PLOS ONE

RESEARCH ARTICLE

Surveillance of communicable diseases using social media: A systematic review

Patrick Pilipiec[1,2], **Isak Samsten**[1], **András Bota**[3]*

1 Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden, **2** School of Business and Economics, Maastricht University, Maastricht, The Netherlands, **3** Embedded Intelligent Systems Lab, Department of Computer Science Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden

* patrick.pilipiec@maastrichtuniversity.nl

OPEN ACCESS

Citation: Pilipiec P, Samsten I, Bota A (2023) Surveillance of communicable diseases using social media: A systematic review. PLoS ONE 18(2): e0282101. https://doi.org/10.1371/journal.pone.0282101

Editor: Luis M. Rocha, Binghamton University Thomas J Watson School of Engineering and Applied Science, UNITED STATES

Received: April 4, 2022

Accepted: February 7, 2023

Published: February 24, 2023

Data Availability Statement: All relevant data are within the paper and Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Background

Communicable diseases pose a severe threat to public health and economic growth. The traditional methods that are used for public health surveillance, however, involve many drawbacks, such as being labor intensive to operate and resulting in a lag between data collection and reporting. To effectively address the limitations of these traditional methods and to mitigate the adverse effects of these diseases, a proactive and real-time public health surveillance system is needed. Previous studies have indicated the usefulness of performing text mining on social media.

Objective

To conduct a systematic review of the literature that used textual content published to social media for the purpose of the surveillance and prediction of communicable diseases.

Methodology

Broad search queries were formulated and performed in four databases. Both journal articles and conference materials were included. The quality of the studies, operationalized as reliability and validity, was assessed. This qualitative systematic review was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Results

Twenty-three publications were included in this systematic review. All studies reported positive results for using textual social media content to surveille communicable diseases. Most studies used Twitter as a source for these data. Influenza was studied most frequently, while other communicable diseases received far less attention. Journal articles had a higher quality (reliability and validity) than conference papers. However, studies often failed to provide important information about procedures and implementation.

Conclusion

Text mining of health-related content published on social media can serve as a novel and powerful tool for the automated, real-time, and remote monitoring of public health and for the surveillance and prediction of communicable diseases in particular. This tool can address limitations related to traditional surveillance methods, and it has the potential to supplement traditional methods for public health surveillance.

1 Introduction

Communicable diseases are a severe threat to public health [1]. These infectious diseases include, among others, dengue, Ebola, malaria, measles, different strains of influenza, and Zika virus. In particular, influenza can result in respiratory symptoms of varying severity, and it can cause high mortality among the vulnerable population of older adults who also have a chronic condition related to the respiratory or immune system [2, 3].

Every year, seasonal influenza causes approximately half a million deaths globally [4, 5]. The 1918 Spanish flu pandemic was estimated to result in the mortality of 40 million people [6]. In addition, communicable diseases can also have catastrophic effects on the economy and society [7]. For example, seasonal influenza was estimated to have a financial burden of $83.3 billion annually in the United States alone [3].

Recently, the large disrupting effect of communicable diseases has again been observed with the outbreak of severe acute respiratory syndrome coronavirus 2 (hereafter SARS-CoV-2), which is assumed to have emerged in the city of Wuhan in China, after which it quickly spread around the globe [8]. Many governments chose a lockdown of society during the outbreak to increase control over this virus, to protect the healthcare system from overload, to mitigate the potential spread and to limit the number of casualties [9]. The International Monetary Fund (IMF) has indicated that this lockdown and the effects of SARS-CoV-2 resulted in a contraction of European economies, by on average, seven percent in 2020 [10].

This illustrates that the context of our society should also be considered. The ongoing fast-paced mobility of people requires a global system for the surveillance of communicable diseases, since the public health in one country can easily and rapidly be impacted that of by another country located on the other side of the planet.

Therefore, there is a need for public health authorities to detect outbreaks of communicable diseases as early as possible, to monitor these diseases and to initiate preventive measures immediately [11–13]. In addition, there exists an ongoing urgency to develop new technologies to forecast communicable disease outbreaks [14–17].

Early detection of communicable diseases is crucial to organize and allocate the required health resources, to control the spread of the disease and to avoid or mitigate further contamination [18]. This urgency is even more significant in the case of epidemic outbreaks, such as the novel SARS-CoV-2 [8], which demand real-time monitoring and rapid initiation of appropriate interventions [19].

However, the traditional methods for public health surveillance have many shortcomings, such as a lag between data collection and reporting [20–22]. To address these drawbacks, a proactive method is needed to automatically detect and monitor disease outbreaks worldwide in real time and to minimize any delays in this process [23, 24].

The emergence and widespread adoption of social media platforms has received a great amount of attention in the literature [25]. People share significant information on health-

related experiences on social media [26, 27]. Various studies have indicated that the analysis of health-related content published to social media has the potential to significantly improve the public health surveillance system [28–30]. In addition, in the preceding years, various studies have been published that utilized health-related textual content from social media for the purpose of public health surveillance of communicable diseases. Furthermore, three reviews [31–33] have been performed on the topic of internet-based public health surveillance, and while these reviews provide new insights about the vast opportunities of using social media content for public health surveillance, these reviews are not systematic reviews.

We acknowledge that only five systematic reviews have been conducted thus far that are somewhat related to this topic. First, Velasco et al. [34] found that although incorporating digital content as a source for public health surveillance has great potential, there is a reluctance among public health authorities to include this content in the systems for public health surveillance. Second, Charles-Smith et al. [35] found that analyzing content published to social media has the potential to increase public health, but they based their findings on only 10 publications. Third, Fung et al. [36] performed a systematic review of 12 studies that utilized social media content published during the 2014–2015 Ebola epidemic in West Africa, and they reported that no study evaluated their utility for any public health organization. The aforementioned systematic literature reviews were, among others, not tailored to communicable diseases [34, 35] and emphasized only one regional epidemic [36]. Fourth, Abad et al. [37] conducted a scoping review to summarize the literature on applications of natural language processing for digital public health surveillance, and they emphasized databases in the field of medicine and public health in their literature search strategy. Not including relevant databases in the field of computer science, data science, and information science is a limitation of their study. Fifth, Gupta and Katarya [38] performed a systematic review on the utilization of social media data in real-time public health surveillance systems, and they concluded that, compared to traditional methods, the analysis of social media data has increased the ability of these systems to predict diseases. However, two differences of their systematic review are that the literature search involved all types of artificial intelligence instead of focusing on the branch of natural language processing. As a consequence, they had a limited emphasis on the technical aspects of natural language processing in detail, such as explaining how preprocessing of natural language was performed and which methods and tools were used. We believe that a new systematic review is required that emphasizes the technical aspects of these applications of natural language processing for the surveillance of communicable diseases. In addition, considering the changes in social media use and advances in the respective fields of science in the foregoing half decade since the reviews above were published, some of these reviews may already be outdated.

Therefore, in this paper, we perform a new and thorough systematic literature review that investigates how textual content published to social media can be used for the purpose of the surveillance and prediction of communicable diseases. A systematic review of the evidence on this topic can greatly benefit public health authorities. In addition to evidence about the effectiveness of specific methods, this systematic review also provides a synthesis of the communicable diseases that were studied, social media platforms that were used, and which software and algorithms were utilized in these studies. If textual content from social media can indeed be used to surveille and predict outbreaks of communicable diseases, then such systems may become a powerful tool and asset for public health authorities and have the potential to address most of the limitations of the methods that are commonly used in traditional public health surveillance systems [28–30].

There is an opportunity to develop a proactive global public health surveillance system [23]. This tool should enable the automated and real-time monitoring of diseases worldwide by

including information from various novel sources containing contextual information about social media users while minimizing the overall processing time from data collection to the reporting of identified findings [24]. This tool could significantly benefit rapid and evidence-based decision-making regarding infectious disease outbreaks [24]. A systematic review could provide more insight into this opportunity.

2 Background

Public health surveillance, also called epidemiologic surveillance, involves the ongoing and systematic collection, management, and monitoring of data about diseases, with the purpose of identifying trends, e.g., [39–41]. The overall objective of public health surveillance is to detect outbreaks of diseases at the earliest possible time so that the required preparatory activities can be planned and performed and sufficient health resources can be allocated to enable high-quality and timely public health interventions intended to mitigate the disease [42, 43]. In addition, once the disease finally appears, the authorities, medical professionals, and the entire society can immediately initiate the planned remediating activities, facilitating an effective and prompt intervention. Therefore, public health surveillance is a crucial system for the identification, prevention, and control of disease outbreaks [44] while enabling a better allocation of health resources [18].

2.1 Traditional system for surveillance

In the traditional system for public health surveillance, the responsible public health authorities continuously collect data on diseases, which are primarily derived from diagnosed cases that are reported by emergency departments, hospitals, laboratories, and other medical professionals [1, 45]. The identified illnesses are predominantly observed from clinical data such as diagnoses and clinical reports [45, 46]. It has, however, been argued that these passive surveillance strategies fail to provide complete and timely overviews of the diseases [47].

In addition, historical data are analyzed to identify and visualize disease-related trends, such as seasonal influenza, which has often been occurring around the same months throughout the preceding decades, and may, therefore, be predicted with a reasonable accuracy [6, 48]. In contrast, other researchers report that the influenza virus continuously evolves into slightly different variations each year, which makes forecasting the timing of influenza outbreaks as well as their impacts on the society very difficult [49]. However, the emergence of many other infectious diseases cannot be forecasted based on historical data [47, 50].

2.2 Limitations of the traditional system for surveillance

The methods that are commonly used in the traditional system for public health surveillance have been practiced for many decades. Although these systems are known to improve public health and reduce mortality, there is no consensus on the degree of usefulness of individual methods or on the best way to support their function [51]. Likewise, other literature has reported that the authorities have been unable to successfully reduce the incidence and prevalence of dengue and other mosquito-related epidemics [52].

Overall, these systems involve two significant limitations.

First, a major drawback is that these methods are inefficient and time-consuming [1, 20]. To identify confirmed cases, the system requires lab work that is very labor intensive to operate and maintain, which significantly increases the time required to process the clinical data [6]. For example, in the United States, the time required to collect and analyze the data about seasonal influenza and to produce and distribute the reports was estimated to be two weeks [49]. As a consequence, once these reports are finally distributed to politicians, medical

professionals, and the general public, the reported findings are very likely to be outdated and may thus no longer accurately represent the current situation [53]. Therefore, these methods are not suited for the surveillance of novel infectious diseases such as SARS-CoV-2, which emerged in late 2019 [8] and involves an urgent need for real-time updates and demands immediate interventions [19]. While contact tracing is able to successfully trace infections, non-symptomatic and mild cases are nearly impossible to track and can easily enter other countries unnoticed.

Second, for communicable diseases such as malaria, disease trends can only be detected and analyzed after the actual outbreak of this disease [47, 50]. A severe limitation is that the outbreak and distribution of such diseases cannot be forecasted reliably [54].

Consequently, in the context of our highly dynamic society, the interconnectivity of all major cities by air travel leads to the very likely scenario that the outbreak of an infectious disease will easily spread around the globe in a matter of a few days, especially in non-symptomatic or mild cases [47, 55, 56].

2.3 Value of understanding text published to social media

Humans spend a significant amount of their time on social media communicating and disseminating information [4]. Social media platforms provide access to an abundance of valuable and public user-generated data that may be useful for public health surveillance and to detect, monitor, and prevent diseases [19, 45, 57, 58]. This makes social media platforms an important source for generating new knowledge [19].

A distinctive feature of social media is that it transforms its users into human sensors, although potentially biased and unreliable, who personally report on a variety of events and who may provide additional contextual information [6]. Furthermore, social media platforms often also collect geographical information about the precise locations of their users, which adds an additional and potentially valuable geographical dimension to these data [19, 59].

The analysis of textual content from social media is, however, not restricted to the field of diseases; an abundance of studies have used data from social media for application in many domains.

For example, user-generated content has been analyzed in a wide variety of domains, such as agriculture [60], business [61], and consumer behavior [62]; it has been used for purposes ranging from the detection of earthquakes [63], emergency and disaster management [64] to understanding migraines [65], presidential elections [66], political campaigns [67], and product design [68], to predicting the revenue of movies [69], forecasting sports events [70], identifying the topical interests of users [71], identifying trending topics [72], and investigating voting patterns in elections [73].

2.4 Natural language processing

The unstructured nature of social media content, compared to structured data, demands much more preprocessing and processing before it can be analyzed [50]. Most data that are generated today have an unstructured format, e.g., text [74]. Only a small fraction of data has a structured format, which can then be analyzed directly with well-established techniques from data mining [74].

In the preceding years, extensive techniques for processing human language have been developed and refined, and the relevant domain that emerged has been named natural language processing (hereafter NLP) [75, 76]. NLP resembles the science of using computers to understand human language, while text mining provides the required methods and algorithms.

The purpose of text mining is to "discover novel information in a timely manner from large-scale text collections by developing high performance algorithms for sourcing and converting unstructured textual data to a machine understandable format and then filtering this according to the needs of its users" [75]. Therefore, text mining is used for the automatic discovery of patterns, relationships, and high-quality insights from textual data [77, 78].

Among others, the domain of text mining includes the following major techniques [79–81]:

- extraction of concepts, entities, and the relationships between them [82];
- clustering text based on a measurement of similarity [83, 84];
- predicting words or other lexical units (as part of a word processor or chatbot) [85, 86];
- summarizing text in documents [82];
- discovering associations between words and other tokens [87];
- classification of text into various categories [78, 88]; and
- assigning affective states to text (sentiment analysis) [82].

These techniques are used abundantly among both researchers and professionals [50].

Sentiment analysis is a popular technique that is frequently used in the domain of text mining. Sentiment analysis involves the identification of attitudes, emotions, and opinions that people have in relation to an entity, which is observed from expressed human language [89, 90]. The opportunity derived from using sentiment analysis on content from social media is that it may enable innovative applications [20]. For example, sentiment analysis enables the identification of content as either a fact or an opinion (also called subjectivity). In addition, for opinions, sentiment analysis can also identify polarity, namely, whether an opinion is positive, neutral, or negative.

Furthermore, because text mining can be used to extract structured data from unstructured content, techniques used in data mining can subsequently be applied to analyze these structured features further [50]. NLP was used in medicine and public health [81] for, among others, allergies [91–93], depression [94–96], to gauge public health concerns [97], marijuana and drug abuse [98–101], obesity [102, 103], suicide-related thoughts and conversations [104–106], and tobacco and e-cigarette use [107–110].

3 Methodology

This qualitative systematic review was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [111, 112] (see S2 Appendix). However, most of the reviewed papers did not contain controlled trials, comparable statistical analysis, or comparable methodologies, making it impossible to apply the entire PRISMA checklist to this review. Therefore, we only applied items on the checklist if they were applicable, and thus, our review does not conform completely to the guidelines.

The following search strategy and procedures for study selection and analysis were used. The study selection, quality assessment of the included studies, and thematic analysis were performed by one author (PP). However, the procedures and findings were discussed by all authors, and potential disagreements were resolved by consensus.

3.1 Information sources

This systematic review is based on literature that was indexed by four large databases, namely, the ACM Digital Library, IEEE Xplore, PubMed, and Web of Science. These databases were selected because of their relevance to this topic.

The ACM Digital Library and IEEE Xplore databases were searched for publications in the fields of computer science, data science, information management, and information technology. IEEE Xplore was also selected because much research on this topic is exclusively published at conferences instead of in peer-reviewed journals. The Institute of Electrical and Electronics Engineering (IEEE) hosts many of these relevant conferences. Furthermore, PubMed was included because of its focus on literature in the domain of medicine and healthcare, while Web of Science is a very broad database that indexes the literature from many relevant disciplines, such as public policy and the social sciences.

3.2 Search strategy

An optimized and broad search strategy was formulated for each of the four databases (see S1 Appendix). Overall, the search strategy consisted of two blocks with search terms related to natural language processing and public health monitoring. In addition, database-specific filters were applied to narrow the search results further.

The first block, natural language processing, contained the search terms artificial intelligence, machine learning, text mining, computational linguistics, natural language processing, sentiment analysis, word embeddings, and Natural Language Toolkit. Abbreviations and wildcards were included to find alternative phrasing of these concepts. The OR operator was used to combine these search terms.

The second block, public health monitoring, contained the search terms public health surveillance, public health monitoring, and health monitoring. Experimental searches have indicated that these broader search terms resulted in the most relevant results. The OR operator was used to combine these search terms.

If supported by the database, subject headings such as MeSH terms for PubMed were also included in the search strategy. Subsequently, the AND operator was used to combine the queries from each block into the final search query.

The literature search was performed in March 2020. After executing the formulated search queries in each database, additional filters were manually applied to narrow the search results further. Although the precise filters were different across the databases, two examples of such filters are that publications were only written in the English language and that these studies were published in journals or presented at conferences.

For each of the four databases, all search results were then exported and subsequently imported into the same EndNote Library. Because these databases partially returned the same results, the deduplication strategy by Bramer et al. [113] was used to eliminate these duplicate publications from the EndNote Library. Consequently, the EndNote Library contained only unique results.

3.3 Process of study selection

The remaining publications were screened and selected using three subsequent phases based on their title, abstract, and full text. To avoid erroneously excluding publications, the screening in these phases was performed with high flexibility. Therefore, if there was any doubt concerning a publication's eligibility or when insufficient information was provided to confidently exclude a manuscript, that publication was retained for further screening in a subsequent phase.

In the first phase, the titles of these publications were screened for their relevance to the topic of this systematic review. The titles of eligible studies indicated the analysis of textual content for the surveillance or monitoring of diseases.

In the second phase, the abstract and keywords of the remaining studies were screened for information indicating the analysis of textual content that was generated by users and published to at least social media, with the purpose of public health surveillance and monitoring of communicable diseases. As a result, studies that only analyzed news articles were considered irrelevant and were eliminated.

Finally, the third phase involved rigorous screening of the full text of the remaining publications. Eligible studies reported original and empirical research analyzing the textual content that the general public published to at least social media, with the purpose of surveilling and monitoring public health with respect to communicable diseases. This phase did not discriminate between geographies, social media platforms, or communicable diseases. However, publications that only investigated non-communicable diseases were eliminated. When studies investigated communicable diseases, this systematic review did not discriminate between the type of disease, i.e., all communicable diseases were included in this systematic review.

This resulted in a remaining subset of the identified publications that was included for further selection in this systematic review.

3.4 Selection criteria

Overall, eligible publications reported original and empirical research that reported findings on the application of analyzing user-generated textual content from social media for the monitoring and prediction of communicable diseases. Reviews, discussion papers, editorials, and papers that only proposed a framework for the analysis of social media content without the actual application and reporting of these findings were eliminated. All peer-reviewed journal articles and publications related to conferences were included.

In addition, although studies were considered relevant if they included textual content that was published to at least social media, this systematic review did not discriminate between the different social media platforms. All social media platforms were considered relevant and were included. Likewise, this systematic review included all papers irrespective of the language of the social media content used, the geography of these users and their content, or the authors of the identified publications.

This study aimed to aggregate the reported findings on the surveillance and monitoring of public health based on the experiences of the population that were published on social media. Therefore, papers were excluded if they only included content that was published on social media by authors other than the general public, such as governments, health professionals, and commercial entities.

3.5 Data analysis

In accordance with Kampmeijer et al. [114] and utilizing the process described by Pilipiec et al. [81], the included studies were first assessed according to their quality, which was operationalized as reliability and validity. A reliable study provided a thorough and complete description of the methods that were used for the data collection and data analysis, and this process was also considered repeatable [114]. A valid study reported results that were consistent with the research objective and the utilized research methods [114]. An ordinal scale was used to grade studies with respect to their reliability and validity as either low, medium, or high. Regardless of the quality level, all studies were included in this systematic review.

Directed qualitative content analysis, also called thematic analysis, was used to analyze the included studies [115]. Thematic analysis is a primary method for qualitative research that is widely used among qualitative researchers [116, 117]. Its popularity may be explained because thematic analysis is a highly flexible method that can produce trustworthy insights [116, 118].

The themes of interest were based on the objective of this systematic review. The following themes were extracted from these publications: authors, year of publication, publication type, name of communicable disease, social media platform used, sample size, language of the data, period of data collection, horizon of data collection, country, software used for natural language processing, methods and techniques used for natural language processing, investigated target, algorithm used for predicting the target, reported result, description of the results, reliability, and validity.

The extracted information from all included publications was used to create an extraction matrix. The results were summarized using tables, and a synthesis of this information was presented narratively. In addition to assessing the quality of the studies that were included in this systematic review, the PRISMA checklist [111, 112] in S2 Appendix was used to assess the quality of this systematic review.

4 Results

The flow diagram in Fig 1 presents the results of the studies that were selected to be included in this systematic review. The execution of the optimized search queries in the four databases (see S1 Appendix) yielded a total of 5,318 hits. Of these results, 250 records were identified through the ACM Digital Library, 2,549 records were found in IEEE Xplore, PubMed yielded 226 records, and Web of Science returned 2,293 records. However, 744 records were identified as duplicates and were, therefore, removed. This resulted in an EndNote Library with 4,574 unique records.

Subsequently, screening was performed in three consecutive phases to exclude irrelevant records, according to the process described in Sections 3.3 and 3.4.

In the first phase, 4,347 records were removed after screening the title, resulting in 227 remaining records. In the second phase, the records were screened based on their abstracts and keywords. The 191 records that were considered irrelevant were eliminated. This resulted in 36 remaining studies. In the third phase, the full texts of the records were screened. However, the full texts of two records could not be retrieved, and these studies were subsequently removed. Of these records, 11 records were considered not to be relevant and were excluded. This resulted in the identification of 23 eligible publications that were included in this systematic review. A detailed description of the characteristics of these studies is presented in S3 Appendix.

4.1 Study characteristics

Table 1 presents an extensive description of the studies that were included in this systematic review. All studies were published between 2010 and 2019. A majority of these studies (65.2%) were published in the last five years [6, 30, 44, 45, 53, 58, 74, 76, 119–125]. Most studies were published in 2015 (17.4%) [119–122] and 2016 (17.4%) [58, 123–125], while no studies were published in 2012.

A majority of the studies (65.2%) were published as a peer-reviewed journal article [6, 28, 30, 44, 45, 49, 53, 58, 74, 119, 120, 122, 123, 125, 126], while the remaining 34.8 percent of the studies were published at a conference [4, 18, 24, 76, 121, 124, 127, 128].

The included studies investigated a total of seven communicable diseases, and publications may have reported findings on multiple diseases. Influenza was studied most frequently (61.5%) [4, 6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 121–124, 126]. This was followed by dengue [76, 125, 127] and measles [24, 74, 119], which were each studied in 11.5 percent of the included studies. Ebola [120], HIV/AIDS [128], listeria [24], and tuberculosis [24] were studied least often (3.9% each). Only one study investigated more than one communicable disease; this study analyzed four diseases (i.e., influenza, listeria, measles, and tuberculosis) [24].

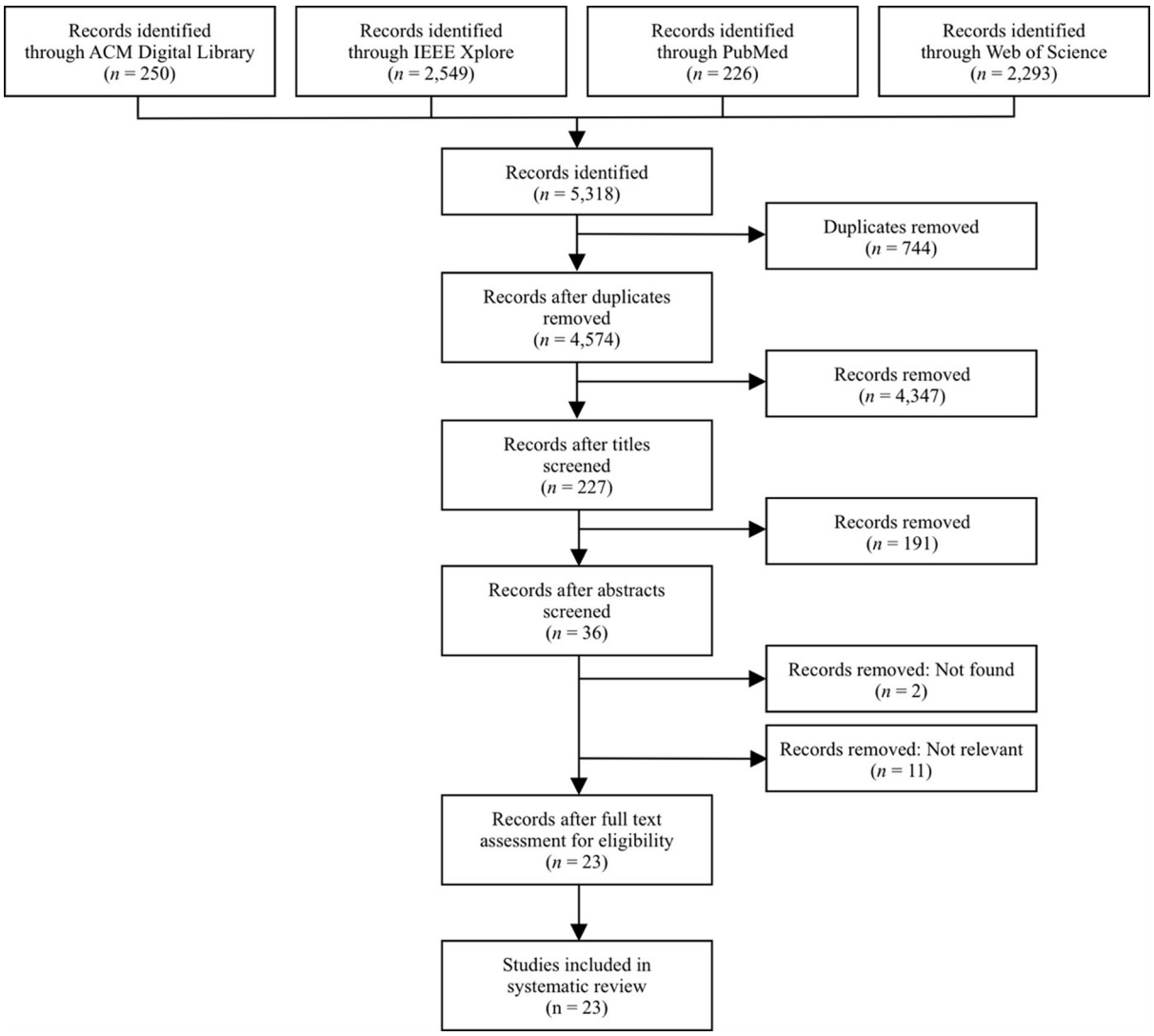

Fig 1. Results of study selection.

https://doi.org/10.1371/journal.pone.0282101.g001

The results in Table 1 for the input sources, employed methods, and study effectiveness are discussed in the subsequent subsections.

4.2 Input sources

User-generated textual content was retrieved from three social media platforms (see Table 1). Content published to Twitter was used most frequently (87.0%) [6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 74, 76, 119–121, 123–127]. The platforms Sina Weibo (8.7%) [4, 122] and Yahoo! Knowledge (4.4%) [128] were studied the least. All studies only included content from one social media platform.

There was a vast difference in the sample size that was included in the studies. This sample size ranged from 667 tweets [76] to 171,027,275 tweets [53]. Overall, in most studies, the sample size was either less than 25,000 (34.8%) [28, 45, 74, 76, 119, 124, 125, 128] or one million or more (30.4%) [4, 6, 18, 30, 53, 123, 126]. In 26.1 percent of the studies, the sample size was

Table 1. Description of studies analyzed (23 studies included).

Category	Sub-category	*n* (%)	Publications
Year	2010	2 (8.7)	[126, 128]
	2011	1 (4.4)	[127]
	2012	0 (0.0)	-
	2013	2 (8.7)	[18, 24]
	2014	3 (13.0)	[4, 28, 49]
	2015	4 (17.4)	[119–122]
	2016	4 (17.4)	[58, 123–125]
	2017	2 (8.7)	[44, 53]
	2018	2 (8.7)	[30, 45]
	2019	3 (13.0)	[6, 74, 76]
Publication type	Conference proceeding	8 (34.8)	[4, 18, 24, 76, 121, 124, 127, 128]
	Journal article	15 (65.2)	[6, 28, 30, 44, 45, 49, 53, 58, 74, 119, 120, 122, 123, 125, 126]
Communicable disease	Dengue	3 (11.5)	[76, 125, 127]
	Ebola	1 (3.9)	[120]
	HIV/AIDS	1 (3.9)	[128]
	Influenza	16 (61.5)	[4, 6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 121–124, 126]
	Listeria	1 (3.9)	[24]
	Measles	3 (11.5)	[24, 74, 119]
	Tuberculosis	1 (3.9)	[24]
Social media platform	Sina Weibo	2 (8.7)	[4, 122]
	Twitter	20 (87.0)	[6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 74, 76, 119–121, 123–127]
	Yahoo! Knowledge	1 (4.4)	[128]
Sample size	Less than 25,000	8 (34.8)	[28, 45, 74, 76, 119, 124, 125, 128]
	25,000 to 99,999	3 (13.0)	[44, 120, 121]
	100,000 to 249,999	1 (4.4)	[49]
	250,000 to 999,999	2 (8.7)	[122, 127]
	1,000,000 or more	7 (30.4)	[4, 6, 18, 30, 53, 123, 126]
	Unknown	2 (8.7)	[24, 58]
Language of data	Arabic	1 (4.2)	[6]
	English	5 (20.8)	[6, 44, 120, 124, 126]
	Japanese	1 (4.2)	[30]
	Mandarin	2 (8.3)	[4, 122]
	Unknown	15 (62.5)	[18, 24, 28, 45, 49, 53, 58, 74, 76, 119, 121, 123, 125, 127, 128]
Horizon of data collection	Less than 1 month	2 (8.7)	[74, 120]
	1 to 6 months	9 (39.1)	[4, 6, 18, 24, 28, 44, 49, 122, 124]
	7 to 12 months	3 (13.0)	[119, 126, 127]
	13 to 18 months	1 (4.4)	[121]
	19 to 24 months	1 (4.4)	[123]
	25 or more months	5 (21.7)	[30, 53, 76, 125, 128]
	Unknown	2 (8.7)	[45, 58]
Country	Australia	1 (4.2)	[121]
	Brazil	3 (12.5)	[76, 125, 127]
	Canada	1 (4.2)	[124]
	China	2 (8.3)	[4, 122]

(*Continued*)

Table 1. (Continued)

Category	Sub-category	*n* (%)	Publications
	India	1 (4.2)	[45]
	Japan	1 (4.2)	[30]
	New Zealand	1 (4.2)	[121]
	Taiwan	1 (4.2)	[128]
	The Netherlands	1 (4.2)	[119]
	United Arab Emirates	1 (4.2)	[6]
	United States	6 (25.0)	[24, 28, 44, 49, 53, 58]
	Unknown	5 (20.8)	[18, 74, 120, 123, 126]
Software for NLP	Apache Lucene's PorterStemFilter	1 (3.7)	[124]
	Apache Lucene's StopFilter	1 (3.7)	[124]
	Datasift service	1 (3.7)	[123]
	Natural Language Toolkit	1 (3.7)	[6]
	OpenNLP	1 (3.7)	[124]
	Stanford CoreNLP	2 (7.4)	[44, 124]
	The Stanford parser	1 (3.7)	[44]
	Unknown	19 (70.4)	[4, 18, 24, 28, 30, 45, 49, 53, 58, 74, 76, 119–122, 125–128]
Processing for NLP	Content analysis	1 (1.7)	[120]
	Detecting URLs	1 (1.7)	[124]
	Dimensionality reduction	1 (1.7)	[45]
	Feature weighting	1 (1.7)	[45]
	Homogenization	2 (3.3)	[74, 124]
	Language categorization	1 (1.7)	[6]
	LDA topics	1 (1.7)	[53]
	Lemmatization	3 (5.0)	[44, 45, 123]
	n-gram generation	6 (10.0)	[24, 44, 53, 74, 76, 120]
	Normalization using frequency-based methods	1 (1.7)	[45]
	Remove symbols and URLs	1 (1.7)	[120]
	Sentiment analysis	5 (8.3)	[24, 119, 124, 126, 127]
	Stemming	7 (11.7)	[6, 24, 45, 74, 76, 123, 124]
	Stop word removal	6 (10.0)	[6, 18, 45, 74, 76, 124]
	Term filtering	1 (1.7)	[74]
	Term Frequency—Inverse Document Frequency (TF-IDF)	8 (13.3)	[4, 24, 44, 49, 53, 58, 74, 76]
	Text embeddings	1 (1.7)	[53]
	Thematic analysis	1 (1.7)	[119]
	Topic detection	1 (1.7)	[120]
	Tokenization	4 (6.7)	[6, 18, 45, 74]
	Tweet filtering	1 (1.7)	[6]
	Unknown	6 (10.0)	[28, 30, 121, 122, 125, 128]
Algorithm for prediction of target	1-gram Term Frequency classifier	1 (2.4)	[44]
	Association rule mapping	1 (2.4)	[127]
	Chi-Square test	1 (2.4)	[126]
	Correlation analysis	2 (4.9)	[53, 119]
	Decision Tree	2 (4.9)	[45, 76]
	Fuzzy Algorithm for Extraction, Monitoring and Classification of infectious Diseases (FAEMC-ID)	1 (2.4)	[74]

(*Continued*)

Table 1. (Continued)

Category	Sub-category	*n* (%)	Publications
	Hidden Markov Model	1 (2.4)	[123]
	k-Means clustering	2 (4.9)	[4, 120]
	k-Nearest Neighbors	1 (2.4)	[4]
	Latent Dirichlet allocation (LDA)	1 (2.4)	[44]
	Linear regression	5 (12.2)	[6, 28, 53, 125, 127]
	Maximum entropy	1 (2.4)	[125]
	Naïve Bayes	5 (12.2)	[24, 45, 76, 124, 125]
	Random Forest	1 (2.4)	[45]
	Recurrent neural networks with Long short-term memory (LSTM)	1 (2.4)	[53]
	ST-DBSCAN	1 (2.4)	[127]
	Support vector machines	10 (24.4)	[4, 24, 30, 44, 45, 49, 53, 58, 125, 128]
	Time series	1 (2.4)	[120]
	Unknown	3 (7.3)	[18, 121, 122]
Result	Negative	0 (0.0)	-
	Positive	23 (100.0)	[4, 6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 74, 76, 119–128]
Reliability	Low	3 (13.0)	[18, 122, 128]
	Medium	16 (69.6)	[4, 24, 28, 30, 45, 49, 53, 58, 74, 76, 120, 121, 124–127]
	High	4 (17.4)	[6, 44, 119, 123]
Validity	Low	3 (13.0)	[18, 122, 128]
	Medium	16 (69.6)	[4, 24, 28, 30, 45, 49, 53, 58, 74, 76, 120, 121, 124–127]
	High	4 (17.4)	[6, 44, 119, 123]

https://doi.org/10.1371/journal.pone.0282101.t001

between 25,000 and 999,999 items [44, 49, 120–122, 127]. However, 8.7 percent of the studies failed to report the sample size [24, 58].

The studies investigated user-generated textual content that was written in different languages. The content was most often written in English (20.8%) [6, 44, 120, 124, 126]. Content written in Mandarin (8.3%) [4, 122], Arabic (4.2%) [6], and Japanese (4.2%) [30] was included less frequently. Only one study (4.4%) investigated content that was written in more than one language, namely, in Arabic and English [6]. However, a vast majority of the studies (65.2%) failed to report the language of the content that was analyzed [18, 24, 28, 45, 49, 53, 58, 74, 76, 119, 121, 123, 125, 127, 128].

The time horizon with respect to the date of publication of the analyzed data was also diverse, ranging from one week [74, 120] to 106 months [76]. However, most of the studies (39.1%) analyzed samples that were published in periods ranging from one to six months [4, 6, 18, 24, 28, 44, 49, 122, 124]. Additionally, more than one-fifth of the studies (21.7%) analyzed content that was published during a period of at least 25 months [30, 53, 76, 125, 128]. Only two studies (8.7%) included content that was published during a period less than one month [74, 120]. However, two studies (8.7%) did not disclose the precise time horizon for the publication dates of the included samples [45, 58].

The included studies analyzed content related to 11 countries. Posts published in the United States were analyzed most often (25.0%) [24, 28, 44, 49, 53, 58], followed by those published in Brazil (12.5%) [76, 125, 127] and China (8.3%) [4, 122]. The remaining countries are Australia [121], Canada [124], India [45], Japan [30], New Zealand [121], Taiwan [128], the Netherlands

[119], and the United Arab Emirates [6], which each were studied in 4.2 percent of the included studies. Five studies (21.7%), however, failed to disclose the geographical locations of the included posts [18, 74, 120, 123, 126]. With the exception of Robinson et al. [121], who analyzed posts from Australia and New Zealand, the remaining studies included content from only one country.

4.3 Employed methods

In the forthcoming synthesis of the methods that publications employed, the reader should be aware that our objective was to investigate the methods that authors utilized and explicitly mentioned in their manuscript. We acknowledge the possibility that authors applied common methods for text analysis, such as stop word removal, tokenization, stemming, and lemmatization, but failed to report this in their manuscript. This may be explained by the fact that these preprocessing methods are highly common in natural language processing.

Although all studies analyzed textual content using some variant of natural language processing, a majority of the studies (82.6%) failed to disclose information on the software that was used [4, 18, 24, 28, 30, 45, 49, 53, 58, 74, 76, 119–122, 125–128] (see Table 1). Only four studies (17.4%) provided information about the utilized software [6, 44, 123, 124]. When studies reported the software utilized, seven software packages were discussed. Stanford CoreNLP was used most often (7.4%) [44, 124], while Apache Lucene's PorterStemFilter [124], Apache Lucene's StopFilter [124], Datasift service [123], Natural Language Toolkit [6], OpenNLP [124], and The Stanford parser [44] were used the least (3.7% each). Additionally, studies could utilize more than one software package. Byrd et al. [124] used four software packages for natural language processing.

The studies reported 21 methods and algorithms for natural language preprocessing. Term Frequency—Inverse Document Frequency (TF-IDF) was used most often (13.3%) [4, 24, 44, 49, 53, 58, 74, 76]. This was followed by stemming (11.7%) [6, 24, 45, 74, 76, 123, 124], *n*-gram generation (10.0%) [24, 44, 53, 74, 76, 120], stop word removal (10.0%) [6, 18, 45, 74, 76, 124], sentiment analysis (8.3%) [24, 119, 124, 126, 127], tokenization (6.7%) [6, 18, 45, 74], and lemmatization (5.0%) [44, 45, 123]. The remaining 14 methods and algorithms were used in 25.0 percent of the studies. Although the majority of studies reported detailed information about these methods and algorithms, more than a quarter (26.1%) of the included studies did not disclose this information [28, 30, 121, 122, 125, 128].

A vast majority of the studies (87.0%) reported information on the algorithms that were used to predict the target, i.e., the outcome estimated using the textual content. A total of 18 algorithms were utilized. Support vector machines (24.4%) [4, 24, 30, 44, 45, 49, 53, 58, 125, 128], linear regression (12.2%) [6, 28, 53, 125, 127], and Naïve Bayes (12.2%) [24, 45, 76, 124, 125] were used most often. These three supervised learning algorithms are highly popular among data mining practitioners. Therefore, their utilization was expected for the prediction of a numerical outcome or a category.

Although a vast majority of studies disclosed information on the algorithm used, 13.0 percent of the studies [18, 121, 122] did not provide such information.

4.4 Study effectiveness

All studies reported positive results on using user-generated textual content from social media to monitor or surveille communicable diseases (see Table 1). Although positive findings were reported, it was explicitly discussed in one study that lower educated males of older age are less likely to disclose information on the infectious disease dengue to Twitter, making this platform less suitable for the monitoring and surveillance of this disease among this group of persons [125].

Furthermore, the quality of the included studies was evaluated based on its reliability and validity. A majority of studies [4, 24, 28, 30, 45, 49, 53, 58, 74, 76, 120, 121, 124–127] had medium reliability (69.6%) and validity (69.6%). Four studies [6, 44, 119, 123] were found to have high reliability (17.4%) and validity (17.4%). This means that these studies not only provided a complete description of the methods that were used for the data collection and data analysis and that this process was considered repeatable, but these studies also reported results that are consistent with the research objective and the utilized research methods [114]. For the remaining studies [18, 122, 128], the reliability (13.0%) and validity (13.0%) were, however, low.

4.5 Analysis of publications by publication type

Furthermore, the included publications were additionally analyzed based on the publication type, i.e., conference proceedings and journal articles (see Table 2). Of these publications, eight studies (34.8%) were presented at a conference [4, 18, 24, 76, 121, 124, 127, 128], and 15 studies (65.2%) were published in a peer-reviewed journal [6, 28, 30, 44, 45, 49, 53, 58, 74, 119, 120, 122, 123, 125, 126]. Overall, the analysis indicates that both conference proceedings and journal articles reported comparable findings.

However, there are a few notable and novel differences regarding the following themes: the type of communicable disease, social media platform, geographical locations of included samples, and the quality of these studies, which was operationalized as reliability and validity.

There were no notable differences between the studies with respect to the communicable diseases that were investigated. In both conference proceedings and journal articles, there was a strong emphasis on monitoring and surveilling influenza [4, 6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 121–124, 126]. However, several diseases were only investigated in journal articles (i.e., Ebola [120]), while HIV/AIDS [128], listeria [24], and tuberculosis [24] were only investigated in conference proceedings.

In addition, both conference proceedings and journal articles placed a strong emphasis on the social media platform Twitter [6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 74, 76, 119–121, 123–127]. Sina Weibo [4, 122], however, received the least attention in both types of publications. Additionally, Yahoo! Knowledge [128] was only studied in one conference proceeding but not in a journal article.

Although both conference proceedings and journal articles investigated content that was published in various countries, journal articles relatively more often included textual content that was published in the United States [28, 44, 49, 53, 58]. However, journal articles were also more likely to lack a disclosure of geographical information [74, 120, 123, 126]. There were, however, no notable differences between the continents.

Last, only journal articles were evaluated as having high reliability and high validity [6, 44, 119, 123]. No conference proceedings were assessed as high on these themes. In contrast, conference proceedings were more likely to have low reliability and low validity [18, 128] relative to journal articles [122]. There were no notable differences between conference proceedings and journal articles that were assessed as having a medium quality. Overall, journal articles, therefore, had a higher quality than conference proceedings.

4.6 Analysis of publications by social media platform

In addition to the analyses above, the included publications were also analyzed based on the social media platforms from which the content was extracted (see Table 3). These social media platforms are Sina Weibo, Twitter, and Yahoo! Knowledge. Overall, comparable findings were reported across the groups of the literature that utilized content from each of the social media platforms.

Table 2. Description of studies analyzed by publication type (23 studies included).

Category	Sub-category	Conference proceeding *n* (%)	Journal article *n* (%)	Publications
Year	2010	1 (4.4)	1 (4.6)	[126, 128]
	2011	1 (4.4)	0 (0.0)	[127]
	2012	0 (0.0)	0 (0.0)	-
	2013	2 (8.7)	0 (0.0)	[18, 24]
	2014	1 (4.4)	2 (8.7)	[4, 28, 49]
	2015	1 (4.4)	3 (13.0)	[119–122]
	2016	1 (4.4)	3 (13.0)	[58, 123–125]
	2017	0 (0.0)	2 (8.7)	[44, 53]
	2018	0 (0.0)	2 (8.7)	[30, 45]
	2019	1 (4.4)	2 (8.7)	[6, 74, 76]
Communicable disease	Dengue	2 (7.7)	1 (3.9)	[76, 125, 127]
	Ebola	0 (0.0)	1 (3.9)	[120]
	HIV/AIDS	1 (3.9)	0 (0.0)	[128]
	Influenza	5 (19.2)	11 (42.3)	[4, 6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 121–124, 126]
	Listeria	1 (3.9)	0 (0.0)	[24]
	Measles	1 (3.9)	2 (7.7)	[24, 74, 119]
	Tuberculosis	1 (3.9)	0 (0.0)	[24]
Social media platform	Sina Weibo	1 (4.4)	1 (4.4)	[4, 122]
	Twitter	6 (26.1)	14 (60.9)	[6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 74, 76, 119–121, 123–127]
	Yahoo! Knowledge	1 (4.4)	0 (0.0)	[128]
Sample size	Less than 25,000	3 (13.0)	5 (21.7)	[28, 45, 74, 76, 119, 124, 125, 128]
	25,000 to 99,999	1 (4.4)	2 (8.7)	[44, 120, 121]
	100,000 to 249,999	0 (0.0)	1 (4.4)	[49]
	250,000 to 999,999	1 (4.4)	1 (4.4)	[122, 127]
	1,000,000 or more	2 (8.7)	5 (21.7)	[4, 6, 18, 30, 53, 123, 126]
	Unknown	1 (4.4)	1 (4.4)	[24, 58]
Language of data	Arabic	0 (0.0)	1 (4.2)	[6]
	English	1 (4.2)	4 (16.7)	[6, 44, 120, 124, 126]
	Japanese	0 (0.0)	1 (4.2)	[30]
	Mandarin	1 (4.2)	1 (4.2)	[4, 122]
	Unknown	6 (25.0)	9 (37.5)	[18, 24, 28, 45, 49, 53, 58, 74, 76, 119, 121, 123, 125, 127, 128]
Horizon of data collection	Less than 1 month	0 (0.0)	2 (8.7)	[74, 120]
	1 to 6 months	4 (17.4)	5 (21.7)	[4, 6, 18, 24, 28, 44, 49, 122, 124]
	7 to 12 months	1 (4.4)	2 (8.7)	[119, 126, 127]
	13 to 18 months	1 (4.4)	0 (0.0)	[121]
	19 to 24 months	0 (0.0)	1 (4.4)	[123]
	25 or more months	2 (8.7)	3 (13.0)	[30, 53, 76, 125, 128]
	Unknown	0 (0.0)	2 (8.7)	[45, 58]
Country	Australia	1 (4.2)	0 (0.0)	[121]
	Brazil	2 (8.3)	1 (4.2)	[76, 125, 127]
	Canada	1 (4.2)	0 (0.0)	[124]
	China	1 (4.2)	1 (4.2)	[4, 122]
	India	0 (0.0)	1 (4.2)	[45]

(*Continued*)

Table 2. (Continued)

Category	Sub-category	Conference proceeding *n* (%)	Journal article *n* (%)	Publications
	Japan	0 (0.0)	1 (4.2)	[30]
	New Zealand	1 (4.2)	0 (0.0)	[121]
	Taiwan	1 (4.2)	0 (0.0)	[128]
	The Netherlands	0 (0.0)	1 (4.2)	[119]
	United Arab Emirates	0 (0.0)	1 (4.2)	[6]
	United States	1 (4.2)	5 (20.8)	[24, 28, 44, 49, 53, 58]
	Unknown	1 (4.2)	4 (16.7)	[18, 74, 120, 123, 126]
Software for NLP	Apache Lucene's PorterStemFilter	1 (3.7)	0 (0.0)	[124]
	Apache Lucene's StopFilter	1 (3.7)	0 (0.0)	[124]
	Datasift service	0 (0.0)	1 (3.7)	[123]
	Natural Language Toolkit	0 (0.0)	1 (3.7)	[6]
	OpenNLP	1 (3.7)	0 (0.0)	[124]
	Stanford CoreNLP	1 (3.7)	1 (3.7)	[44, 124]
	The Stanford parser	0 (0.0)	1 (3.7)	[44]
	Unknown	7 (25.9)	12 (44.4)	[4, 18, 24, 28, 30, 45, 49, 53, 58, 74, 76, 119–122, 125–128]
Processing for NLP	Content analysis	0 (0.0)	1 (1.7)	[120]
	Detecting URLs	1 (1.7)	0 (0.0)	[124]
	Dimensionality reduction	0 (0.0)	1 (1.7)	[45]
	Feature weighting	0 (0.0)	1 (1.7)	[45]
	Homogenization	1 (1.7)	1 (1.7)	[74, 124]
	Language categorization	0 (0.0)	1 (1.7)	[6]
	LDA topics	0 (0.0)	1 (1.7)	[53]
	Lemmatization	0 (0.0)	3 (5.0)	[44, 45, 123]
	n-gram generation	2 (3.3)	4 (6.7)	[24, 44, 53, 74, 76, 120]
	Normalization using frequency-based methods	0 (0.0)	1 (1.7)	[45]
	Remove symbols and URLs	0 (0.0)	1 (1.7)	[120]
	Sentiment analysis	3 (5.0)	2 (3.3)	[24, 119, 124, 126, 127]
	Stemming	3 (5.0)	4 (6.7)	[6, 24, 45, 74, 76, 123, 124]
	Stop word removal	3 (5.0)	3 (5.0)	[6, 18, 45, 74, 76, 124]
	Term filtering	0 (0.0)	1 (1.7)	[74]
	Term Frequency—Inverse Document Frequency (TF-IDF)	3 (5.0)	5 (8.3)	[4, 24, 44, 49, 53, 58, 74, 76]
	Text embeddings	0 (0.0)	1 (1.7)	[53]
	Thematic analysis	0 (0.0)	1 (1.7)	[119]
	Topic detection	0 (0.0)	1 (1.7)	[120]
	Tokenization	1 (1.7)	3 (5.0)	[6, 18, 45, 74]
	Tweet filtering	0 (0.0)	1 (1.7)	[6]
	Unknown	2 (3.3)	4 (6.7)	[28, 30, 121, 122, 125, 128]
Algorithm for prediction of target	1-gram Term Frequency classifier	0 (0.0)	1 (2.4)	[44]
	Association rule mapping	1 (2.4)	0 (0.0)	[127]
	Chi-Square test	0 (0.0)	1 (2.4)	[126]
	Correlation analysis	0 (0.0)	2 (4.9)	[53, 119]
	Decision Tree	1 (2.4)	1 (2.4)	[45, 76]
	Fuzzy Algorithm for Extraction, Monitoring and Classification of infectious Diseases (FAEMC-ID)	0 (0.0)	1 (2.4)	[74]

(*Continued*)

Table 2. (Continued)

Category	Sub-category	Conference proceeding *n* (%)	Journal article *n* (%)	Publications
	Hidden Markov Model	0 (0.0)	1 (2.4)	[123]
	k-Means clustering	1 (2.4)	1 (2.4)	[4, 120]
	k-Nearest Neighbors	1 (2.4)	0 (0.0)	[4]
	Latent Dirichlet allocation (LDA)	0 (0.0)	1 (2.4)	[44]
	Linear regression	1 (2.4)	4 (9.8)	[6, 28, 53, 125, 127]
	Maximum entropy	0 (0.0)	1 (2.4)	[125]
	Naïve Bayes	3 (7.3)	2 (4.9)	[24, 45, 76, 124, 125]
	Random Forest	0 (0.0)	1 (2.4)	[45]
	Recurrent neural networks with Long short-term memory (LSTM)	0 (0.0)	1 (2.4)	[53]
	ST-DBSCAN	1 (2.4)	0 (0.0)	[127]
	Support vector machines	3 (7.3)	7 (17.1)	[4, 24, 30, 44, 45, 49, 53, 58, 125, 128]
	Time series	0 (0.0)	1 (2.4)	[120]
	Unknown	2 (4.9)	1 (2.4)	[18, 121, 122]
Result	Negative	0 (0.0)	0 (0.0)	-
	Positive	8 (34.8)	15 (65.2)	[4, 6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 74, 76, 119–128]
Reliability	Low	2 (8.7)	1 (4.4)	[18, 122, 128]
	Medium	6 (26.1)	10 (43.5)	[4, 24, 28, 30, 45, 49, 53, 58, 74, 76, 120, 121, 124–127]
	High	0 (0.0)	4 (17.4)	[6, 44, 119, 123]
Validity	Low	2 (8.7)	1 (4.4)	[18, 122, 128]
	Medium	6 (26.1)	10 (43.5)	[4, 24, 28, 30, 45, 49, 53, 58, 74, 76, 120, 121, 124–127]
	High	0 (0.0)	4 (17.4)	[6, 44, 119, 123]

https://doi.org/10.1371/journal.pone.0282101.t002

Despite the overall comparability of the results, there are several notable and novel differences for the following themes: type of communicable disease and the quality of studies. The latter was operationalized as reliability and validity.

First, the studies that analyzed content from Twitter were most likely to investigate the communicable disease influenza (53.9%) [6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 121, 123, 124, 126], followed by dengue (11.5%) [76, 125, 127] and measles (11.5%) [24, 74, 119]. Ebola [120], listeria [24], and tuberculosis [24] received far less attention (3.9% each), while HIV/AIDS was not at all investigated using content from Twitter.

Second, the quality of the included studies, which was measured as reliability and validity, overall was higher for publications that utilized content from Twitter. More specifically, half of the studies that used Sina Weibo [122] and all studies that utilized Yahoo! Knowledge [128] had low reliability and low validity, and relatively more studies that used Twitter were medium quality [24, 28, 30, 45, 49, 53, 58, 74, 76, 120, 121, 124–127]. In addition, all studies that were evaluated to be high quality also analyzed content from Twitter [6, 44, 119, 123].

5 Discussion

Overall, our results indicate that textual content from social media can be used reliably to monitor and surveille communicable diseases and to predict the trends of these diseases. This consistency of the evidence indicates that text mining of social media content may be a

Table 3. Description of studies analyzed by social media platform (23 studies included).

Category	Sub-category	Sina Weibo *n* (%)	Twitter *n* (%)	Yahoo! Knowledge *n* (%)	Publications
Year	2010	0 (0.0)	1 (4.4)	1 (4.4)	[126, 128]
	2011	0 (0.0)	1 (4.4)	0 (0.0)	[127]
	2012	0 (0.0)	0 (0.0)	0 (0.0)	-
	2013	0 (0.0)	2 (8.7)	0 (0.0)	[18, 24]
	2014	1 (4.4)	2 (8.7)	0 (0.0)	[4, 28, 49]
	2015	1 (4.4)	3 (13.0)	0 (0.0)	[119–122]
	2016	0 (0.0)	4 (17.4)	0 (0.0)	[58, 123–125]
	2017	0 (0.0)	2 (8.7)	0 (0.0)	[44, 53]
	2018	0 (0.0)	2 (8.7)	0 (0.0)	[30, 45]
	2019	0 (0.0)	3 (13.0)	0 (0.0)	[6, 74, 76]
Publication type	Conference proceeding	1 (4.4)	6 (26.1)	1 (4.4)	[4, 18, 24, 76, 121, 124, 127, 128]
	Journal article	1 (4.4)	14 (60.9)	0 (0.0)	[6, 28, 30, 44, 45, 49, 53, 58, 74, 119, 120, 122, 123, 125, 126]
Communicable disease	Dengue	0 (0.0)	3 (11.5)	0 (0.0)	[76, 125, 127]
	Ebola	0 (0.0)	1 (3.9)	0 (0.0)	[120]
	HIV/AIDS	0 (0.0)	0 (0.0)	1 (3.9)	[128]
	Influenza	2 (7.7)	14 (53.9)	0 (0.0)	[4, 6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 121–124, 126]
	Listeria	0 (0.0)	1 (3.9)	0 (0.0)	[24]
	Measles	0 (0.0)	3 (11.5)	0 (0.0)	[24, 74, 119]
	Tuberculosis	0 (0.0)	1 (3.9)	0 (0.0)	[24]
Sample size	Less than 25,000	0 (0.0)	7 (30.4)	1 (4.4)	[28, 45, 74, 76, 119, 124, 125, 128]
	25,000 to 99,999	0 (0.0)	3 (13.0)	0 (0.0)	[44, 120, 121]
	100,000 to 249,999	0 (0.0)	1 (4.4)	0 (0.0)	[49]
	250,000 to 999,999	1 (4.4)	1 (4.4)	0 (0.0)	[122, 127]
	1,000,000 or more	1 (4.4)	6 (26.1)	0 (0.0)	[4, 6, 18, 30, 53, 123, 126]
	Unknown	0 (0.0)	2 (8.7)	0 (0.0)	[24, 58]
Language of data	Arabic	0 (0.0)	1 (4.2)	0 (0.0)	[6]
	English	0 (0.0)	5 (20.8)	0 (0.0)	[6, 44, 120, 124, 126]
	Japanese	0 (0.0)	1 (4.2)	0 (0.0)	[30]
	Mandarin	2 (8.3)	0 (0.0)	0 (0.0)	[4, 122]
	Unknown	0 (0.0)	14 (58.3)	1 (4.2)	[18, 24, 28, 45, 49, 53, 58, 74, 76, 119, 121, 123, 125, 127, 128]
Horizon of data collection	Less than 1 month	0 (0.0)	2 (8.7)	0 (0.0)	[74, 120]
	1 to 6 months	2 (8.7)	7 (30.4)	0 (0.0)	[4, 6, 18, 24, 28, 44, 49, 122, 124]
	7 to 12 months	0 (0.0)	3 (13.0)	0 (0.0)	[119, 126, 127]
	13 to 18 months	0 (0.0)	1 (4.4)	0 (0.0)	[121]
	19 to 24 months	0 (0.0)	1 (4.4)	0 (0.0)	[123]
	25 or more months	0 (0.0)	4 (17.4)	1 (4.4)	[30, 53, 76, 125, 128]
	Unknown	0 (0.0)	2 (8.7)	0 (0.0)	[45, 58]
Country	Australia	0 (0.0)	1 (4.2)	0 (0.0)	[121]
	Brazil	0 (0.0)	3 (12.5)	0 (0.0)	[76, 125, 127]
	Canada	0 (0.0)	1 (4.2)	0 (0.0)	[124]
	China	2 (8.3)	0 (0.0)	0 (0.0)	[4, 122]
	India	0 (0.0)	1 (4.2)	0 (0.0)	[45]
	Japan	0 (0.0)	1 (4.2)	0 (0.0)	[30]

(*Continued*)

Table 3. (Continued)

Category	Sub-category	Sina Weibo *n* (%)	Twitter *n* (%)	Yahoo! Knowledge *n* (%)	Publications
	New Zealand	0 (0.0)	1 (4.2)	0 (0.0)	[121]
	Taiwan	0 (0.0)	0 (0.0)	1 (4.2)	[128]
	The Netherlands	0 (0.0)	1 (4.2)	0 (0.0)	[119]
	United Arab Emirates	0 (0.0)	1 (4.2)	0 (0.0)	[6]
	United States	0 (0.0)	6 (25.0)	0 (0.0)	[24, 28, 44, 49, 53, 58]
	Unknown	0 (0.0)	5 (20.8)	0 (0.0)	[18, 74, 120, 123, 126]
Software for NLP	Apache Lucene's PorterStemFilter	0 (0.0)	1 (3.7)	0 (0.0)	[124]
	Apache Lucene's StopFilter	0 (0.0)	1 (3.7)	0 (0.0)	[124]
	Datasift service	0 (0.0)	1 (3.7)	0 (0.0)	[123]
	Natural Language Toolkit	0 (0.0)	1 (3.7)	0 (0.0)	[6]
	OpenNLP	0 (0.0)	1 (3.7)	0 (0.0)	[124]
	Stanford CoreNLP	0 (0.0)	2 (7.4)	0 (0.0)	[44, 124]
	The Stanford parser	0 (0.0)	1 (3.7)	0 (0.0)	[44]
	Unknown	2 (7.4)	16 (59.3)	1 (3.7)	[4, 18, 24, 28, 30, 45, 49, 53, 58, 74, 76, 119–122, 125–128]
Processing for NLP	Content analysis	0 (0.0)	1 (1.7)	0 (0.0)	[120]
	Detecting URLs	0 (0.0)	1 (1.7)	0 (0.0)	[124]
	Dimensionality reduction	0 (0.0)	1 (1.7)	0 (0.0)	[45]
	Feature weighting	0 (0.0)	1 (1.7)	0 (0.0)	[45]
	Homogenization	0 (0.0)	2 (3.3)	0 (0.0)	[74, 124]
	Language categorization	0 (0.0)	1 (1.7)	0 (0.0)	[6]
	LDA topics	0 (0.0)	1 (1.7)	0 (0.0)	[53]
	Lemmatization	0 (0.0)	3 (5.0)	0 (0.0)	[44, 45, 123]
	n-gram generation	0 (0.0)	6 (10.0)	0 (0.0)	[24, 44, 53, 74, 76, 120]
	Normalization using frequency-based methods	0 (0.0)	1 (1.7)	0 (0.0)	[45]
	Remove symbols and URLs	0 (0.0)	1 (1.7)	0 (0.0)	[120]
	Sentiment analysis	0 (0.0)	5 (8.3)	0 (0.0)	[24, 119, 124, 126, 127]
	Stemming	0 (0.0)	7 (11.7)	0 (0.0)	[6, 24, 45, 74, 76, 123, 124]
	Stop word removal	0 (0.0)	6 (10.0)	0 (0.0)	[6, 18, 45, 74, 76, 124]
	Term filtering	0 (0.0)	1 (1.7)	0 (0.0)	[74]
	Term Frequency—Inverse Document Frequency (TF-IDF)	1 (1.7)	7 (11.7)	0 (0.0)	[4, 24, 44, 49, 53, 58, 74, 76]
	Text embeddings	0 (0.0)	1 (1.7)	0 (0.0)	[53]
	Thematic analysis	0 (0.0)	1 (1.7)	0 (0.0)	[119]
	Topic detection	0 (0.0)	1 (1.7)	0 (0.0)	[120]
	Tokenization	0 (0.0)	4 (6.7)	0 (0.0)	[6, 18, 45, 74]
	Tweet filtering	0 (0.0)	1 (1.7)	0 (0.0)	[6]
	Unknown	1 (1.7)	4 (6.7)	1 (1.7)	[28, 30, 121, 122, 125, 128]
Algorithm for prediction of target	1-gram Term Frequency classifier	0 (0.0)	1 (2.4)	0 (0.0)	[44]
	Association rule mapping	0 (0.0)	1 (2.4)	0 (0.0)	[127]
	Chi-Square test	0 (0.0)	1 (2.4)	0 (0.0)	[126]
	Correlation analysis	0 (0.0)	2 (4.9)	0 (0.0)	[53, 119]
	Decision Tree	0 (0.0)	2 (4.9)	0 (0.0)	[45, 76]
	Fuzzy Algorithm for Extraction, Monitoring and Classification of infectious Diseases (FAEMC-ID)	0 (0.0)	1 (2.4)	0 (0.0)	[74]
	Hidden Markov Model	0 (0.0)	1 (2.4)	0 (0.0)	[123]

(*Continued*)

Table 3. (Continued)

Category	Sub-category	Sina Weibo *n* (%)	Twitter *n* (%)	Yahoo! Knowledge *n* (%)	Publications
	k-Means clustering	1 (2.4)	1 (2.4)	0 (0.0)	[4, 120]
	k-Nearest Neighbors	1 (2.4)	0 (0.0)	0 (0.0)	[4]
	Latent Dirichlet allocation (LDA)	0 (0.0)	1 (2.4)	0 (0.0)	[44]
	Linear regression	0 (0.0)	5 (12.2)	0 (0.0)	[6, 28, 53, 125, 127]
	Maximum entropy	0 (0.0)	1 (2.4)	0 (0.0)	[125]
	Naïve Bayes	0 (0.0)	5 (12.2)	0 (0.0)	[24, 45, 76, 124, 125]
	Random Forest	0 (0.0)	1 (2.4)	0 (0.0)	[45]
	Recurrent neural networks with Long short-term memory (LSTM)	0 (0.0)	1 (2.4)	0 (0.0)	[53]
	ST-DBSCAN	0 (0.0)	1 (2.4)	0 (0.0)	[127]
	Support vector machines	1 (2.4)	8 (19.5)	1 (2.4)	[4, 24, 30, 44, 45, 49, 53, 58, 125, 128]
	Time series	0 (0.0)	1 (2.4)	0 (0.0)	[120]
	Unknown	1 (2.4)	2 (4.9)	0 (0.0)	[18, 121, 122]
Result	Negative	0 (0.0)	0 (0.0)	0 (0.0)	-
	Positive	2 (8.7)	20 (87.0)	1 (4.4)	[4, 6, 18, 24, 28, 30, 44, 45, 49, 53, 58, 74, 76, 119–128]
Reliability	Low	1 (4.4)	1 (4.4)	1 (4.4)	[18, 122, 128]
	Medium	1 (4.4)	15 (65.2)	0 (0.0)	[4, 24, 28, 30, 45, 49, 53, 58, 74, 76, 120, 121, 124–127]
	High	0 (0.0)	4 (17.4)	0 (0.0)	[6, 44, 119, 123]
Validity	Low	1 (4.4)	1 (4.4)	1 (4.4)	[18, 122, 128]
	Medium	1 (4.4)	15 (65.2)	0 (0.0)	[4, 24, 28, 30, 45, 49, 53, 58, 74, 76, 120, 121, 124–127]
	High	0 (0.0)	4 (17.4)	0 (0.0)	[6, 44, 119, 123]

https://doi.org/10.1371/journal.pone.0282101.t003

powerful and novel tool for public health authorities [28, 30]. This proactive and real-time tool addresses most of the limitations that are common among the traditional methods used for public health surveillance [1, 20, 21, 47, 50, 54]. In addition, this tool can be used for the remote sensing of user-generated experiences that were published to social media [19, 45, 57, 58]. This finding is consistent with the literature, which suggests that text mining of social media content has the potential to supplement the traditional methods for public health surveillance, such as the reporting of diagnosed cases by medical professionals [1, 45, 46].

Furthermore, Twitter was used most frequently as a source of user-generated health content. This finding is consistent with other studies that indicated that users publicly publish their own health-related information to Twitter, making Twitter a relevant social media platform [26, 27, 129, 130]. Some studies indicate, however, that Twitter may not be a reliable source for health-related content, and alternative sources should be identified that include content from this population [125].

Various techniques were used to process textual content. For example, sentiment analysis can be used to establish the subjectivity of content, such that news, which contains predominantly facts, can be distinguished from personal experiences that contain opinions. Because the included publications predominantly studied personal experiences and, therefore, excluded news, sentiment analysis or perhaps alternative strategies were used to classify this content.

Last, a discussion of our findings would not be complete without a reflection on Google Flu Trends. With the emergence of the internet, novel applications have been developed that

collect and analyze data for the purpose of public health surveillance [19]. To address some of the challenges of the traditional methods for public health surveillance, the software company Google built Google Flu Trends, which utilizes influenza-related search queries and search patterns from its users to estimate regional seasonal influenza outbreaks [6, 131, 132]. The underlying presumption of using search queries to predict influenza is that people, when they experience changes in their health status, search the internet for symptoms, treatments, and other medical advice for self-diagnosis [50]. The influenza-related search queries may then be analyzed for early indications of a seasonal influenza outbreak [19]. Therefore, increases or decreases in these search patterns may indicate the outbreak or the end of the seasonal flu season, respectively [19]. This made Google Flu Trends a novel real-time and global tool for remote sensing [123]. To enable researchers and public health authorities to perform their own analyses, Google also publishes these historical datasets online [58].

Some studies reported that Google Flu Trends achieves a higher accuracy for the prediction of seasonal influenza outbreaks than traditional methods [15]. For example, these search queries were used to predict seasonal influenza rates two weeks in advance at a 90 percent accuracy [127]. Similarly, influenza-related hospital visits were also analyzed using Google Flu Trends [133].

However, many researchers have reported that Google Flu Trends still faces many drawbacks related to its accuracy [58, 124]. For example, Google Flu Trends was found to be inaccurate with respect to variations in seasonal influenza patterns that occur on an annual basis [134]. In addition, it did not predict the 2009 A(H1N1) pandemic and performed suboptimal in forecasting subsequent seasonal influenza seasons [134–138]. Predominantly, the reliability of Google Flu Trends has been seriously questioned since 2013, when it failed to predict the intensity of the seasonal influenza outbreak [139]. Others have also reported that that Google Flu Trends has suboptimal performance [140].

Although Google Flu Trends has remediated several limitations of traditional health surveillance methods, additional innovations that provide improvements are required to enable better public health surveillance [6]. Furthermore, due to the repeated failure to detect infectious disease outbreaks and the shortcomings described above, Google Flu Trends was discontinued in 2015 [54]. Therefore, there exists a need for alternative and more suitable surveillance methods [134, 140], which we aimed to address using the present systematic review.

5.1 Limitations

This systematic review had six noteworthy limitations.

First, study selection, information extraction, quality assessment of publications, and analysis were performed by one researcher (PP). This may have introduced bias. However, the procedures and results were discussed by all authors, and disagreements were resolved by consensus.

Second, in the included publications, there was an unequal distribution of the analyzed communicable diseases. For example, studies most often reported on the effectiveness of using social media to monitor and surveille influenza, while fewer studies analyzed the effectiveness in relation to dengue and measles. Likewise, Ebola, HIV/AIDS, listeria, and tuberculosis received the least attention. Therefore, most studies reported findings on the same diseases, but it remains unknown to what extent these positive findings also hold for infectious diseases that were studied least often.

Third, Twitter was investigated in a vast majority of studies. However, Sina Weibo and Yahoo! Knowledge received very little attention. Additionally, other social media platforms

exist, such as Facebook, which were not investigated at all. Therefore, it remains unknown whether content from other source media platforms can also be used effectively for the public health surveillance of communicable diseases, especially because these platforms may be targeted to different populations and, thus, may enable the monitoring of specific subgroups in this population.

Fourth, it is common and unavoidable that user-generated content published to social media is inherently noisy and biased. Most users are unqualified to assess their medical symptoms and may exaggerate mild or unrelated symptoms. Users may also be malicious and intentionally publish fake content and seek to discredit competition. We suggest the consideration of these factors when evaluating the effectiveness of the data sources and proposed tools.

Fifth, a majority of the publications structurally failed to report important information. For example, many publications did not explicitly disclose the language and geographical origin of the included content, although this could sometimes be implicitly inferred. This is particularly relevant because a vast majority of studies used Twitter, which does record the geographical location of its users. Similarly, the software, as well as specific methods and techniques used for natural language processing, were often omitted. In addition to a lack of information about the implementation in the included studies, the authors often failed to reflect on their collaboration with the authorities, such as public health institutes. All studies investigated how text can be processed and understood, and the reporting of such crucial information is, therefore, essential for replicability.

Sixth, this qualitative systematic review followed the PRISMA guidelines. As discussed in the methodology section, due to the interdisciplinary nature of the reviewed studies and their limitations, we acknowledge that it was not possible to complete every item from the PRISMA checklist (see S2 Appendix). For the same reason, no PROSPERO registration was made.

5.2 Theoretical recommendations

In the following, four recommendations are suggested to researchers.

First, although a vast number of the researchers included in this study investigated influenza, which clearly makes influenza a popular disease on this topic, and to a lesser extent dengue and measles have also been studied, it is essential that other communicable diseases also receive more attention in the literature. Indeed, many infectious diseases exist that pose a threat to public health, and it remains unknown whether these diseases can be monitored and predicted effectively using textual content. Therefore, we recommend that other infectious diseases be studied more frequently to produce more evidence on this topic.

Second, similarly, Twitter is clearly a popular social media platform for text mining. Some of its popularity is also related to the public accessibility of its content. However, many other popular platforms exist that have received far less or even no attention in the literature. It is, therefore, recommended that future research also account for those platforms. This is particularly relevant because only then can it be established whether certain platforms are more useful than others to surveille and predict infectious diseases, or perhaps these platforms may yield contradicting findings.

Third, a majority of studies failed to report critical information, such as the language and geographical origin of their content and the software, methods, and techniques used for natural language processing. Including such information is essential to establish the reliability and validity of findings and because it enables other researchers to replicate the study. It is, therefore, recommended that researchers disclose such information. In addition, it is highly recommended that the software that was developed to collect and analyze the data in studies is well

documented and published for reuse by the community and that authors thoroughly describe the application of their NLP analysis.

Fourth, the included journal articles were overall of higher quality than conference proceedings. This difference may be partly explained by the peer review involved, which may be more elaborate for journals than for conferences. However, another explanation is related to the limited amount of important information that was disclosed about the included data, methodologies, and analyses. Therefore, it is highly recommended that researchers provide more of the information needed to establish the reliability and validity of their studies and the reported findings.

6 Conclusion

Our findings in this work indicate that text mining of health-related content published to social media can serve as a novel and powerful tool for the automated, real-time, and remote monitoring of public health and for the surveillance and prediction of communicable diseases in particular.

According to our results, practitioners at public health authorities may benefit from utilizing natural language processing applied to social media data for the surveillance of communicable diseases as a supplement to their traditional methods. Natural language processing provides an automated, real-time tool to analyze user-generated content that includes contextual information to surveille and predict communicable diseases worldwide. This systematic review indicates that textual content from social media can be an important source of this knowledge. Another benefit of social media content is that it enables remote sensing via the internet by collecting public information. There is, however, no need to replace traditional methods, such as the collection of information about diagnosed cases from medical practitioners. Nevertheless, practitioners are highly recommended to include textual content from social media as a supplementary source for their data in their public health surveillance efforts to monitor and predict communicable diseases.

Supporting information

S1 Appendix. Search queries.
(DOCX)

S2 Appendix. PRISMA checklist.
(DOCX)

S3 Appendix. Characteristics of studies.
(DOCX)

Author Contributions

Conceptualization: Patrick Pilipiec.

Data curation: Patrick Pilipiec.

Formal analysis: Patrick Pilipiec.

Investigation: Patrick Pilipiec.

Methodology: Patrick Pilipiec.

Project administration: Patrick Pilipiec.

Resources: Patrick Pilipiec.

Software: Patrick Pilipiec.

Supervision: Patrick Pilipiec.

Validation: Patrick Pilipiec.

Visualization: Patrick Pilipiec.

Writing – original draft: Patrick Pilipiec.

Writing – review & editing: Patrick Pilipiec, Isak Samsten, András Bota.

References

1. Denecke K, Krieck M, Otrusina L, Smrz P, Dolog P, Nejdl W, et al. How to Exploit Twitter for Public Health Monitoring? Methods Inf Med. 2013; 52(4):326–39. https://doi.org/10.3414/ME12-02-0010 PMID: 23877537
2. Centers for Disease Control and Prevention. Flu Symptoms & Complications 2019 [Available from: https://www.cdc.gov/flu/symptoms/symptoms.htm.
3. Molinari N-AM, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, et al. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. Vaccine. 2007; 25:5086–96. https://doi.org/10.1016/j.vaccine.2007.03.046 PMID: 17544181
4. Yang N, Cui X, Hu C, Zhu W, Yang C, editors. Chinese Social Media Analysis for Disease Surveillance. International Conference on Identification, Information and Knowledge in the Internet of Things; 2014; Beijing, China: IEEE.
5. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu BY. Online Social Networks Flu Trend Tracker: A Novel Sensory Approach to Predict Flu Trends. In: Gabriel J, Schier J, Van Huffel S, Conchon E, Correia C, Fred A, et al., editors. Biomedical Engineering Systems and Technologies. 357. Berlin, Germany: Springer; 2013. p. 353–68.
6. Alkouz B, Al Aghbari Z, Abawajy JH. Tweetluenza: Predicting Flu Trends from Twitter Data. Big Data Mining and Analytics. 2019; 2(4):273–87.
7. Horimoto T, Kawaoka Y. Influenza: Lessons from Past Pandamics, Warnings from Current Incidents. Nature Reviews Microbiology. 2005; 3:591–600.
8. World Health Organization. Coronavirus 2020 [Available from: https://www.who.int/health-topics/coronavirus.
9. Sun Reporter. How other European countries including Spain, Italy and France are coming out of coronavirus lockdown 2020 [Available from: https://www.thesun.co.uk/news/11380418/european-countries-lockdown-exit-strategy/.
10. Henley J. EU countries take first cautious steps out of coronavirus lockdown 2020 [Available from: https://www.theguardian.com/world/2020/apr/14/eu-countries-coronavirus-lockdown-italy-spain.
11. Fairbrother G, Cassedy A, Ortega-Sanchez IR, Szilagyi PG, Edwards KM, Molinari N-A, et al. High costs of influenza: Direct medical costs of influenza disease in young children. Vaccine. 2010; 28:4913–9. https://doi.org/10.1016/j.vaccine.2010.05.036 PMID: 20576536
12. Li S, Leader S. Economic burden and absenteeism from influenza-like illness in healthy households with children (5–17 years) in the US. Respiratory Medicine. 2007; 101:1244–50. https://doi.org/10.1016/j.rmed.2006.10.022 PMID: 17156991
13. Ryan J, Zoellner Y, Gradl B, Palache B, Medema J. Establishing the health and economic impact of influenza vaccination within the European Union 25 countries. Vaccine. 2006; 24:6812–22. https://doi.org/10.1016/j.vaccine.2006.07.042 PMID: 17034909
14. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLoS Comput Biol. 2015; 11(8):e1004382. https://doi.org/10.1371/journal.pcbi.1004382 PMID: 26317693
15. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza Forecasting with Google Flu Trends. PLoS ONE. 2013; 8(2):e56176. https://doi.org/10.1371/journal.pone.0056176 PMID: 23457520
16. Polgreen PM, Nelson FD, Neumann GR. Use of Prediction Markets to Forecast Infectious Disease Activity. Healthcare Epidemiology. 2007; 44(2):272–9. https://doi.org/10.1086/510427 PMID: 17173231
17. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012–2013 season. Nat Commun. 2013;4. https://doi.org/10.1038/ncomms3837 PMID: 24302074

18. Lee K, Agrawal A, Choudhary A. Real-Time Disease Surveillance Using Twitter Data: Demonstration on Flu and Cancer. New York City, NY, USA: Association for Computing Machinery; 2013. 1474–7 p.

19. Bello-Orgaz G, Hernandez-Castro J, Camacho D. A Survey of Social Web Mining Applications for Disease Outbreak Detection. In: Camacho D, Braubach L, Venticinque S, Badica C, editors. Intelligent Distributed Computing VIII. Studies in Computational Intelligence. 570. Berlin: Springer-Verlag; 2015. p. 345–56.

20. Ali K, Dong H, Bouguettaya A, Erradi A, Hadjidj R, editors. Sentiment Analysis as a Service: A social media based sentiment analysis framework. IEEE International Conference on Web Services (ICWS); 2017; Honolulu, HI, USA: IEEE.

21. Blendon RJ, Benson JM, Desroches CM, Weldon KJ. Using Opinion Surveys to Track the Public's Response to a Bioterrorist Attack. Journal of Health Communication. 2003; 8:83–92. https://doi.org/10.1080/713851964 PMID: 14692573

22. Ngugi BK, Harrington B, Porcher EN, Wamai RG. Data quality shortcomings with the US HIV/AIDS surveillance system. Health Inform J. 2019; 25(2):304–14. https://doi.org/10.1177/1460458217706183 PMID: 28486860

23. Chen H, Zeng D, Yan P. Infectious Disease Informatics: Syndromic Surveillance for Public Health and BioDefense. Boston, MA, USA: Springer; 2010.

24. Ji X, Chun SA, Geller J. Monitoring Public Health Concerns Using Twitter Sentiment Classifications. IEEE International Conference on Healthcare Informatics; Philadelphia, PA, USA: IEEE; 2013. p. 335–44.

25. Khatua A, Khatua A, editors. Immediate and long-term effects of 2016 Zika Outbreak: A Twitter-based study. 18th International Conference on e-Health Networking, Applications and Services (Healthcom); 2016 14–16 Sept. 2016; Munich, Germany: IEEE.

26. Al-garadi MA, Khan MS, Varathan KD, Mujtaba G, Al-Kabsi AM. Using online social networks to track a pandemic: A systematic review. J Biomed Inform. 2016; 62:1–11. https://doi.org/10.1016/j.jbi.2016.05.005 PMID: 27224846

27. Shah M. Disease Propagation in Social Networks: A Novel Study of Infection Genesis and Spread on Twitter. JMLR: Workshop and Conference Proceedings. 2016; 53:1–17.

28. Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A Case Study of the New York City 2012–2013 Influenza Season With Daily Geocoded Twitter Data From Temporal and Spatiotemporal Perspectives. J Med Internet Res. 2014; 16(10):e236. https://doi.org/10.2196/jmir.3416 PMID: 25331122

29. Santos JC, Matos S. Analysing Twitter and web queries for flu trend prediction. Theoretical Biology and Medical Modelling. 2014; 11:S6. https://doi.org/10.1186/1742-4682-11-S1-S6 PMID: 25077431

30. Wakamiya S, Kawai Y, Aramaki E. Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study. JMIR Public Health and Surveillance. 2018; 4(3):e65. https://doi.org/10.2196/publichealth.8627 PMID: 30274968

31. Aiello AE, Renson A, Zivich PN. Social Media–and Internet-Based Disease Surveillance for Public Health. Annual Review of Public Health. 2020; 41:101–18. https://doi.org/10.1146/annurev-publhealth-040119-094402 PMID: 31905322

32. Jordan SE, Hovet SE, Fung IC, Liang H, Fu K, Tse ZTH. Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. Data. 2019; 4(1):6.

33. 'Conway M, Hu M, Chapman WW. Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and Consumer-Generated Data. Yearbook of Medical Informatics. 2019; 28(1):208–17. https://doi.org/10.1055/s-0039-1677918 PMID: 31419834

34. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social Media and Internet-Based Data inGlobal Systems for Public HealthSurveillance: A Systematic Review. The Mlibank Quarterly. 2014; 92(1):7–33.

35. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM, et al. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PLoS ONE. 2015; 10(10):e0139701. https://doi.org/10.1371/journal.pone.0139701 PMID: 26437454

36. Fung IC, Duke CH, Finch KC, Snook KR, Tseng P, Hernandez AC, et al. Ebola virus disease and social media: A systematic review. Am J Infect Control. 2016; 44(12):1660–71. https://doi.org/10.1016/j.ajic.2016.05.011 PMID: 27425009

37. Abad ZSH, Kline A, Sultana M, Noaeen M, Nurmambetova E, Lucini F, et al. Digital public health surveillance: a systematic scoping review. npj Digit Med. 2021;4.

38. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: A systematic review. J Biomed Inform. 2020;108. https://doi.org/10.1016/j.jbi.2020.103500 PMID: 32622833

39. Porta M. A Dictionary of Epidemiology. New York City, NY, USA: Oxford University Press; 2014.

40. Thacker SB, Berkelman RL. Public Health Surveillance in the United States. Epidemiologic Reviews. 1988; 10:164–90. https://doi.org/10.1093/oxfordjournals.epirev.a036021 PMID: 3066626

41. Lee LM, Teutsch SM, Thacker SB, St. Louis ME. Principles and Practice of Public Health Surveillance. New York City, NY, USA: Oxford University Press; 2010.

42. Runge-Ranzinger S, Horstick O, Marx M, Kroeger A. What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? Tropical Medicine and International Health. 2008; 13(8):1022–41. https://doi.org/10.1111/j.1365-3156.2008.02112.x PMID: 18768080

43. Centers for Disease Control and Prevention. CDC's Vision for Public Health Surveillance in the 21st Century. Atlanta, GA, USA: Centers for Disease Control and Prevention; 2012.

44. Kagashe I, Yan Z, Suheryani I. Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data. J Med Internet Res. 2017; 19(9):e315. https://doi.org/10.2196/jmir.7393 PMID: 28899847

45. Jain VK, Kumar S. Rough set based intelligent approach for identification of H1N1 suspect using social media. Kuwait J Sci. 2018; 45(2):8–14.

46. Lombardo JS, Buckeridge DL. Disease Surveillance: A Public Health Informatics Approach. Hoboken, NJ, USA: John Wiley & Sons; 2007.

47. Gu X, Chen H, Yang B. Heterogeneous Data Mining for Planning Active Surveillance of Malaria. ASE Big Data & Social Informatics 2015; Kaohsiung, Taiwan: Association for Computing Machinery; 2015. p. Article 34.

48. Bóta A, Holmberg M, Gardner L, Rosvall M. Socioeconomic and environmental patterns behind H1N1 spreading in Sweden. Scientific Reports. 2021; 11:22512. https://doi.org/10.1038/s41598-021-01857-4 PMID: 34795338

49. Aslam AA, Tsou M-H, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The Reliability of Tweets as a Supplementary Method of Seasonal Influenza Surveillance. J Med Internet Res. 2014; 16(11):e250. https://doi.org/10.2196/jmir.3532 PMID: 25406040

50. Corley CD, Cook DJ, Mikler AR, Singh KP. Text and Structural Data Mining of Influenza Mentions in Web and Social Media. Int J Environ Res Public Health. 2010; 7(2):596–615. https://doi.org/10.3390/ijerph7020596 PMID: 20616993

51. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Van Tong MPH. Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks 2004 [Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5305a1.htm.

52. Centers for Disease Control and Prevention. Dengue 2020 [Available from: https://www.cdc.gov/dengue/.

53. Volkova S, Ayton E, Porterfield K, Corley CD. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. PLoS ONE. 2017; 12(12):e0188941. https://doi.org/10.1371/journal.pone.0188941 PMID: 29244814

54. Ertem Z, Raymond D, Meyers LA. Optimal multi-source forecasting of seasonal influenza. PLoS Comput Biol. 2018; 14(9):e1006236. https://doi.org/10.1371/journal.pcbi.1006236 PMID: 30180212

55. Gardner LM, Bóta A, Gangavarapu K, Kraemer MUG, Grubaugh ND. Inferring the risk factors behind the geographical spread and transmission of Zika in the Americas. Plos Neglect Trop Dis. 2018; 12(1):e0006194. https://doi.org/10.1371/journal.pntd.0006194 PMID: 29346387

56. Colizza V, Barrat A, Barthélemy M, Vespignani A. The role of the airline transportation network in the prediction and predictability of global epidemics. Proceedings of the National Academy of Sciences of the United States of America. 2016; 103(7):2015–20.

57. Dingli A, Mercieca L, Spina R, Galea M, editors. Event detection using social sensors. 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM); 2015; Rennes, France: IEEE.

58. Allen C, Tsou M-H, Aslam A, Nagel A, Gawron J-M. Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. PLoS One. 2016; 11(7):e0157734. https://doi.org/10.1371/journal.pone.0157734 PMID: 27455108

59. Capdevila J, Cerquides J, Torres J, editors. Recognizing warblers: a probabilistic model for event detection in Twitter. ICML2016 Anomaly Detection Workshop; 2016; New York City, NY, USA.

60. Rivera SJ, Minsker BS, Work DB, Roth D. A text mining framework for advancing sustainability indicators. Environmental Modelling & Software. 2014; 62:128–38.

61. Mahdavi Anari SM, Bakri A, Ibrahim R, editors. Understanding factors on the customer intention behavior through Facebook commerce: a conceptual model. International Symposium on Technology Management and Emerging Technologies (ISTMET 2014); 2014; Bandung, Indonesia.

62. Liu B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. 2012; 5(1):1–167.

63. Sakaki T, Okazaki M, Matsuo Y, editors. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. International conference on World Wide Web; 2010; Raleigh, NC, USA.

64. Gao H, Barbier G, Goolsby R. Harnessing the Crowdsourcing Power of Social Media for Disaster Relief. IEEE Intelligent Systems. 2011; 26(3):10–4.

65. Nascimento TD, DosSantos MF, Danciu T, DeBoer M, Van Holsbeeck H, Lucas SR, et al. Real-Time Sharing and Expression of Migraine Headache Suffering on Twitter: A Cross-Sectional Infodemiology Study. J Med Internet Res. 2014; 16(4):e96. https://doi.org/10.2196/jmir.3265 PMID: 24698747

66. Tsou M-H, Lusher D, Han S, Spitzberg B, Gawron JM, Gupta D, et al. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. Cartography and Geographic Information Science. 2013; 40 (4):337–48.

67. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM, editors. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. 4th International AAAI Conference on Weblogs and Social Media; 2010.

68. Liu Y, Han W, Tian Y, Que X, Wang W, editors. Trending topic prediction on social network. 5th IEEE International Conference on Broadband Network & Multimedia Technology; 2013; Guilin, China.

69. Asur S, Huberman BA. Predicting the Future With Social Media 2010 [Available from: https://arxiv.org/abs/1003.5699.

70. Zhao S, Zhong L, Wickramasuriya J, Vasudevan V. Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games 2011 [Available from: https://arxiv.org/abs/1106.4300.

71. Spasojevic N, Yan J, Rao A, Bhattacharyya P, editors. LASTA: Large Scale Topic Assignment on Multiple Social Networks. 20th ACM SIGKDD international conference on Knowledge discovery and data mining; 2014; New York City, NY, USA.

72. Yang S, Kolcz A, Schlaikjer A, Gupta P, editors. Large-Scale High-Precision Topic Modeling on Twitter. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2014; New York City, NY, USA.

73. Contractor D, Chawda BS, Mehta S, Subramaniam LV, Faruquie TA, editors. Tracking political elections on social media: applications and experience. 24th International Conference on Artificial Intelligence; 2015; Buenos Aires, Argentina.

74. Jahanbin K, Rahmanian F, Rahmanian V, Jahromi AS. Application of Twitter and web news mining in infectious disease surveillance systems and prospects for public health. GMD Hyg Infect Control. 2019; 14:12. https://doi.org/10.3205/dgkh000334 PMID: 32047718

75. Collier N. Uncovering text mining: A survey of current work on web-based epidemic intelligence. Glob Public Health. 2012; 7(7):731–49. https://doi.org/10.1080/17441692.2012.699975 PMID: 22783909

76. Saire JEC, editor Building Intelligent Indicators to Detect Dengue Epidemics in Brazil using Social Networks. Colombian Conference on Applications in Computational Intelligence (ColCACI); 2019; Barranquilla, Colombia: IEEE.

77. Feldman R, Sanger J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge, UK: Cambridge University Press; 2006.

78. Tan AH, editor Text Mining: The State of the Art and the Challenges. 3rd Pacific-Asia Conference PAKDD'99; 1999; Beijing, China.

79. Cohen AM, Hersh WR. A Survey of Current Work in Biomedical Text Mining. Brief Bioinform. 2005; 6 (1):57–71. https://doi.org/10.1093/bib/6.1.57 PMID: 15826357

80. McCaig D, Bhatia S, Elliott MT, Walasek L, Meyer C. Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of fitness tracking technology across online communities. International Journal of Eating Disorders. 2018; 51:647–55. https://doi.org/10.1002/eat.22882 PMID: 29734478

81. Pilipiec P, Liwicki M, Bota A. Using Machine Learning for Pharmacovigilance: A Systematic Review. Pharmaceutics. 2022; 14:266. https://doi.org/10.3390/pharmaceutics14020266 PMID: 35213998

82. Tekin M, Etlioğlu M, Koyuncuoğlu Ö, Tekin E. Data Mining in Digital Marketing. In: Durakbasa NM, Gencyilmaz MG, editors. Proceedings of the International Symposium for Production Research 2018. Cham, Switzerland: Springer; 2018. p. 44–61.

83. Aggarwal CC, Zhai C. A Survey of Text Clustering Algorithms. In: Aggarwal CC, Zhai C, editors. Mining Text Data. Boston, MA, USA: Springer; 2012. p. 77–128.

84. Lu Y, Zhang P, Liu J, Li J, Deng S. Health-Related Hot Topic Detection in Online Communities Using Text Clustering. PLoS One. 2013; 8(2):e56221. https://doi.org/10.1371/journal.pone.0056221 PMID: 23457530

85. Nassirtoussia AK, Aghabozorgia S, Wah TY, Ngo DCL. Text mining for market prediction: A systematic review. Expert Syst Appl. 2014; 41(16):7653–70.

86. Oberreuter G, Velásquez JD. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. Expert Syst Appl. 2013; 40(9):3756–63.

87. Sunikka A, Bragge J. Applying text-mining to personalization and customization research literature–Who, what and where? Expert Syst Appl. 2012; 39(11):10049–58.

88. Sadiq AT, Abdullah SM, editors. Hybrid Intelligent Technique for Text Categorization. 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT); 2012 November 26–28; Kuala Lumpur, Malaysia: IEEE.

89. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. 2014; 5:1093–113.

90. Serrano-Guerrero J, Olivas JA, Romero FP, Herrera-Viedma E. Sentiment analysis: A review and comparative analysis of web services. Inf Sci. 2015; 311:18–38.

91. Lee K, Agrawal A, Choudhary A, editors. Mining Social Media Streams to Improve Public Health Allergy Surveillance. Advances in Social Networks Analysis and Mining 2015; 2015 August 25–28; Paris, France: Association for Computing Machinery.

92. Nargund K, Natarajan S, editors. Public health allergy surveillance using micro-blogs. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2016 September 21–24; Jaipur, India: IEEE.

93. Rong J, Michalska S, Subramani S, Du J, Wang H. Deep learning for pollen allergy surveillance from twitter in Australia. BMC Med Inform Decis Mak. 2019; 19:19.

94. Leis A, Ronzano F, Mayer MA, Furlong L, Sanz F. Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. J Med Internet Res. 2019; 21(6):e14199. https://doi.org/10.2196/14199 PMID: 31250832

95. Mowery D, Smith H, Cheney T, Stoddard G, Coppersmith G, Bryan C, et al. Understanding Depressive Symptoms and Psychosocial Stressors on Twitter: A Corpus-Based Study. J Med Internet Res. 2017; 19(2):e48. https://doi.org/10.2196/jmir.6895 PMID: 28246066

96. Zucco C, Calabrese B, Cannataro M, editors. Sentiment analysis and affective computing for depression monitoring. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017 November 13–17; Kansas City, MO, USA: IEEE.

97. Ji X, Chun SA, Wei Z, Geller J. Twitter sentiment classification for measuring public health concerns. Soc Netw Anal Min. 2015; 5:13. https://doi.org/10.1007/s13278-015-0253-5 PMID: 32226558

98. Cavazos-Rehg PA, Zewdie K, Krauss MJ, Sowles SJ. "No High Like a Brownie High": A Content Analysis of Edible Marijuana Tweets. Am J Health Promot. 2018; 32(4):e0223318. https://doi.org/10.1177/0890117116686574 PMID: 29214836

99. Phan N, Chun SA, Bhole M, Geller J, editors. Enabling Real-Time Drug Abuse Detection in Tweets. 2017 IEEE 33rd International Conference on Data Engineering (ICDE); 2017 April 19–22; San Diego, CA, USA: IEEE.

100. Tran T, Nguyen D, Nguyen A, Golen E, editors. Sentiment Analysis of Marijuana Content via Facebook Emoji-Based Reactions. 2018 IEEE International Conference on Communications (ICC); 2018 May 20–24; Kansas City, MO, USA: IEEE.

101. Ward PJ, Rock PJ, Slavova S, Young AM, Bunn TL, Kavuluru R. Enhancing timeliness of drug overdose mortality surveillance: A machine learning approach. PLoS One. 2019; 14(10):e0223318. https://doi.org/10.1371/journal.pone.0223318 PMID: 31618226

102. Cesare N, Dwivedi P, Nhuyen QC, Nsoesie EO. Use of social media, search queries, and demographic data to assess obesity prevalence in the United States. Palgr Commun. 2019; 5:106. https://doi.org/10.1057/s41599-019-0314-x PMID: 32661492

103. Kent EE, Prestin A, Gaysynsky A, Galica K, Rinker R, Graff K, et al. "Obesity is the New Major Cause of Cancer": Connections Between Obesity and Cancer on Facebook and Twitter. J Cancer Educ. 2016; 31:453–9. https://doi.org/10.1007/s13187-015-0824-1 PMID: 25865399

104. Brown RC, Bendig E, Fischer T, Goldwich AD, Baumert M, Plener PL. Can acute suicidality be predicted by Instagram data? Results from qualitative and quantitative language analyses. PLoS One. 2019; 14(9):e0220623. https://doi.org/10.1371/journal.pone.0220623 PMID: 31504042

105. Kavuluru R, Ramos-Morales M, Holaday T, Williams AG, Haye L, Cerel J, editors. Classification of Helpful Comments on Online Suicide Watch Forums. BCB '16: ACM International Conference on

Bioinformatics, Computational Biology, and Health Informatics; 2016 October 2–5; Seattle, WA, USA: Association for Computing Machinery.

106. Song J, Song TM, Seo D, Jin JH. Data Mining of Web-Based Documents on Social Networking Sites That Included Suicide-Related Words Among Korean Adolescents. J Adolesc Health. 2016; 59 (6):668–73. https://doi.org/10.1016/j.jadohealth.2016.07.025 PMID: 27693129

107. Cole-Lewis H, Pugatch J, Sanders A, Varghese A, Posada S, Yun C, et al. Social Listening: A Content Analysis of E-Cigarette Discussions on Twitter. J Med Internet Res. 2015; 17(10):e243. https://doi.org/10.2196/jmir.4969 PMID: 26508089

108. Kim A, Miano T, Chew R, Eggers M, Nonnemaker J. Classification of Twitter Users Who Tweet About E-Cigarettes. J Med Internet Res. 2017; 3(3):e63. https://doi.org/10.2196/publichealth.8060 PMID: 28951381

109. Kostygina G, Tran H, Shi Y, Kim Y, Emery S. 'Sweeter Than a Swisher': amount and themes of little cigar and cigarillo content on Twitter. Tob Control. 2016; 25:i75–i82. https://doi.org/10.1136/tobaccocontrol-2016-053094 PMID: 27697951

110. Myslín M, Zhu S, Chapman W, Conway M. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. J Med Internet Res. 2013; 15(8):e174. https://doi.org/10.2196/jmir.2534 PMID: 23989137

111. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009; 339:b2700. https://doi.org/10.1136/bmj.b2700 PMID: 19622552

112. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Medicine. 2009; 6(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097 PMID: 19621072

113. Bramer WM, Giustini D, De Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. Journal of the Medical Library Association: JMLA. 2016; 104 (3):240–3. https://doi.org/10.3163/1536-5050.104.3.014 PMID: 27366130

114. Kampmeijer R, Pavlova M, Tambor M, Golinowska S, Groot W. The use of e-health and m-health tools in health promotion and primary prevention among older adults: a systematic literature review. BMC Health Serv Res. 2016; 16(Suppl 5):290. https://doi.org/10.1186/s12913-016-1522-3 PMID: 27608677

115. Hsieh H-F, Shannon SE. Three Approaches to Qualitative Content Analysis. Qualitative Health Research. 2005; 15(9):1277–88. https://doi.org/10.1177/1049732305276687 PMID: 16204405

116. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006; 3:77–101.

117. Nowell LS, Norris JM, White DE, Moules NJ. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. International Journal of Qualitative Methods. 2017; 16:1–13.

118. King N. Using templates in the thematic analysis of text. In: Cassell C, Symon G, editors. Essential guide to qualitative methods in organizational research. London, UK: SAGE; 2004. p. 257–70.

119. Mollema L, Harmsen IA, Broekhuizen E, Clijnk R, De Melker H, Paulussen T, et al. 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. J Med Internet Res. 2015; 17(5):e128. https://doi.org/10.2196/jmir.3863 PMID: 26013683

120. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control. 2015; 43(6):563–71. https://doi.org/10.1016/j.ajic.2015.02.023 PMID: 26042846

121. Robinson B, Sparks R, Power R, Cameron M, editors. Social Media Monitoring for Health Indicators. 21st International Congress on Modelling and Simulation; 2015; Gold Coast, Australia: Modelling & Simulation Soc Australia & New Zealand Inc.

122. Zhang EX, Yang YY, Di Shang R, Simons JJP, Quek BK, Yin XF, et al. Leveraging social networking sites for disease surveillance and public sensing: the case of the 2013 avian influenza A(H7N9) outbreak in China. West Pa Surveill Response. 2015; 6(2):66–72. https://doi.org/10.5365/WPSAR.2015.6.1.013 PMID: 26306219

123. Chen L, Hossain KSMT, Butler P, Ramakrishnan N, Prakash BA. Syndromic Surveillance of Flu on Twitter Using Weakly Supervised Temporal Topic Models. Data Min Knowl Discov. 2016; 30(3):681–710.

124. Byrd K, Mansurov A, Baysal O, editors. Mining Twitter Data For Influenza Detection and Surveillance. International Workshop on Software Engineering in Healthcare Systems; 2016; Austin, TX, USA: IEEE.

125. Nsoesie EO, Flor L, Hawkins J, Maharana A, Skotnes T, Marinho F, et al. Social Media as a Sentinel for Disease Surveillance: What Does Sociodemographic Status Have to Do with It? PLoS Currents Outbreaks. 2016;8.

126. Chew C, Eysenbach G. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS ONE. 2010; 5(11):e14118. https://doi.org/10.1371/journal.pone.0014118 PMID: 21124761

127. Gomide J, Veloso A, Meira W, Almeida V, Benevenuto F, Ferraz F, et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. 3rd International Web Science Conference; Koblenz, Germany: Association for Computing Machinery; 2011. p. Article 3.

128. Ku Y, Chiu C, Zhang Y, Fan L, Chen H, editors. Global Disease Surveillance using Social Media: HIV/AIDS Content Intervention in Web Forums. IEEE International Conference on Intelligence and Security Informatics; 2010; Vancouver, BC, Canada.

129. Ram S, Zhang WL, Williams M, Pengetnze Y. Predicting Asthma-Related Emergency Department Visits Using Big Data. IEEE J Biomed Health Inform. 2015; 19(4):1216–23. https://doi.org/10.1109/JBHI.2015.2404829 PMID: 25706935

130. Achrekar H, Gandhe A, Lazarus R, Yu S-H, Liu B, editors. Predicting Flu Trends using Twitter Data. 1st International Workshop on Cyber-Physical Networking Systems; 2011; Shanghai, China.

131. Carneiro HA, Mylonakis E. Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. Clinical Infectious Diseases. 2009; 49(10):1557–64. https://doi.org/10.1086/630200 PMID: 19845471

132. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457:1012–4. https://doi.org/10.1038/nature07634 PMID: 19020500

133. Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012–2013 Influenza Epidemic. PLoS ONE. 2013; 8(12):e83672. https://doi.org/10.1371/journal.pone.0083672 PMID: 24349542

134. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. PLoS Comput Biol. 2013; 9(10):e1003256. https://doi.org/10.1371/journal.pcbi.1003256 PMID: 24146603

135. Baker MG, Wilson N, Huang QS, Paine S, Lopez L, Bandaranayake D, et al. Pandemic influenza A (H1N1)v in New Zealand: the experience from April to August 2009. Eurosurveillance. 2009; 14 (34):1–6.

136. Ortiz JR, Zhou H, Shay DK, Neuzil KM, Goss CH, editors. Does Google Influenza Tracking Correlate With Laboratory Tests Positive For Influenza? B25 H1N1, Seasonal Influenza and Other Viral Pneumonia: Clini- cal and Mechanistic Insights; 2010; New Orleans, LA, USA: American Thoracic Society.

137. Scarpino SV, Dimitrov NB, Meyers LA. Optimizing Provider Recruitment for Influenza Surveillance Networks. PLoS Comput Biol. 2012; 8(4):e1002472. https://doi.org/10.1371/journal.pcbi.1002472 PMID: 22511860

138. Valdivia A, López-Alcalde J, Vicente M, Pichiule M, Ruiz M, Ordobas M. Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks–results for 2009–10. Eurosurveillance. 2010; 15(29):1–6. https://doi.org/10.2807/ese.15.29.19621-en PMID: 20667303

139. Butler D. When Google got flu wrong. Nature. 2013; 494:155–6. https://doi.org/10.1038/494155a PMID: 23407515

140. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014; 343:1203–5.