This is a repository copy of *Using twitter for public health surveillance from monitoring and prediction to public response*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/157301/

Version: Published Version

Article:

Jordan, Sophie E., Hovet, Sierra E., Fung, Isaac Chun Hai et al. (3 more authors) (2018) Using twitter for public health surveillance from monitoring and prediction to public response. Data. 6. ISSN 2052-4463

https://doi.org/10.3390/data4010006

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

data

MDPI

Review

Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response

Sophie E. Jordan [1], Sierra E. Hovet [2], Isaac Chun-Hai Fung [3], Hai Liang [4], King-Wa Fu [5] and Zion Tsz Ho Tse [2,*]

[1] School of Chemical, Materials, and Biomedical Engineering, College of Engineering, University of Georgia, Athens, GA 30602, USA; sophie.jordan25@uga.edu

[2] School of Electrical and Computer Engineering, College of Engineering, University of Georgia, Athens, GA 30602, USA; hovetsia@uga.edu

[3] Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA 30460, USA; cfung@georgiasouthern.edu

[4] School of Journalism and Communication, Chinese University of Hong Kong, Hong Kong, China; hailiang@cuhk.edu.hk

[5] Journalism and Media Studies Centre, The University of Hong Kong, Hong Kong, China; kwfu@hku.hk

* Correspondence: ziontse@uga.edu; Tel.: +1-706-542-4189

Received: 14 December 2018; Accepted: 22 December 2018; Published: 29 December 2018

Abstract: Twitter is a social media platform where over 500 million people worldwide publish their ideas and discuss diverse topics, including their health conditions and public health events. Twitter has proved to be an important source of health-related information on the Internet, given the amount of information that is shared by both citizens and official sources. Twitter provides researchers with a real-time source of public health information on a global scale, and can be very important in public health research. Classifying Twitter data into topics or categories is helpful to better understand how users react and communicate. A literature review is presented on the use of mining Twitter data or similar short-text datasets for public health applications. Each method is analyzed for ways to use Twitter data in public health surveillance. Papers in which Twitter content was classified according to users or tweets for better surveillance of public health were selected for review. Only papers published between 2010–2017 were considered. The reviewed publications are distinguished by the methods that were used to categorize the Twitter content in different ways. While comparing studies is difficult due to the number of different methods that have been used for applying Twitter and interpreting data, this state-of-the-art review demonstrates the vast potential of utilizing Twitter for public health surveillance purposes.

Keywords: public health; Twitter; classification; data mining; Zika; Ebola

1. Introduction

Promoting and protecting the health of communities is the goal of public health. Public health professionals strive to prevent illness and injury. They track outbreaks and shed light on why poor health affects some more than others. Epidemiology is concerned with the dynamics of health conditions in populations. Research in epidemiology aims to identify the distribution, incidence, and etiology of diseases in order to facilitate their prevention, control, and treatment [1]. Public health surveillance is the practice of public health agencies that collect, manage, analyze, and interpret data in a systematic and ongoing manner, and disseminate such data to programs that will facilitate actions in public health [2]. Disease outbreaks are increasingly becoming more frequent and diverse around the world due to a combination of ecological, environmental, and socio-economic factors. New tools for detecting outbreaks are being developed, including those that analyze digital data ("digital disease detection") [3]. Rapid response through improved surveillance is important to combat emerging infectious diseases such as Ebola and Zika [4]. Technological innovation and its applications in routine surveillance for other diseases, such as influenza, may enable nowcasts and the prediction of disease trends [5,6].

Twitter is a social media platform for sharing short-text updates that may contain public health information. Since platforms such as Twitter are in real-time and can be mined as such, they are promising for widespread implementation in public health applications [7–9].

Twenty-eight research papers discussing the uses of Twitter data for the field of public health are summarized in this review (Section 3). A focus is placed on emerging studies that use data-mining techniques such as tweet classification, sentiment analysis, and user classification for public health surveillance. The goals of this work are to present researchers with the most current accomplishments in this field, and provide them with an understanding of the challenges and potential of using Twitter for public health. Currently, filtering and categorizing tweets into useful data is accomplished via numerous different methods, many of which rely heavily on human intelligence to verify results. There is also a lack of a standard method of comparison for the results of tweet classifications. Before Twitter can be accepted as a reliable source of data for health surveillance, these areas must be improved. Sections 4 and 5 summarize the existing challenges and the future directions in the field.

2. Methods

Twitter data has been found to be useful for several different public health applications, including: (1) monitoring diseases, (2) public reaction, (3) outbreak/emergency situations, (4) prediction, (5) lifestyle, (6) geolocation, and (7) general applications. The articles cited in this paper were found using keyword searches on the bibliographic databases *Google Scholar* [10], *IEEE Xplore* [11], and *Science Direct* [12] from 1 January 2010 to 31 December 2017. All online library access, conference proceedings, and other literature repositories for each author or expert were also researched for relevant articles. The initial search yielded 864 research papers. All of the papers that were found were then processed, and their abstracts were reviewed for relevance and filtered as per the specific selection criteria. Articles that were not pertinent to public health and Twitter were identified through a manual analysis of each abstract, and were excluded. A rating scale of 0 (clearly irrelevant) to 10 (clearly relevant) was applied to the remaining articles. The authors performed a manual scan of each article to assess the scale. To eliminate duplicate or similar research papers published by the same authors, the article that was published in a journal was selected, and the others were eliminated. The articles were then organized into a prioritized list and reviewed in depth in order to write the corresponding sections. After the elimination process was complete, a total of 28 research articles remained for review (see Figure 1).

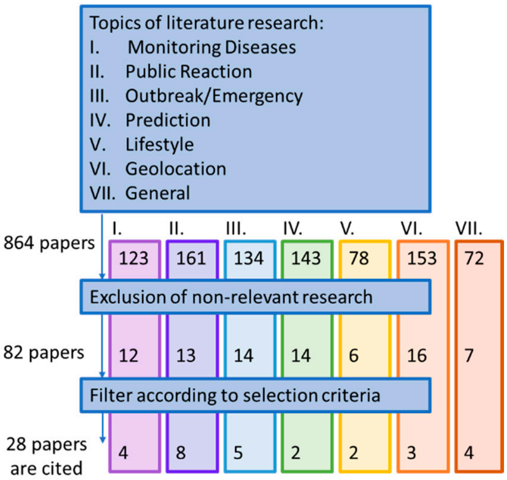


Figure 1. Diagram of research elimination process.

3. Public Health Applications

3.1. Monitoring Diseases

Paul and Dredze proposed a new associative topic model for identifying tweets regarding ailments (Table 1) [13]. This model, called the Ailment Topic Aspect Model (ATAM), identifies relevant tweets by using a combination of keywords and associated topics. ATAM learns the symptoms and treatments that are associated with specific ailments, and organizes the health terms into ailment groups. It then separates the coherent ailment groups from the more general topics. ATAM identifies latent topic information from a large dataset and enables browsing frequently co-occurring words [14]. In testing, both ATAM and latent Dirichlet allocation (LDA) methods were applied to the same dataset. Human intelligence was used to review the ATAM and LDA labels for ailment-related tweets. For the LDA method, 45% agreed with the labels; for the ATAM method, 70% agreed with the labels. The ATAM method produces more detailed ailment information through the inclusion of symptoms and treatments as well. The data from this method was compared to influenza-like illness (ILI) data from the Centers for Disease Control and Prevention (CDC). The Pearson's correlation coefficient between the ATAM frequencies and the CDC data was 0.934 (Google Flu Trends yielded a correlation of 0.932 with the CDC). These results show that the ATAM method is capable of monitoring disease and providing detailed information on occurring ailments.

Gesualdo et al. designed and tested a minimally-trained algorithm for identifying ILI on Twitter [15]. Using the definition of an ILI case from the European Centre for Disease Prevention and Control, the authors created a Boolean search query for Twitter data. This query identifies all of the tweets reporting a combination of symptoms that satisfies the query. The algorithm learns technical and naïve terms to identify all of the jargon expressions that are related to a specific technical term. It was trained based on pattern generalization using term pairs (one technical and one naïve; e.g., emesis–vomiting). After training, the algorithm was able to extract basic health-related term patterns from the web. The performance of this algorithm was manually evaluated by experts. One hundred tweets satisfying the query were selected along with 500 random symptom-containing tweets. These were evaluated by three of the authors independently, and the overall rate of precision was 0.97. When compared to influenza trends reported by the United States (U.S.) Outpatient ILI Surveillance Network (ILINet), the trends that the query found yielded a correlation coefficient of 0.981. The tweets were also selected for geolocation purposes by identifying those with GPS, time zone, place code, etc. The geolocated tweets were compared to the ILINet data to return a correlation coefficient of 0.980.

Coppersmith, Dredze, and Harman analyzed mental health phenomena on Twitter through simple natural language processing methods [16]. The focus of their study was on four mental health conditions: (1) post-traumatic stress disorder (PTSD), (2) depression, (3) bipolar disorder, and (4) seasonal affective disorder (SAD). Self-expressions of mental illness diagnosis were used to identify the sample of users for this study. Diagnosis tweets were manually assessed and labeled as genuine or not. Three methods of analysis were conducted. The first was pattern-of-life. This method looked at social engagement and exercise as positive influences and insomnia as a sign of negative outcomes. Sentiment analysis was also used to determine positive or negative outlooks. Pattern-of-life analysis performs especially poorly in detecting depression, but surprisingly, it performs especially well in detecting SAD. Another analysis method utilized was linguistic inquiry word count (LIWC), which is a tool for the psychometric analysis of language data. LIWC is able to provide quantitative data regarding the state of a patient from the patient's writing. LIWC generally performed on par with pattern-of-life analysis. A third means of analysis was language models (LMs). LMs were used to estimate the likelihood of a given sequence of words. The LMs had superior performance compared to the other analysis methods. The purpose of this study was to generate proof-of-concept results for the quantification of mental health signals through Twitter.

Table 1. Summary of social media mining for disease-monitoring applications in Section 3.1. CDC: Centers for Disease Control and Prevention, ILI: influenza-like illness, ILINet: ILI Surveillance Network, LDA: latent Dirichlet allocation.

Author	Data	Mining Technique	Sample Size	Location/Language	Software	Merits	Limitations
Paul and Dredze [13]	Google Flu Trends, CDC data	High recall word filter with list of 20,000 key phrases	1.6 million tweets	(English)	N/A	Discovers larger number of ailments than LDA with more detailed information.	Lacks geolocation and temporal information for tracking diseases.
Gesualdo et al. [15]	ILINet	APIs (application programming interfaces)	N/A	United States (English)	N/A	Tweets satisfying query have correlation coefficient of 0.981 with ILINet. Geolocation yielded correlation coefficient of 0.980.	Twitter not representative of population; U.S. definition of ILI differs from the one used in this study; only one influenza season studied.
Coppersmith et al. [16]	N/A	Twitter API	3200 tweets per diagnosed user; ~10 k users	(English)	N/A	Indications of different language patterns between control and diagnosed.	Only identifies users who identify themselves as diagnosed with a mental health disorder.
Denecke et al. [17]	N/A	Web crawling, APIs	continuous	(English and German)	M-Eco, Tree Tagger, Stanford Parser	Reduction of monitoring effort.	Only 5–20% of signals were found to be relevant.

Denecke et al. presented a prototype implementation of a disease surveillance system called M-Eco that processes social media data for relevant disease outbreak information [17]. The M-Eco system uses a pool of data from Twitter, blogs, forums, television, and radio programs. The data is continuously filtered for keywords. Texts containing keywords are further analyzed to determine their relevance to disease outbreaks, and signals are automatically generated by unexpected behaviors. Signals are only generated when the threshold for the number of texts with the same word or phrase has been exceeded. These signals, which are mainly generated from news agencies' tweets, are again analyzed for false alarms and visualized through geolocation, tag clouds, and time series. The M-Eco system allows for searching and filtering the signals by various criteria.

3.2. Public Reaction

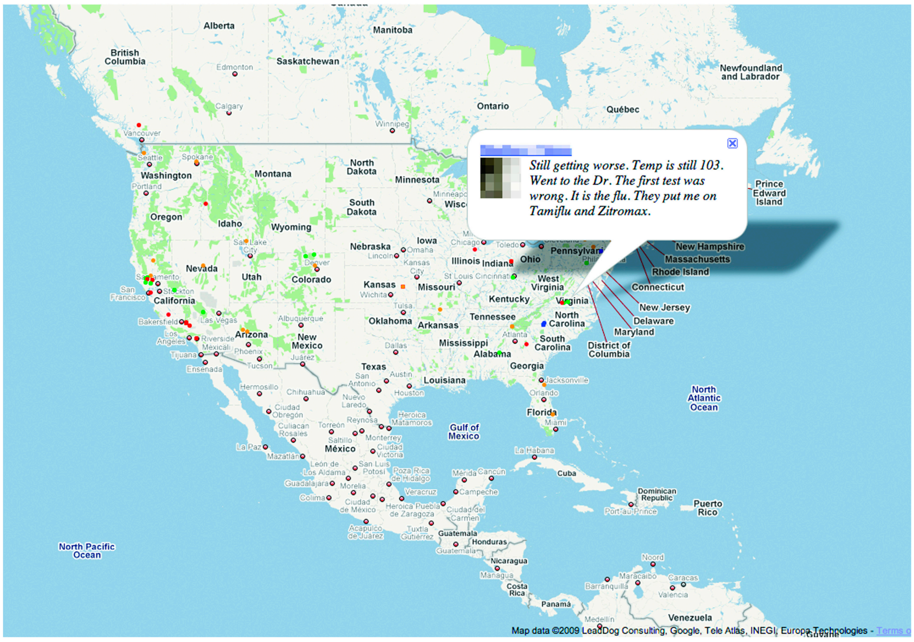
Adrover et al. [18] attempted to identify Twitter users who have HIV and determine if drug treatments and their associated sentiments could be detected through Twitter (Table 2). Beginning with a dataset of approximately 40 million tweets, they used a combination of human and computational approaches, including keyword filtering, crowdsourcing, computational algorithms, and machine learning, to filter the noise from the original data. The narrowed sample consisted of only 5443 tweets. The small sample size and extensive manual hours dedicated to filtering, tagging, and processing the data limited this method. However, the analysis of this data led to the identification of 512 individual users who self-reported HIV and the effects of HIV treatment drugs, as well as a community of 2300 followers with strong, friendly ties. Around 93% of tweets provided information on adverse drug effects. It was found that 238 of the 357 tweets were associated with negative sentiment, with only 78 positive and 37 neutral tweets.

Ginn et al. presented a corpus of 10,822 tweets mentioning adverse drug reactions (ADRs) for training Twitter mining tools [19]. These tweets were mined from the Twitter application programming interface (API) and manually annotated by experts with medical and biological science backgrounds. The annotation was a two-step process. First, the original corpus of tweets was processed through a binary annotation system to identify mentions of ADRs. ADRs, which are defined as "injuries resulting from medical drug use", were carefully distinguished from the disease, symptom, or condition that caused the patient to use the drug initially. The Kappa value for binary classification was 0.69. Once the ADR-mentioning tweets were identified, the second step, full annotation, began. The tweets were annotated for identification of the span of expressions regarding ADRs and labeled with the Unified Medical Language System for IDs. The final annotated corpus of tweets was then used to train two different machine learning algorithms: Naïve Bayes and support vector machines (SVMs). Analysis was conducted by observing the frequency and distribution of ADR mentions, the agreement between the two annotators, and the performance of the text-mining classifiers. The performance was modest, setting a baseline for future development.

Sarker and Gonzalez proposed a method of classifying ADRs for public health data by using advanced natural language processing (NLP) techniques [20]. Three datasets were developed for the task of identifying ADRs from user-posted internet data: one consisted of annotated sentences from medical reports, and the remaining two were built in-house on annotated posts from Twitter and the DailyStrength online health community, respectively. The data from each of the three corpora were combined into a single training set to utilize in machine learning algorithms. The ADR classification performance of the combined dataset was significantly better than the existing benchmarks with an F-score of 0.812 (compared to the previous 0.77). Semantic features such as topics, concepts, sentiments, and polarities were annotated in the dataset as well, providing a basis for the high performance levels of the classifiers.

Behera and Eluri proposed a method of sentiment analysis to monitor the spread of diseases according to location and time [21]. The goal of their research was to measure the degree of concern in tweets regarding three diseases: malaria, swine flu, and cancer. The tweets were subjected to a two-step sentiment classification process to identify negative personal tweets. The first step of classification consisted of a subjectivity clue-based algorithm to determine which tweets were personal and which were non-personal (e.g., advertisements and news sources) The second step involved applying lexicon-based and Naïve Bayes classifiers to the dataset. These classifiers distinguished negative sentiment from non-negative (positive or neutral) sentiment. To improve the performance of these classifiers, negation handling and Laplacian Smoothing techniques were combined with the algorithms. The best performance came from the combination of Naïve Bayes and negation handling for a precision of 92.56% and an accuracy of 95.67%. After isolating the negative personal tweets, the degree of concern was measured.

Signorini, Segre, and Polgreen studied the usefulness of Twitter data in tracking the rapidly evolving public sentiment regarding H1N1 influenza and the actual disease activity [22]. Using keywords to filter the Twitter API and obtain a dataset of over 950,000 tweets, they time-stamped and geolocated each tweet using the author's self-declared home location. A JavaScript application was developed to display a continuously updating Google map of influenza and H1N1-related tweets according to their geographical context. The tweets and sentiments are depicted as color-coded dots on the map, as shown in Figure 2. Users can scroll over the dots to read the related tweets (see Figure 2). Estimates of ILI occurrence rates performed with an average error of 0.28%. When the geolocations of the tweets were factored in, the dataset was reduced due to the rarity of geotagged tweets. The average error for regional ILI estimates was slightly higher at 0.37%. This study demonstrated the concept that Twitter traffic can be used to track public sentiment and concern, and potentially estimate the real-time disease activity of H1N1 and ILIs.



Still getting worse. Temp is still 103. Went to the Dr. The first test was wrong. It is the flu. They put me on Tamiflu and Zitromax.

Figure 2. Influenza Google Map with highlighted tweet. Dots on the map depict different flu-related tweets that have been geolocated [22].

Myslín et al. studied the public sentiment toward tobacco and tobacco-related products through Twitter data [23]. Tweets were manually classified by two annotators to identify genre, theme, and sentiment. From a cohort of 7362 tweets mined through the Twitter API, 57.3% (4215) were classified as tobacco-related. The tweets were then used to train machine learning classifiers to distinguish between tobacco-related and irrelevant tweets as well as positive, negative, or neutral sentiment in tweets. Three machine learning algorithms were tested in this study: SVM, Naïve Bayes, and K-Nearest Neighbors (KNN). The F-score for discriminating between tobacco-related and irrelevant tweets was 0.85. The SVMs yielded the highest performance. Overall, sentiment toward tobacco was found to be more positive (1939/4215, 46%) than negative (1349/4215, 32%) or neutral (see Figure 3). These values were found even after the advertising tweets (9%) were excluded. Words relating to hookah or e-cigarettes were highly predictive of positive sentiment, while more general terms related to tobacco were predictive of negative sentiment. This suggests gaps in public knowledge regarding newer tobacco products. This study was limited by the number of keywords that were used to find tobacco-related tweets. While the novelty effects of hookah and e-cigarettes were not considered in the analysis, this work demonstrated the capabilities of machine learning classifiers trained on Twitter data to determine public sentiment and identify areas to direct public health information dissemination.

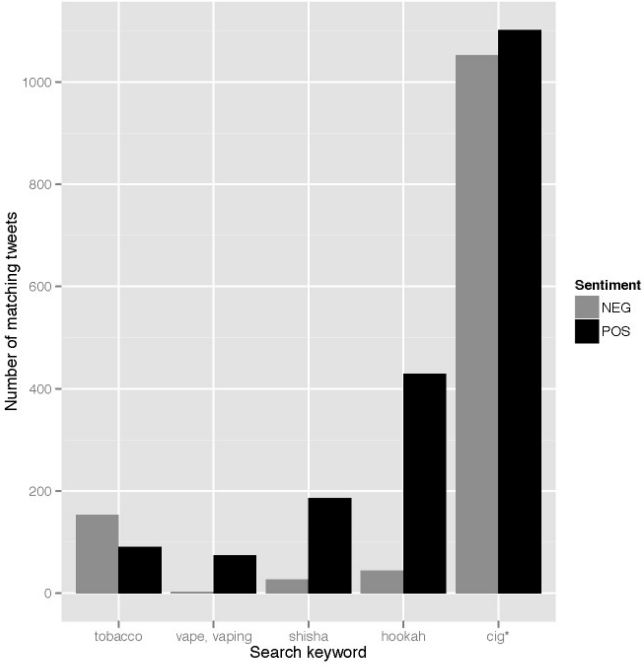


Figure 3. Public sentiment of tweets by search keyword [23].

Ji et al. used Twitter to track the spread of public concern regarding epidemics [24]. Their methods included separating tweets into personal and news (non-personal) categories to focus on public concern. The personal tweets were further classified into personal negative and personal non-negative, depending on the sentiment detected. Training data auto-generated from an emotion-oriented, clue-based method was used to train and test three different machine learning models. The tweets that were classified as personal negative were used to generate a Measure of Concern (MOC) and format the MOC to a timeline. The MOC timeline was compared to a news timeline. The peaks were compared to find a Jaccard correlation coefficient with a range of 0.2–0.3. These results are insufficient for predictions. However, some MOC peaks aligned with news peaks on the timeline, suggesting that the general public expresses negative emotions when news activity increases.

Colleta et al. studied the public sentiment classification of tweets using a combination of SVM and cluster ensemble techniques [25]. This algorithm, named the C^{3}E-SL, is capable of combining classifiers with cluster ensembles to refine tweet classifications from additional information provided by the clusters. Four different categories of tweets were used to train and test the C^{3}E-SL algorithm. The first set consisted of 621 training tweets (215 positive and 406 negative) related to the topic of health care reform. The second set, the Obama–McCain debate, was made up of 3238 tweets. Neutral tweets were removed, leaving only 1906 (710 positive and 1196 negative) to be used for training. The third set contained 1224 tweets (570 positive and 654 negative) related to Apple, Google, Microsoft, and Twitter. The final set consisted of 359 manually annotated tweets (182 positive and 177 negative) from a study completed at Stanford [26]. The results demonstrated that the C^{3}E-SL algorithm performed better than the SVM classifier alone and was competitive with the highest performances found in the literature.

Table 2. Summary of social media mining for public reaction applications as in Section 3.2. ADR: adverse drug reactions, API: application programming interface, NLP: natural language processing, SVM: Support Vector Machines.

Author	Data	Mining Technique	Sample Size	Location/ Language	Software	Merits	Limitations
Adrover et al. [18]	N/A	Gnip Inc.	~40 million tweets	(English)	N/A	Accurate reporting of toxicities	Few users tweet about HIV and adverse treatment effects.
Ginn et al. [19]	N/A	Twitter API	187,450 tweets	(English)	N/A	Able to train Naïve Bayes and SVMs algorithms	Modest performance. Many manual annotations for accuracy.
Sarker and Gonzalez [20]	ADE (medical case reports)	Ongoing ADR research	10,822 tweets, 10,617 daily-strength comments, 6821 mentions in medical reports	(English)	N/A	Use of advanced NLP techniques significantly improves F-scores	Subject to error through misinterpretation.
Behera and Eluri, [21]	N/A	Twitter API	~4500 tweets	unknown	N/A	Classifies tweets as positive, negative, or neutral with high precision and accuracy	No comparison of results to known disease epidemic data.
Signorini et al. [22]	CDC data	Twitter API	951,697 tweets	United States (English)	JavaScript	Average error for ILI estimates is 0.28%, and for regional ILI estimates is 0.37%	Few tweets are geotagged, making regional estimates harder to obtain. No comparable data; results cannot be verified
Myslin et al. [23]	N/A	Twitter API	7362 tweets	(English)	N/A	Identified overall positive sentiment toward tobacco, specifically hookah, and e-cigarettes	Small annotated dataset, limited keywords in identifying tobacco-related tweets, novelty effects not considered.
Ji et al. [24]	News data	Twitter API	unknown	(English)	N/A	Some measure of concern timeline peaks correlated with news timeline peaks	Irony, sarcasm, and profanity make machine learning difficult. All correlation results inconclusive.
Coletta et al. [25]	Literature results	Twitter API	Dataset (1) 621, (2) 3238, (3) 1224, (4) 359	United States (English)	N/A	High accuracy ratings in all four datasets (1. 79.62%, 2. 75.18%, 3. 82.15%, 4. 77.69–81.84%)	Other combinations of classifiers and cluster ensembles not considered.

3.3. Outbreak and Emergency

France and Christopher Cheong used Twitter to conduct a social network analysis case study for the floods of Queensland, New South Wales, and Victoria, Australia, from March 2010 to February 2011 (Table 3) [27]. The research goal was to identify the main active users during these events, and determine their effectiveness in disseminating critical information regarding the crisis. Two types of networks were generated for each of the three flood-affected sites: a "user" network based on the responses of users to certain tweets, and a "user-resources" network connecting user tweets to the included resource links. The most active users were found to be local authorities, political personalities, social media volunteers, traditional media reporters, and nonprofit, humanitarian, and community organizations.

Odlum and Yoon collected over 42,000 tweets related to Ebola during the outbreak in summer 2014 [28]. This Twitter data was analyzed to monitor the trends of information spread, examine early epidemic detection, and determine public knowledge and attitudes regarding Ebola. Throughout the summer, a gradual increase was detected in the rate of information dissemination. An increase in Ebola-related Twitter activity occurred in the days prior to the official news alert. This increase is indicative of Twitter's potential in supporting early warning systems in the outbreak surveillance effort. The four main topics found in Ebola-related tweets during the epidemic were risk factors, prevention education, disease trends, and compassion toward affected countries and citizens. The public concern regarding Ebola nearly doubled on the day after the CDC's health advisory.

Missier et al. studied the performance of two different approaches to detecting Twitter data relevant to dengue and other *Aedes*-borne disease outbreaks in Brazil [29]; both supervised classification and unsupervised clustering using topic modeling performed well. The supervised classifier identified four different classes of topics: (1) mosquito focus was the most directly actionable class; (2) sickness was the most informative class; (3) news consisted of indirectly actionable information; and (4) jokes made up approximately 20% of the tweets studied, and were regarded as noise. It was difficult to distinguish jokes from relevant tweets due to the prevalence of common words and topics. A training set of 1000 tweets was manually annotated and used to train the classifier. Another set of 1600 tweets was used to test the classifier, and resulted in an accuracy range of 74–86% depending on the class. Over 100,000 tweets were harvested for the LDA-based clustering. A range of two to eight clusters were formed, and interclustering and intraclustering were calculated to determine the level of distinction between clusters. The intraclustering was found to be over double that of interclustering, indicating that the clusters were well separated. Overall, clustering using topic modeling was found to offer less control over the content of the topics than a traditional classifier. However, the classifier required a lot of manual annotations, and was thus costlier than the clustering method.

Schulz et al. presented an analysis of a multi-label learning method for classification of incident-related Twitter data [30]. Tweets were processed using three different methods (binary relevance, classifier chains, and label powerset) to identify four labels: (S) Shooting, (F) Fire, (C) Crash, and (I) Injury. Each approach was analyzed for precision, recall, exact match, and h-loss. Keyword-based filtering yielded poor results in each evaluation category, indicating that it is inadequate for multi-label classification. It was found that the correlation between labels needs to be taken into account for classification. The classifier chains method is able to outperform the other methods if a cross-validation is performed on the training data. Overall, it was found that multiple labels were able to be detected with an exact match of 84.35%.

Table 3. Summary of social media mining for outbreak and emergency applications as in Section 3.3. NSW: New South Wales, QLD: Queensland, SNA: social network analysis.

Author	Data	Mining Technique	Sample Size	Location/ Language	Software	Merits	Limitations
Cheong and Cheong [27]	N/A	In-house script with list of hashtags (#)	6014 QLD tweets, 384 NSW tweets, and 1122 Vic tweets	Australia (English)	N/A	SNA capable of identifying most active users, patterns, and tweet frequencies	Queensland flood data was only collected post-flood. Nature and quality of tweets undetermined.
Odlum andYoon [28]	N/A	Google Chrome NCapture	42,236 tweets	Global, mainly Africa, Europe, and America (English)	NCapture	Collected useful data during Ebola outbreak of 2014	Single language. Used self-reported location information.
Missier et al. [29]	Manual annotations	N/A	1000, 1600, and 100,000 tweets	Brazil (Portuguese)	N/A	Classifier accuracy of 84.4%. Clear distinction between clusters	Manual annotations limit the classifier. Clear segregation difficult to achieve in the clustering method.
Schulz et al. [30]	N/A	unknown	unknown	(English)	Mulan	Exact match of 84.35% for identification of multiple labels	Keyword-based filtering inadequate. Injury difficult to identify. Misclassifications due to assignment to "no incident" label.
Gomide et al. [31]	Brazilian Health Ministry	Twitter API	465,444 tweets	Brazil (Portuguese)	N/A	Number of tweets and official data correlated to R^2 = 0.9578	No discussion or analysis of results presented.

Gomide et al. proposed a four-dimensional active surveillance methodology for tracking dengue epidemics in Brazil using Twitter [31]. The four dimensions were volume (the number of tweets mentioning "dengue"), time (when these tweets were posted), location (the geographic information of the tweets), and public perception (overall sentiment toward dengue epidemics). The number of dengue-related tweets was compared to official statistics from the same time period obtained from the Brazilian Health Ministry, and an R2 value of 0.9578 was obtained. The time and location information were combined to predict areas of outbreak. A clustering approach was used to find cities in close proximity to each other with similar dengue incidence rates at the same time. The Rand index value was found to be 0.8914.

3.4. Prediction

Santos and Matos investigated the use of tweets and search engine queries to estimate the incidence rate of influenza (Table 4) [32]. In this study, tweets regarding ILI were manually classified as positive or negative according to whether the message indicated that the author had the flu. These tweets were then used to train machine learning models to make the positive or negative classification for the entire set of 14 million tweets. After classification, the Twitter-generated influenza incidence rate was compared to epidemiological results from Influenzanet, which is a European-wide network for flu surveillance. In addition to the Twitter data, 15 million search queries from the SAPO ((Online Portuguese Links Server)) search platform were included in the analysis. A linear regression model was applied to the predicted influenza trend and the Influenzanet data to result in a correlation value of approximately 0.85.

To test the accuracy of the models in predicting influenza incidence from one flu season to the next, more linear regression models were implemented. The data generated was then compared to the weekly incidence rate reported by the European Influenza Surveillance Network (EISN). The predicted trend appeared to be a week ahead of the EISN report. Interestingly, in this comparison, the flu trend was overestimated by the model in week nine. The EINS did not show the exaggerated rate of influenza; however, media reports and the National Institute of Health demonstrate a high incidence rate in Portugal at the time. This study demonstrated the ability of the models to correlate as well as 0.89 to Influenzanet and across seasons, with a Pearson correlation coefficient (r) value of 0.72.

Kautz and Sadilek proposed a model to predict the future health status ("sick" or "healthy") of an individual with accuracy up to 91% [33]. This study was conducted using 16 million tweets from one month of collection in New York City. Users who posted more than 100 GPS-tagged tweets in the collection month (totaling 6237 individual users) were investigated by data mining regarding their online communication, open accounts, and geolocated activities to describe the individual's behavior. Specifically, the locations, environment, and social interactions of the users were identified. Locations were determined through GPS monitoring, and used to count visits to different 'venues' (bars, gyms, public transportation, etc.), physical encounters with sick individuals (defined as co-located within 100 m), and the ZIP code of the individual (found by analyzing the mean location of a user between the hours of 01:00–06:00). The environment of the user was also determined through GPS, as well as the relative distance of the user to pollution sources (factories, power plants, transportation hubs, etc.). The social interactions of a user were determined through their online communication (Figure 4). Social status was analyzed using the number of reciprocated 'follows' on Twitter, mentions of the individual's name, number of 'likes' and retweets, and through the PageRank calculation. Applying machine learning techniques to mined data, researchers were able to find the feature that was most strongly correlated with poor health: the proximity to pollution sources. Higher social status was strongly correlated with better health, while visits to public parks was also positively correlated with improved health. Overall, the model explained more than 54% of the variance in people's health.

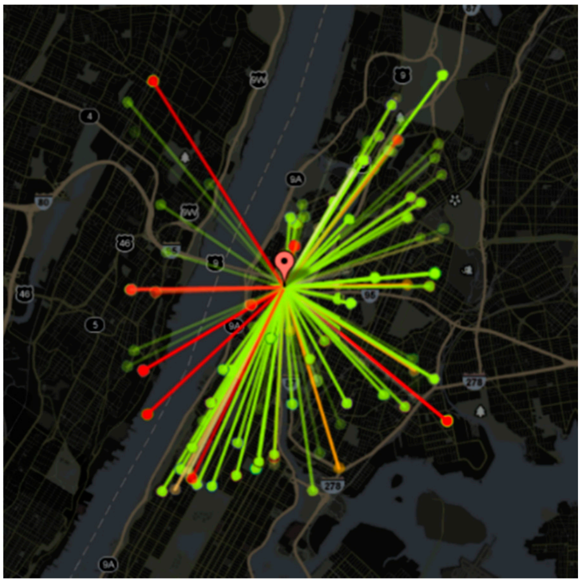


Figure 4. Social network of user *u* (center) overlaid on a map of New York City. Each line is a friend of user *u*; red lines are sick friends; green lines are healthy [33].

The methods used in this study infer "sick" versus "healthy" from brief messages, leaving room for misinterpretation. The visits to certain venues and interactions with sick individuals may be false positives. In addition, some illness may be overreported or underreported via social media. Thus, controlling for misrepresentations of the occurrence of illnesses must be improved through cross-referencing social media reports with other sources of data.

3.5. Public Lifestyle

Pennacchiotti and Popescu proposed a system for user classification in social media (Table 5) [34]. This team focused on classifying users according to three criteria: political affiliation (Democrat or Republican), race (African American or other, in this case), and potential as a customer for a particular business (Starbucks). Their machine learning framework relied on data from user profile accounts, user tweeting behavior (i.e., number of tweets per day, number of replies, etc.), linguistic content (main topics and lexical usage), and the social network of the user. The combination of all the features is more successful in classifying users than any individual feature. This framework was most successful in identifying the political affiliation of users. The features that were most accurate for this task were the social network and followers of the user, followed by the linguistic and profile features. The most difficult category was race, with values near 0.6–0.7. Linguistic features were most accurate for this task.

Prier et al. proposed the use of LDA for topic modeling Twitter data [14]. LDA was used to analyze terms and topics from a dataset of over two million tweets. The topic model identified a series of conversational topics related to public health, including physical activity, obesity, substance abuse, and healthcare. Unfortunately, the LDA method of analysis was unable to detect less common topics, such as the targeted topic of tobacco use. Instead, the researchers built their own query list by which to find tweets. The query list included terms such as "tobacco", "smoking", "cigarette", "cigar", and "hookah". By topic modeling this tobacco data subset, they were able to gain understanding of how Twitter users are discussing tobacco usage.

Table 4. Summary of social media mining for prediction applications as in Section 3.4.

Author	Data	Mining Technique	Sample Size	Location/ Language	Software	Merits	Limitations
Santos and Matos [32]	Influenzanet	SVM, Scikit-learn toolkit	~14 million tweets, ~15 million queries	Portugal (Portuguese)	NLP Toolkit, Scikit-learn Toolkit	Trained regression models applied from one season to the next.	The narrow scope of the Portuguese language. Correlation ratios could be improved.
Sadilek and Kautz [33]	Census data	SVM, regression decision trees, least-squares regression	~16 million tweets	New York City, New York, U.S. (English)	PageRank	Predicted future health of individuals with 91% accuracy.	Subject to false positives from both interpretation of messages as well as GPS data.

Table 5. Summary of social media mining for public lifestyle applications as in Section 3.5.

Author	Data	Mining Technique	Sample Size	Location/ Language	Software	Merits	Limitations
Pennacchiotti and Popescu [34]	N/A	Twitter API	15,000 users	United States (English)	Opinion Finder 1.5	>0.8 accuracy in identifying political affiliation.	Linguistic features used to identify ethnicity adopted by other ethnicities.
Prier et al. [14]	N/A	Twitter API	2,231,712 tweets for comprehensive, 1963 tweets for tobacco subset	United States: GA, ID, IN, KS, LA, MA, MS, OR, and PA. (English)	N/A	Extracts valuable topics from large datasets.	Twitter demographics may not be representative.

3.6. Geolocation

Dredze et al. introduced a system to determine the geographic location of tweets through the analysis of "Place" tags, GPS positions, and user profile data (Table 6) [35]. The purpose of the proposed system, called Carmen, was to assign a location to each tweet from a database of structured location information. "Place" tags on tweets associate a location with the message. These tags may include information such as the country, city, geographical coordinates, business name, or street address. Other tweets are GPS-tagged, and include the latitude and longitude coordinates of the location. The user profile contains a field where the user can announce their primary location. However, the profiles are subject to false information or nonsensical entries (i.e., "Candy Land"). User profiles are insufficient in accounting for travel as well. Carmen uses a combination of factors to infer the origin of the tweet. This system analyzes the language of the tweet, the "Place" and GPS tags, and the profile of the user. This information can provide the country, state, county, and city from which the tweet originated. Health officials may utilize Carmen's geolocation to track the occurrence of disease rates and prevent and manage outbreaks. Traditional systems rely on patient clinical visits, which take up to two weeks to publish. However, with this system, officials can use Twitter to find the possible areas of outbreaks in real time, improving reaction time.

Yepes et al. proposed a method for analyzing Twitter data for health-related surveillance [36]. To conduct their analysis, this group obtained 12 billion raw tweets from 2014. These tweets were filtered to include tweets only in the English language and excluded all retweets. Prior to filtering, heuristics were applied to the dataset. An in-domain medical named entity recognizer, called Micromed, was used to identify all of the relevant tweets. Micromed uses supervised learning, having been trained on 1300 manually annotated tweets. This system was able to recognize three medical entities: diseases, symptoms, and pharmacological substances. After filtering the tweets, MALLET (machine learning for language toolkit) was used to group the tweets by topic. An adapted geotagging system (LIW-meta) was used to determine geographic information from the posts. LIW-meta uses a combination of explicit location terms, implicit location-indicative words (LIW), and user profile data to infer geolocations from the tweets that lack GPS labels. The results of their work yielded geotagging with 0.938 precision. Yepes also observed that tweets mentioning terms such as "heart attack" are frequently used in the figurative sense more than in the medical sense when posting on social media. Other figurative usage of terms includes the use of "tired" to mean bored or impatient rather than drowsiness as a symptom. However, the usage of some pharmacological substance words, such as "marijuana" and "caffeine" are more likely to be indicative of the frequency of people using these substances.

Prieto et al. proposed an automated method for measuring the incidence of certain health conditions by obtaining Twitter data that was relevant to the presence of the conditions [37]. A two-step process was used to obtain the tweets. First, the data was defined and filtered according to specially crafted regular expressions. Secondly, the tweets were manually labeled as positive or negative for training classifiers to recognize the four health states. The health conditions that were studied were influenza, depression, pregnancy, and eating disorders. To begin the filtering, tweets originating in Portugal and Spain were selected using Twitter search API and geocoding information from the Twitter metadata. A language detection library was used to filter tweets that were not in Portuguese or Spanish. Once the tweets of the correct origin and language were identified, machine learning was applied to the data in order to filter out tweets that were not indicative of the person having the health condition. Finally, feature selection was applied to the data. Classification results of 0.7–0.9 in the area under the receiver operating characteristic (ROC) curve (AUC) and F-measure were obtained. The number of features was reduced by 90% by feature selection algorithms such as correlation-based feature selection (CFS), Pearson correlation, Gain Ration, and Relief. Classification results were improved with the feature selection algorithms by 18% in AUC and 7% in F-measure.

3.7. General

Tuarob et al. proposed a combination of five heterogeneous base classifiers to address the limitations of the traditional bag-of-words approach to discover health-related information in social media (Table 7) [38]. The five classifiers that were used were random forest, SVM, repeated incremental pruning to produce error reduction, Bernoulli Naïve Bayes, and multinomial Naïve Bayes. Over 5000 hand-labeled tweets were used to train the classifiers and cross-validate the models. A small-scale and a large-scale evaluation were performed to investigate the proposed model's abilities. The small-scale evaluation used a 10-fold cross-validation to tune the parameters of the proposed model and compare it with the state-of-the-art method. The proposed model outperformed the traditional method by 18.61%. The large-scale evaluation tested the trained classifiers on real-world data to verify the ability of the proposed model. This evaluation demonstrated a performance improvement of 46.62%.

Sriram developed a new method of classifying Twitter messages using a small set of authorship features that were included to improve the accuracy [39]. Tweets were classified into one of five categories focused on user intentions: news, events, opinions, deals, and private messages. The features extracted from the author's profile and the text were used to classify the tweets through three different classifiers. The fourth classifier, bag-of-words (BOW), was used to process tweets without the authorship features. It was considered a baseline because of its popularity in text classification. Compared to the BOW approach, each classifier that used the authorship features had significantly improved accuracy and processing time. The greatest number of misclassified tweets was found between News and Opinions categories.

Lee et al. proposed a method of classification of tweets based on Twitter Trending Topics [40]. Tweets were analyzed using text-based classification and network-based classification to fit into one of 18 categories such as sports, politics, technology, etc. For text-based classification, the BOW approach was implemented. In network-based classification, the top five similar topics for a given topic were identified through the number of common influential users. Each tweet could only be designated as falling into one category, which led to increased errors.

Parker et al. proposed a framework for tracking public health conditions and concerns via Twitter [41]. This framework uses frequent term sets from health-related tweets, which were filtered according to over 20,000 keywords or phrases, as search queries for open-source resources such as Wikipedia, Mahout, and Lucene. The retrieval of medical-related articles was considered an indicator of a health-related condition. The fluctuating frequent term sets were monitored over time to detect shifts in public health conditions and concern. This method was found to identify seasonal afflictions. However, no quantitative data was reported.

Table 6. Summary of social media mining for geolocation applications as in Section 3.6.

Author	Data	Mining Technique	Sample Size	Location/Language	Software	Merits	Limitations
Dredze et al. [35]	N/A	Yahoo's PlaceFinder API	N/A	N/A	Carmen, Java	Predicts location with up to 90% accuracy.	Geolocates each tweet individually. Room for improved efficiency.
Yepes et al. [36]	N/A	Gnip Decahose	28 million tweets	New York, London, Chicago (English)	Micromed, LIW-META	High precision in geotagging tweets.	Figurative usage of terms hinders ability to detect medical entities.
Prieto et al. [37]	N/A	Twitter API	10 million tweets	Portugal and Spain (Portuguese, Spanish)	Waikato Environment for Knowledge Analysis (WEKA)	Reduced features, improved classification.	Manual training for each new language.

Table 7. Summary of social media mining for general applications as in Section 3.7. ATAM: Ailment Topic Aspect Model.

Author	Data	Mining Technique	Sample Size	Location/ Language	Software	Merits	Limitations
Tuarob et al. [38]	Manually annotated	unknown	(1) 5000 tweets, (2) 700 million tweets, (3) 1,348,800 statuses	unknown	N/A	Improved performance in mitigating limitations and baseline.	High rate of false positives.
Sriram [39]	Mannual annotations	unknown	5407 tweets	(English)	Weka	Only Twitter data cosidered.	Outperforms Bag-of-Words approach in classification accuracy as well as time.
Lee et al. [40]	N/A	Twitter API	768 topics from What the Trend	(English)	Weka, SPSS modeler	Tweets could only be assigned to one group.	Highest accuracy (70.96%) found from network-based classifiers.
Parker et al. [41]	Mahout, Lucene, and Wikipedia	ATAM	1.6 million health-related tweets	(English)	N/A	Used on health-related tweets, not unfiltered tweets.	Clearly identifies seasonal afflictions.

4. Conclusions

Literature on data mining for public health surveillance were summarized, and their different uses for better public health were discussed. Twitter proves to be a vast source of information for public health. Twitter data is real-time, and is available from a large number of users across different geographic regions via API or Firehose. However, since tweets are very short and written for different purposes, various methods have to be employed to extract the desired information. Unfortunately, due to the lack of standardization, each article used different methods of analysis, making comparisons difficult to draw. Furthermore, the reliance on human intelligence for verifying the results of each method is cumbersome, and bears the potential for error. Many of these studies lack definitively successful results as well. In order to improve the potential for using Twitter and other social media platforms, there should be a standard method of determining the accuracy and effectiveness of the proposed models.

Overall, this review provides an in-depth overview of the techniques for utilizing Twitter for public health purposes. The number of expert research groups dedicating their resources and intelligence to this field is by itself a strong indicator of the potential of this field. While there remain a great many opportunities for improvement, the work described in this review sets a strong foundation for the development of machine-learning techniques in utilizing the untapped wealth of data in social media to improve different aspects of public health.

5. Future Directions

The text classification of Twitter data serves the purpose of information retrieval for public health surveillance. Nevertheless, most of the work done to date are similar to proprietary systems and are not applicable to other use cases. Thus, work toward an open-source algorithm or system that can be used across researchers and practitioners to categorize tweets in real-time is very important in order to better track information that is pertinent to outbreaks and facilitate real-time decision making.

Studies using data mining techniques with Twitter data are growing rapidly in public health research. However, some common challenges exist. First, most studies use non-probabilistic sampling to collect Twitter data. Many studies used the Twitter search API or streaming API, which has been demonstrated to be unreliable and incomplete [42]. Furthermore, retrieving relevant health tweets using a list of keywords might be problematic. Many tweets are relevant, but do not mention the predefined words, whereas many tweets including the keywords may be irrelevant. King et al. developed a new method for keyword searching to reduce this bias [43]. Many studies using geolocation to filter tweets can cause additional selection bias. Embedding geolocation information or not in tweets is not a random process. The geotagged tweets are a biased sample of the population. Liang et al. proposed using propensity score weighting to adjust the bias [44]. Nevertheless, Twitter users cannot represent all social media users or the general population of a society. Collecting data from multiple platforms can improve the representativeness.

Second, most studies focused on tweets, such as the topics and sentiment of the selected tweets. However, users are more important than tweets in public health research. According to Liang and Fu, most tweets are posted by a few extremely active users [45]. The occurrence of many tweets containing the word "sick" does not imply that there are many sick people. Future studies need to switch their unit of analysis from tweets to individual users. Additionally, generating a random sample of users is easier than generating a random sample of tweets [45].

Third, most studies focused on a single language (English) in a single area (U.S.). Twitter is a global platform with users from various cultures. Meanwhile, many public health problems are global in nature. Comparative studies dealing with multilingual texts will be helpful. For example, it is possible to analyze how Twitter users from different countries react differently to the same outbreak (e.g., MERS outbreak in Korea) [46].

Finally, in addition to the content data (texts), Twitter provides rich data on images, behaviors, and networks (following relationships). Even though text mining techniques are dominant in previous studies, future studies can benefit from other types of data. From example, the number of tweets might not be the best indicator of health concern. Instead, how many users searched or viewed the relevant information might be better indicators. The network data can also be used to study how information spreads through social networks [47].

Author Contributions: Writing and original draft preparation, S.E.J.; Review and editing, S.E.H., Conceptualization and Methodology, I.C.-H.F., H.L., K.-W.F. and Z.T.H.T.; Resources and Funding Acquisition, Z.T.H.T.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Friss, R.H.; Sellers, T.A. Epidemiology for public health practice. *Epidemiology* **1999**, *20*, 25.
2. Thacker, S.B.; Qualters, J.R.; Lee, L.M. Centers for Disease Control and Prevention, Public health surveillance in the United States: Evolution and challenges. *MMWR Surveill. Summ.* **2012**, *61*, 3–9.
3. Brownstein, J.S.; Freifeld, C.C.; Madoff, L.C. Digital disease detection—Harnessing the Web for public health surveillance. *N. Engl. J. Med.* **2009**, *360*, 2153–2157. [CrossRef]
4. Christian, K.A.; Iuliano, A.D.; Uyeki, T.M.; Mintz, E.D.; Nichol, S.T.; Rollin, P.; Staples, J.E.; Arthur, R.R. What We Are Watching—Top Global Infectious Disease Threats, 2013–2016: An Update from CDC's Global Disease Detection Operations Center. *Health Secur.* **2017**, *15*, 453–462. [CrossRef] [PubMed]
5. Biggerstaff, M.; Alper, D.; Dredze, M.; Fox, S.; Fung, I.C.H.; Hickmann, K.S.; Lewis, B.; Rosenfeld, R.; Shaman, J.; Tsou, M.-H. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infect. Dis.* **2016**, *16*, 357. [CrossRef] [PubMed]
6. Lewis, S.H.; Burkom, H.S.; Babin, S.; Blazes, D.L. Promising advances in surveillance technology for global health security. In *Disease Surveillance: Technological Contributions to Global Health Security*; CRC Press: Boca Raton, FL, USA, 2016; p. 179.
7. Griffis, H.M.; Kilaru, A.S.; Werner, R.M.; Asch, D.A.; Hershey, J.C.; Hill, S.; Ha, Y.P.; Sellers, A.; Mahoney, K.; Merchant, R.M. Use of social media across US hospitals: Descriptive analysis of adoption and utilization. *J. Med. Int. Res.* **2014**, *16*, 264. [CrossRef] [PubMed]
8. Kim, S.J.; Marsch, L.A.; Hancock, J.T.; Das, A.K. Scaling up Research on Drug Abuse and Addiction Through Social Media Big Data. *J. Med. Int. Res.* **2017**, *19*, 353. [CrossRef]
9. Martinez-Millana, A.; Fernandez-Llatas, C.; Bilbao, I.B.; Salcedo, M.T.; Salcedo, V.T. Evaluating the Social Media Performance of Hospitals in Spain: A Longitudinal and Comparative Study. *J. Med. Int. Res.* **2017**, *19*, 181. [CrossRef]
10. Google. Available online: https://scholar.google.com/ (accessed on 28 November 2018).
11. IEEE. Available online: http://ieeexplore.ieee.org/Xplore/home.jsp (accessed on 28 November 2018).
12. Direct, S. Available online: http://www.sciencedirect.com/ (accessed on 28 November 2018).
13. Paul, M.J.; Dredze, M. A model for mining public health topics from Twitter. *Health* **2012**, *11*, 16.
14. Prier, K.W.; Smith, M.S.; Giraud-Carrier, C.; Hanson, C.L. Identifying health-related topics on twitter. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, College Park, MD, USA, 30–31 March 2011; pp. 18–25.
15. Gesualdo, F.; Stilo, G.; Gonfiantini, M.V.; Pandolfi, E.; Velardi, P.; Tozzi, A.E. Influenza-like illness surveillance on Twitter through automated learning of naïve language. *PLoS ONE* **2013**, *8*, e82489. [CrossRef]
16. Coppersmith, G.; Dredze, M.; Harman, C. Quantifying mental health signals in twitter. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Baltimore, MA, USA, 27 June 2014; pp. 51–60.
17. Denecke, K.; Krieck, M.; Otrusina, L.; Smrz, P.; Dolog, P.; Nejdl, W.; Velasco, E. How to exploit twitter for public health monitoring. *Methods Inf. Med.* **2013**, *52*, 326–339. [PubMed]
18. Adrover, C.; Bodnar, T.; Huang, Z.; Telenti, A.; Salathé, M. Identifying adverse effects of HIV drug treatment and associated sentiments using twitter. *JMIR Public Health Surveill.* **2015**, *1*, 7. [CrossRef] [PubMed]

19. Ginn, R.; Pimpalkhute, P.; Nikfarjam, A.; Patki, A.; O'Connor, K.; Sarker, A.; Gonzalez, G. Mining Twitter for adverse drug reaction mentions: A corpus and classification benchmark. In Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, Reykjavík, Iceland, 31 May 2014.
20. Sarker, A.; Gonzalez, G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Inform.* **2015**, *53*, 196–207. [CrossRef] [PubMed]
21. Behera, P.N.; Eluri, S. Analysis of Public Health Concerns using Two-step Sentiment Classification. *Int. J. Eng. Res. Technol.* **2015**, *4*, 606–610.
22. Signorini, A.; Segre, A.M.; Polgreen, P.M. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS ONE* **2011**, *6*, e19467. [CrossRef] [PubMed]
23. Myslín, M.; Zhu, S.-H.; Chapman, W.; Conway, M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *J. Med. Int. Res.* **2013**, *15*, 174. [CrossRef] [PubMed]
24. Ji, X.S.; Chun, A.; Wei, Z.; Geller, J. Twitter sentiment classification for measuring public health concerns. *Soc. Netw. Anal. Min.* **2015**, *5*, 13. [CrossRef]
25. Coletta, L.F.S.; da Silva, N.F.F.; Hruschka, E.R.; Hruschka, E.R. Combining classification and clustering for tweet sentiment analysis. In Proceedings of the 2014 Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, Brazil, 18–22 October 2014; pp. 210–215.
26. Go, A.; Bhayani, R.; Huang, L. Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report Stanford. 2009, Volume 1. Available online: https://bit.ly/2Aj2rXi (accessed on 28 December 2018).
27. Cheong, F.; Cheong, C. Social Media Data Mining: A Social Network Analysis of Tweets during the 2010–2011 Australian Floods. In Proceedings of the 2011–15th Pacific Asia Conference on Information Systems: Quality Research in Pacific, Brisbane, Australia, 7–11 July 2011; p. 46.
28. Odlum, M.; Yoon, S. What can we learn about the Ebola outbreak from tweets? *Am. J. Infect. Control* **2015**, *43*, 563–571. [CrossRef]
29. Missier, P.; Romanovsky, A.; Miu, T.; Pal, A.; Daniilakis, M.; Garcia, A.; da Silva Sousa, L. Tracking dengue epidemics using twitter content classification and topic modelling. In Proceedings of the 16th International Conference on Web Engineering, Lugano, Switzerland, 6–9 June 2016; pp. 80–92.
30. Schulz, A.; Mencía, E.L.; Dang, T.T.; Schmidt, B. Evaluating multi-label classification of incident-related tweets. In Proceedings of the Making Sense of Microposts (# Microposts 2014), Seoul, Korea, 7–11 April 2014; p. 7.
31. Gomide, J.; Veloso, A.; Meira, W., Jr.; Almeida, V.; Benevenuto, F.; Ferraz, F.; Teixeira, M. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In Proceedings of the 3rd International Web Science Conference, Koblenz, Germany, 15–17 June 2011; p. 3.
32. Santos, J.C.; Matos, S. Analysing Twitter and web queries for flu trend prediction. *Theor. Biol. Med. Model.* **2014**, *11*, S6. [CrossRef]
33. Sadilek, A.; Kautz, H. Modeling the impact of lifestyle on health at scale. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, Rome, Italy, 4–8 February 2013; pp. 637–646.
34. Pennacchiotti, M.; Popescu, A.-M. A Machine Learning Approach to Twitter User Classification. In Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, 17–21 July 2011; pp. 281–288.
35. Dredze, M.; Paul, M.J.; Bergsma, S.; Tran, H. Carmen: A twitter geolocation system with applications to public health. In Proceedings of the AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI), Bellevue, DC, USA, 14–18 July 2013; pp. 20–24.
36. Yepes, A.J.; MacKinlay, A.; Han, B. Investigating public health surveillance using Twitter. *ACL-IJCNLP* **2015**, *2015*, 164.
37. Prieto, V.M.; Matos, S.; Alvarez, M.; Cacheda, F.; Oliveira, J.L. Twitter: A good place to detect health conditions. *PLoS ONE* **2014**, *9*, e86191. [CrossRef] [PubMed]
38. Tuarob, S.; Tucker, C.S.; Salathe, M.; Ram, N. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *J. Biomed. Inform.* **2014**, *49*, 255–268. [CrossRef] [PubMed]
39. Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; Demirbas, M. Short text classification in twitter to improve information filtering. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, 19–23 July 2010; pp. 841–842.

40. Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M.M.A.; Agrawal, A.; Choudhary, A. Twitter trending topic classification. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), Vancouver, BC, Canada, 11 December 2011; pp. 251–258.
41. Parker, J.; Wei, Y.; Yates, A.; Frieder, O.; Goharian, N. A framework for detecting public health trends with twitter. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara, ON, Canada, 25–29 August 2013; pp. 556–563.
42. González-Bailón, S.; Wang, N.; Rivero, A.; Borge-Holthoefer, J.; Moreno, Y. Assessing the bias in samples of large online networks. *Soc. Netw.* **2014**, *38*, 16–27. [CrossRef]
43. King, G.; Lam, P.; Roberts, M.E. Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *Am. J. Political Sci.* **2017**, *61*, 971–988. [CrossRef]
44. Liang, H.; Shen, F.; Fu, K.-W. Privacy protection and self-disclosure across societies: A study of global Twitter users. *New Media Soc.* **2017**, *19*, 1476–1497. [CrossRef]
45. Liang, H.; Fu, K.-W. Testing propositions derived from Twitter studies: Generalization and replication in computational social science. *PLoS ONE* **2015**, *10*, e0134270. [CrossRef] [PubMed]
46. Fung, I.C.H.; Zeng, J.; Chan, C.H.; Liang, H.; Yin, J.; Liu, Z.; Fu, K.W. Twitter and Middle East respiratory syndrome, South Korea, 2015: A multi-lingual study. *Infect. Dis. Health* **2018**, *23*, 10–16. [CrossRef]
47. Liang, H. How Did Ebola Information Spread on Twitter? In Proceedings of the National Conference on Health Communication, Marketing, and Media (NCHCMM), Atlanta, GA, USA, 23–25 August 2016.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).