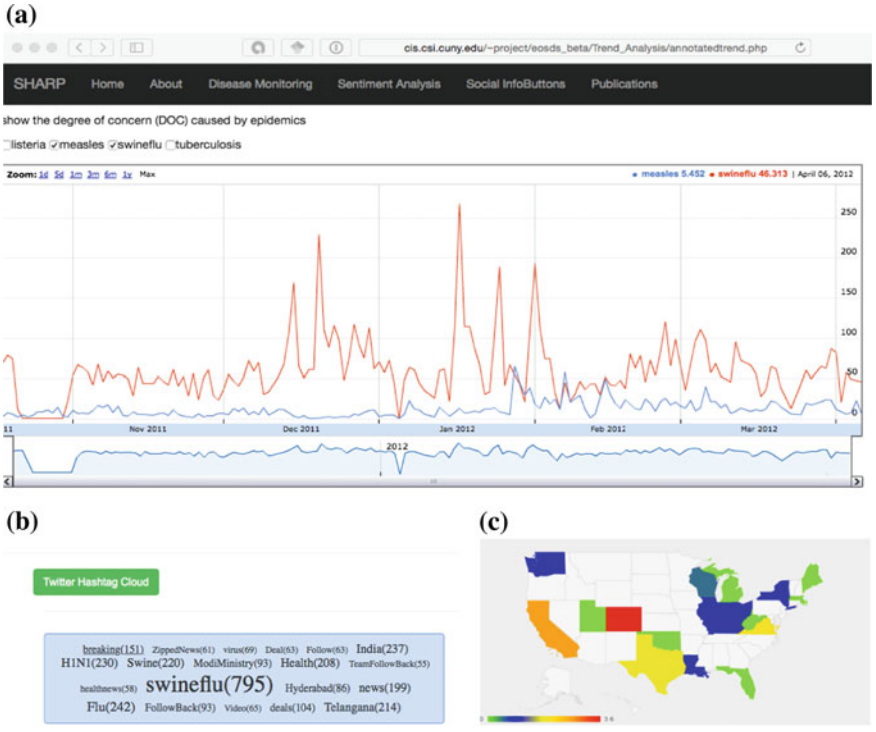

Fig. 8 ESMOS visual analytics tools for public concern monitoring: **a** Sentiment timeline chart, **b** Topics cloud and **c** Public concern distribution map

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. 8b to browse the top topics within a certain time period for different diseases, and individual tweets. The public health concern heat map in Fig. 8c shows the state-level public concern levels.

The ESMOS prototype is currently implemented to monitor limited diseases (c.f. Table 5), but our proposed model can use a disease ontology, such as a dedicated epidemic ontology or a UMLS ontology, to monitor any disease of interest.

5 Chapter Summary

In this chapter, we explored the potential of mining social network data, such as tweets, to provide a tool for public health specialists and government decision makers to gauge a 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 distinguished Personal from News (Non-Personal) tweets. In the second stage, the sentiment analysis was applied only to Personal tweets to distinguish Negative from Non-Negative tweets. In order to 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, respectively. These moderate agreements illustrate the complexity of the sentiment classification task, since even humans exhibit relatively low agreement on the labels of tweets.

Our contributions are summarized as follows.

(1) We developed a two-step sentiment classification method by combining clue-based labeling and Machine Learning (ML) methods by first automatically labeling the training datasets, and then building classifiers for Personal tweets and classifiers for tweet sentiments. The two-step classification method shows 10 % and 22 % increase of accuracy over the clue-based method on epidemic and mental health dataset, respectively in Personal versus Non-Personal classification. In Negative versus Non-Negative classification, the frequency-based method FPEA, which uses the AFINN lexicon, increases the accuracy of the best classifier by 7 % for the epidemic dataset and by 2 % for the mental health dataset compared with our previously used method EPE which had a list of profanities and a list of emoticons as the only sentiment clues. Thus, the use of AFINN resulted in a measurable improvement over our previous work.

(2) We quantified the MOC using the results of sentiment classification, and used it to reveal the timeline trends of sentiments of tweets. The peaks of MOC and the peaks of NN (Non-Negative) correlated with the peaks of News with Jaccard Coefficients of 0.25–0.3.

(3) We applied our sentiment classification method and the Measure of Concern 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.

Future work involves the following.

(1) The Measure of Concern (MOC) is currently based on the number of Personal Negative tweets and total number of tweets on the same day. The Measure of Concern 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. To achieve this goal, we need to extend the simplistic Negative/Non-Negative categories to a wider range of well recognized emotions, such as "concern", "surprise", "disgust", or "confusion".

We plan to employ an ontology engineering approach to construct an emotion ontology from our collected Twitter messages. The emotion ontology will contain the basic emotions such as anger, confusion, disgust, fear, concern, sadness, etc. along with their representative words or phrases. With the constructed emotion ontology, we will be able to detect the tweets expressing real concern and to more accurately quantify the trend of the Measure of Concern.

(2) 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. In Personal versus 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 likely 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 plan to perform a quantitative comparison of these knowledge-based approaches with our ML approach in the future.

(3) The prototype ESMOS implementation needs to be scaled to detect and monitor diseases on a large scale using a full scale Disease Ontology. The sentiment analysis should be performed as the data is captured so that the tracking of public concerns happens in real-time. Public concern about health in general is not limited on infectious diseases but concerns may also be expressed about particular drugs or treatments, and even current and proposed health policies. To promote an understanding of all the contexts related to a disease outbreak, the system needs to consider many different data sources.

(4) 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 HealthMap and CDC reports. The sentiment trends for topics will also be studied by combining the sentiment analysis algorithms with topic modeling algorithms.

(5) All our work so far used epidemics, mental health and clinical science as domains. Thus, all of our experiments were health-related. However, there are other areas where tweets and the traditional news compete with each other. These areas include politics, e.g. presidential candidate debates, the economy, e.g., a precipitous fall of the Dow Jones index, natural disasters, e.g. typhoons and hurricanes, acts of terrorism and war, e.g., roadside bombs, and spontaneous protests, as they were common during the Arab Spring. All these areas are excellent targets for testing theories about the interplay between news and social media.

Acknowledgments This research has been partially funded by the PSC-CUNY Research Foundation under the awards 42-64266 and 43-65232, and by the Leir Charitable Foundations through the School of Management at NJIT.

References

1. Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D.: Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. **5**, e151 (2008)
2. Collier, N., Doan, S.: Syndromic classification of Twitter messages. Electron. Healthc. **91**, 186–195 (2012)
3. Signorini, A., Segre, A.M., Polgreen, P.M.: 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**, e19467 (2011)
4. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1568–1576 (2011)
5. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the Social Web. In: Proceedings of 2nd International Workshop on Cognitive Information Processing, pp. 411–416 (2010)
6. Reuters News. http://www.reuters.com/article/2014/10/18/us-health-ebola-usa-idUSKCN0I61BO20141018
7. Zhu, X., Wu, S., Miao, D., Li, Y.: Changes in emotion of the Chinese public in regard to the SARS period. Soc. Behav. Personal. **36**, 447–454 (2008)
8. Guardian News. http://www.guardian.co.uk/world/2011/mar/17/chinese-panic-buy-salt-japan
9. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**, 1012–1014 (2009)
10. Twitter. http://www.twitter.com
11. Twitter Documentation. https://dev.twitter.com/docs
12. Ji, X., Chun, S.A., Wei, Z., Geller, J.: Twitter sentiment classification for measuring public health concerns. Soc. Netw. Anal. Min. **5**, 1–25 (2015)
13. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining lexicon-based and learning-based methods for Twitter sentiment analysis (2011)
14. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. Mining Text Data, pp. 415–463 (2012)
15. Mohammad, S.M., Kiritchenko, S., Zhu, X.: NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets (2013). arXiv preprint arXiv:1308.6242
16. Saif, H., Fernandez, M., He, Y., Alani, H.: Evaluation datasets for twitter sentiment analysis. In: Proceedings of 1st Workshop on Emotion and Sentiment in Social and Expressive Media (2013)

17. Sha, Y., Yan, J., Cai, G.: Detecting public sentiment over PM2.5 pollution hazards through analysis of Chinese microblog. In: The 11th International Conference on Information Systems for Crisis Response and Management, pp. 722–726 (2014)
18. Ji, X., Chun, S.A., Geller, J.: Monitoring public health concerns using Twitter sentiment classifications. In: Proceedings of IEEE International Conference on Healthcare Informatics, pp. 335–344 (2013)
19. Liben Nowell, D., Kleinberg, J.: The link prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**, 1019–1031 (2007)
20. Hollnagel, E., Woods, D.D.: Cognitive systems engineering: new wine in new bottles. Int. J. Man Mach. Stud. **18**, 583–600 (1983)
21. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. **40**(D1), D940–D946 (2012)
22. List of ICD-9 Codes. https://en.wikipedia.org/wiki/List_of_ICD-9_codes_001%E2%80%93139:_infectious_and_parasitic_diseases
23. Chun, S., Geller, J.: Evaluating ontologies based on the naturalness of their preferred terms. In: Proceedings of the 41st Annual International Conference on System Sciences, pp. 238–238 (2008)
24. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**, 1–167 (2012)
25. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 105–112 (2003)
26. Hansen, L.K., Arvidsson, A., Nielsen, F.Å., Colleoni, E., Etter, M.: Good friends, bad news-affect and virality in twitter. Future information technology, pp. 34–43. Springer, Berlin (2011)
27. Esuli, A., Sebastiani, F.: Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of LREC, pp. 417–422 (2006)
28. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**, 1–135 (2008)
29. Pang, B., Lee, L., Vaithyanathan, S.: 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 (2003)
30. Mishne, G.: Experiments with mood classification in blog posts. In: Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access (2005)
31. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of Human Language Technologies Conference on Empirical Methods in Natural Language Processing, pp. 347–354 (2005)
32. Johansson, F., Brynielsson, J., Quijano, M.N.: Estimating citizen alertness in crises using social media monitoring and analysis. In: Proceedings of European Intelligence and Security Informatics Conference, pp. 189–196 (2012)
33. Brynielsson, J., Johansson, F., Jonsson, C., Westling, A.: Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. Secur. Inf. **3**, 1–11 (2014)
34. Saif, H., Fernández, M., Alani, H.: Automatic stopword generation using contextual semantics for sentiment analysis of Twitter. In: Proceedings of 13th International Semantic Web Conference (2014)
35. Refaee, E., Rieser, V.: An Arabic twitter corpus for subjectivity and sentiment analysis. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, pp. 2268–2273 (2014)
36. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of Twitter. The Semantic Web-ISWC, pp. 508–524 (2012)
37. Barbosa, L., Feng, J.: Robust sentiment detection on Twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 36–44 (2010)

38. Bifet, A., Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Proceedings of the 13th International Conference on Discovery Science, pp. 1–15. Springer (2010)
39. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation, pp. 1320–1326 (2010)
40. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: 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 (2011)
41. Zhou, Z., Zhang, X., Sanderson, M.: Sentiment analysis on twitter through topic-based lexicon expansion. Databases Theory and Applications, pp. 98–109. Springer International Publishing, Switzerland (2014)
42. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, New York (1990)
43. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
44. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Interact. Intell. Syst. **2**, 1–27 (2011)
45. Salathe, M., Khandelwal, S.: Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput. Biol. **7**, e1002199 (2011)
46. Zhuang, L., Jing, F., Zhu, X.-Y.: Movie review mining and summarization. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 43–50 (2006)
47. Chew, C., Eysenbach, G.: Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS One **5**, e14118 (2010)
48. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: 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 (2010)
49. Wiebe, J., Riloff, E.: Creating subjective and objective sentence classifiers from unannotated texts. In: Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing, pp. 486–497 (2005)
50. Wilson, T., Wiebe, J.: Annotating opinions in the world press. In: Proceedings of 4th SIGdial Meeting on Discourse and Dialogue, pp. 13–22 (2003)
51. Twitter 4J. http://twitter4j.org/en/
52. Profanity List. http://web.njit.edu/xj25/eosds_beta/files/profanity_list.txt
53. FCC Guide. http://www.fcc.gov/guides/obscenity-indecency-and-profanity
54. News Stopwords. http://web.njit.edu/xj25/eosds_beta/files/news_stopwords.txt
55. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. **11**, 10–18 (2009)
56. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**, 378–382 (1971)
57. Bruns, A., Stieglitz, S.: Twitter data: what do they represent? Inf. Technol. **56**, 240–245 (2014)
58. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose (2013). arXiv preprint arXiv:1306.5204
59. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., Rosenquist, J.N.: Understanding the demographics of Twitter users. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 554–557 (2011)
60. Broersma, M., Graham, T.: Twitter as a news source: how Dutch and British newspapers used tweets in their news coverage, 2007–2011. Journal. Pract. **7**, 446–464 (2013)