Title

Developing a standardized protocol for computational sentiment analysis research using health-related social media data.

Journal

Journal of the American Medical Informatics Association, 28(6)

ISSN

1067-5027

Authors

He, Lu
Yin, Tingjue
Hu, Zhaoxian
et al.

Publication Date

2021-06-12

DOI

10.1093/jamia/ocaa298

Journal of the American Medical Informatics Association, 28(6), 2021, 1125–1134
doi: 10.1093/jamia/ocaa298
Advance Access Publication Date: 22 December 2020
Research and Applications

Research and Applications

Developing a standardized protocol for computational sentiment analysis research using health-related social media data

Lu He[1], Tingjue Yin[1], Zhaoxian Hu[1], Yunan Chen[1], David A. Hanauer[2,3], and Kai Zheng[1,4]

[1]Department of Informatics, Donald Bren School of Information and Computer Science, University of California, Irvine, Irvine, California, USA, [2]Department of Learning Health Sciences, School of Medicine, University of Michigan, Ann Arbor, Michigan, USA, [3]Department of Pediatrics, School of Medicine, University of Michigan, Ann Arbor, Michigan, USA and [4]Department of Emergency Medicine, School of Medicine, University of California, Irvine, Irvine, California, USA

Corresponding Author: Kai Zheng, PhD, FACMI, Department of Informatics, Donald Bren School of Information and Computer Science, University of California, 6095 Donald Bren Hall, Irvine, CA 92697-3440, USA; kai.zheng@uci.edu

Received 1 September 2020; Editorial Decision 27 October 2020; Accepted 4 December 2020

ABSTRACT

Objective: Sentiment analysis is a popular tool for analyzing health-related social media content. However, existing studies exhibit numerous methodological issues and inconsistencies with respect to research design and results reporting, which could lead to biased data, imprecise or incorrect conclusions, or incomparable results across studies. This article reports a systematic analysis of the literature with respect to such issues. The objective was to develop a standardized protocol for improving the research validity and comparability of results in future relevant studies.
Materials and Methods: We developed the Protocol of Analysis of senTiment in Health (PATH) based on a systematic review that analyzed common research design choices and how such choices were made, or reported, among eligible studies published 2010-2019.
Results: Of 409 articles screened, 89 met the inclusion criteria. A total of 16 distinctive research design choices were identified, 9 of which have significant methodological or reporting inconsistencies among the articles reviewed, ranging from how relevance of study data was determined to how the sentiment analysis tool selected was validated. Based on this result, we developed the PATH protocol that encompasses all these distinctive design choices and highlights the ones for which careful consideration and detailed reporting are particularly warranted.
Conclusions: A substantial degree of methodological and reporting inconsistencies exist in the extant literature that applied sentiment analysis to analyzing health-related social media data. The PATH protocol developed through this research may contribute to mitigating such issues in future relevant studies.

Key words: sentiment analysis, reference standard [E05.978.808], social media [L01.178.75], user-generated content, Web 2.0, Facebook, Twitter, Instagram, natural language processing [L01.224.050.375.580], computing methodologies [L01.224], machine learning [G17.035.250.500]

INTRODUCTION

Background and Significance

Social media platforms such as Twitter and Facebook provide a public forum for anyone to create and disseminate content related to health, health care, or public health. For example, patients commonly share their disease journeys and exchange informational and emotional support with others who have similar conditions.[1,2] Social media is also commonly used by the general public to voice their opinions on issues such as important health policies, such as the Affordable Care Act[3] and the lockdown orders due to the COVID-19 (coronavirus disease 2019) pandemic,[4] and controversial medical interventions and treatments, such as human papillomavirus vaccination[5,6] and the use of hydroxychloroquine for treating COVID-19.[7]. Because social media data are generally publicly available, relatively easy to obtain (eg, through platform-provided application programming interface), and contributed by geographically and demographically diverse user populations,[8] they have become an increasingly important source of information used by researchers to investigate a wide range of health-related topics. In fact, prior studies have demonstrated that public opinions expressed on social media platforms are highly correlated with poll results based on conventional surveys, confirming the feasibility of using such data for rigorous scientific research. [3]

The sheer amount of user-generated social media data makes the data difficult to manually analyze. Qualitative studies on small, selective samples preclude generalization to larger datasets. With the recent advances in natural language processing (NLP) and the increasing computing capability to process big data, researchers have now been able to use cutting-edge NLP techniques to efficiently analyze large volumes of free-text data with minimal manual effort. Sentiment analysis, in particular, has received increasing attention. Sentiment analysis, also referred to as opinion mining,[9] is "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions toward entities such as products, services, organizations, individuals, issues, events, topics, and their attributes."[10] A simple keyword search using "sentiment analysis" or "opinion mining" in PubMed yielded 348 articles; most were published in the recent 10 years, and the majority were based on computational methods. For instance, Davis et al.[3] used NLP to study the general public's sentiments toward the Affordable Care Act; and Huppertz and Otto developed a machine learning model to analyze Facebook posts to assess patient opinions regarding their healthcare providers.[11]

To date, numerous computational sentiment analysis methods (hereafter referred to as sentiment analyzers) have been developed, ranging from lexicon-based dictionary lookups to machine learning algorithms.[12–14] These methods have demonstrated satisfactory performance across many research domains; even though studies have commonly acknowledged the challenges to analyzing sentiments embedded in social media data due to their unique characteristics such as frequent use of short text, informal expressions and layperson terms for medical concepts, and special communication gimmicks such as hashtags and emojis.[15,16]

While computational sentiment analysis is an invaluable tool for understanding health-related opinions expressed on social media platforms, in our prior work, we noticed multiple issues in how existing studies were conducted and how their results were reported.[17] For example, the keywords used to retrieve social media content often do not take into account the unique characteristics of consumer language used in social media posts, and some studies made rather arbitrary research design choices such as whether to filter out content contributed by nonlaypersons (eg, advocate groups and pharmaceutical companies) or whether to retain special types of data (eg, images/videos, hashtags, emojis, hyperlinks). Many also appeared to simply borrow existing sentiment analyzers developed in nonhealth domains (eg, movie review) without validating their appropriateness for the particular study context, even though it has been repeatedly reported that the poor cross-domain transferability of sentiment analyzers could lead to inaccurate interpretations of data, or completely wrong conclusions.[9,18] These issues may diminish the validity of the research. Indeed, in the literature, several studies have pointed out that they may have a significant impact on research results and conclusions. For example, a recent study found that organizational accounts posted more tweets expressing a positive attitude toward e-cigarettes than individual users.[19] Similarly, another study found that organizational tweets, which comprise more than 70% of the tweets related to the side effects of chemotherapies, tend to be more neutral, compared with tweets posted by individual users.[20] Social bots (ie, computer programs that generate tweets automatically) exhibit similar behavior. For example, Allem et al[21] showed that social bots were more likely to post pro-cannabis tweets than nonbot users. These findings suggest that the study design decision on whether to, or whether not to, differentiate social media content based on content contributors could lead to different findings and conclusions when conducting sentiment analysis research. Further, in our previous work,[17] we evaluated 3 commonly used sentiment analyzers by applying them to 2 manually annotated social media health datasets. We found that all of these tools demonstrated poor performance, incorrectly classifying the neutrality of the posts in over 50% of the cases, compared with the sentiment labels assigned by human annotators. Further, inconsistencies in how different methods and tools were chosen and applied make it difficult to compare and synthesize results across studies, hindering our ability to accumulate knowledge as a community. These observations motivated this work, through which we characterized common methodological and results reporting issues found in this body of literature, in order to develop a standardized protocol, which we refer to as the Protocol of Analysis of senTiment in Health (PATH), that may contribute to improving the quality and results comparability of future sentiment analysis research using health-related social media data, and other social media data analyses more broadly.

OBJECTIVE

The objectives of this study were 2-fold: (1) to conduct a systematic review of the literature to identify common issues in research design and results reporting among studies that applied computational sentiment analysis to social media data on topics related to health, health care, or public health; and (2) to develop the PATH based on the analysis of the relevant literature.

MATERIALS AND METHODS

Systematic literature review

We conducted the search in January 2020 using 3 literature databases: PubMed, IEEE Xplore, and the ACM Digital Library. We included articles published in English and in peer-reviewed journals or conferences over a 10-year period between January 1, 2010, and December 31, 2019. Development of the search query (Table 1) was informed by previous literature reviews on the use of computational

Table 1. Search query

((social media) OR (social network*) OR (social web*) OR (online social network*) OR (support group*) OR (Web 2.0) OR (Facebook) OR (Twitter) OR (MySpace) OR (Instagram) OR (YouTube) OR (Tumblr) OR (MedHelp) OR (WebMD) OR (online health communit*) OR (online forum*) OR (message board*) OR (discussion group*)) AND ((sentiment analysis) OR (opinion mining)) AND health*

methods for analyzing health-related social media text.[22–26] We also supplemented our literature search results with articles referenced in these existing reviews.

Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline,[27] we first screened titles and abstracts of the retrieved articles by applying the following exclusion criteria (1) studies conducted in topic areas not relevant to health, health care, or public health; (2) studies that analyzed non-English social media content; (3) studies that only performed manual review (eg, qualitative content analysis) of the data, as this article concerns sentiment analysis research that uses computational methods; and (4) studies that focused on development of new sentiment analyzers (eg, to report the algorithmic or mathematical underpinning of a new sentiment analysis algorithm), or development of new software architectures (eg, to provide real-time sentiment analysis through cloud-based services), instead of analyzing social media data to generate empirical insights. Two authors (L.H. and Z.H.) independently screened the titles and abstracts of a random set of 50 articles. The screening results were discussed, and disagreements were resolved through consensus development research meetings. The remaining titles and abstracts were evenly split into 2 sets and separately reviewed. Then, full texts of the articles meeting the inclusion and exclusion criteria were retrieved and independently screened for eligibility by 2 authors (L.H. and T.Y.), who also independently extracted data from the final set of articles included in the review. Interrater reliability was assessed whenever applicable.

Development of the PATH protocol

We developed the PATH protocol through the following 3 steps. First, using the qualitative deductive coding and constant comparison approach,[28] we identified and categorized distinctive research design choices that needed to be commonly made in relevant studies (eg, how to retrieve social media data and what sentiment analyzer to use). Then, we analyzed inconsistencies among the existing studies on these design choices and, when applicable, whether the rationale for a made choice was reported in the article. Finally, we synthesized the results from the analyses above to produce the PATH, the objective of which is to minimize such inconsistencies in order to improve the validity and results comparability of future sentiment analysis research in health.

RESULTS

The PRISMA flow diagram exhibiting the screening process is reported in Figure 1. The literature search returned 417 results; 409 remained after duplicated entries were removed. The first-round of screening based on titles and abstracts yielded 158 potentially relevant articles. The interrater agreement ratio was 0.88. Of these, 75 were deemed relevant upon a review of their full texts. The interrater agreement ratio was 0.94. We then conducted a citation analysis to identify additional relevant articles, which resulted in 14 more articles added. The final set selected for qualitative synthesis thus consisted of a total of 89 articles. Excluded articles and reasons for their exclusion are provided in the Supplementary Appendix 1.

Overall statistics

Of the 89 articles included in the review, most (n = 58) were published between 2017 and 2019. More than half (n = 51) analyzed Twitter data. The second and third most popularly studied platforms were Facebook (n = 5) and YouTube (n = 4), respectively. Besides these general-purpose social media sites, some studies (n = 17) also examined health-specific online communities such as MedHelp (n = 3), CancerSurvivorNetwork (n = 3), JuiceDB (n = 2), BreastCancer.org (n = 2), WebMD (n = 1), QuitNet (n = 1), TalkLife (n = 1), LiveJournal (n = 1), Drug.com (n = 1), GLOBALink (n = 1), and BecomeAnEX.org (n = 1).

Most of the studies reviewed (n = 47 of 89) used keywords to retrieve social media data, 6 used hashtags in addition, 6 identified users through questionnaire surveys or manual review of user posts, and 5 were based on geo-tagged locations. Among the studies that employed off-the-shelf sentiment analyzers, the LIWC (Linguistic Inquiry and Word Count) tool was most popular, used in 8 studies, followed by SentimentStrength (n = 5), LabMT (n = 4), and TextBlob (n = 3). Among those that developed their own machine learning models, support vector machine and naïve Bayes were most popular, each used in 13 studies, respectively, followed by logistic regression (n = 5), AdaBoost (n = 4), and k-nearest neighbors (n = 3). Among these machine learning studies, common features selected to train the model included bag of words (n = 7), word embeddings (n = 5), and linguistics features (n = 5) such as post length and part-of-speech. Common evaluation metrics for assessing model performance were accuracy (n = 16), F score (n = 12), precision (n = 7), recall (n = 7), and receiver-operating characteristic curve (n = 4).

Research design choices

Based on the articles reviewed, we first identified a list of distinctive research design choices that needed to be commonly made in conducting health-related computational sentiment analysis research using social media data. We then organized these design choices, reported in Table 2, according to the following 3 dimensions: (1) platform selection, (2) data curation; and (3) sentiment analysis method. The first dimension concerns how studies choose the appropriate social media platform that would be most informative for the research questions at hand; the second dimension concerns how to retrieve and curate relevant data that do not introduce unwanted biases (eg, whether to retain or remove advisements posted by pharmaceutical companies), or loss of critical information (eg, whether to retain, remove, or substitute hashtags and emojis). The third dimension concerns how to select the appropriate analytical tool best suited for the particular study context, and whether and how to validate the tool before applying it to the study data.

Methodological and reporting inconsistencies among the existing studies

Next, we assessed inconsistencies in how the existing studies reviewed made the aforementioned research design choices, and how they reported the rationale of making such choices, or the lack thereof. The results are shown in Table 3. A more detailed data analysis sheet is provided in Supplementary Appendix 2.

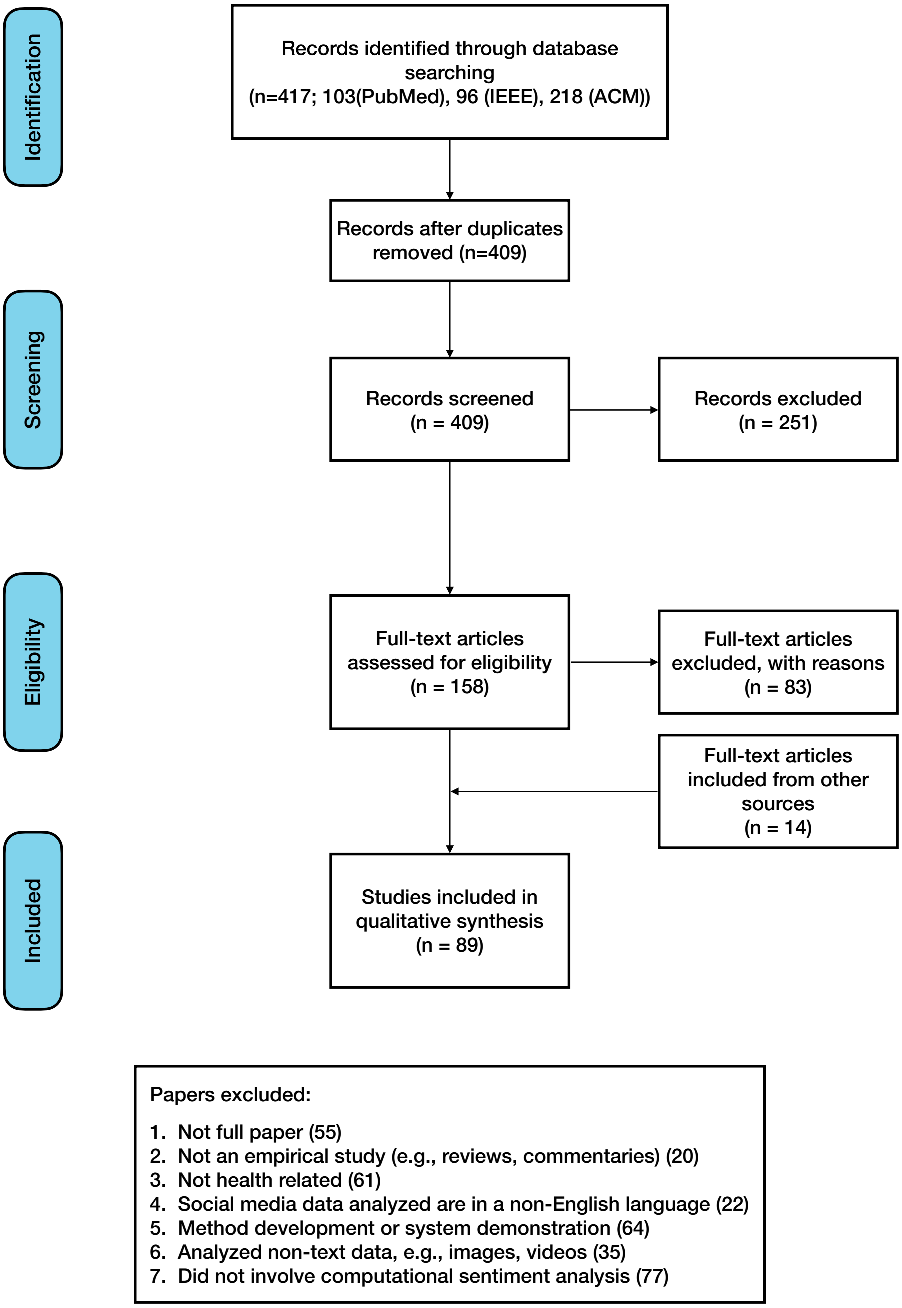
Identification
Screening
Eligibility
Included
Records identified through database searching (n=417; 103(PubMed), 96 (IEEE), 218 (ACM))
Records after duplicates removed (n=409)
Records screened (n = 409)
Records excluded (n = 251)
Full-text articles assessed for eligibility (n = 158)
Full-text articles excluded, with reasons (n = 83)
Full-text articles included from other sources (n = 14)
Studies included in qualitative synthesis (n = 89)
Papers excluded:
1. Not full paper (55)
2. Not an empirical study (e.g., reviews, commentaries) (20)
3. Not health related (61)
4. Social media data analyzed are in a non-English language (22)
5. Method development or system demonstration (64)
6. Analyzed non-text data, e.g., images, videos (35)
7. Did not involve computational sentiment analysis (77)

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Flow Diagram.

As shown in Table 3, a number of studies (n = 24 of 63 applicable) did not use, or did not report, any method for determining the relevance of the research data retrieved. Most (n = 45 of 58 applicable) did not differentiate the data based on content creator. While about one-third of the studies (n = 32 of 89) reported how special types of data such as hashtags and emojis were handled (ie, retain, remove, or substitute), less than half (n = 12) provided a rationale for the choice made. For the last dimension, sentiment analysis method, the majority of the studies (n = 49 of 89) did not provide any justification as to why the particular sentiment analyzer or the machine learning model was chosen. Additionally, 53 studies of 89 did not validate the sentiment analyzer for their particular study context. Among those that did use manually annotated data for sentiment analyzer validation, many (n = 11 of 26 applicable) did not involve multiple coders. Last, among the studies that used machine learning (n = 29), 10 did not describe the features selected. Among those that did, several (n = 7 of 19) did not justify the feature selection process.

Development of PATH

Based on the results of systematically analyzing the relevant literature (Tables 2 and 3), we developed the PATH, exhibited in Figure 2, which similarly contains 3 dimensions, namely platform selection, data curation, and sentiment analysis method, and a total of 16 distinctive research design choices. We believe that these research design choices, when applicable, must be carefully considered and

Table 2. Distinctive research design choices

Dimension	Design choice	Description	Example
Platform selection	Which social media platform provides data that are most informative to answer the research questions of the study?	As different social media platforms attract different types of users and foster different forms of communication, studies may want to evaluate available options and decide which one(s) would provide the best information for studying the research topic of interest.	“*WebMD.com hosts one of the few online communities that offer moderators in patient forums. Their diabetes community shows the most active participation of both patients and moderators among other WebMD communities.*”[29]
Data curation	What is the strategy used to retrieve relevant data?	Procedures for identifying potentially relevant social media posts based on keywords, user characteristics, or other means of information retrieval; and procedures for determining the relevance and comprehensiveness of the data retrieved.	“*We started with a set of relevant seed keywords (eg, ‘lynch syndrome’). Then, we searched on Twitter with these keywords to retrieve a sample of tweets, evaluated whether the retrieved tweets were indeed relevant to Lynch syndrome, and identified additional keywords to be used for the next rounds of searches.*”[30]
	Whether to differentiate data based on content creator?	Social media data can be contributed by different entities such as laypersons, healthcare providers, health systems, government agencies, advocacy groups, and pharmaceutical companies. Depending on the research objective, studies may want to treat data differently based on the creator of the content.	“*In order to gain insights into the opinion and experience of cancer patients about chemotherapy, these cancer-related user accounts were classified into two groups: individual accounts and organization accounts. The individual accounts belonged to cancer patients as well as their families, whereas the organization accounts include organizations, oncologists, news sources, and personnel who are neither patients nor family members.*”[20]
	How to handle special types of data.	As social media data frequently contain elements such as images or videos, hashtags, emojis, and hyperlinks, studies should determine whether to retain, remove, or substitute such data at the preprocessing stage, and explain the rationale for the approach chosen and its implications for study results.	“*We cleaned out contents such as emoji icons, urls, ‘#’, and ‘@’ from each tweet. By observing the data, we noticed that hashtags tended to store very important content. For instance, a lot of the anti-vaccine tweets contained ‘#CDCwhistleblower’. Therefore, instead of deleting the content of hashtags, we only deleted the ‘#’ symbols and used the hashtag content as part of the content of tweets to train the models.*”[31]
Sentiment analysis method	Which sentiment analyzer is most suited for the study context, particularly the characteristics of the social media data to be analyzed?	Among many options available, which sentiment analyzer to choose that would maximize the quality of the study analysis.	“*In this study, we use SentiStrength as (i) it has been used to measure the emotional content in online ED communities and shown good inter-rater reliability; (ii) it is designed for short informal texts with abbreviations and slang, and thus suitable to process tweets.*”[32]
	Whether to validate the selected sentiment analyzer on the study data.	Even if the selected sentiment analyzer has been applied by others to similar datasets in the past, it may still be worthwhile to conduct prestudy validation to ensure it performs satisfactorily on the data collected for the particular study.	“*In addition to the already mentioned evaluation of the accuracy and performance of EMOTIVE, a brief qualitative manual review of a sample of EMOTIVE’s output showed a consistent and correctly categorized set of emotions among the seven basic emotions.*”[33]
	If prestudy validation is to be performed, whether to obtain a manually annotated dataset for training or evaluation purposes.	To validate the performance of the selected sentiment analyzer, studies may want to obtain manual annotations of a subset of the study data, ideally with multiple coders so that interrater reliability can be assessed.	“*To identify and calibrate the classification model, 298 randomly selected posts were manually labeled by two independent annotators as belonging to either the positive or negative sentiment class. Cohen’s κ statistics ($\kappa = 0.82$) suggested high inter-annotator agreement. Then the two annotators discussed posts whose sentiment they initially disagreed on until they reached a consensus on sentiment labels.*”[34]

(continued)

Table 2. continued

Dimension	Design choice	Description	Example
	If prestudy validation is performed, whether the validation results are computed and reported using established quantitative metrics.	Studies should report the validation results based on commonly used quantitative evaluation metrics such as F score, or receiver-operating characteristic curve.	*"For this dataset, classifiers performed reasonably well, with F1 scores ranging from 0.48 to 0.68. However, the logistic regression classifier used with the n-gram model performed the best with an F1 score of 0.68. This performance is comparable with that in similar studies."*[35]
	Design choices specifically related to developing or training machine learning–based models.	In developing or training machine learning–based sentiment analyzers, studies should evaluate different competing models (eg, support vector machine, decision trees), as well as different features that may be selected to train the model (eg, bag of words, word vectors).	*"The n-gram model performed slightly better than the word-embedding model. For this dataset, classifiers performed reasonably well, with F1 scores ranging from 0.48 to 0.68."*[35]

Table 3. Methodological and reporting inconsistencies among the existing studies (N = 89)

Dimension	Item	Reported	Not Reported	Not applicable[a]
Platform selection (PS)	PS1. Description of the social media platform studied	89	0	0
	PS1-A. Justifications for selecting the social media platform	79	10	0
Data curation (DC)	DC1. Methods for retrieving study data	85	4	0
	DC2. Methods for determining the relevance and comprehensiveness of the data retrieved[b]	39	24	26[c]
	DC3. Differentiated treatment based on content creator[b]	13	45	31[c]
	DC4. Handling of special types of data (eg, images/videos, hashtags, emojis, hyperlinks)[b]	32	57	0
	DC4-A. Justifications for how special types of data are handled (n = 32)[b]	12	20	0
Sentiment analysis method (SAM)	SAM1. Description of the sentiment analyzer used	83	6	0
	SAM1-A. Justifications for selecting the sentiment analyzer[b]	40	49	0
	SAM2. (If machine learning) Description of the features selected (n = 29)[b]	19	10	0
	SAM2-A. (If machine learning) Justifications for selecting the features (n = 19)[b]	12	7	0
	SAM3. Validation of the sentiment analyzer before applying it to study data[b]	36	53	0
	SAM3-A. Annotated data used for validation or training (n = 36)	30	6	0
	SAM3-B. Whether multiple coders were involved in independently annotating the data (n = 30)[b]	15	11	4[d]
	SAM3-B-1. If multiple coders were involved, whether inter-rater reliability was quantitatively assessed and reported (n = 15)	13	2	0
	SMA3-C. Use of quantitative evaluation metrics for reporting the validation results (n = 36)	32	4	0

[a] Includes articles to which the particular research design choice did not apply. For example, the "methods for determining the relevance of the data retrieved" design choice might not be applicable to studies that focused on a disease-specific online patient forum, where all user posts were presumably relevant.

[b] A substantial degree of inconsistencies exists, defined as the research design choice or the rationale for making the choice being reported in fewer than two-thirds of the articles reviewed.

[c] These studies analyzed user posts in health forums, the data from which were presumably all relevant.

[d] Used annotated data from existing sources.

thoroughly reported in order to ensure the research validity and comparability of results in studies that apply sentiment analysis to analyzing user-generated health text in social media.

DISCUSSION

Social media has become an important resource of information for researchers to better understand patient journeys, their interactions with health systems and healthcare providers, as well as patients' and the general public's opinion toward important health policies and controversial medical interventions and treatments. A large number of such studies have been published in recent years, most of which used computational methods to analyze the sentiments expressed in the data. However, based on our systematic analysis of the literature, we found that there is a substantial degree of inconsistency in how such studies were conducted and how their results

PROTOCOL OF ANALYSIS OF SENTIMENT IN HEALTH (PATH)

	Design or Reporting Considerations	Description
Platform Selection	□ **PS1. Description of the social media platform studied**	Characteristics of the social media platform studied such as intended audience and interaction modality.
	□ **PS1-A. Justifications for selecting the social media platform**	Why is the chosen social media platform provide data that are most informative to answer the research questions of the study?
Data Curation	□ **DC1. Methods for retrieving study data**	What is the strategy used to retrieve study data, e.g., by keywords search or by targeting particular users with certain characteristics?
	□ **DC2. Methods for determining the relevance, and comprehensiveness (if applicable), of the data retrieved**	What are the methods used to ensure that the data retrieved are pertinent to the research topic(s) of interest, and are reasonably compete?
	□ **DC3. Differentiated treatment based on content creator**	Are data contributed by different entities such as laypersons, healthcare providers, and pharmaceutical companies treated differently?
	□ **DC4. Handling of special types of data (e.g., images/videos, hashtags, emojis, hyperlinks)**	Are special types of data retained, removed, or substituted in the analysis?
	□ **DC4-A. Justifications for how special types of data are handled**	Why are special types of data handled in the particular way, and what are the implications?
Sentiment Analysis Methods	□ **SAM1. Description of the sentiment analyzer**	What is the sentiment analysis tool or machine-learning model used in the study?
	□ **SAM1-A. Justifications for selecting the sentiment analyzer**	Why is the chosen sentiment analyzer most suited for the study context?
	□ **SAM2. (If machine learning) Description of the features selected**	What are the features used in the machine-learning model, and how are they selected?
	□ **SAM2-A. (If machine learning) Justifications for selecting the features**	Why are the chosen features most suited for the study context?
	□ **SAM3. Validation of the sentiment analyzer before applying it to study data**	How is the performance of the chosen sentiment analyzer assessed against the study data?
	□ **SAM3-A. Annotated data used for validation or training**	What are the training or evaluation data used, and how are these data obtained?
	□ **SAM3-B. Whether multiple coders were involved in independently annotating the data**	If applicable, are multiple coders involved in independently annotating the training or evaluation data?
	□ **SAM3-B-1. If multiple coders were involved, whether inter-rater reliability was quantitively assessed and reported**	What is the quantitative inter-rater reliability between the multiple coders?
	□ **SMA3-C. Use of quantitative evaluation metrics for reporting validation results**	What are the quantitative metrics (e.g., F-score, ROC) used to assess the analyzer performance?

Figure 2. Protocol of Analysis of senTiment in Health (PATH).

were reported, which may diminish the quality of research in addition to making it difficult to conduct meta-analyses to accumulate generalizable knowledge as a field. Subsequently, we discuss some of these methodological or reporting inconsistencies identified through this work and how they may affect research validity and comparability of results across studies.

First, some studies did not at all describe the process of sifting through available social media platforms to choose the ones that were most informative, in comparison with other competing social media outlets, to best answer the research questions at hand. Many simply stated that the chosen platform was a popularly used one, or commonly studied in prior research, or provided the easiest access to data. We believe such justifications, while may be reasonable due to practical reasons (eg, difficulties in accessing patient conversations in private Facebook groups), could potentially threaten the validity of the study, and researchers should use all means necessary to minimize possible data biases and improve the generalizability of their research results and conclusions. Indeed, previous studies that compared multiple social media platforms did find that different venues afforded different health content,[36,37] appealed to different user populations with distinctive characteristics,[38] or featured different interaction modality (eg, moderated vs not moderated) that may affect the nature of the discourses.[37] All of these factors could have significant implications on the results and conclusions of sentiment analysis research using health-related social media data.

Second, many existing studies did not conduct, or did not report, the data curation process for determining the relevance and comprehensiveness, if applicable, of the study data. This is particularly concerning in the analysis of health-related social media content because of the frequent use of ambiguous acronyms and abbreviations (eg, SOB for shortness of breath), similar medical concepts that may not be generally differentiated by laypersons (eg, dementia and Alzheimer's disease), and mixed usage of consumer language vs professional terms (eg, heart attack vs myocardial infarction). Further, very few studies treated their study data differently based on content creator, being laypersons, healthcare providers, health systems, government agencies, advocacy groups, or pharmaceutical companies. Depending on the objective of the study, this could result in "contaminated" data that did not truly reflect the sentiments of the target study population, and could consequently lead to imprecise or incorrect conclusions.[20,21,39,40] Future studies may consider adopting the methods proposed by Kim et al[41] and Adams et al[42] on how to develop and iteratively refine search keywords (eg, through word embeddings) for retrieving the content of interest from social media platforms, and how to thoroughly evaluate the relevance, and comprehensiveness (if applicable), of the information retrieved using manually annotated data. Furthermore, few studies described how they handled special types of data such as images or videos, hashtags, emojis, and hyperlinks, which are commonly used in social media discourses and can in fact convey important information about the sentiments being expressed.[17,43] However, this process was omitted from most existing studies, or was only causally mentioned (eg, all special types of data were removed) without providing any justification as to how the particular handling method used might affect the study results.

Third, most studies did not provide a rationale for choosing among many different sentiment analyzers available. Only a small number of the studies validated the selected tool to assess its performance (ie, precision and recall) against the study data. This can be problematic, as prior research has repeatedly demonstrated that different sentiment analyzers, especially those general-purpose ones developed or trained on datasets from nonhealth domains (eg, movie reviews), could produce substantially different results due to their poor domain transferability and the idiosyncrasies of health-related social media conversations.[17,44] Among the studies that did perform validation, only half involved multiple coders to annotate the training or evaluation data. This could also raise questions into research validity because previous studies have shown that annotating social media sentiments in general, and of health-related content in particular, is a challenging task even among experienced domain experts.[45,46] Therefore, having a single pair of eyes would not be considered sufficient for assuring the quality of annotations of such data.

The research design and reporting recommendation that we developed through this study, PATH, represents an initial step to address each of these issues. Applying a standardized protocol such as PATH in future health-related social media sentiment analysis research may also produce a higher level of consistencies in research design, conduct, and reporting, leading toward better comparability of results across studies. We believe that some elements of PATH, such as platform selection, data curation, and tool validation, also apply broadly for other studies that use computational methods to analyze health-related social media content, beyond just sentiment analysis. Therefore, we hope that this work will stimulate more critical reflection and development of standardized research protocols in a broader scope of computational analysis of social media data.

This study has several limitations. While sentiment analysis is an important tool for analyzing social media data, other methods such as topic modeling, spatiotemporal analysis, and social network analysis are also popularly used, which are not addressed in this article. Further, while we hope all elements proposed in the PATH protocol should be adhered to in future relevant studies, we understand that some desired research design choices may not be attainable due to resource constraints (eg, cost-prohibitive to involve multiple coders to annotate training or evaluation data) or practical reasons (eg, impossible to get data from the social media platform that provides the ideal user mix and the ideal content). The PATH protocol should therefore be interpreted as a set of recommended steps rather than mandatory requirements.

CONCLUSION

In this article, we systematically analyzed the body of literature that applied computational sentiment analysis to studying health-related social media content. The results highlighted a substantial degree of inconsistencies in how existing studies were conducted or how their results were reported. These findings led to the development of a recommended research design and reporting guideline, PATH. We believe that application of PATH in future sentiment analysis studies could lead to better research validity and comparability of results. The elements in the PATH protocol may also provide insights more broadly into other genres of research studies that use computational methods to analyze health-related social media data.

FUNDING

This work was supported in part by the National Center for Research Resources and the National Center for Advancing Translational Sciences of the National Institutes of Health through grant UL1TR001414. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

LH, TY, and ZH conducted the systematic review of the literature. LH and KZ designed the study and drafted the manuscript. YC and DAH assisted in designing the study and reviewed and revised the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

DATA AVAILABILITY STATEMENT

The data underlying this article is available in the online supplementary information files.

CONFLICT OF INTERESTS STATEMENT

The authors have no competing interests to declare.

REFERENCES

1. Pruksachatkun Y, Pendse SR, Sharma A. Moments of change: analyzing peer-based cognitive support in online mental health forums. In: *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019: 64: 1–13. doi: 10.1145/3290605.3300294
2. Cabling ML, Turner JW, Hurtado-de-Mendoza A, *et al.* Sentiment analysis of an online breast cancer support group: communicating about tamoxifen. *Health Commun* 2018; 33 (9): 1158–65.
3. Davis MA, Zheng K, Liu Y, Levy H. Public response to Obamacare on Twitter. *J Med Internet Res* 2017; 19 (5): e167.
4. Thelwall M, Thelwall S. Retweeting for COVID-19: consensus building, information sharing, dissent, and lockdown life. *arXiv*: 2004.02793; 2020.
5. Du J, Xu J, Song H, Liu X, Tao C. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *J Biomed Semantics* 2017; 8 (1): 9.
6. Du J, Xu J, Song H-Y, Tao C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Med Inform Decis Mak* 2017; 17 (S2): 69.
7. Shepherd M. Does the public response to the latest viral covid-19 cure video mark the death of critical thinking? *Forbes*. https://www.forbes.com/sites/marshallshepherd/2020/07/28/public-response-to-latest-viral-covid-19-cure-videois-it-the-death-of-critical-thinking/ Accessed July 31, 2020.
8. Pew Research Center Internet and Technology. Demographics of social media users and adoption in the United States. Social media fact sheet. https://www.pewinternet.org/fact-sheet/social-media/ Accessed August 1, 2019.
9. Pang B, Lee L. Opinion mining and sentiment analysis. *FNT Inf Retriev* 2008; 2 (1–2): 1–135.
10. Liu B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. Vol. 5, No. 1. San Rafael, CA: Morgan & Claypool; 2012: 1–67.
11. Huppertz JW, Otto P. Predicting HCAHPS scores from hospitals' social media pages: a sentiment analysis. *Health Care Manage Rev* 2018; 43 (4): 359–67.
12. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol* 2010; 29 (1): 24–54.
13. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*; 2010.
14. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2014: 55–60.
15. Stieglitz S, Mirbabaie M, Ross B, Neuberger C. Social media analytics – challenges in topic discovery, data collection, and data preparation. *Int J Inf Manag* 2018; 39: 156–68.
16. Denecke K, Deng Y. Sentiment analysis in medical settings: new opportunities and challenges. *Artif Intell Med* 2015; 64 (1): 17–27.
17. He L, Zheng K. How do general-purpose sentiment analyzers perform when applied to health-related online social media data? *Stud Health Technol Inform* 2019; 264: 1208–12.
18. Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics; 2007: 440–447. https://www.aclweb.org/anthology/P07-1056. Accessed July 9, 2019.
19. Martinez LS, Hughes S, Walsh-Buhi ER, Tsou M-H. "Okay, we get it. You vape": an analysis of geocoded content, context, and sentiment regarding e-cigarettes on Twitter. *J Health Commun* 2018; 23 (6): 550–62.
20. Zhang L, Hall M, Bastola D. Utilizing Twitter data for analysis of chemotherapy. *Int J Med Inform* 2018; 120: 92–100.
21. Allem J-P, Escobedo P, Dharmapuri L. Cannabis surveillance with twitter data: emerging topics and social bots. *Am J Public Health* 2020; 110 (3): 357–62.
22. Yin Z, Sulieman LM, Malin BA. A systematic literature review of machine learning in online personal health data. *J Am Med Inform Assoc* 2019; 26 (6): 561–76.
23. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *J Med Internet Res* 2013; 15 (4): e85.
24. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. *JMIR Public Health Surveill* 2018; 4 (2): e43.
25. Chancellor S, Baumer EPS, De Choudhury M. Who is the "human" in human-centered machine learning: the case of predicting mental health from social media. *Proc ACM Hum-Comput Interact* 2019; 3 (CSCW): 1–32.
26. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. *JMIR Med Inform* 2020; 8 (1): e16023.
27. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009; 151 (4): 264–9.
28. Corbin J, Strauss A. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks, CA: Sage; 2014.
29. Huh J, Yetisgen-Yildiz M, Pratt W. Text classification for assisting moderators in online health communities. *J Biomed Inform* 2013; 46 (6): 998–1005.
30. Bian J, Zhao Y, Salloum RG, *et al.* Using social media data to understand the impact of promotional information on laypeople's discussions: a case study of lynch syndrome. *J Med Internet Res* 2017; 19 (12): e414.
31. Yuan X, Crooks AT. Examining online vaccination discussion and communities in Twitter. In: *SMSociety '18: Proceedings of the 9th International Conference on Social Media and Society*; 2018: 197–206.
32. Wang T, Brede M, Ianni A, Mentzakis E. Social interactions in online eating disorder communities: a network perspective. *PLoS One* 2018; 13 (7): e0200800.
33. Gruebner O, Lowe SR, Sykora M, Shankardass K, Subramanian SV, Galea S. A novel surveillance approach for disaster mental health. *PLoS One* 2017; 12 (7): e0181233.
34. Zhao K, Yen J, Greer G, Qiu B, Mitra P, Portier K. Finding influential users of online health communities: a new metric based on sentiment influence. *J Am Med Inform Assoc* 2014; 21 (e2): e212–8.
35. Mamidi R, Miller M, Banerjee T, Romine W, Sheth A. Identifying key topics bearing negative sentiment on Twitter: insights concerning the 2015-2016 zika epidemic. *JMIR Public Health Surveill* 2019; 5 (2): e11036.
36. Roccetti M, Casari A, Marfia G. Inside chronic autoimmune disease communities: a social networks perspective to Crohn's patient behavior and medical information. In: *ASONAM '15: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*; 2015: 1089–96.
37. Wiley MT, Jin C, Hristidis V, Esterling KM. Pharmaceutical drugs chatter on Online Social Networks. *J Biomed Inform* 2014; 49: 245–54.
38. Haimson OL. Mapping gender transition sentiment patterns via social media data: toward decreasing transgender mental health disparities. *J Am Med Inform Assoc* 2019; 26 (8–9): 749–58.
39. Allem J-P, Ferrara E. The importance of debiasing social media data to better understand e-cigarette-related attitudes and behaviors. *J Med Internet Res* 2016; 18 (8): e219.
40. Broniatowski DA, Jamison AM, Qi S, *et al.* Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *Am J Public Health* 2018; 108 (10): 1378–84.
41. Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *J Med Internet Res* 2016; 18 (2): e41.

42. Adams N, Artigiani EE, Wish ED. Choosing your platform for social media drug research and improving your keyword filter list. *J Drug Issues* 2019; 49 (3): 477–92.
43. Hogenboom A, Bal D, Frasincar F, Bal M, de Jong F, Kaymak U. Exploiting emoticons in sentiment analysis. In: *SAC '13: Proceedings of the 28th Annual ACM Symposium on Applied Computing*; 2013: 703–10.
44. Lu Y, Hu X, Wang F, Kumar S, Liu H, Maciejewski R. Visualizing social media sentiment in disaster scenarios. In: *WWW '15: Proceedings of the 24th International Conference on World Wide Web*; 2015: 1211–5.
45. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. When 'bad' is 'good'": identifying personal communication and sentiment in drug-related tweets. *JMIR Public Health Surveill* 2016; 2 (2): e162.
46. Alvaro N, Conway M, Doan S, Lofi C, Overington J, Collier N. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. *J Biomed Inform* 2015; 58: 280–7.