Journal of Biomedical Informatics 108 (2020) 103500

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Social media based surveillance systems for healthcare using machine learning: A systematic review

Aakansha Gupta, Rahul Katarya*

Department of Computer Science & Engineering, Delhi Technological University, Delhi 110042, India

ARTICLE INFO

Keywords:
Health informatics
Machine learning
Outbreak detection
Surveillance systems
Social media

ABSTRACT

Background: Real-time surveillance in the field of health informatics has emerged as a growing domain of interest among worldwide researchers. Evolution in this field has helped in the introduction of various initiatives related to public health informatics. Surveillance systems in the area of health informatics utilizing social media information have been developed for early prediction of disease outbreaks and to monitor diseases. In the past few years, the availability of social media data, particularly Twitter data, enabled real-time syndromic surveillance that provides immediate analysis and instant feedback to those who are charged with follow-ups and investigation of potential outbreaks. In this paper, we review the recent work, trends, and machine learning(ML) text classification approaches used by surveillance systems seeking social media data in the healthcare domain. We also highlight the limitations and challenges followed by possible future directions that can be taken further in this domain.

Methods: To study the landscape of research in health informatics performing surveillance of the various health-related data posted on social media or web-based platforms, we present a bibliometric analysis of the 1240 publications indexed in multiple scientific databases (IEEE, ACM Digital Library, ScienceDirect, PubMed) from the year 2010–2018. The papers were further reviewed based on the various machine learning algorithms used for analyzing health-related text posted on social media platforms.

Findings:: Based on the corpus of 148 selected articles, the study finds the types of social media or web-based platforms used for surveillance in the healthcare domain, along with the health topic(s) studied by them. In the corpus of selected articles, we found 26 articles were using machine learning technique. These articles were studied to find commonly used ML techniques. The majority of studies (24%) focused on the surveillance of flu or influenza-like illness (ILI). Twitter (64%) is the most popular data source to perform surveillance research using social media text data, and Support Vector Machine (SVM) (33%) being the most used ML algorithm for text classification.

Conclusions: The inclusion of online data in surveillance systems has improved the disease prediction ability over traditional syndromic surveillance systems. However, social media based surveillance systems have many limitations and challenges, including noise, demographic bias, privacy issues, etc. Our paper mentions future directions, which can be useful for researchers working in the area. Researchers can use this paper as a library for social media based surveillance systems in the healthcare domain and can expand such systems by incorporating the future works discussed in our paper.

1. Introduction

Syndromic surveillance systems aim for the collection of data, which can help in providing a basic scenario for all communicable infectious diseases. These systems may vary depending on the source of data, their planned duration, and how data is recorded and acquired. These systems may use traditional data and real-time data from various social media platforms. In the field of healthcare, these systems usually focus on the early identification of illness clusters and the symptom period before the confirmation of a particular disease by any clinical unit or laboratory and to mobilize the rapid response. Surveillance systems are usually concerned with the systematic collection, analysis, and interpretation of the collected data along with the detection, confirmation, and reporting of disease, and also considering the public health response. Objectivized definitions, algorithmic diagnosis, and electronic databases have made surveillance systems more user-friendly and effective over time [1]. Traditional biosurveillance relies on clinical encounters to collect information, which is a time-consuming process.

* Corresponding author.

https://doi.org/10.1016/j.jbi.2020.103500
Received 10 April 2019; Received in revised form 21 June 2020; Accepted 26 June 2020
Available online 02 July 2020

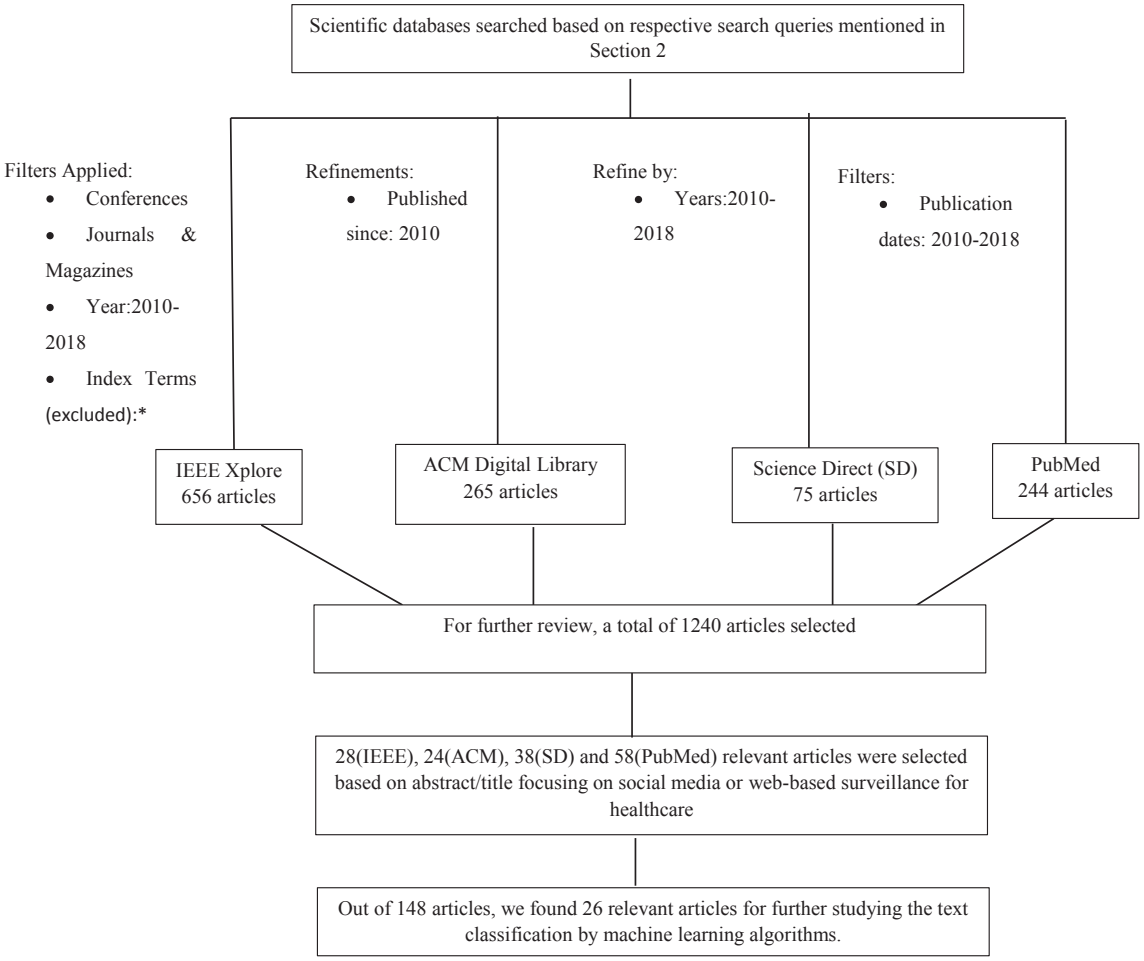


Fig. 1. Search methodology for the selection of relevant articles. * Cloud computing, patient diagnosis, video surveillance, mobile computing, microorganisms, patient monitoring, wireless sensor networks.

Also, traditional pandemic surveillance is mostly a manual process that causes a delay of one to two weeks in the availability of the data by clinical diagnosis [2,3]. In the last few years, the availability of web-based data sources emerged as an extension to traditional surveillance systems [4] and has sustainably contributed to infectious disease surveillance by providing real-time statistics and reducing the cost of public health [5]. It can be noted how rapidly events can be detected in real-time using internet-based surveillance data when an ordinary individual's social media post has led to a tremendous increase in public engagement with skin cancer prevention [6]. Despite the other uses of social media [7], the role of monitoring social media can be explored in making healthcare decisions [8]. Also, the early detection of diseases and immediate public health response has led to the need for new approaches and technologies to reinforce the capacity of traditional syndromic surveillance systems. Several research papers/articles have been appeared in reviewing the various surveillance systems using social-media data [5,9,10,11,13,14,15,16,17,18]. These papers have covered the applications, technologies, algorithms, data sources, and their evaluation. However, a recent review of surveillance systems in the health informatics domain through social media is not available to the best of our knowledge. In this paper, we review machine learning technologies and approaches published in this domain, mainly in the past few years, and also mention the challenges and future directions.

This review paper examines a set of research questions that would allow us to get the latest trends followed by social media based surveillance systems in the field of healthcare. Moreover, it also helps us to get an overview of recently used machine learning algorithms for analyzing the data used by these systems. These research questions (RQ) are stated as follows:

- RQ1: Which machine learning techniques are popular among authors of research papers when developing a social media based surveillance systems in the health sector?
- RQ2: What are the most commonly used sources of social media data for the surveillance of health-related topics?
- RQ3: What can be the applications of social media based surveillance systems in the area of health informatics?
- RQ4: Are there any limitations or challenges faced by syndromic surveillance systems with the inclusion of social media data?

To answer the research questions mentioned above, we extracted 1240 research articles that published studies related to our research, from various scientific digital libraries, and inspected them. The methodology for the selection of these articles is discussed in Section 2.

When looking at the surveillance systems in the field of health informatics, various machine learning algorithms like Deep Neural Network (DNN), Naive Bayes (NB), Multinomial Naive Bayes (NBM), SVM, etc. have been chosen and proposed for epidemic prediction classification approach. We will answer RQ1 by discussing and explaining a few machine learning approaches in Section 3, though the actual work is much broader. To answer RQ2, Section 4 will identify the social media sources involved in the collection of data, followed by the applications of such surveillance systems in the next section, i. e. Section 5. This section holds the answer to question RQ3. Section 6 aims to answer RQ4 by discussing the limitations and challenges, and Section 7 mentions the original contribution of this writing, followed by

the conclusion, results, and future work in Section 8.

2. Articles selection method

The scope of this paper involves studies of the social media-based surveillance systems that predict the disease in real-time or near real-time using machine learning approaches. The selection criteria for the research articles were set to incorporate papers published in the year 2010–2018.

The following scientific databases were explored to provide a comprehensive bibliography of research papers on social media based surveillance systems in the healthcare domain:

- ACM Portal
- IEEE Xplore
- Science Direct
- PubMed

IEEE Xplore database was advanced searched to form the following query: ((("Abstract": surveillance) OR "Document Title": surveillance OR "Abstract": outbreak) AND ("Abstract": health OR "Abstract": disease)) and 656 articles (Conferences and Journals & Magazines) were retrieved when filters mentioned in Fig. 1 were applied. Similarly, an ACM Digital Library searched for query: recordAbstract:(((outbreak OR surveillance) OR acmdlTitle:(+surveillance)) AND (health* OR disease)) retrieved 265 articles.Also, we advanced searched ScienceDirect database for query: (surveillance OR outbreak) AND (health* OR disease) AND "social media" in title, abstract, and keywords. As a result, 75 articles were extracted based on the search terms. Lastly, PubMed, which accesses the MEDLINE database, was searched for publications. On searching this resource with query: (((surveillance[Title/Abstract]) OR outbreak[Title/Abstract]) AND((health[Title/Abstract]) OR disease [Title/Abstract]) AND social media[Title/Abstract]), we get 244 articles . For further analysis, a total of 1240 articles were identified in the initial query of the knowledge sources.

Each of the 1240 articles, were screened independently by each author of the paper, based on abstract and title. If the abstract or title or both are explaining social media or web-based surveillance, then we considered them for further research else they were rejected. While following this step, an article was included in the corpus only when both authors agreed it was relevant; disagreements were handled using consensus. The next step we performed was to consider the articles that had utilized machine learning approaches in their methodology. Fig. 1 describes the steps followed in selecting the relevant articles for this study.

Besides, we searched Google Scholar to get the statistics related to our research area, reflecting the trends of the past few years. Fig. 2 shows the count of recent research papers and patents published since 2010 till 2018. The terms involved for research were the "surveillance system", "social media", "machine learning" and "health informatics". This plot clearly shows an increase in the number of publications about surveillance systems involving social media data and machine learning algorithms in the healthcare domain over time.

3. RQ1: machine learning methods used by surveillance systems for processing social media data

In this section, we will explain the most commonly used machine learning based classification methods employed to analyze health-related text from social media platforms. As mentioned in section 2, a total of 26 articles were found relevant for studying the ML-based text classification algorithms.

In recent years, machine learning has gained much attention, especially in analyzing the patterns in images or raw data. M. Bates [19] addressed how the progress in machine learning allows epidemiologists to mine through a broad set of digital data. A. Mike and C. Daniel [11] reviewed the conjunction of natural language processing and machine learning with social media platforms to support the analysis of massive dataset for population-level mental health research. Among different methodological variations of machine learning, some architecture stands out in popularity. For instance, we noted that the k-Nearest Neighbor (k-NN) classifier's precision was superior to several other machine learning classifiers such as NBM Modal, NB, and SVM [20], for classifying the tweets between two classes-i.e., real occurrences of allergy or awareness tweets. Similarly, K. Lee, A. Agrawal, A. Choudhary et al. [21] showed that best text classification performance was obtained using Multinomial Naive Bayes Modal with F-measure of 0. 811 when compared against the other classifiers such as NB, Random Forest (RF), and SVM. The author [22] exhibited a model using the multilayer perceptron with backpropagation algorithm on Twitter data to predict the weekly status of the US population infected with ILI. Even several supervised machine learning algorithms were studied for detecting the personal health experience tweets [23] and used the deep gramulator approach to improve precision when applied to independent test sets.

The unsupervised classification algorithms do not require labeled data sets to predict the output, like supervised algorithms. Because of this reason, the unsupervised classification methods seem to be a more attractive alternative in the process of analyzing the text, but they could be more challenging in achieving a similar accuracy as supervised methods. The same can be observed [24] when L. Sousa, R. de Mello, D. Cedrim et al. performed the classification of tweets using supervised and unsupervised methods. They concluded that the topic modeling (LDA), one of the unsupervised techniques, presents less control over the content of topics in comparison to a traditional classifier, particularly on a naturally noisy media channel. Hence, Multinomial Naive

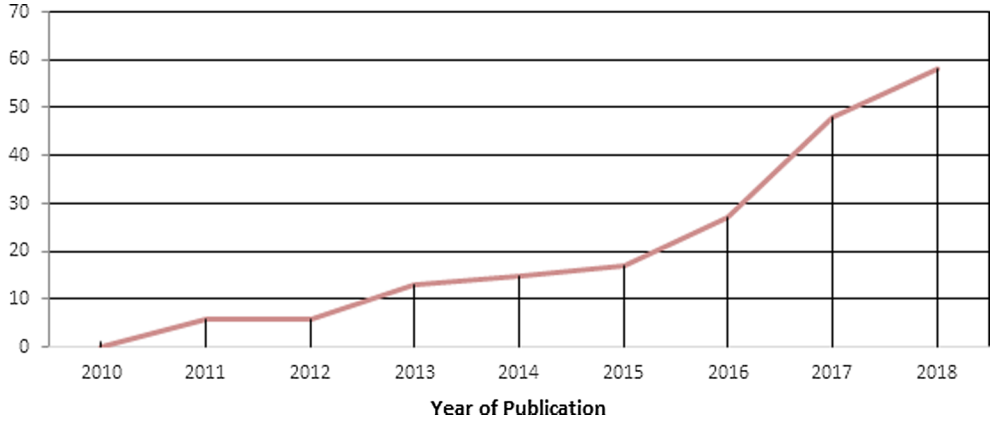
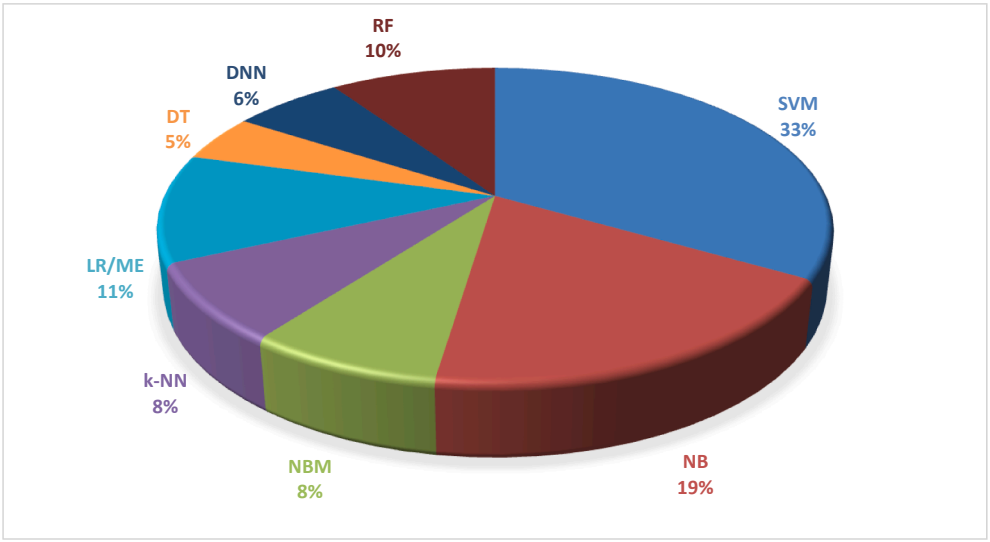


Fig. 2. Distribution per year of articles on Google Scholar.



NB: Naive Bayes, NBM: Multinomial Naive Bayes, k-NN: k-Nearest Neighbor, ME: Maximum Entropy, LR: Logistic Regression, DT: Decision Tree, DNN: Deep Neural Network, RF: Random Forest, SVM: Support Vector Machine

Fig. 3. Distribution of ML methods for health-related text classification by social media based surveillance systems. NB: Naive Bayes, NBM: Multinomial Naive Bayes, k-NN: k-Nearest Neighbor, ME: Maximum Entropy, LR: Logistic Regression, DT: Decision Tree, DNN: Deep Neural Network, RF: Random Forest, SVM: Support Vector Machine.

Bayes, a supervised classification approach, was considered for classifying the Twitter content.

Extensive use of machine learning algorithms to process and distinguish health-related social media data includes [20,21,23,25].

Fig. 3 represents the commonly used machine learning methods for health-related text classification in selected papers.

3.1. Support vector machine

SVM is a popular binary classifier built upon the concept of decision planes that define decision boundaries. In this approach, original training data is transformed into a higher dimension using a nonlinear mapping. Within this new dimension, a linear optimal separating hyperplane is searched to minimize the distance between hyperplane points and maximize the margin between the classes [26,27]. It has been known for its superior performance in text classification with word features. Although the performance of classification algorithms highly depends on the input parameters and application, yet for binary classification tasks, SVM was observed to be highly suitable. V. K. Jain and S. Kumar [28] reported the SVM classifier as a good performer in terms of accuracy in predicting the class of tweets(disease-related tweets/irrelevant tweets). Similarly, the SVM algorithm was able to achieve an accuracy of 90.09% when tweets were classified as infodemiological or non-infodemiological [29]. N. Yang, X. Cui, C. Hu C et al. [30] were able to classify 'sick microblog' and 'not sick microblog' posts using the SVM classification model. They also showed that time consumption by SVM for classification task did not get much affected when the micro-blogs needed to be arranged increased by 100 times, though there is a huge increase in the time consumption by KNN to complete the classification task. SVM turned out to be the best tweet classification method when compared with other machine learning techniques and has been utilized to classify social media data on a range of physical health issues: [20,21,23,24,25,31,32,33,34,35,36,37,38,39].

3.2. Naive Bayes

It is a classification algorithm for binary and multiclass classification problems. This algorithm makes a naive assumption that there are no predictors i. e. features are independent of each other, and one feature's impact on predicting class does not depend on the presence of another feature [27]. This method is based on the Bayes Theorem, shown as below:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Where:
H: Hypothesis that data X belongs to a specific class C
X: Data with the class yet known
P(H|X): Posterior probability of hypothesis H conditioned on X
P(H): Prior probability of hypothesis H
P(X|H): Posterior probability of X conditioned based hypothesis H
P(X): Prior Probability of X.

V. K. Jain and S. Kumar [28] classified datasets into mosquito-borne disease-relevant and irrelevant tweets using SVM and Naive Bayes, and further relevant tweets were classified into three classes: symptoms, fear, and prevention using same classifiers. V. Kumar and S. Kumar [25] performed the classification task on tweets to differentiate swine flu-related text from noise or irrelevant tweets using various ML techniques such as Decision Tree, Naive Bayes, SVM, and Random Forest. They considered every relevant swine flu-related word as a feature and found both Naive Bayes and SVM provided the best classification result, with F-measure of 0. 77. The Naive Bayes classifier gave the best performance when dengue suspected tweets were classified as irrelevant or relevant, considering emojis, location information, unigrams, bigrams, and trigrams [40]. Naive Bayes is a popular classification algorithm and is used by many authors for text classification: [13,20,21,24,36,41,39] showing an average performance when compared to other classifiers.

3.3. Multinomial Naive Bayes:

Multinomial Naive Bayes networks, a variant of Naive Bayes networks, are better for text documents. Multinomial networks consider the frequency of words, and underlying calculations of probability are adjusted accordingly, while NB networks do not consider the frequency count.

X. Ji, S.A. Chun and J. Geller [36] used different ML methods (NB, NBM, and SVM) to classify tweets as either personal or news-related. They further classified personal tweets into two categories: negative or neutral tweets. Among all the methods used for classification and sentiment analysis, NBM accomplished overall the best outcome and turned out to be better than the other two machine learning classifiers regarding time consumed to build the classifier. Similarly, K. Lee, A. Agrawal, and A. Choudhary [21] developed a real-time allergy surveillance system that distinguished tweets as either positive or negative, where a positive tweet mentions about the author or someone around

the author having allergy symptoms. And, the tweet is labeled negative if it talks about news, advertisement, or general awareness of allergies. They showed that best text classification performance was obtained using Naive Bayes Multinomial Modal with F-measure of 0. 811 when compared against the other classifiers such as NB, RF, and SVM. Similarly, this classifier gave the best recall and f-measure of 0. 859 and 0. 857 when the tweets were classified among the multiple allergy hidden classes [20].

3.4. k-Nearest neighbor

k-Nearest-Neighbor is an instance-based statistical analysis method to perform classification. Its implementation requires an integer k, a set of training data and measure closeness metric. The given training set is used as an input vector to form different regions for different classes. When given an unlabeled object, this classifier searches for k training sets in pattern space that are closest to the unlabeled object. These k training sets are the k "nearest neighbors" of the unlabeled object. "Closeness" is defined in terms of a distance metric, such as Euclidean distance, which is given by:

$$dist(X1, X2) = \sqrt{\sum_{i=1}^{n} (x1i - x2i)^2} \quad (2)$$

where, X1 = ($x11$, $x12$, …, $x1n$) and $\boldsymbol{X}2$ = (. $x21$, $x22$, …, $x2n$) representing two objects or points [27,42]. Nargund K, Natarajan S. [20] used k-NN alongside Naive Bayes, SVM, and Naive Bayes Multinomial to identify and recognize messages reporting and discussing different types of allergies. They noted that k-NN has better precision than other approaches in the identification and assignment of tweets as either actual incident of allergy or awareness tweets. This classifier was observed as having a good precision for text classification than other classifiers such as k-means, but not the best [30]. Other research papers that have used the k-NN are: [23,33].

3.5. Logistic regression

Logistic regression was proposed in the late 1960s and early 1970s and became routinely available in statistical packages in the early 1980s [43]. LR is a statistical technique for analyzing a dataset for a binary classification problem. It helps in discovering the relationship between a dependent binary variable and at least one independent variable. Each independent variable is multiplied with weights and summed up. This outcome will sum up to a sigmoid function to get the result in the range of 0 and 1. The values below 0.5 are considered as 0, and those above 0.5 are considered as 1. In this manner, optimization techniques aim to find the best regression coefficients and weights. Logistic regression is mathematically constrained to produce probabilities in the range [0,1]. Also, it can converge on parameter estimates relatively easily.

Along with different classification algorithms, logistic regression is also preferred for the data classification task. For instance, the logistic regression gave a better recall and F1 measure than SVM in the classification of asthma relevant and irrelevant tweets [32]. Logistic regression showed excellent precision for the analysis of personal and non-personal tweets [23]. Among the research papers, we have reviewed [33,44] also utilized logistic regression classifier. Maximum Entropy classifier, sometimes called Multinomial Logistic Regression [45], is also used for the text classification task. Tweets related to illness were identified using Maximum Entropy [46]. Another study that used Maximum Entropy for tweet classification includes [40].

3.6. Decision tree

Decision Tree is a flowchart-like tree structure, where each non-leaf node represents a test on an attribute, each branch denotes an outcome of the test, and each terminal node holds a class label. Attributes values of an unlabeled sample, X, are tested against the decision tree to predict its class. A unique path is traced from root(topmost node) to a terminal node based on attributes' values, which holds the predicted class for the unlabeled sample [47,27]. DTs are easy to assimilate and have good accuracy. They can handle real-valued items, categorical features items, and items with a mixture of both. They are flexible enough to handle items with some missing features. Unfortunately, decision trees are poor at handling changes as a minor change in input data may lead to massive changes in the constructed tree. They are good at naturally supporting classification problems with more than two classes and capable of handling regression problems. Finally, once constructed, new items can be classified quickly.

J-48-Decision Trees classifier performed well in predicting positive and negative tweets related to personal health experience [33]. Similarly, R. A. Calix and A. General achieved an average result using Decision Trees classifiers for classifying Personal health experience tweets [23]. Even, DT classifier was experimented for distinguishing tweets related to swine flu [25].

3.7. Deep Neural Network:

A standard neural network (NN) consists of many simple, connected neurons, whereas Deep Neural Network(DNN) employs a deep architecture in NNs with a certain level of complexity and an increased number of layers in a single layer [48,49]. In the past few years, Deep Neural Networks (DNNs) have gained much popularity in text classification. DNN classifiers outperform every other conventional classifier experimented such as IB1-k-Nearest Neighbor, J48-Decision Tree, LR, and SVM when tweets were classified as personal and non– personal health experience tweets [33]. CNNs have achieved remarkable performance in computer vision and deep learning. It is a class of Neural Network that is proven very useful in the areas of text processing, image recognition, and classification. Recently, CNNs are actively exploited for text classification in the health domain. J. Du, L. Tang, Y. Xiang et al. have also used different types of DNN, i.e., Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM), along with other machine learning approaches for measles-related tweet classification tasks, where CNN has shown a remarkable performance [50]. To study the prediction of chickenpox and eliminate delays in disease reporting of existing surveillance systems, S. Chae, S. Kwon and D. Lee [51] optimized parameters of DNN and LSTM (a special kind of RNN) algorithms.

3.8. Random forest

Random Forest is an ensemble learner which improves the accuracy of the model by combining a collection of decision tree classifiers (forest) to generate the aggregated result. It uses classification and regression trees (CART) methodology to grow the trees. At each node, attributes are randomly selected to determine the split and generate individual decision trees. The values of the random vector sampled are responsible for determining each tree. During classification, votes are cast by each tree, and the class with maximum votes is returned [52]. RF can handle multidimensional data and is capable of estimating missing data. It also considers the importance of the variables used in classification. RFs consider many fewer attributes for each split, and they are efficient on vast databases. Random Forest approach is used along with other conventional machine learning approaches for social media text classification, such as [21,23–25].

Some authors have explored other popular machine learning approaches for text mining, such as k-means [30,35], clustering [13], etc. Dai X, Bikdash M and Meyer B. [13] proposed a word embedding based clustering technique to classify health-related tweets. A tweet can be grouped on the grounds of similar words and can be classified based on the similarity measure. They compared their proposed method with Naive Bayes classifier and found the former is superior to the Naive

Bayes method. Topic modeling, a well-known algorithm of semantic clustering, has indicated a valuable outcome for classification [24].

4. RQ2: popular social media data sources for data collection

Since there is a considerable increase in social media users to share information, there is significant traction among researchers to analyze social media activities for public health purposes. Also, social media can cover various topics in addition to those covered by traditional data sources of public-health.

Social media has emerged as a feasible source for health communication [10,53,54,55]. Although some research has cast doubt on whether social media data would have utility for detecting outbreak [56,46], analysis of social media content for healthcare data has been a topic of broad interest [57,58,59,60]. To track and forecast health events, [61] explains an urgent need for including social media data sources while disseminating epidemic outbreak advisories using their Facebook or Twitter pages to address a more extensive public base in no time. Hence, social media posts and online search behavior could be useful sources of information about health outbreaks.

4.1. Twitter

Twitter is one of the leading micro-blogging services, where registered users can post tweets or retweet other posts that can be read by unregistered users. With over 300 million monthly active users, Twitter has become a reliable and fast source to evaluate the incidence of diseases in a population. H. Kwak, C. Lee, H. Park et al. [62] conducted a study to examine the potential of Twitter as a new source of information sharing. As social media postings from Twitter have become a reliable and fast source to evaluate the incidence of diseases in a population, effective and efficient methods must be developed to process and examine health-related tweets. Usually, dimensions like location, volume, time [63], and public perceptions are considered for disease surveillance. In a recent work [37], data collected from Twitter was utilized to find information during different epidemics that were useful for various health organizations. C. Bosley et al. used a set of seven terms to collect more than 60 thousand tweets and then examine and classify them, focusing on cardiac arrest and resuscitation [64]. For early detection of influenza activity, [22] proposed a model that uses real-time Twitter data streams and U.S. Centers for Disease Control and Prevention (CDC) historical datasets to foresee future influenza activities. Lee K, Agrawal A, Choudhary A. [21] examined the allergy activities. In their work, they collected tweets that mentioned allergy-related tweets. A natural language processing approach is adopted [65] to reach Ebola-related tweets considering four main topics based on clusters of keywords: risk factors, prevention education, disease trends, and compassion. An unsupervised method was used [13] to partition vectors represented tweets into clusters of similar words. The tweet is then classified as disease-related or unrelated based on the clusters' similarity measures. N. El-bathy, C. Gloster, M. El-bathy et al. [66] proposed a surveillance lifecycle architecture using a novel genetic algorithm to get relevant data from the large set of online data accessible faster at a lower cost. Health problems such as respiratory, gastrointestinal, heat-related illness, and ILI symptoms circulating among the population during mass gathering were also detected using Twitter [67]. The use of Twitter data has been done to examine a variety of public health incidences such as allergy:[20], mosquito-borne disease: [24], dengue: [29,68], flu/influenza: [13,69,70,71,38,72,73], H1N1: [25,74,75,76] and other disease [77,78,79,80,81,82,83,84,85].

4.2. Instagram:

Instagram is a photo and video-sharing service founded in 2010, with over 800 million registered users [86]. Guidry JPD, Jin Y, Orr CA et al. [87] examined the Ebola-related social media posts on Twitter and Instagram, and results suggested that Instagram can be an optimal platform for communication and reaching the public in times of worldwide health crises. Studies have also examined Zika-focused messages on Instagram [88]. As Instagram is a photo and video sharing platform, the image and video data available at this platform can be a promising source for disease surveillance.

4.3. Crowdsourcing

Crowdsourcing is a process in which a large, undefined group of volunteers or part-time workers are involved to provide services, ideas, or content through a flexible open call [89]. Usually, a large group of people with varying degrees of knowledge and experience are involved in contributing to a common goal. Most well-known methods of crowdsourcing include Google Consumer Surveys, Amazon Mechanical Turk, and proprietary websites. N. EO. , K. SA. and B. JS [90] studied the impact of reviews on foodservice on Yelp.com, a business review site on foodborne illness surveillance efforts, and observed that sickness reports generated online could complement traditional surveillance systems. M. O. Lwin, S. Vijayakumar, O. Noel et al. [91] have collected the crowdsourced information regarding mosquito bites, symptoms, and suspected mosquito breeding sites and reported the same to help health authorities by early warning them of dengue outbreaks. A tool combining the use of crowdsourcing and text classification technique was proposed to track misinformation about the Zika virus on Twitter [92,93].

4.4. Other microblogs

A microblog represents a stream of text written by an individual comprising periodic and brief updates presented to the readers in reverse-chronological order. Compared with other social media, the information on microblogs is transferred in a truncated manner as the length of the post is limited. The short posts reduce users' time, and on an average, a microblogger can post many updates in a day. Various popular microblog services other than Twitter also exist. For example, Tumblr, Reddit, Sina Weibo, Pinterest, etc. N. Yang, X. Cui, C. Hu C et al. [30] analyzed the content of a famous Chinese social medium (Sina micro-blog) and then used them to predict the flu outbreak in a region of Beijing. To forecast seasonal flu, Z. Ertem, D. Raymond and L.A. Meyers collected data from WordPress flu-related blogs (WordPress Flu), along with other online data sources such as Wikipedia and Twitter [94]. Most of the papers that have collected data from microblogs, other than Twitter, used Chinese microblogs such as [95] (Sina), [96](Sina), [97] (Sina, Tencent), [98] (Sina).

4.5. Internet search query

One of the most well-known sources of data for Internet-based surveillance in the field of healthcare is Internet Search Query analysis, especially Google Trends (GT). Google Flu Trends was one of the earliest examples, which started providing real-time data to the public in 2008. It observed flu outbreaks around the world, based on flu-related terms people searched on the Internet. It hailed in providing the data one to two weeks faster and almost as precise as the CDC's [99]. J.D. Sharpe, R.S. Hopkins, R.L. Cook et al. compared the predictive performance of Google, Wikipedia, and Twitter-based surveillance with each other. As a result, Google Flu Trends showed superiority in sensitivity rates and positive predictive values [100].

Interestingly, online search behaviour successfully predicted the sudden increment in asthma-related emergency visits [101]. Apart from using Twitter data, this paper has combined internet users' search interests from Google with the environmental data to collect information on asthma-related visits. Google Trends' data has also been utilized to track infectious diseases such as tuberculosis [102], influenza [103,104,105]. Some studies considered web-based search queries

[56,106,107] to analyze and predict the dissemination of an outbreak.

Some studies have searched other social media platforms for data collection. For example, S. Chaudhary and S. Naaz have accessed the digital health data for various diseases from Facebook pages of Practo, National Portal of India, and the Integrated Disease Surveillance Program (IDSP) [61]. Y. A. Strekalova [108] studied the Facebook users' characteristics, commenting about the Ebola outbreak on CDC's Facebook posts for more than seven months. S. Gittelman, V. Lange, C. Crawford et al. [109] employed Facebook “likes”, which proved to be an effective predictor of mortality, diseases, and lifestyle behaviors. YouTube is another popular source for analyzing videos on various health topics such as Ebola virus disease [110]. Another study analyzed public responses from YouTube videos related to the Zika virus to determine the video content [111]. S. Choi, J. Lee, M. Kang et al. [112] have mentioned that short-text comments on news articles are top-rated in Korea and behaves as a platform to express personal emotions and thoughts. They studied the public emotions using the comments on these news articles about the Middle East Respiratory Syndrome (MERS) outbreak.

Fig. 4 represents the most searched social media sources for data collection by the selected research articles, and it can be easily noted that Twitter is the most popular social media source for health-related data collection.

Fig. 5 shows the share of the health topic or diseases studied by the authors of the selected research papers with ILI or flu as the most common disease for research related experiments.

Table 1 represents the machine learning techniques adopted for classification of health-related text and also the social media sources that were used along with corresponding research papers from the set of the selected relevant articles.

5. RQ3: applications of social media based surveillance systems

This section discusses the various recent applications of surveillance systems in the area of health informatics. These include disease prediction, tracking misinformation, global awareness, etc.

5.1. Syndromic surveillance-based disease prediction

Syndromic surveillance has emerged as a potential tool to predict outbreaks for public health purposes through data gathered from various sources before clinically confirmed data is available. The desired result is to minimize the spread in the population and take preventive measures. In the past few years, social media information has been widely used to estimate disease incidences and to detect disease outbreaks. Such data would be beneficial for public health officials in earlier detection of outbreaks than the traditional methods. Usually, the data is in the form of self-reported symptoms. Studies have shown that surveillance systems in the healthcare domain can be used to predict the diseases for public health concerns. Twitter data was used as a tool for early warning and outbreak detection, such as to predict syphilis [113], swine flu [75], tuberculosis [102], flu [78], and Ebola [56]. Another study that was able to examine disease incidences of dengue and typhoid fever in a region was suggested by the Philippines [29]. Similarly, there is another study [41] where surveillance systems have used social media data for disease detection.

5.2. Magnitude estimation of disease over some time

Surveillance systems can be used to estimate the magnitude of the problem. The planning, resource allocation, treatments, and prevention can be done by estimating the future of disease levels. The analysis provided by the surveillance systems can be useful to determine the level of the disease over some time, and assessments can be made accordingly.

5.3. Event-based surveillance and disease prediction

Event-based surveillance involves fast capturing of data in an organized manner about events that are at potential risk to public health. The data can be from diverse Internet sources such as media reports, online discussion platforms, routine reporting systems, personal information, or rumors. In the web forum context, an event is characterized as excessive news postings. The significance of the event can be considered proportional to the number of postings about it.

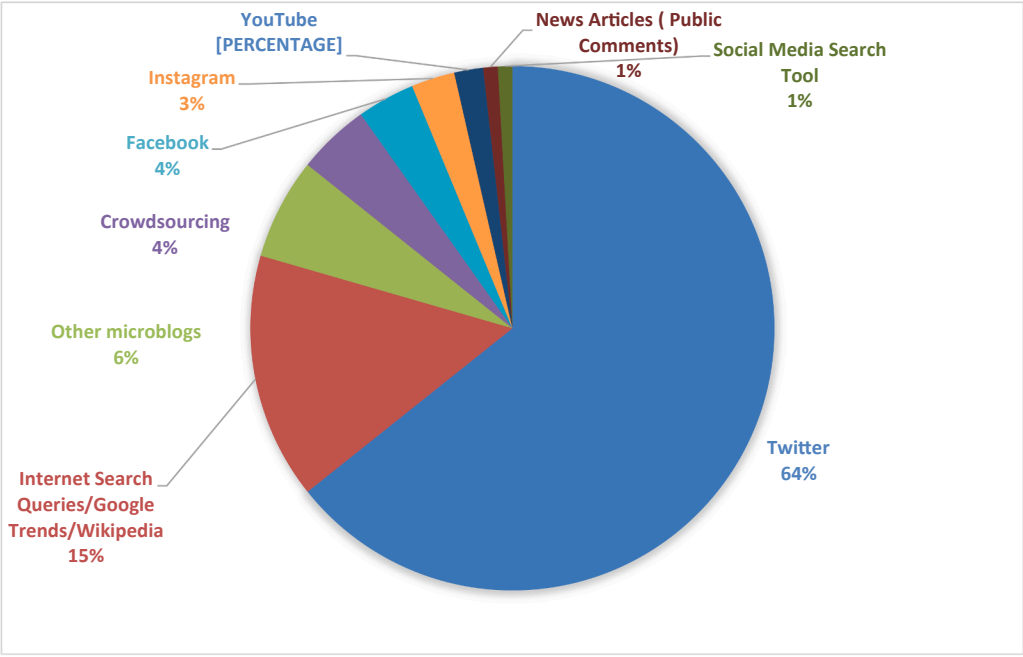


Fig. 4. Types of social media platforms searched for health-related data collection.

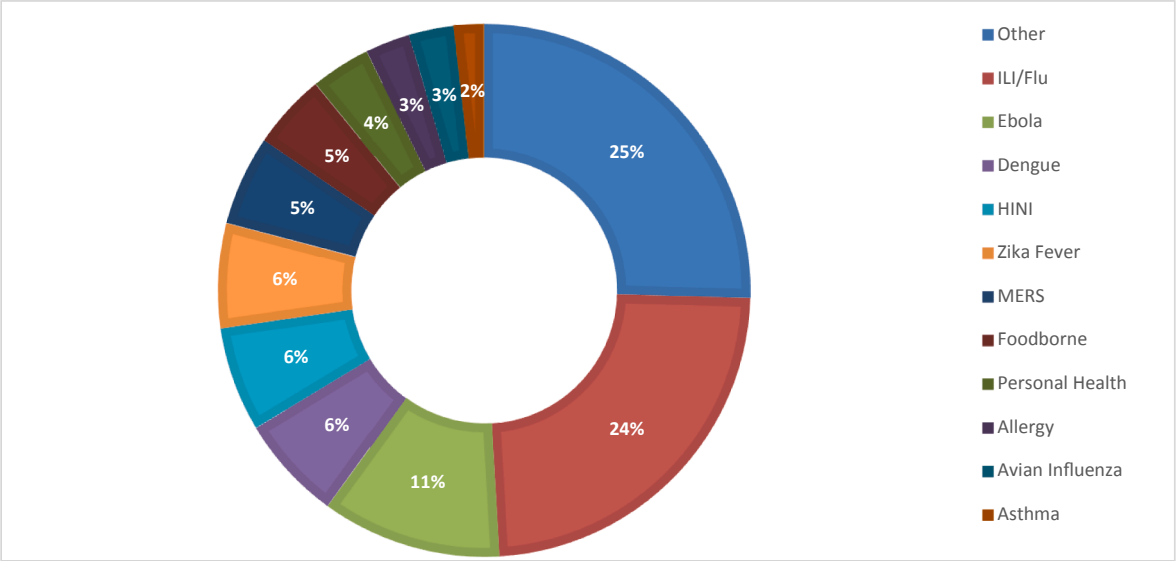


Fig. 5. Categorization of most common diseases on social media platforms.

Therefore, the event effect can be identified on topic diffusion from the amount of posting on the topic. When the number of postings on a topic surpasses the usual number of postings, we can expect that there is an event at that time. The well-known examples of event-based surveillance systems are HealthMap, a system that tracks news related to health events, EpiSPIDER, etc. The use of social media-based public health intelligence monitoring techniques to provide situational awareness of potential health threats to support surveillance activities has grown tremendously over the last few decades [58,14]. There is a study that analyzed the Zika virus epidemic using Twitter corpus [114]. N. Thapen, D. Simmie, C. Hankin et al. [39] proposed DEFENDER, a software system with potential health events detection functionality that monitors Twitter stream and then outputs the generated events to the users in Front-end UI.

5.4. Analyze user's reactions to health events:

Surveillance systems in the field of healthcare can be used to measure social media users' reactions to health promotion messages or events. T. Tran and K. Lee [115] have gathered 2 billion tweets in 90 languages from Twitter between August 2014 and December 2014 to understand citizen reactions towards Ebola by extracting geotagged Ebola-related tweets. Another study collected Twitter data to express emotions during different stages of an outgoing health event [50]. S. Choi, J. Lee, M. Kang et al. [112] analyzed the relationship between mass media and emotional public reactions during a nation-wide outbreak of MERS in 2015 in Korea. E. K. Seltzer, E. Horst-Martz, M. Lu et al. [88] proposed that public sentiments can be described using Instagram and also highlighted areas of concern for public health. Another study [116] that reflected the interests of Twitter users to share

Table 1
Summary of ML Classification Approaches used in Health Surveillance Systems based on Social Media.

Health Topic	ML Approach*	Social Media Data Source	Article Citation	Year
Chickenpox	DNN, RNN	Internet Search Query	[51]	2018
Measles	CNN, RNN, SVM, k-NN, NB, RF	Twitter	[50]	2018
Mosquito-borne diseases (Dengue, Zika, Chikungunya)	SVM, NB	Twitter	[24]	2018
Dengue	SVM, k-means	Twitter	[35]	2018
Personal-Health Experience	SVM, LR, k- NN, DT, DNN	Twitter	[23]	2017
MERS, Ebola	SVM,NB.LR	Twitter	[37]	2017
Personal-Health Experience	k- NN, DT, LR, SVM, DNN	Twitter	[33]	2017
Mosquito-borne diseases (Dengue, Malaria, Chikungunya)	SVM, NB	Twitter	[28]	2017
Flu	NB	Twitter	[13]	2017
Dengue, Typhoid	SVM	Twitter	[29]	2017
Influenza	ME	Twitter	[46]	2016
Dengue	SVM,NB,ME	Twitter	[40]	2016
Flu	SVM	Twitter	[38]	2016
Asthma	SVM, LR	Twitter	[32]	2016
Allergy	SVM, NB, NBM, k- NN	Twitter	[20]	2016
Influenza	SVM	Twitter	[34]	2016
Health-Related	SVM NB	Twitter	[39]	2016
Flu/Influenza, H1N1	LR	Twitter	[44]	2016
Influenza-A (H1N1)	NB, SVM,RF, DT	Twitter	[25]	2015
Allergy	NBM, NB, RF, SVM	Twitter	[21]	2015
Asthma	Artificial Neural Network (ANN)	Twitter	[101]	2015
ILI	SVM	Twitter	[73]	2014
Flu	k-means, k-NN, SVM	Sina Microblog	[30]	2014
Influenza	SVM	Twitter	[72]	2013
Personal health	NB, NBM, SVM	Twitter	[36]	2013
Health Related	SVM,NB,NBM,RF	Twitter	[31]	2013

*Abbreviations are defined in the body of the paper

Zika virus news such as symptoms, stories of Zika infected pregnant women, and worries of parents to be.

Moreover, R. Gaspar, S. Gorjão, B. Seibt et al. have shown that, during the EHEC/Escherichia coli bacteria outbreak in Europe in 2011, amplification of public reactions over Twitter has a severe political and economic impact [117]. K. Liu, L. Li, T. Jiang et al. analyzed variations in responses to different disease outbreaks by analyzing the public's search behavior [95]. In another study, I.C. Fung, K. Fu, Y. Ying et al. gained insight into Chinese people's reactions to different outbreaks by analyzing social media data [96].

5.5. Global awareness of events

Once the event has been detected, surveillance systems can be used to monitor the general public's awareness and perception towards health events. User-generated sentiments on social media platforms towards an outbreak situation reflect their knowledge, attitudes, and perception. Social media allow the sharing of public sentiments, opinions, and responses during outbreaks [118]. The high uncertainty related to Zika may not only lead individuals to social media to search for essential information [119], but also resulted in creating and spreading conspiracy theories.

5.6. Tracking misinformation

The capability of social media to provide the information in real-time also allows the fast flow of misinformation among people during an ongoing epidemic that can have severe reactions. However, in our knowledge, limited research has been done to detect the spread of misinformation about such events, yet few researchers have worked on widespread misinformation posted on social media platforms [98]. Ghenai A, Mejova Y. [92] have used machine learning techniques to track misinformation regarding the Zika virus on the Twitter platform. Also, another study [12] examined ways to correct misinformation about the measles vaccination by studying the reactions of two social media groups. Later, they also examined the effect of correcting the misinformation on the users who reacted to measles vaccination. There is a study [120] that presented an analysis of hoax medical news in social media. Another study based on YouTube data indicated that the spreading of both informational and conspiracy theory took place in a similar manner [111], which was a striking result for health organizations. To counter the spread of such misinformation, more attention is required towards the content posted on YouTube or any other online platform. Another study analyzed 1680 microblog posts and found that more than 20% of posts were misleading messages [97]. To fill the gap on how health organizations should get a response and correct misinformation, E. Hagg, V. S. Dahinten, and L. M. Currie [59] discussed that the potential for misinformation could be a barrier to use social media data in the area of healthcare.

5.7. Uncovering the topics popular during the outbreak

One more application of health-related surveillance systems can be to extract the topics popular during the outbreak. The data collected for the surveillance purpose can be used to obtain multiple concurrent events during the outbreak [121]. Even different topics, including government/politics and economy, were also captured to analyze public reactions during the MERS outbreak in Korea in 2015 [112]. K. Rudra, A. Sharma, N. Ganguly et al. mentioned that during epidemics, opinions, and sentiments related information is mostly contained in the non-disease tweets (tweets that do not convey any information about the epidemic situation) [37]. Similarly, data collected for public health surveillance purposes can be utilized to uncover ongoing topics during health events.

6. RQ4: limitations and challenges of social media based surveillance systems

This section aims to highlight some of the limitations and challenges faced in using social media data by surveillance systems.

While the surveillance systems based on social media data for outbreak detection or health events has led to the advancement in early detection of epidemics and associated events, still some studies have challenged the outcome of these surveillance systems for some of the following reasons:

6.1. Noise

One of the challenges faced while collecting the data is Noise. The data collected from social media sites may contain data that are irrelevant to the task. Such data referring to illness terms have no relation to health. For instance, a large number of activities related to flu may be caused by posts containing the term “Irish Flu.” Unfortunately, sometimes the user may post a status, and they are suspected to be infected when they are not. In this manner, such deceptive information can influence disease management for the public health department. To prevent such noise in data essential text processing techniques such as feature weighting, tokenization, stemming, frequency-based methods, etc. are available, still more training would be to get the relevant data for further analysis.

6.2. Data validation

Another challenge associated with social media data is its validation. The use of unofficial data from social media may lead to the problem of standardization, verification, and control [122]. T. Bodnar and M. Salathé focused on validating the ground truth associated with a large amount of heterogeneous social media data [123]. The dataset is one of the most critical components while creating prediction models. The outcomes of the prediction models highly depend on the dataset. The dataset for the prediction models includes historical information, training data, and testing data. A large amount of training data is required to train the forecasting models and testing data for valid testing of predictions based on the training of models. Hence, the correctness of data is essential as much as the volume and quality of data from reliable sources [124].

6.3. Low confidence

One more challenge that arises with social media-based data is low confidence. There is a study [125] showing that government websites are a more trustworthy resource for obtaining vaccination information than social media. Online data related to health is of varying quality. Many social media data analysis methods may show spikes showing something noticeable is going on, but that can be a reflection of panic and not the real incidences of a disease outbreak. Also, users may post that they have flu when they have a common cold, or people might talk about disease due to increased media hype. There is a study presenting conspiracy about the Zika virus outbreak on Reddit during a public health crisis [126]. Such behaviors are the reasons behind the low confidence associated with online data. To prevent this issue, more training would be required as an approach to modify the classification algorithms.

6.4. Demographic bias

Although information such as location, username, number of followers, user profiles, etc. can be collected along with the tweet content, demographic information such as age, gender, and race are hard to figure out through tweets, making it difficult to determine who is posting about the outbreak and to whom public health efforts should be

directed. Among the very few studies which have analyzed the users' profile [108], studied the Facebook comments in terms of sex and found that men composed a greater number of posts per person than women. Another bias is that younger people are the most common users of social media. Also, the bias is strongly supported by the observation that social media data are skewed towards active users who are often young adults, well-educated, and persons having a good income [40]. This paper used machine learning techniques to support that socio-demographic factors have a significant role in reporting disease-related posts on social media. Therefore, it is considered that social media users do not represent the whole population [127]. Any outcome based on social media data excludes people who do not use such platforms, who are probably going to be the most vulnerable in population, and even those who often do not want to share their health experiences publicly. Also, it can be speculated that those who are in discomfort, sick, elderly, or disabled would be less likely to be active users.

Language is another bias due to the demographics of a population. Most studies focus on a single language (English) even though the epicenter of the epidemic/event is a country where English is not an official language [128]. Very few studies have exhibited research relying on languages other than English to deal with disease tracking. For instance, M. U. Ilyas [129] has geographically filtered disease names in tweets both in Arabic and English language. Moreover, the author observed that tweets in English were in a very small fraction. Their approach was verified by having a high degree of correlation between the actual occurrence of MERS-Coronavirus cases and disease-related tweets in the Arabic language. Another study [83] has shown that there might be a difference in the reactions of people belonging to different linguo-cultural backgrounds towards the same outbreak.

6.5. Privacy issues

There is also discussion about ethical concerns while retrieving data from social media [130], including the privacy of the datasets collected using social media for health purposes [131]. Although there is the public availability of social media data, users may not want their posts or data to be used for research. Social media platform hosts must consider users' privacy expectations, such as diagnoses from public data or concern over algorithms that reveal unstated user demographics. Only a few studies [130,132] have shown interest so far. And hence, there is a broader scope for considering privacy while performing research on social media data.

6.6. Lexical and linguistic variability

Though communication over social media helps extract healthcare information, it is challenging in terms of semantic interpretation of the language. In particular, social media texts are informal and ambiguous, and hence, just matching keywords may not capture their semantic interpretation, resulting in achieving the incomplete result. N. Limsopatham and N. Collier [133] have discussed and researched this limitation.

7. Original contributions of the paper

Our paper has the following contributions that have not been considered carefully in the previous papers:

- Provides the article selection queries from different digital libraries databases for selecting the relevant articles.
- Presents an overview of popular ML classification algorithms in the context of social media based surveillance systems in the healthcare domain.
- Statistical analysis of most scrutinized social media platforms and health topics studied by the selected articles.
- Provides a summarized view of the most used online participatory media platforms for data collection, along with the publication year of the research paper, health topic studied, and machine learning methods used for analysis.

8. Findings, conclusions, and future directions

Among all the types of social media platforms studied by the selected articles, Fig. 4 has clearly shown that Twitter (64%) was undoubtedly the most searched platform. Also, SVM (33%) was the most used classification technique among all the recently used machine learning based classifiers. In addition, it was observed that SVM was the most promising classifier when the data needs to be classified between two classes.

This paper aims to study the latest trends in surveillance systems using social media and machine learning algorithms in the area of public health. We have noted that in comparison to traditional surveillance systems, social media based surveillance systems show superiority. Also, we have discussed the applications of using dynamic natured social media data for the advancement of surveillance systems in the area of public health. Also, we can expect the challenges mentioned in this paper will lead to the development of new technologies, new capabilities for public health research.

Some of the future works that we have observed can be:

(1) The conjunction of online data with other physical conditions like weather conditions, demographic information, etc. to get better prediction results.
(2) Incorporating the input features such as sentiment content, comments, locations, etc. of users' posts along with the text content for the better analysis and prediction of health events and other related events.
(3) Classification of users' posts into different categories such as health, news, ads, etc. and assigning different weights to different categories so that forecasting accuracy can be improved.
(4) The text analysis and image analysis approach can be extended to video content analysis to enhance the disease prediction performance.
(5) We observed there is a wide scope of improvement in the area of topic modeling to get more accurate results.
(6) Implementation of predictive models that can use various social media platforms simultaneously to get accurate and timely predictions of epidemics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Mathur, Surveillance systems for health care associated infections, J. Patient Saf. Infect. Control. 3 (2015) 4–11, https://doi.org/10.1016/j.jpsic.2015.02.002.
[2] K. Lee, Real-Time Disease Surveillance Using Twitter Data : Demonstration on Flu and Cancer, (n.d.) 1474–1477.
[3] R.W. Newkirk, J.B. Bender, C.W. Hedberg, The Potential Capability of Social Media as a Component of Food Safety and Food Terrorism Surveillance Systems, Foodborne Pathog. Dis. 9 (2012) 120–124, https://doi.org/10.1089/fpd.2011.0990.
[4] D.B. Neill, New directions in artificial intelligence for public health surveillance, IEEE Intell. Syst. 27 (2012) 56–59, https://doi.org/10.1109/MIS.2012.18.
[5] S.J. Yan, A.A. Chughtai, C.R. Macintyre, Utility and potential of rapid epidemic intelligence from internet-based sources, Int. J. Infect. Dis. 63 (2017) 77–87, https://doi.org/10.1016/j.ijid.2017.07.020.
[6] S.M. Noar, E. Leas, B.M. Althouse, M. Dredze, D. Kelley, J.W. Ayers, Can a selfie promote public engagement with skin cancer ? Prev. Med. (Baltim). (2017) 1–4, https://doi.org/10.1016/j.ypmed.2017.10.038.
[7] I.C.-H. Fung, Z.T.H. Tse, K.-W. Fu, The use of social media in public health surveillance, 6 (2015) 10–13. 10.5365/wpsar.2015.6.1.019.
[8] L. Mollema, I.A. Harmsen, E. Broekhuizen, R. Clijnk, H. De Melker, T. Paulussen,

G. Kok, R. Ruiter, E. Das, Disease Detection or Public Opinion Reflection ? Content Analysis of Tweets , Other Social Media , and Online Newspapers During the Measles Outbreak in the Netherlands in 2013, 17 (2013) 1–12. 10.2196/jmir.3863.

[9] S.J. Park, S. Hong, D. Kim, Y. Seo, J.H. Hur, W. Jin, D. Precision, Development of a Real-Time Stroke Detection System for Elderly Drivers Using Quad-Chamber Air Cushion and IoT Devices, (2018) 1–5. 10.4271/2018-01-0046.Abstract.

[10] I.C.H. Fung, C.H. Duke, K.C. Finch, K.R. Snook, P.L. Tseng, A.C. Hernandez, M. Gambhir, K.W. Fu, Z.T.H. Tse, Ebola virus disease and social media: A systematic review, Am. J. Infect. Control. 44 (2016) 1660–1671, https://doi.org/10.1016/j.ajic.2016.05.011.

[11] A. Mike, C. Daniel, Social media, big data, and mental health: current advances and ethical implications, Curr. Opin. Psychol. (2016), https://doi.org/10.1016/j.copsyc.2016.01.004.

[12] L. Tang, B. Bie, S.P. Ma, D. Zhi, Social media and outbreaks of emerging infectious diseases: A systematic review of literature, AJIC Am. J. Infect. Control. (2018), https://doi.org/10.1016/j.ajic.2018.02.010.

[13] X. Dai, M. Bikdash, B. Meyer, From Social Media to Public Health Surveillance: Word Embedding based Clustering Method for Twitter Classification, 2017.

[14] J. O'Shea, Digital disease detection: A systematic review of event-based internet biosurveillance systems, Int. J. Med. Inform. 101 (2017) 15–22, https://doi.org/10.1016/j.ijmedinf.2017.01.019.

[15] L. Fernandez-luque, M. Imran, Humanitarian Health Computing using Artificial Intelligence and Social Media: A Narrative Literature Review, Int. J. Med. Inform. (2018), https://doi.org/10.1016/j.ijmedinf.2018.01.015.

[16] A. Alessa, M. Faezipour, A review of influenza detection and prediction through social networking sites, (2018) 1–27. 10.1186/s12976-017-0074-5.

[17] H.A. Park, H. Jung, J. On, S.K. Park, H. Kang, Digital epidemiology: Use of digital data collected for non-epidemiological purposes in epidemiological studies, Healthc. Inform. Res. (2018), https://doi.org/10.4258/hir.2018.24.4.253.

[18] T. Eckmanns, K. Denecke, E. Velasco, T. Agheneza, G. Kirchner, Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review, Milbank Q. 92 (2014) 7–33, https://doi.org/10.1111/1468-0009.12038.

[19] M. Bates, Tracking Disease: Digital Epidemiology Offers New Promise in Predicting Outbreaks, IEEE Pulse. 8 (2017) 18–22, https://doi.org/10.1109/MPUL.2016.2627238.

[20] K. Nargund, S. Natarajan, Public health allergy surveillance using micro-blogs, in: 2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016. (2016) 1429–1433. 10.1109/ICACCI.2016.7732248.

[21] K. Lee, A. Agrawal, A. Choudhary, Datasets, Mining Social Media Streams to Improve Public Health Allergy Surveillance, (2015) 815–822.

[22] K. Lee, A. Agrawal, A. Choudhary, Forecasting Influenza Levels Using Real-Time Social Media Streams, in: Proc. - 2017 IEEE Int. Conf. Healthc. Informatics, ICHI 2017. (2017) 409–414. 10.1109/ICHI.2017.68.

[23] R.A. Calix, R. Gupta, M. Gupta, K. Jiang, Deep gramulator: Improving precision in the classification of personal health-experience tweets with deep learning, in: Proc. - 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017. 2017-Janua (2017) 1154–1159. 10.1109/BIBM.2017.8217820.

[24] L. Sousa, R. de Mello, D. Cedrim, A. Garcia, P. Missier, A. Uchôa, A. Oliveira, A. Romanovsky, VazaDengue: An information system for preventing and combating mosquito-borne diseases with social networks, Inf. Syst. 75 (2018) 26–42, https://doi.org/10.1016/j.is.2018.02.003.

[25] V. Kumar, S. Kumar, An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter, 70 (2015) 801–807. 10.1016/j.procs.2015.10.120.

[26] S. Saini, S. Kohli, Machine Learning Techniques for Effective Text Analysis of Social Network E-health Data, (2016) 3783–3788.

[27] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2011, (n.d.).

[28] V.K. Jain, S. Kumar, Effective surveillance and predictive mapping of mosquito-borne diseases using social media, J. Comput. Sci. 25 (2018) 406–415, https://doi.org/10.1016/j.jocs.2017.07.003.

[29] K. Espina, M. Regina, J.E. Estuar, Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines, Proc. Comput. Sci. 121 (2017) 554–561, https://doi.org/10.1016/j.procs.2017.11.073.

[30] N. Yang, X. Cui, C. Hu, W. Zhu, C. Yang, Chinese Social Media Analysis for Disease Surveillance, (2014) 17–21. 10.1109/IIKI.2014.11.

[31] S. Tuarob, C.S. Tucker, M. Salathe, N. Ram, Discovering Health-Related Knowledge in Social Media Using Ensembles of Heterogeneous Features, (2013) 1685–1690.

[32] W. Zhang, S. Ram, M. Burkart, Y. Pengetnze, Extracting Signals from Social Media for Chronic Disease Surveillance, (2016) 79–83. 10.1145/2896338.2897728.

[33] K. Jiang, R. Gupta, M. Gupta, R.A. Calix, G.R. Bernard, Identifying Personal Health Experience Tweets with Deep Neural Networks* HHS Public Access, Conf. Proc. IEEE Eng. Med. Biol. Soc. 2017 (2017) 1174–1177, https://doi.org/10.1109/EMBC.2017.8037039.

[34] S. Wakamiya, After the Boom No One Tweets : Microblog-based Influenza Detection Incorporating Indirect Information, (2016) 1–9.

[35] M.A. Carlos, M. Nogueira, R.J. Machado, Analysis of dengue outbreaks using big data analytics and social networks, in: 2017 4th Int. Conf. Syst. Informatics, ICSAI 2017. 2018-Janua (2018) 1592–1597. 10.1109/ICSAI.2017.8248538.

[36] X. Ji, S.A. Chun, J. Geller, Monitoring public health concerns using twitter sentiment classifications, in: Proc. - 2013 IEEE Int. Conf. Healthc. Informatics, ICHI 2013. (2013) 335–344. 10.1109/ICHI.2013.47.

[37] K. Rudra, A. Sharma, N. Ganguly, M. Imran, Classifying Information from Microblogs during Epidemics, in: Proc. 2017 Int. Conf. Digit. Heal. - DH '17. (2017) 104–108. 10.1145/3079452.3079491.

[38] C. Allen, M.H. Tsou, A. Aslam, A. Nagel, J.M. Gawron, Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza, PLoS One. 11 (2016) 1–10, https://doi.org/10.1371/journal.pone.0157734.

[39] N. Thapen, D. Simmie, C. Hankin, J. Gillard, DEFENDER : Detecting and Forecasting Epidemics Using Novel Data-Analytics for Enhanced Response, (2016) 1–19. 10.1371/journal.pone.0155417.

[40] E.O. Nsoesie, L. Flor, J. Hawkins, A. Maharana, T. Skotnes, F. Marinho, J.S. Brownstein, Social Media as a Sentinel for Disease Surveillance: What Does Sociodemographic Status Have to Do with It? PLoS Curr. (2016), https://doi.org/10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6.

[41] K. Byrd, A. Mansurov, O. Baysal, Mining Twitter data for influenza detection and surveillance, Proc. Int. Work. Softw. Eng. Healthc. Syst. - SEHS '16. (2016) 43–49, https://doi.org/10.1145/2897683.2897693.

[42] K. Koutroumbas, N. Kalouptsidis, Nearest neighbor pattern classification neural networks, (2002) 2911–2915. 10.1109/icnn.1994.374694.

[43] C.Y.J. Peng, K.L. Lee, G.M. Ingersoll, An introduction to logistic regression analysis and reporting, J. Educ. Res. 96 (2002) 3–14, https://doi.org/10.1080/00220670209598786.

[44] L. Zhao, J. Chen, F. Chen, W. Wang, C.T. Lu, N. Ramakrishnan, SimNest: Social media nested epidemic simulation via online semi-supervised deep learning, Proc. - IEEE Int. Conf. Data Mining, ICDM. 2016 (2016-Janua) 639–648, https://doi.org/10.1109/ICDM.2015.39.

[45] I. Korkontzelos, D. Piliouras, A.W. Dowsey, S. Ananiadou, Boosting drug named entity recognition using an aggregate classifier, Artif. Intell. Med. 65 (2015) 145–153, https://doi.org/10.1016/j.artmed.2015.05.007.

[46] J. Mowery, Twitter Influenza Surveillance: Quantifying Seasonal Misdiagnosis Patterns and their Impact on Surveillance Estimates, Online J Public Heal. Inf. (2016), https://doi.org/10.5210/ojphi.v8i3.7011.

[47] S. Rasoul Safavian, D. Landgrebe, A Survey of Decision Tree Classifier Methodology, 2017.

[48] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F.E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing. 234 (2017) 11–26, https://doi.org/10.1016/j.neucom.2016.12.038.

[49] J. Schmidhuber, Deep Learning in neural networks: An overview, Neural Networks. 61 (2015) 85–117, https://doi.org/10.1016/j.neunet.2014.09.003.

[50] J. Du, L. Tang, Y. Xiang, D. Zhi, J. Xu, H.Y. Song, C. Tao, Public perception analysis of tweets during the 2015 measles outbreak: Comparative study using convolutional neural network models, J. Med. Internet Res. 20 (2018) 1–11. 10.2196/jmir.9413.

[51] S. Chae, S. Kwon, D. Lee, Predicting Infectious Disease Using Deep Learning and Big Data, Int. J. Environ. Res. Public Health. (2018), https://doi.org/10.3390/ijerph15081596.

[52] L.E.O. Breiman, Random Forests LEO, (2001) 5–32. 10.1023/A:1010933404324.

[53] W.D. Jenkins, B. Wold, Use of the Internet for the surveillance and prevention of sexually transmitted diseases, Microbes Infect. 14 (2012) 427–437, https://doi.org/10.1016/j.micinf.2011.12.006.

[54] G.D. Haddow, K.S. Haddow, G.D. Haddow, K.S. Haddow, Chapter Eleven – Communicating During a Public Health Crisis, Disaster Commun. a Chang. Media World. (2014) 195–209. 10.1016/B978-0-12-407868-0.00011-2.

[55] K. Denecke, P. Dolog, P. Smrz, Making Use of Social Media Data in Public Health, (2012) 243–246.

[56] E. Yom-tov, Ebola data from the Internet: An Opportunity for Syndromic Surveillance or a News Event? Categories and Subject Descriptors, (n.d.) 115–119.

[57] T. Nguyen, M.E. Larsen, B.O. Dea, D.T. Nguyen, J. Yearwood, D. Phung, S. Venkatesh, H. Christensen, Kernel-based features for predicting population health indices from geocoded social media data, (2017). 10.1016/j.dss.2017.06.010.

[58] P. Kostkova, A Roadmap to Integrated Digital Public Health Surveillance : the Vision and the Challenges, (2013) 687–693.

[59] E. Hagg, V.S. Dahinten, L.M. Currie, The emerging use of social media for health-related purposes in low and middle-income countries: A scoping review, Int. J. Med. Inform. 115 (2018) 92–105, https://doi.org/10.1016/j.ijmedinf.2018.04.010.

[60] T.H. Van De Belt, P.T. Van Stockum, L.J.L.P.G. Engelen, J. Lancee, R. Schrijver, J. Rodríguez-baño, E. Tacconelli, K. Saris, M.M.H.J. Van Gelder, A. Voss, Social media posts and online search behaviour as early-warning system for MRSA outbreaks, (2018) 1–10.

[61] S. Chaudhary, S. Naaz, Use of Big Data in Computational Epidemiology for Public Health Surveillance, (2017) 150–155.

[62] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a Social Network or a News Media? Arch. Zootec. 60 (2011) 297–300, https://doi.org/10.4321/S0004-05922011000200015.

[63] A. Stefanidis, E. Vraga, G. Lamprianidis, J. Radzikowski, P.L. Delamater, K.H. Jacobsen, D. Pfoser, A. Croitoru, A. Crooks, Zika in Twitter: Temporal Variations of Locations , Actors , and Concepts Corresponding Author:, 3 (n.d.). 10.2196/publichealth.6925.

[64] J.C. Bosley, N.W. Zhao, S. Hill, F.S. Shofer, D.A. Asch, L.B. Becker, R.M. Merchant, Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication, Resuscitation 84 (2013) 206–212, https://doi.org/10.1016/j.resuscitation.2012.10.017.

[65] M.O. Edd, S.Y. Rn, What can we learn about the Ebola outbreak from tweets ? Am. J. Infect. Control. 43 (2015) 563–571, https://doi.org/10.1016/j.ajic.2015.02.023.

[66] N. El-bathy, C. Gloster, M. El-bathy, G. Stein, R. Stevenson, Intelligent Surveillance Lifecycle Architecture for Epidemiological Data Clustering using Twitter and Novel Genetic Algorithm, (2014) 149–155.

[67] Y. Khan, G.J. Leung, P. Belanger, E. Gournis, D.L. Buckeridge, L. Liu, Y. Li, I.L. Johnson, Comparing Twitter data to routine data sources in public health surveillance for the 2015 Pan / Parapan American Games : an ecological study, 2018.
[68] A. Veloso, F. Ferraz, Dengue surveillance based on a computational model of spatio-temporal locality of Twitter, 2011.
[69] J. Zaldumbide, R.O. Sinnott, Identification and Validation of Real-Time Health Events through Social Media, in: Proc. - 2015 IEEE Int. Conf. Data Sci. Data Intensive Syst. 8th IEEE Int. Conf. Cyber, Phys. Soc. Comput. 11th IEEE Int. Conf. Green Comput. Commun. 8th IEEE Inte. (2015) 9–16. 10.1109/DSDIS.2015.27.
[70] K. Talvis, K. Chorianopoulos, K.L. Kermanidis, Real-time monitoring of flu epidemics through linguistic and statistical analysis of twitter messages, in: Proc. - 9th Int. Work. Semant. Soc. Media Adapt. Pers. SMAP 2014. (2014) 83–87. 10.1109/SMAP.2014.38.
[71] V. Lampos, N. Cristianini, Tracking the flu pandemic by monitoring the social web, in: 2010 2nd Int. Work. Cogn. Inf. Process. CIP2010. (2010) 411–416. 10.1109/CIP.2010.5604088.
[72] 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 (2013), https://doi.org/10.1371/journal.pone.0083672.
[73] A.A. Aslam, M.-H. Tsou, B.H. Spitzberg, L. An, J.M. Gawron, D.K. Gupta, K.M. Peddecord, A.C. Nagel, C. Allen, J.-A. Yang, S. Lindsay, The reliability of tweets as a supplementary method of seasonal influenza surveillance, n.d.
[74] P. Kostkova, S. Garbin, J. Moser, W. Pan, Integration and Visualization Public Health Dashboard : The medi + board Pilot Project, (2014) 657–662.
[75] P. Kostkova, M. Szomszor, C. St, Luis, #Swineflu: The Use of Twitter as an EarlyWarning and Risk Communication, ACM Trans. Manag. Inf. Syst. 5 (2014) 1–25, https://doi.org/10.1145/2597892.
[76] C. Chew, G. Eysenbach, Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak, PLoS One. 5 (2010) 1–13. 10.1371/journal.pone.0014118.
[77] C. Study, E. Hus, E. Diaz-aviles, Tracking Twitter for Epidemic Intelligence Case Study: EHEC/HUS Outbreak in Germany, 2011, (2011) 82–85.
[78] L. Chen, K.S.M.T. Hossain, P. Butler, N. Ramakrishnan, B.A. Prakash, Flu Gone Viral : Syndromic Surveillance of Flu on Twitter using Temporal Topic Models, (2014) 755–760. 10.1109/ICDM.2014.137.
[79] C. Comito, C. Pizzuti, Twitter-based Influenza Surveillance : An Analysis of the 2016-2017 and 2017-2018 Seasons in Italy, 2018.
[80] B. Zou, R. Gorton, I.J. Cox, On Infectious Intestinal Disease Surveillance using Social Media Content, (2016) 157–161. 10.1145/2896338.2896372.2.1.
[81] W. Yang, L. Mu, GIS analysis of depression among Twitter users, Appl. Geogr. 60 (2015) 217–223, https://doi.org/10.1016/j.apgeog.2014.10.016.
[82] O.B. Da'ar, F. Yunus, N.Md. Hossain, M. Househ, Impact of Twitter intensity, time, and location on message lapse of bluebird's pursuit of fleas in Madagascar, J. Infect. Public Health. 10 (2017) 396–402, https://doi.org/10.1016/j.jiph.2016.06.011.
[83] I.C.H. Fung, J. Zeng, C.H. Chan, H. Liang, J. Yin, Z. Liu, Z.T.H. Tse, K.W. Fu, Twitter and Middle East respiratory syndrome, South Korea, 2015: A multi-lingual study, Infect. Dis. Heal. 23 (2018) 10–16. 10.1016/j.idh.2017.08.005.
[84] M.S. Deiner, T.M. Lietman, S.D. McLeod, J. Chodosh, T.C. Porco, Surveillance tools emerging from search engines and social media data for determining eye disease patterns, JAMA Ophthalmol. (2016), https://doi.org/10.1001/jamaophthalmol.2016.2267.
[85] A.B. Seidenberg, S.L. Pagoto, T.A. Vickey, E. Linos, M.R. Wehner, R.D. Costa, A.C. Geller, Tanning bed burns reported on Twitter: over 15,000 in 2013, Transl. Behav. Med. 6 (2016) 271–276. 10.1007/s13142-016-0388-6.
[86] K. Systrom, Strengthening Our Commitment to Safety and Kindness for 800 Million, 2017. < https://instagram.tumblr.com/post/165759350412/170926-news > .
[87] J.P.D. Guidry, Y. Jin, C.A. Orr, M. Messner, S. Meganck, Ebola on Instagram and Twitter: How health organizations address the health crisis in their social media engagement, Public Relat. Rev. 43 (2017) 477–486, https://doi.org/10.1016/j.pubrev.2017.04.009.
[88] E.K. Seltzer, E. Horst-Martz, M. Lu, R.M. Merchant, Public sentiment and discourse about Zika virus on Instagram, Public Health. 150 (2017) 170–175, https://doi.org/10.1016/j.puhe.2017.07.015.
[89] E.E. Arolas, F.G. Ladrón-de-Guevara, Towards an integrating crowdsourcing definition, 32 (2016) 189–200. 10.1177/0165551500000000.
[90] N. EO., K. SA., B. JS., Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports., Prev. Med. (Baltim). 67 (2014) 264–269. 10.1016/j.ypmed.2014.08.003.
[91] M.O. Lwin, S. Vijaykumar, O.N.N. Fernando, S.A. Cheong, V.S. Rathnayake, G. Lim, Y.L. Theng, S. Chaudhuri, S. Foo, A 21st century approach to tackling dengue: Crowdsourced surveillance, predictive mapping and tailored communication, Acta Trop. 130 (2014) 100–107, https://doi.org/10.1016/j.actatropica.2013.09.021.
[92] A. Ghenai, Y. Mejova, Catching Zika Fever : Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter, (2017) 5090. 10.1109/ICHI.2017.58.
[93] P. Quade, E.O. Nsoesie, P. Quade, A Platform for Crowdsourced Foodborne Illness Surveillance : Description of Users and Reports, 3 (2017) 1–9. 10.2196/publichealth.7076.
[94] Z. Ertem, D. Raymond, L.A. Meyers, Optimal multi-source forecasting of seasonal influenza, PLoS Comput. Biol. 14 (2018) 1–16, https://doi.org/10.1371/journal.pcbi.1006236.
[95] K. Liu, L. Li, T. Jiang, B. Chen, Z. Jiang, Z. Wang, Y. Chen, J. Jiang, H. Gu, Chinese public attention to the outbreak of ebola in west africa: Evidence from the online big data platform, Int. J. Environ. Res. Public Health. (2016), https://doi.org/10.3390/ijerph13080780.
[96] I.C. Fung, K. Fu, Y. Ying, B. Schaible, Y. Hao, C. Chan, Z.T.-H. Tse, Chinese social media reaction to the MERS-CoV and avian influenza A (H7N9) outbreaks, 2013.
[97] B. Chen, J. Shao, K. Liu, G. Cai, Z. Jiang, Y. Huang, H. Gu, J. Jiang, Does Eating Chicken Feet With Pickled Peppers Cause Avian Influenza? Observational Case Study on Chinese Social Media During the Avian Influenza A (H7N9) Outbreak, 4 (n.d.) 1–10. 10.2196/publichealth.8198.
[98] I.C.H. Fung, K.W. Fu, C.H. Chan, B.S.B. Chan, C.N. Cheung, T. Abraham, Z.T.H. Tse, Social media's initial reaction to information and misinformation on ebola, august 2014: Facts and rumors, Public Health Rep. (2016), https://doi.org/10.1177/003335491613100312.
[99] H.A. Carneiro, E. Mylonakis, Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks, Clin. Infect. Dis. 49 (2009) 1557–1564, https://doi.org/10.1086/630200.
[100] J.D. Sharpe, R.S. Hopkins, R.L. Cook, C.W. Striley, Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis, JMIR Public Heal. Surveill. 2 (2016) e161, , https://doi.org/10.2196/publichealth.5901.
[101] S. Ram, W. Zhang, M. Williams, Y. Pengetnze, Predicting asthma-related emergency department visits using big data, IEEE J. Biomed. Heal. Informatics. 19 (2015) 1216–1223, https://doi.org/10.1109/JBHI.2015.2404829.
[102] X. Zhou, J. Ye, Y. Feng, Tuberculosis surveillance by analyzing google trends, IEEE Trans. Biomed. Eng. 58 (2011) 2247–2254, https://doi.org/10.1109/TBME.2011.2132132.
[103] T.J. Bruno, K.H. Wertz, Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data, J. Chromatogr. A. 736 (1996) 175–184, https://doi.org/10.1109/BIBE.2015.7367640.
[104] H. Xue, Y. Bai, H. Hu, H. Liang, Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network, IEEE Access. 6 (2017) 563–575, https://doi.org/10.1109/ACCESS.2017.2771798.
[105] D.W. Seo, S.Y. Shin, Methods using social media and search queries to predict infectious disease outbreaks, Healthc. Inform. Res. (2017), https://doi.org/10.4258/hir.2017.23.4.343.
[106] Y. Luo, D. Zeng, Z. Cao, X. Zheng, Y. Wang, Q. Wang, H. Zhao, Using multi-source web data for epidemic surveillance: A case study of the 2009 Influenza A (H1N1) pandemic in Beijing, in: Proc. 2010 IEEE Int. Conf. Serv. Oper. Logist. Informatics, SOLI 2010. (2010) 76–81. 10.1109/SOLI.2010.5551614.
[107] C. CD, C. DJ, M. AR, S. KP, Using Web and Social Media for Influenza Surveillance, (2010) 531–535. 10.1007/978-1-4419-5913-3_61.
[108] Y.A. Strekalova, Emergent health risks and audience information engagement on social media, Am. J. Infect. Control. 44 (2016) 363–365, https://doi.org/10.1016/j.ajic.2015.09.024.
[109] S. Gittelman, V. Lange, C.A.G. Crawford, C.A. Okoro, E. Lieb, S.S. Dhingra, E. Trimarchi, A New Source of Data for Public Health Surveillance: Facebook Likes, 17 (n.d.) 1–10. 10.2196/jmir.3970.
[110] C.H. Basch, C.E. Basch, K.V. Ruggles, R. Hammond, Coverage of the Ebola Virus Disease Epidemic on YouTube, Disaster Med. Public Health Prep. (2015), https://doi.org/10.1017/dmp.2015.77.
[111] A. Nerghes, P. Kerkhof, I. Hellsten, Early Public Responses to the Zika-Virus on YouTube: Prevalence of and Differences Between Conspiracy Theory and Informational Videos, in: 10th ACM Conf. OnWeb Sci. (2018) 127–134. 10.1145/3201064.3201086.
[112] S. Choi, J. Lee, M. Kang, H. Min, Y. Chang, S. Yoon, Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks, Methods 129 (2017) 50–59, https://doi.org/10.1016/j.ymeth.2017.07.027.
[113] S.D. Young, N. Mercer, R.E. Weiss, E.A. Torrone, S.O. Aral, Using social media as a tool to predict syphilis, Prev. Med. (Baltim) 109 (2018) 58–61, https://doi.org/10.1016/j.ypmed.2017.12.016.
[114] D. Nolasco, J. Oliveira, Subevents Detection through Topic Modeling in Social Media Posts, Futur. Gener. Comput. Syst. (2018), https://doi.org/10.1016/j.future.2018.09.008.
[115] T. Tran, K. Lee, Understanding Citizen Reactions and Ebola-Related Information Propagation on Social Media, (2016) 106–111.
[116] K.W. Fu, H. Liang, N. Saroha, Z.T.H. Tse, P. Ip, I.C.H. Fung, How people react to Zika virus outbreaks on Twitter? A computational content analysis, Am. J. Infect. Control. 44 (2016) 1700–1702, https://doi.org/10.1016/j.ajic.2016.04.253.
[117] R. Gaspar, S. Gorjão, B. Seibt, L. Lima, J. Barnett, A. Moss, J. Wills, Tweeting during food crises: A psychosocial analysis of threat coping expressions in Spain, during the 2011 European EHEC outbreak, Int. J. Hum. Comput. Stud. 72 (2014) 239–254. 10.1016/j.ijhcs.2013.10.001.
[118] L. Tang, B. Bie, D. Zhi, Tweeting about measles during stages of an outbreak: A semantic network approach to the framing of an emerging infectious disease, Am. J. Infect. Control. 46 (2018) 1375–1380, https://doi.org/10.1016/j.ajic.2018.05.019.
[119] K.H. Pine, Y. Chen, W. Lafayette, Managing Uncertainty : Using Social Media for Risk Assessment during a Public Health, Crisis (2017).
[120] M.H. Purnomo, S. Sumpeno, E.I. Setiawan, D. Purwitasari, Keynote Speaker II: Biomedical Engineering Research in the Social Network Analysis Era: Stance Classification for Analysis of Hoax Medical News in Social Media, Proc. Comput. Sci. 116 (2017) 3–9, https://doi.org/10.1016/j.procs.2017.10.049.
[121] C. Robertson, L. Yee, Avian influenza risk surveillance in North America with online media, PLoS One. (2016), https://doi.org/10.1371/journal.pone.0165688.
[122] G. Blouin-Genest, A. Miller, The politics of participatory epidemiology: Technologies, social media and influenza surveillance in the US, Heal. Policy

Technol. 6 (2017) 192–197, https://doi.org/10.1016/j.hlpt.2017.02.001.
[123] T. Bodnar, M. Salathé, Validating Models for Disease Detection Using Twitter Regression on Tweet Count, (2012) 699–702.
[124] A.A. Bharambe, D.R. Kalbande, Techniques and Approaches for Disease Outbreak Prediction, (2016) 100–102. 10.1145/2909067.2909085.
[125] J.R. Cataldi, A.F. Dempsey, S.T. O'Leary, Measles, the media, and MMR: Impact of the 2014–15 measles outbreak, Vaccine. 34 (2016) 6375–6380, https://doi.org/10.1016/j.vaccine.2016.10.048.
[126] Y. Kou, X. Gui, Y. Chen, K. Pine, Conspiracy Talk on Social Media: Collective Sensemaking during a Public Health Crisis, Proc. ACM Human-Computer Interact. 1 (2017) 1–21, https://doi.org/10.1145/3134696.
[127] L.E. Charles-Smith, T.L. Reynolds, M.A. Cameron, M. Conway, E.H.Y. Lau, J.M. Olsen, J.A. Pavlin, M. Shigematsu, L.C. Streichert, K.J. Suda, C.D. Corley, Using social media for actionable disease surveillance and outbreak management: A systematic literature review, PLoS One. (2015), https://doi.org/10.1371/journal.pone.0139701.
[128] G. Barata, K. Shores, J.P. Alperin, Local chatter or international buzz? Language differences on posts about Zika research on Twitter and Facebook, PLoS One. (2018), https://doi.org/10.1371/journal.pone.0190482.
[129] M.U. Ilyas, Disease Tracking in GCC Region Using Arabic Language Tweets, (2018) 417–421.
[130] R. Mckee, Ethical issues in using social media for health and health care research, Health Policy (New. York) 110 (2013) 298–301, https://doi.org/10.1016/j.healthpol.2013.02.006.
[131] E.M. Eggleston, E.R. Weitzman, Innovative uses of electronic health records and social media for public health surveillance, Curr. Diab. Rep. (2014), https://doi.org/10.1007/s11892-013-0468-7.
[132] M.A. Mayer, L. Fernández-Luque, A. Leis, Big Data For Health Through Social Media, Elsevier Inc., 2016. 10.1016/B978-0-12-809269-9.00005-0.
[133] N. Limsopatham, N. Collier, Towards the Semantic Interpretation of Personal Health Messages from Social Media, (2015) 27–30.