Journal of Biomedical Informatics 62 (2016) 1–11

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Methodological Review

Using online social networks to track a pandemic: A systematic review

Mohammed Ali Al-garadi [a,*], Muhammad Sadiq Khan [a], Kasturi Dewi Varathan [a], Ghulam Mujtaba [a], Abdelkodose M. Al-Kabsi [b]

[a] Department of Information System, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur, Malaysia
[b] Medical Microbiology Cyberjaya University College of Medical Sciences (CUCMS), Cyberjaya, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 13 January 2016
Revised 21 April 2016
Accepted 14 May 2016
Available online 17 May 2016

Keywords:
Infectious disease surveillance
Online social network
Machine learning
Systematic review

ABSTRACT

Background: The popularity and proliferation of online social networks (OSNs) have created massive social interaction among users that generate an extensive amount of data. An OSN offers a unique opportunity for studying and understanding social interaction and communication among far larger populations now more than ever before. Recently, OSNs have received considerable attention as a possible tool to track a pandemic because they can provide an almost real-time surveillance system at a less costly rate than traditional surveillance systems.

Methods: A systematic literature search for studies with the primary aim of using OSN to detect and track a pandemic was conducted. We conducted an electronic literature search for eligible English articles published between 2004 and 2015 using PUBMED, IEEExplore, ACM Digital Library, Google Scholar, and Web of Science. First, the articles were screened on the basis of titles and abstracts. Second, the full texts were reviewed. All included studies were subjected to quality assessment.

Result: OSNs have rich information that can be utilized to develop an almost real-time pandemic surveillance system. The outcomes of OSN surveillance systems have demonstrated high correlations with the findings of official surveillance systems. However, the limitation in using OSN to track pandemic is in collecting representative data with sufficient population coverage. This challenge is related to the characteristics of OSN data. The data are dynamic, large-sized, and unstructured, thus requiring advanced algorithms and computational linguistics.

Conclusions: OSN data contain significant information that can be used to track a pandemic. Different from traditional surveys and clinical reports, in which the data collection process is time consuming at costly rates, OSN data can be collected almost in real time at a cheaper cost. Additionally, the geographical and temporal information can provide exploratory analysis of spatiotemporal dynamics of infectious disease spread. However, on one hand, an OSN-based surveillance system requires comprehensive adoption, enhanced geographical identification system, and advanced algorithms and computational linguistics to eliminate its limitations and challenges. On the other hand, OSN is probably to never replace traditional surveillance, but it can offer complementary data that can work best when integrated with traditional data.

1. Introduction

Infodemiology (i.e., information epidemiology) refers to "the set of methods, which study the data specifically, health data on the internet for the purpose of public health studies and policies [1,2]." Infoveillance (i.e., information surveillance) can be defined as a syndromic surveillance that analyzes online data to detect disease outbreaks at a shorter time than traditional surveillance [3,4]. Social media have been used in various health applications [5]. Many researchers have built prediction models of disease spread outbreak using health-related data extracted from the Internet. The utilization of Google Search queries has introduced a Web-based tool for real-time surveillance of disease outbreaks [6–8]. Several studies have used search engine search queries to develop a prediction model of disease spread outbreak [9–11]. Study [12] investigates the potential use of HealthMap to query, filter, integrate, and visualize unstructured reports on disease outbreaks from sources such as Google News and ProMED Mail. HealthMap is found to be a suitable tool that uses text-processing algorithms to identify significant disease outbreak information through a user-friendly interface. Another research [13] was inspired by efforts in

* Corresponding author.
E-mail address: mohammedali@siswa.um.edu.my (M.A. Al-garadi).

http://dx.doi.org/10.1016/j.jbi.2016.05.005
1532-0464/

disease surveillance using social-media-developed belief surveillance. This type of surveillance is used in healthcare to analyze the belief level of a user about the spread of health information in social media. A comparative study between data derived from the Internet and those derived from a formal clinical study confirmed the potential of mining unstructured, text-based, and online forum data for supplementing and validating structured quantitative data collected from clinical studies [4]. The time delay caused by necessary data processing is the main drawback of these methodologies, but the information available on the Internet can provide additional means to address this problem [14]. Similarly, online social networks (OSNs) have been used for real-time surveillance of disease outbreaks. An additional advantage of OSN data is that the data are posted by individuals, as individual users of OSNs are free to post information about their health conditions. OSNs are described as a new type of laboratory and tool for peer-to-peer communication and interaction in the health sector [15]. They can serve as a source of human data to study the health condition of users.

OSN communication is a revolutionary trend that utilizes Web 2.0, which introduces a new feature that enables users to become active. Users can freely express what they feel and share their health condition. By contrast, users only passively read the content in websites based on Web 1.0. The popularity and the proliferation of OSNs have created an extensive social interaction among users and generated a large amount of social data. They offer a unique opportunity to study and understand social interaction and communication among far larger populations now more than before [16]. OSNs have received considerable attention as a possible tool for tracking a pandemic. The increasing attention on using OSN as a surveillance system to track a pandemic is due to the real-time user-generated data provided by social media. OSNs are a perfect source for early-stage pandemic detection because of their real-time nature. This characteristic also enables for fast communication between health agencies and the public in the early stage of pandemic outbreak detection.

Traditional methods are used to estimate the number of people affected by specified pandemic diseases. These methods are based on data collected from phone-based surveys [17] or from health agencies and centers, such as the United States Centers for Disease Control and Prevention (US CDC), the European Influenza Surveillance Scheme (EISS), and Japan's Infection Disease Surveillance Center (IDSC). These centers mainly rely on both virology and clinical data. Visualization and analytical tools for infectious disease epidemiology are discussed in [18]. Traditional pandemic surveillance such as influenza pandemic is entirely manual, and thus it causes one to two weeks of time lag between the time of medical diagnosis and the time when the data become available [19]. OSNs have the potential to eliminate the time lag in traditional surveillance by enabling the extraction of millions of real-time text data, which include geographical location and information regarding one's personal well-being. However, the accuracy of OSN surveillance systems depends on the quality of the algorithms used to distinguish between pandemic-related data and other social media communication data. OSN-based surveillance systems are limited to public data only. The privacy of users should be considered when extracting data from OSNs.

An analysis on the current OSN pandemic surveillance systems is required to understand further their accuracy, challenges and limitations, and efficiency as future surveillance systems. Therefore, we conducted a systematic review of current empirical studies that used OSNs for tracking a pandemic.

This systematic review aims to conduct a critical appraisal of pandemic surveillance based on the data extracted from OSNs. By doing so, we aim to investigate the adequacy of this surveillance as a tool in tracking a pandemic and in identifying the extent of using OSNs to track a pandemic, including their challenges and limitations. The rest of the paper is organized as follows. Section 2 represents a systematic review methodology, including a discussion of the types of studies, search strategy for the identification of studies, screening and selection of criteria, and quality assessment of the included studies. Section 3 explains the result of the paper (search results) and describes all included studies. This section also summarizes the studies listed in Table 2 and analyzes the classifier performance and features used by different studies, as shown in Table 3. Section 4 presents a discussion of the main findings and challenges. Section 5 discusses the limitations of this systematic review. Section 6 reviews future implications. And Section 7 concludes the paper.

2. Methods

This study aims to investigate the adequacy and limitations of pandemic surveillances based on OSN data.

2.1. Types of studies

Most studies that used social network data as their source data for tracking a pandemic are included. OSN can be described as a structure that enables information exchange and dissemination, as well as social interaction among individuals. Therefore, OSN data are the data generated by interactions and online communication among individuals in the network [16]. A pandemic refers to an epidemic that has spread over a wide geographical region and has affected a large number of the population [20].

2.2. Search strategy for the identification of studies

We conducted an electronic literature search for eligible articles using PUBMED, IEEExplore, ACM Digital Library, Google Scholar, and Web of Science. The search keywords were as follows: "online social network," "microblog," "Facebook," "Twitter," "Myspace," "YouTube," "LinkedIn," "Google+," "Friendster," "social media," "social website," "flu," "pandemic," "epidemic," "infectious disease," "seasonal flu," "H1N1," "HIV," "influenza," "Influenza-like Illnesses," "Ebola," and "Zika." Additional search keywords were derived from the synonyms of the aforementioned words. All English articles were examined regardless of the language of the analyzed data. No country was restricted. The references of all articles selected for full-text evaluation were reviewed for potential eligible studies.

2.3. Screening and selection of criteria

In the first stage of the screening, the total number of articles collected was 4105. Subsequently, duplicate articles were

Table 1
Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
English Articles (AND) Articles published between 2004 and 2015 (AND) Articles that used OSN data (user-generated data) (AND) Articles that studied an epidemic that had spread over a wide geographical region and affected a large number of population (AND) Articles that analyzed user-generated data while preserving users' privacy	Studies not primarily aimed to use OSN data (OR) Studies not primarily aimed to track a pandemic (OR) Studies that do not fulfill any inclusion criterion

Table 2
Summary of included studies.

Author	Data source and volume, year and duration of data collection, and geographical location	Data analysis method	Study aims and outcomes
Chew and Eysenbach [35]	**Data source and volume:** Data were collected from Twitter. Authors collected two million tweets related to "swine flu" and "H1N1" **Year and duration:** Between May 1 and December 31, 2009 **Geographical location of data:** Not specified	**Data analysis method:** Manual classification and preliminary automated analyses	**Aim:** To extract the tweets to monitor and analyze the use of the terms "H1N1" versus "swine flu" over time and to validate whether Twitter can be used as a pandemic tracking tool **Outcomes:** Twitter can be used for almost real-time content and sentiment analyses. Using Twitter content, health authorities can be aware of real or perceived concerns raised by the public
Broniatowski et al. [24]	**Data source and volume:** Data were collected from Twitter. One billion tweets were general, and 0.3 billion tweets were health related **Year and duration:** Between September 30, 2012 and May 31, 2013 **Geographical location of data:** This experiment focused on multiple geographic, national, and regional levels. They considered the United States and New York City	**Data analysis method:** The authors used four filters. The first one was the health filter that used the support vector machine (SVM) to classify the tweets into relevant or irrelevant to health. The second and third ones were the influenza and infection filters, respectively, used to classify tweets into whether they indicated influenza or not. If yes, they were further classified if they pertained to influenza infection or awareness of influenza only. These two filters used separate logistic regression models. The fourth one was the location filter used as the geographic location filter	**Aim:** To develop an influenza surveillance system based on data extracted from tweets and to analyze its performance at multiple levels of geographic granularity **Outcomes:** The developed influenza surveillance system was correlated with the surveillance data from the CDC ($r = 0.93$, $p < 0.001$) and the surveillance data from the Department of Health and Mental Hygiene of New York City ($r = 0.88$, $p < 0.001$). The result implies the effectiveness of using Twitter as an influenza surveillance system
Signorini et al. [10]	**Data source and volume:** The data of this study contained two data sets collected from Twitter. The first data set consisted of 951,697 tweets, containing pandemic-related keywords. The second data set consisted of 4,199,166 tweets selected from roughly eight million influenza-related tweets **Year and duration:** The first and second data sets were collected between April 29 and June 1 and between October 1, 2009 and end of December 2012, respectively **Geographical location of data:** Only within the United States	**Data analysis method:** SVM was used as a classifier	**Aim:** To use tweets related to H1N1 or swine flu for building a model for tracking the levels of disease activity and to prove that Twitter can be a faster tool for estimating health-related events than the traditional tools **Outcomes:** The study indicates that Twitter data can be used not only to track users' interest and concerns related to H1N1 influenza but also to estimate disease activity in real time (i.e., one to two weeks faster than the current practice allows)
Kim et al. [25]	**Data source and volume:** Data were collected from Twitter. The authors collected 287 million Korean tweets **Year and duration:** Data were collected for a 51-week period from October 2011 to September 2012 **Geographical location of data:** Not specified, but the study was limited to Korean tweets	**Data analysis method:** The linear least squares regression algorithm was adopted for the daily estimation of influenza spreading score using the influenza-like illness (ILI) daily reports from KCDC for influenza and the generated daily marker frequency matrix from Twitter data	**Aim:** To detect rapidly evolving public influenza transmission awareness and develop a regression model that can track actual disease activity level and predict the ILI-activity level in a population using a delay mode **Outcomes:** The study demonstrated that Twitter data could be used for the real-time prediction of influenza infection
Lampos and Cristianini [26]	**Data source and volume:** Data were collected daily from Twitter, with a daily average of 160,000 tweets over a 24-week period **Year and duration:** Between June 12 and December 6, 2009 **Geographical location of data:** United Kingdom	**Data analysis method:** This study analyzed symptom-related data from a social network (tweets) and converted the statistical information into a flu score. The flu score was compared with data from UK HPA. LASSO was used to rank or determine the weights for the candidate features	**Aim:** To track the spread of epidemic diseases, such as seasonal or pandemic influenza **Outcomes:** This study analyzed symptom-related data from a social network (tweets) and converted the statistical information into a flu score. The flu score was compared with data from the UK HPA. The statistically significant linear correlation was greater than 95%
Aramaki et al. [23]	**Data source and volume:** Data were collected from Twitter. The authors collected 300 million tweets **Year and duration:** Between November 2008 and June 2010 **Geographical location of data:** Japan	**Data analysis method:** First, all tweets were extracted using influenza-related keywords. Second, SVM was used to classify positive and negative tweets. Positive tweets refer to confirmed influenza-related tweets, whereas negative tweets refer to tweets not related to influenza, although they contain influenza-related keywords, such as news or general medical advice	**Aim:** To detect an influenza epidemic using Twitter data in Japan and to investigate the potential of using tweets to detect an influenza epidemic compared with the Google flu system and the traditional influenza system **Outcomes:** The experimental correlation with the gold standard of IDSC was 0.89. In the early stage (i.e., early epidemic stage), the proposed method indicated a high correlation (0.97)

(*continued on next page*)

Table 2 (*continued*)

Author	Data source and volume, year and duration of data collection, and geographical location	Data analysis method	Study aims and outcomes
Culotta [27]	**Data source and volume:** Data were collected from Twitter. Authors collected 574,643 tweets **Year and duration:** 10 weeks between February 12 and April 24, 2010 **Geographical location of data:** Not specified	**Data analysis method:** A classification technique was used to remove all unrelated messages from data	**Aim:** To detect an influenza epidemic by analyzing Twitter data **Outcomes:** Several methods were proposed to identify influenza-related messages and a number of regression models were compared to correlate these messages with CDC statistics. The finding indicates that the classification-hand system provided the best model to achieve a correlation with CDC statistics of 0.78
Achrekar et al. [28]	**Data source and volume:** Data were collected from Twitter, i.e., 4.7 million tweets from 1.5 million unique users **Year and duration:** Between 2009 and 2010 **Geographical location of data:** United States	**Data analysis method:** Machine learning (i.e., SVM) was used to classify the tweets into related and non-related	**Aim:** To monitor the streaming flu-related tweets from Twitter to predict the spread of an influenza epidemic **Outcomes:** Flu prediction using tweets demonstrated a high correlation with the data provided by CDC for ILI. Auto-regression models were developed to predict ILI-activity level in a population using data from CDC. The developed model was tested with and without the measure of Twitter data. The data extracted from Twitter could improve the overall accuracy of the model
Lamb et al. [29]	**Data source and volume:** This study used data containing two billion tweets from previous studies [42]. The second data set collected had 1.8 tweets **Year and duration:** The first and second data sets were from May 2009 and October 2010 and from August 2011 to November 2012, respectively **Geographical location of data:** Not specified	**Data analysis method:** This study used binary classification	**Aim:** To track flu infections on Twitter **Outcomes:** This study discussed how the differences among the categories of flu tweets could provide significant improvements for influenza surveillance. Similarly, analyzing the information of a tweet's author can significantly improve flu prediction
Li and Cardi [30]	**Data source and volume:** Data were collected from Twitter. The authors collected 3.6 million flu-related tweets from 0.9 million Twitter users **Year and duration:** Between June 2008 and June 2010 **Geographical location of data:** Tweet locations were selected within the United States	**Data analysis method:** The detection of flu was based on the detected transition time from the non-epidemic phase to the epidemic phase. Spatio-temporal analysis using unsupervised Bayesian algorithm based on a four-phase Markov Network identified the flu breakout by detecting the transition time (i.e., non-epidemic phase, rising epidemic phase, declining epidemic phase, and stationary-epidemic phase). Supervised SVM classifier was trained to identify flu-related data from the extracted data	**Aim:** To detect flu in the early stage by detecting the current phase of the flu and predicting future phases **Outcomes:** A real-time flu detection system was able to detect a flu breakout in the early stage and predict future epidemic phases. The study introduced an unsupervised Bayesian model based on a Markov Network, which was effectively used for early-stage flu detection using tweets
Diaz-Aviles et al. [31]	**Data source and volume**: Data were collected from Twitter. The authors collected 456,226 tweets related to enterohemorrhagic *Escherichia coli* (EHEC) **Year and duration:** Between May and June 2011 **Geographical location of data:** Germany	**Data analysis method:** A personalized ranking approach was used to discover the relationship and ranks based on the computation of latent semantic topics using LDA and observations of hash-tagging behavior on Twitter	**Aim:** To prove how tracking Twitter data can be used to detect a pandemic EHEC outbreak before the traditional medical system surveillance or other early alarm systems. In contrast to previous studies, this study aimed to detect the sudden outbreak of a disease without a seasonal pattern **Outcomes:** This study introduced the tracking of a pandemic on Twitter using two stages. The first and second stages were early outbreak detection and outbreak analysis and control, respectively
Gomide et al. [32]	**Data source and volume:** The data consisted of two sets. The first set contained tweets from previous studies, and the second set was collected by the author. The first data set contained 27,658 tweets related to dengue, and the second data set contained 465, 444 posted tweets **Year and duration:** The first data set was collected from January 2009 to May 2009. The second data set was collected during the dengue period in 2011 **Geographical location of data:** Brazil, but only the first data set was geographically analyzed. A total of 332 cities in Brazil were selected for the spatio-temporal analysis	**Data analysis method:** This study adopted three steps of analysis. First, the content analysis categorized the tweets into five types, and the association rule was used to map each tweet to its related categories. Second, the cases reported by analyzing Twitter were correlated with cases from the Health Ministry of Brazil using linear regression. Third, the spatio-temporal analysis investigated the tweet locations and time to predict the spread of dengue timely and geographically	**Aim:** To detect a dengue epidemic using the content and the spatio-temporal features of tweets **Outcomes:** This study demonstrated how Twitter could be used for dengue surveillance. Tweet information, such as content, time, and location, were efficiently utilized to predict the spread of the pandemic geographically and timely

Sadilek et al. [33]	**Data source and volume**: 2.5 million geo-tagged Twitter messages **Year and duration:** One month long beginning May 18, 2010 **Geographical location of data:** New York City	**Data analysis method:** The SVM model was developed and used to classify tweets into "sick" or "other." Over 700,000 "sick" messages and three million "other" tweets were used for training the SVM. The final SVM was evaluated using 700,000 tweets. The classifier result indicated 0.98 precision and 0.97 recall	**Aim:** To model the spread of infectious diseases by analyzing tweets and social relationships among Twitter users **Outcomes:** Modeling the spread of disease was achieved by self-reported symptoms using Twitter. The SVM model, which was developed to classify tweets into "sick" or "others," provided a powerful result despite the presence of imbalanced data
Huang et al. [36]	**Data source and volume**: More than 35.3 million tweets from Sina Weibo **Year and duration:** Between August 2009 and April 2012 **Geographical location of data:** China	**Data analysis method:** First, the extracted data were filtered with keywords according to the definition of ILI. The correlation of extracted data with China CDC was conducted using cosine similarity and the standard deviation metrics. The outbreak transmission of flu between cities was detected using the dynamic Bayesian network with the aid of data location and social ties	**Aim**: To provide a strategy using social network data to detect the transmission of a contagion **Outcomes:** The method developed in this study illustrated a high correlation with China CDC; it used the relationship among friends to study the spread of flu. However, the study used keyword filtering, which led to less classifying accuracy of user data. Keyword filtering was unable to distinguish whether the user's post pertains to flu awareness in general or refers to an infectious case
Szomszor et al. [38]	**Data source and volume**: Data were collected from Twitter. Three million tweets were collected **Year and duration:** Between May and December 2009 **Geographical location of data:** United Kingdom	**Data analysis method:** The extracted tweets were first classified into three categories, namely, tweets containing a link, retweets, and self-reporting flu. The self-reporting tweets were correlated with the data from UK HPA	**Aim**: To develop an early warning system to track swine flu and predict disease outbreaks **Outcomes:** The correlation between Twitter data with UK HPA data indicated that the former could be used as an early warning system because Twitter could predict the latter one week in advance
Santos and Matos [34]	**Data source and volume**: Data were collected from Twitter. The authors collected approximately 14 million tweets **Year and duration:** Between March 2011 and February 2012 **Geographical location of data:** Portugal	**Data analysis method:** Naïve Bayes classifier was used to classify tweets to distinguish flu-related ones from other unrelated tweets. The correlation between the classification result and the Influenzanet project was conducted using the multiple regression model	**Aim**: To evaluate whether Twitter can be used for predicting flu spread in Portugal **Outcomes:** This study, which was conducted in Portuguese, demonstrated that tracking a pandemic system using Twitter could be also adapted by other languages. The limitation of this study is in the small number of available tweets caused by data limitation. An over-fitting problem was reported
Young et al. [41]	**Data source and volume**: Data were collected from Twitter. Nearly 14 million tweets were collected	**Data analysis method:** About 553,186,061 tweets were collected and then filtered to include only tweets containing keywords that were HIV-risk related and originated from the United States. The final data set contained 9880 tweets. The tweets were classified into drug risk-related and sex risk-related tweets. Subsequently, both sets of tweets were combined to create an overall category of HIV-related tweets	**Aim**: To detect and remotely monitor HIV outcome using social network data (i.e., Twitter)
	Year and duration: Between May 26 and December 9, 2012 **Geographical location of data:** United States	Univariate regression was used to study the association between the HIV-related tweets (drug-related or sexually related or both) and HIV cases	**Outcomes:** The result of the study indicated a significant positive relationship between the number of HIV-related tweets and HIV cases. The finding supported the use of social network data to detect and remotely monitor HIV outcomes
Wegrzyn-Wolska et al. [40]	**Data source and volume**: Data were collected from Twitter.	**Data analysis method:** First, tweets containing keywords related to chickenpox infection were extracted from Twitter. Second, all tweets were papered to categorize them and include only tweets that directly indicate chickenpox infection to chickenpox. Third, the tweets were correlated with the official data by the French GPs Sentinelles network	**Aim**: To develop a prediction model using Twitter data for epidemic spread (i.e., chickenpox) in France
	Year and duration: During 2013 **Geographical location of data:** France		**Outcomes:** The developed system presented a good correlation with the official data by the French GPs Sentinelles network. However, the developed system did not involve a large distribution analysis of tweets that could have affected the final result. The study also did not examine the correlation of the proposed system with other traditional pandemic surveillance systems to ensure the accuracy of the system

(*continued on next page*)

Table 2 (*continued*)

Author	Data source and volume, year and duration of data collection, and geographical location	Data analysis method	Study aims and outcomes
de Quincey [39]	**Data source and volume:** A total of 130,233 tweets were collected from Twitter **Year and duration:** From June 2012 to April 2013 **Geographical location of data:** United Kingdom	**Data analysis method:** First, all tweets related to hay fever were collected. Second, textual and geographical analyses of tweets were performed to detect and visualize the geographical distribution of tweets	**Aim**: To investigate the feasibility of OSN data in detecting a hay fever pandemic in the United Kingdom based on analyzing tweet content and location **Outcomes:** This study showed that Twitter could be used for hay fever detection. However, issues related to noise in data and sampled population were found
Odlum and Yoon [37]	**Data source and volume:** Data were collected from Twitter. The authors collected 42,236 tweets (16,499 unique and 25,737 retweet) containing the word "Ebola" **Year and duration:** Between July 24, and August 1, 2014 **Geographical location of data:** Tweets that included the geographic location with latitude and longitude codes, but were not limited to a specific country. The author then chose Nigeria as the case study	**Data analysis method:** This study used unigram, bigram, and trigram of tweet content with *K*-means algorithm to cluster (unsupervised learning) the tweets based similarities of content to identify the topic	**Aim:** To detect the outbreak of Ebola virus in the early stage by tracking tweets related to Ebola **Outcomes:** Analysis of the tweets captured the early stage of Ebola outbreak. Ebola-related tweets in Nigeria were chosen three to seven days before the official declaration of the first probable Ebola case. Additional conclusions show the effectiveness of Twitter in spreading health alerts

removed, and the total number of articles decreased to 3704. The articles were analyzed individually by two reviewers. Based on the title, only studies that were clearly relevant to the aims of our systematic review were included. The qualified articles ($n = 697$) were then divided among three reviewers for screening. Papers from the first stage were forwarded to the second stage if they had met the following criteria:

- The paper must be published in English between 2004 and 2015.
- The data source of the study should be extracted from any OSN.
- The paper should study an epidemic that had spread over a wide geographical region and had affected a large number of the population.
- Data extracted from an OSN should preserve the privacy of the users; therefore, only authorized public data should be used in the study.

Articles forwarded to the second stage of screening ($n = 52$) were assessed by four reviewers based on the inclusion and exclusion criteria (Table 1). The four reviewers then compared and discussed until a consensus was reached. Finally, 20 papers were included for detailed analyses.

2.4. Quality assessment

The methodological quality of included papers was assessed using the Critical Appraisal Skills Program (CASP) systematic review checklist [21,22]. This critical appraisal was used because of its comprehensiveness in determining methodological quality. In assessing the results, the focus of each study, the quality of the methodology and the methods used should be considered because they could influence the results. A comparison among studies was conducted to describe the main characteristics, strengths, weaknesses, and limitations of each study based on data extraction (e.g., means of data extraction, volume of extracted data, and geographical area of the data), data quality (e.g., what were the techniques used to classify and whether the extracted data was related to a pandemic or not), study design (e.g., whether an appropriate methodology was clearly implemented), and results (e.g., how the result was presented and whether the final result was a reflection of the analysis).

3. Results

3.1. Search results

A total of 4105 papers were collected during the search process. The majority of the studies identified by this search strategy were excluded because of duplicates ($n = 401$). After the duplicates were removed, the number of papers decreased to 3704. Titles and abstracts were then further screened, and irrelevant ones were discarded. A total of 697 papers were gathered for further screening. The full texts of the 697 remaining articles were examined in detail against the inclusion and exclusion criteria presented in Table 1. Finally, only 20 potential articles were included in this systematic review.

All of the 20 included studies were published between 2004 and 2015. A flow diagram of the search strategy is illustrated in Fig. 1. The common characteristics of the papers and the summary of content analysis are presented in Table 2. The OSN-based system for tracking a pandemic has been adopted in different countries. Six studies limited their geographical tracking within the United States, and it was achieved by restricting the data search to a specified longitude and altitude of the required region. Three studies from the United Kingdom and one study each from Japan,

Table 3
Features and classifier performance of supervised machine learning studies.

Studies	Features	Classifier performance
Broniatowski et al. [24]	This research considered the *n*-gram word features (2–3 words), linguistic style, syntax, and writing style of the tweets	The first classifier determined whether the tweets are relevant or irrelevant to health. It utilized a combination of keyword filtering and a support vector machine (SVM) trained on 5128 tweets. This classifier obtained 90% precision and 32% recall through 10-fold cross-validation In the second and third classifiers, the two classifiers are influenza and infection filters. These classifiers had an estimated precision of 67% and 74% respectively and 87% recall both through 10-fold cross validation
Signorini et al. [10]	In this study, the features are a collection of terms in a dictionary that appear more than 10 times per week	The results produced by the SVM regression classifier were accurate, with an average error of 0.28% (min = 0.04%, max = 0.93%) and a standard deviation of 0.23%
Kim et al. [25]	This research used 500 words as features with phonological and morphological features, including homonyms to the term "influenza," honorifics that could be confused with "influenza," and words with the same stem	The prediction accuracy of the linear least squares regression algorithm was 0.987 when the delay was up to seven days. However, the error rate of prediction also increased with the increase of delay
Lampos and Cristianini [26]	This study (*n*-gram) featured (*n*-gram) words that express illness symptoms	This study used the LARS algorithm and LASSO ranking method. The statistically significant linear correlation was greater than 95%
Aramaki et al. [23]	Word features (Windows Size = 6) were used as features	This study compared SVM (RBF kernel and polynomial kernel), AdaBoost, DT, Bagging, LR, NB, and RF. SVM with the polynomial kernel showed feasibility from both viewpoints of accuracy (75.6% *F*-Measure) and training time (13.256 s)
Culotta [27]	*N*-gram, synonyms, most frequently used words were used as features	The mean accuracy of the message classifier after 10-fold cross validation on the 206 labeled messages was 84.29%. The following conclusions are summarized: (a) Multiple regression outperforms simple regression. (b) Keyword selection is prone to overfitting. (c) Classification is an effective method to remove erroneous messages. (d) Classification is sensitive to the set of training messages
Achrekar et al. [28]	This research used bag-of-words as a feature	The results in this study showed that the number of flu-related tweets is highly correlated with ILI activity in CDC data with a Pearson correlation coefficient of 0.9846 when using auto-regression with the exogenous input prediction (Arx) model
Lamb et al. [29]	In this research, 3-grams with additional features, such as POS Tagger, phrase segmentations using punctuation tags, and stylometry, were used as features	The results obtained by the proposed binary classifier showed that Awareness or Infection (A/I) is the most useful word class feature with 77% *F*-Measure. Self versus Other (S/O), which is the stylometry and pro-drop feature, was the most important after *n*-grams (85% *F*-Measure) obtained using the log linear model
Li and Cardie [30]	This study incorporated three types of features: (a) collocational features, representing words before and after the query word within a window size of three; (b) unigrams, denoting the presence or absence of terms from the dataset; (c) tweet length in tokens; and (d) position of the keyword within the tweet	The performance of features using SVM was evaluated, and the accuracy of a + b (87.30%) and a + b + d (87.26%) were better than a + b + c and a + b + c + d respectively. This result illustrated that the consideration of tweet length (feature c) would largely affect the performance of a classifier. The position of a keyword (feature d) does not significantly contribute to classification performance
Diaz-Aviles et al. [31]	This study used five binary features for each tweet: (a) if a medical condition is present in the tweet, (b) if a location is present in the tweet, (c) if a hashtag is present in the tweet, (d) if a complementary context term is present in the tweet, and (e) if a URL is present in the tweet	The precision using the Personalized Tweet Ranking Algorithm for Epidemic Intelligence (PTR4EI) was 96%. PTR4EI was compared with RankMC (69%) and RankMCL (85%)
Gomide et al. [32]	This study used word features related to dengue, such as words expressing personal experiences about dengue, opinion about dengue, informative resources about dengue, ironic opinion about dengue, and marketing campaigns about dengue	This study used the association rule classifier to classify tweets into personal experiences, opinion, ironic content, resources, marketing. However, the accuracy of this classifier was not specified
Sadilek et al. [33]	This research used unigram, bigram, and trigram word tokens	The evaluation of SVM on a held-out test set of 700,000 tweets showed 0.98 precision and 0.97 recall
Santos and Matos [34]	This study used 650 textual features, bag-of-words, and character bigrams	The naive Bayes (NB) classifier obtained a precision of 0.78 and an *F*-measure of 0.83 based on the cross-validation of the annotated set

Germany, France, Portugal, China, Brazil, and South Korea targeted data to track the spread of a pandemic.

3.2. Description of included studies

A summary of the data sources and volume, geographical location of data, data analysis methods, comparison, and outcome measures for each included study is provided in Table 2. Through the common system processing steps, first, all data related to a pandemic were collected from OSNs through their application programming Interface (API). Second, textual and geographical analyses of data were performed to identify pandemic-related and non-pandemic-related data to detect and track the specified pandemic. Third, the evaluation of the study was conducted by correlating the outcomes between the analysis of OSNs and those gathered from health agencies and centers, such as US CDC, EISS, IDSC, Korea Centers for Disease Control and Prevention (KCDC), and the United Kingdom's Health Protection Agency (UK HPA). These centers rely on both virology and clinical data in [23].

Most of the studies used Twitter as their data source because Twitter is a popular social network website that could cover large populations. Obtaining suitable public data through its API is easy, and the tweets are rich in information that can be used to have a powerful pandemic tracking system.

Machine-learning approaches are commonly used to classify whether or not extracted data are related to a pandemic. Super-

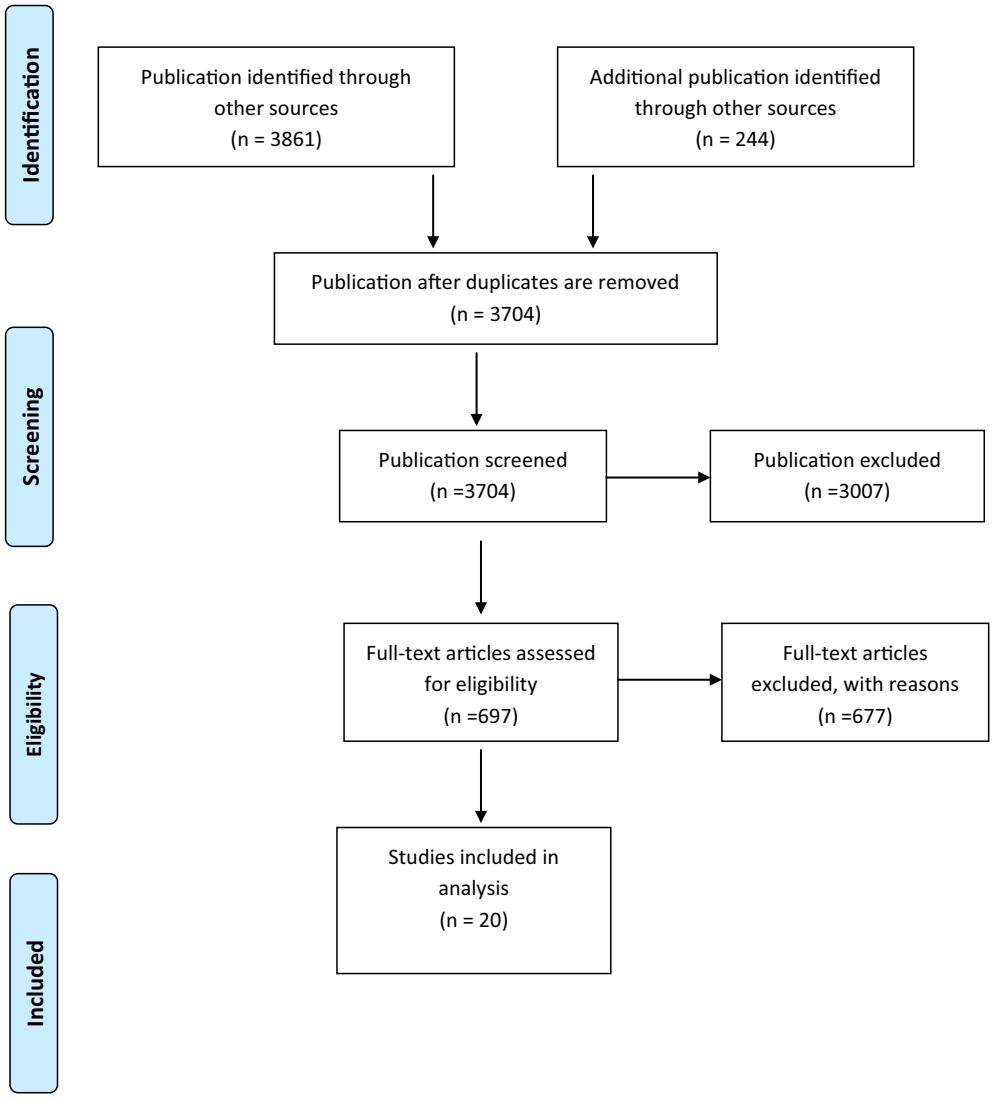

Fig. 1. Flow chart of citations from the initial search to the final number of included publications.

vised machine learning approaches are commonly used to classify related and non-related data. These approaches are preferred when labeled training data are available. A robust learning approach for classifying related and non-related data requires large amounts of labeled training data to facilitate efficient learning and effective discrimination between model classes. Previous studies [10,23–34] used a classification technique to ensure that the data included in their research are relevant to indicate disease spread. Unsupervised Bayesian algorithm based on a four-phase Markov Network was used in [30] to identify a flu breakout by detecting the transition time (i.e., non-epidemic phase, rising epidemic phase, declining epidemic phase, and stationary epidemic phase), and then a support vector machine classifier was trained to identify flu-related data from the extracted data. Study [26] analyzed symptom-related data from a social network website (i.e., tweets) and converted the statistical information into a flu score. In [32], tweets were categorized into five types, and the association rule was used to map each tweet to its related categories. Study [31] adopted the personalized ranking approach that determines relationships and ranks based on the computation of latent semantic topics using Latent Dirichlet Allocation (LDA) and observations of hash-tagging behavior in Twitter. Study [35] used manual classifications and preliminary automated analyses. Study [36] filtered data using keywords according to the definition of a pandemic and related keyword indicators. Research [37] aimed detect the outbreak of Ebola virus in the early stage by tracking tweets related to Ebola. Analysis of the tweets captured the early stage of Ebola outbreak. Ebola-related tweets in Nigeria were chosen three to seven days before the official declaration of the first probable Ebola case. Additional conclusions show the effectiveness of Twitter in spreading health alerts.

To test the accuracy of OSN as a mechanism for tracking a pandemic, official surveillance data were collected from the health agencies and centers and then correlated with OSN data. Studies [10,24,27–30,33,35] used official data from US CDC. Studies [26,38,39] used UK HPA data. Study [25] used official reports from KCDC. Research [23] used official data from IDSC, and [32] studied data from the Health Ministry in Brazil. Research [36] used official data from the Chinese Center for Disease Control and Prevention (China CDC). Study [37] used CDC official Nigeria case report. Research [40] used official data from the French General Practitioners (GPs) Sentinelles network. Study [31] used cases reported by the Robert Koch Institute. Research [34] used Influenzanet data. Study [41] used human immunodeficiency virus (HIV) cases in the United States to correlate with the outcome of their OSN surveillance for tracking a pandemic to ensure the accuracy of their systems.

The common steps followed by the most of included studies can be concluded as following: first stage is data collection. The data

collection variables are data source, data volume, data collected time (year and duration) and data location. Most of the included papers in this systematic review have collected their data related to specified pandemic keywords within specified duration and location. In data collection, stage of most included studies a simple collection method is used such as filtering based on keywords. Nevertheless, in this stage, not all collected are related to the specified pandemic, and more efficient filtering/classification algorithms are needed. Therefore, second stage following data collection is to use the efficient data filtering/classification. Manual filtering/classification and machine learning classification are commonly used to classify collected data. However, manual filtering/classification are time consuming consequently most of the included studies used machine-learning classification. Final stage is to test the adequacy of these surveillances in term of predication accuracy and time-efficiency compared to traditional surveillances.

Most studies [10,23–34] applied supervised machine learning to track pandemics. However applying supervised machine learning may provide successful or failed predication results [43] because building a successful supervised machine model is dependent on various factors. The most important factor is the features used and whether the model has many independent features that correlate well with the class. Most of the effort in building a supervised machine learning model is devoted to this task [43]. Table 3 discusses different features and classifier performances used in supervised machine learning studies.

Table 3 shows that the majority of studies utilized *n*-gram as word features [23,24,26,27,33,34]. However, in [29], the authors utilized part-of-speech (POS) tagging features. In [30], the authors used a combination of various features, such as collocation features, unigram, tweet length, and keyword position, to determine the context of the tweet for deep learning. They achieved 87% accuracy using the SVM classifier. Hence, the analysis shown in the table indicates that selecting significant features is a vital step in supervised classification. Such approach can also improve the performance of the classifier in terms of speed and classification accuracy.

4. Discussion

The popularity and the proliferation of OSNs have created massive social interaction among users and generated a large amount of data. OSNs can be used as platforms to track the spread of infectious diseases. Many researchers have built prediction models of spreading diseases using health-related data extracted from public users' data on social network websites, such as Twitter. The time delay caused by the necessary data processing is the main drawback of those methodologies. However, time delay due to data processing of OSNs-based surveillance takes several hours, whereas time delay caused by syndromic surveillance takes weeks. Moreover, time delay due to data preprocessing of OSN-based surveillance can be reduced by implementing an efficient pre-processing algorithm and by using sophisticated hardware equipment. Time delay in traditional syndromic surveillance is difficult to reduce because it is by caused the inherent limitations of data collection through this surveillance.

The information available in OSNs can provide an additional means to address this problem. To the best of our knowledge, we conducted the first systematic review of empirical studies using OSN data for tracking a pandemic. This research field is still in its infancy, with only 20 studies meeting the criteria for this review.

The results indicate that OSN data can be used to track and estimate users' concerns related to pandemic disease activity in real time. Although OSN data contain a large volume of noise (in the form of links, news, and spams), they remain effective for tracking pandemic diseases. By comparing the data collected from an OSN with the data from health agencies and centers (e.g., US CDC, EISS, IDSC, KCDC for influenza, and UK HPA), an OSN can be used as an early warning detection system. Most of the findings of the studies indicated that OSN data could predict official data up to one week in advance. Therefore, OSNs can be used for faster response than the official data because official data usually take one to two weeks to collect and process.

4.1. What are the challenges in developing a robust OSN pandemic tracking system?

The extraction, analysis, and storage of a large volume of dynamic data require sophisticated infrastructure for high performance and accuracy. An infrastructure including high bandwidth, low latency network, and effective learning algorithm and machines with high computational power and storage resource can solve these technical challenges. Studies [44,45] have demonstrated that cloud computing can provide both high and powerful computation and storage. The data extracted from an OSN are unstructured and require conversion into a suitable format for further processing. They also contain social network abbreviations and slang. Therefore, analyzing the content can be difficult, as it requires powerful data preprocessing. Additionally, OSNs are dynamic, and they grow with time quickly beyond the capabilities of conventional geographical information systems for visualizing the location of data [46]. Designing algorithms that can analyze large, dynamic, and unstructured data remains a challenge. The algorithm should provide high accuracy with less time delay to ensure real-time and accurate pandemic tracking. The other challenges are data access and data privacy preservation. OSN API only enables access to public data, and thus the data collection is limited. However, changing private data to public is not recommended because it may cause attacks (e.g., reputation slander and spamming) [47]. These challenges require the combination of epidemiologic expertise, analytical expertise, and advanced computational skills [48].

The extensive adoption of OSN creates a large volume of user-generated data from millions of users worldwide. The analysis of such large data can be compared with large-scale observational population-based epidemiologic studies [49]. However, OSN data are subject to limitations (e.g., user selection bias and keywords selection bias). Only public users' data can be analyzed. In addition, only a set of keywords related directly to the disease is used to extract the data from OSN API, and thus many disease-related posts that do not contain the specified keywords can be overlooked. Furthermore, the information extracted from OSNs is not interpreted by specialists for relevance before it is sent to a surveillance system [50]. Additionally, data posted in social network websites are difficult to verify. Moreover, the inherent limitation of using OSNs to track a pandemic is in the collection of representative data of sufficient population coverage and in the lack of a well-defined study population. For example, small cities or villages may not have sufficient social network users and data to produce a robust tracking system. However, OSNs are increasing everyday considering their extensive adoption by the society. With a growing number of OSN users, these challenges may diminish with time. Aside from the aforementioned limitations of using data from social network websites, although the analysis is limited to the publicly available data, public data often contain sensitive information that may affect user privacy.

A robust methodology that can deal with OSN challenges and limitations is required. The scientific methodology should be able to deal with the API system efficiently, manage the large streaming data, handle noisy data, and reduce data bias.

A robust methodology is not only a scientific [51] requirement but also an ethical challenge because users falsely detected to be

affected by an infectious disease may harm others. The damage can take several forms, including personal harm and financial loss.

Current advances in the data-processing capabilities of machines and machine learning research present the opportunity for using this huge data sources for many applications, including tracking a pandemic. However, future research on OSN-based surveillance must not solely emphasize the development of new detection approaches or the different applications of these methods to new types of diseases, such as Zika; future research on OSN-based surveillance should also examine other possible methods for incorporating these approaches into existing traditional systems [52,53]. OSN can probably never replace traditional surveillance, but it can offer complementary data that can work best when integrated with traditional data. Therefore, future studies should focus on how to use OSN-based surveillance to complement existing systems.

5. Limitation

Only articles published in the English language were included. Consequently, a language bias could exist. Moreover, the focus of this systematic review was on studies that collected data from OSNs to develop an online information surveillance system to detect and track a pandemic. This study excluded information surveillance-based studies that primarily aimed to refrain from using data from social network as the data source. Moreover, this review focused on studies that used collected OSN public data only to ensure privacy preservation. The exclusion of unpublished articles could also constitute as a limitation.

6. Future implications

This systematic review discusses the use of OSN for tracking a pandemic. The effectiveness of an OSN-based surveillance system depends on the quality of the data and the analysis method. Future studies should focus on how to improve the quality of data by developing techniques that can utilize available profile information and language used in the text to enhance the detection of user geographical location and address the limitations of population coverage.

An OSN-based surveillance system uses primitive algorithms to detect data that indicate an infectious disease. However, with the extensive adoption of social network and rapid data growth, future OSN-based surveillance systems require advanced algorithms that can extract, analyze, detect, and track the spread of a pandemic based on real time with high accuracy and precision. Advanced computational linguistics is also needed to effectively analyze dynamic, large, and unstructured data. The approach of OSN-based surveillance systems is systematic and general; therefore, they may be applicable to a wide range of infectious diseases.

Future work in OSN-based surveillance should focus on improving detection approaches with high filter methods, testing OSN-based surveillance for unexplored applications such as Zika, and integrating OSN-based surveillance with existing traditional systems [52,53]. Future research should introduce flexible OSN-based surveillance with frameworks that can integrate a number of data sources from different OSN platforms to obtain a comprehensive understanding of pandemic spread from different angles.

7. Conclusion

This systematic review aims to conduct a critical appraisal of pandemic surveillances based on the data extracted from OSNs to investigate their adequacy as a tool in tracking a pandemic and to identify the extent of using OSNs to track a pandemic and their challenges and limitations. According to this systematic review, OSN data contain rich information that can be used for tracking a pandemic. The data collection of traditional data, which are collected from traditional surveys and clinical reports, are time consuming and costly. For example, the traditional surveillance of influenza pandemic, which is entirely manual, causes a time lag of one to two weeks between the time of medical diagnosis and the time when the data become available. However, OSN data can be collected in almost real time at a cheaper cost. Additionally, the geographical and temporal information can provide an exploratory analysis of the spatiotemporal dynamics of the spread of infectious disease.

The approaches of an OSN-based surveillance system are systematic and general, and they may be applicable to a wide range of infectious diseases. Data from social networks can be combined with traditional epidemiological data [24,35,54] to improve the detection and prediction accuracy at a rapid and less costly rate.

This systematic review concludes that OSNs are probably to never replace traditional surveillance, but they offer complementary data that can work best when integrated with traditional data. This systematic review highlights that although certain OSN-based surveillance systems provide effective results, collecting representative data with a sufficient population coverage, which leads to the lack of well-defined study population, remains a challenge. Moreover, bias in the collected data may affect the final result of these systems. The accuracy and precision of future OSN-based surveillance systems depend on the quality of data and the analysis methods, which require advanced algorithms and computational linguistics.

Conflict of interest statement

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

We would like to take this opportunity to thank University of Malaya UMRG (RP028D-14AET) for funding this Research.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2016.05.005.

References

[1] G. Eysenbach, Infodemiology: tracking flu-related searches on the web for syndromic surveillance, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2006.
[2] P. Velardi et al., Twitter mining for fine-grained syndromic surveillance, Artif. Intell. Med. 61 (3) (2014) 153–163.
[3] G. Eysenbach, Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet, J. Med. Internet Res. 11 (1) (2009).
[4] J.S. Brownstein, C.C. Freifeld, L.C. Madoff, Digital disease detection—harnessing the Web for public health surveillance, N. Engl. J. Med. 360 (21) (2009) 2153–2157.
[5] M. Sampson et al., A systematic review of methods for studying consumer health YouTube videos, with implications for systematic reviews, PeerJ 1 (2013) e147.
[6] H.A. Carneiro, E. Mylonakis, Google trends: a web-based tool for real-time surveillance of disease outbreaks, Clin. Infect. Dis. 49 (10) (2009) 1557–1564.
[7] A.F. Dugas et al., Influenza forecasting with Google flu trends, PLoS ONE 8 (2) (2013) e56176.
[8] V.M. Dukic, H.F. Lopes, N. Polson, Tracking flu epidemics using Google Flu Trends and Particle Learning, available at SSRN 1513705, 2009.
[9] J. Ginsberg et al., Detecting influenza epidemics using search engine query data, Nature 457 (7232) (2009) 1012–1014.
[10] A. Signorini, A.M. Segre, P.M. Polgreen, The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic, PLoS ONE 6 (5) (2011) e19467.

[11] A. Hulth, G. Rydevik, A. Linde, Web queries as a source for syndromic surveillance, PLoS ONE 4 (2) (2009) e4378.
[12] C.C. Freifeld et al., HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports, J. Am. Med. Inform. Assoc. 15 (2) (2008) 150–157.
[13] S. Bhattacharya et al., Belief surveillance with twitter, in: Proceedings of the 4th Annual ACM Web Science Conference, ACM, 2012.
[14] V. Lampos, T. De Bie, N. Cristianini, Flu detector-tracking epidemics on Twitter, in: Machine Learning and Knowledge Discovery in Databases, Springer, 2010, pp. 599–602.
[15] D. Centola, The spread of behavior in an online social network experiment, Science 329 (5996) (2010) 1194–1197.
[16] J. Ratkiewicz et al., Detecting and tracking political abuse in social media, in: ICWSM, 2011.
[17] A.J. Elliot et al., Monitoring the emergence of community transmission of influenza A/H1N1 2009 in England: a cross sectional opportunistic survey of self sampled telephone callers to NHS direct, BMJ 2009 (2009) 339.
[18] L.N. Carroll et al., Visualization and analytics tools for infectious disease epidemiology: a systematic review, J. Biomed. Inform. 51 (2014) 287–298.
[19] K. Lee, A. Agrawal, A. Choudhary, Real-time disease surveillance using Twitter data: demonstration on flu and cancer, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013.
[20] M.S. Porta et al., A Dictionary of Epidemiology, Oxford University Press, 2014.
[21] K. Daine et al., The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people, PLoS ONE 8 (10) (2013) e77555.
[22] J. Singh, Critical appraisal skills programme, J. Pharmacol. Pharmacother. 4 (1) (2013) 76.
[23] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: detecting influenza epidemics using Twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011.
[24] D.A. Broniatowski, M.J. Paul, M. Dredze, National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic, PLoS ONE 8 (12) (2013).
[25] E.-K.K. Kim et al., Use of hangeul twitter to track and predict human influenza infection, PLoS ONE 8 (7) (2013).
[26] V. Lampos, N. Cristianini, Tracking the Flu Pandemic by Monitoring the Social Web, 2010.
[27] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, in: Proceedings of the First Workshop on Social Media Analytics, ACM, 2010.
[28] H. Achrekar et al., Predicting flu trends using twitter data, in: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2011.
[29] A. Lamb, M.J. Paul, M. Dredze, Separating fact from fear: tracking flu infections on twitter, in: Proceedings of NAACL-HLT, 2013.
[30] J. Li, C. Cardie, Early Stage Influenza Detection from Twitter arXiv preprint arXiv:1309.73402013.
[31] E. Diaz-Aviles et al., Epidemic intelligence for the crowd, by the crowd, in: ICWSM, 2012.
[32] J. Gomide et al., Dengue surveillance based on a computational model of spatio-temporal locality of Twitter, in: Proceedings of the 3rd International Web Science Conference, ACM, 2011.
[33] A. Sadilek, H.A. Kautz, V. Silenzio, Modeling spread of disease from social interactions, in: ICWSM, 2012.
[34] J.C. Santos, S. Matos, Analysing Twitter and web queries for flu trend prediction, Theor. Biol. Med. Modell. 11 (Suppl. 1) (2014) S6.
[35] C. Chew, G. Eysenbach, Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak, PLoS ONE 5 (11) (2010).
[36] J. Huang, H. Zhao, J. Zhang, Detecting flu transmission by social sensor in China, in: Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, IEEE, 2013.
[37] M. Odlum, S. Yoon, What can we learn about the Ebola outbreak from tweets?, Am J. Infect. Control 43 (6) (2015) 563–571.
[38] M. Szomszor, P. Kostkova, E. De Quincey, # swineflu: Twitter predicts swine flu outbreak in 2009, in: Electronic Healthcare, Springer, 2009, pp. 18–26.
[39] E. de Quincey, Potential of Social Media to Determine Hay Fever Seasons and Drug Efficacy, Planet@ Risk 2 (4) (2014).
[40] K. Wegrzyn-Wolska, L. Bougueroua, G. Dziczkowski, Infodemiology by Tweet mining methods, Stud. Inform. Univ. 11 (3) (2013) 65–79.
[41] S.D. Young, C. Rivers, B. Lewis, Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes, Prev. Med. 63 (2014) 112–115.
[42] B. O'Connor et al., From tweets to polls: linking text sentiment to public opinion time series, in: ICWSM, vol. 11, 2010, pp. 122–129.
[43] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (10) (2012) 78–87.
[44] M. Armbrust et al., A view of cloud computing, Commun. ACM 53 (4) (2010) 50–58.
[45] B. Tograph, Y.R. Morgens, Cloud computing, Commun. ACM 51 (7) (2008).
[46] A. Padmanabhan et al., FluMapper: an interactive CyberGIS environment for massive location-based social media data analysis, in: Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery, ACM, 2013.
[47] G. Hogben, Security issues and recommendations for online social networks, ENISA Position Paper, no. 1, 2007.
[48] M. Salathe et al., Digital epidemiology, PLoS Comput. Biol. 8 (7) (2012) e1002616.
[49] I.C.-H. Fung, Z.T.H. Tse, K.-W. Fu, The use of social media in public health surveillance, West. Pac. Surveillance Resp. J.: WPSAR 6 (2) (2015) 3.
[50] E. Velasco et al., Social media and internet-based data in global systems for public health surveillance: a systematic review, Milbank Q. 92 (1) (2014) 7–33.
[51] E. Vayena et al., Ethical challenges of big data in public health, PLoS Comput. Biol. 11 (2) (2015) e1003904.
[52] D.C. Pattie et al., A public health role for Internet search engine query data?, Military Med 174 (8) (2009) xi–xii.
[53] G.J. Milinovich et al., Internet-based surveillance systems for monitoring emerging infectious diseases, Lancet. Infect. Dis. 14 (2) (2014) 160–168.
[54] M.A. Stoové, A.E. Pedrana, Making the most of a brave new world: opportunities and considerations for using Twitter as a public health monitoring tool, Prev. Med. 63 (2014) 109–111.