Journal of Clinical and Translational Science

www.cambridge.org/cts

Implementation, Policy and Community Engagement

Cite this article: Zolnoori M, Huang M, Patten CA, Balls-Berry JE, Goudarzvand S, Brockman TA, Sagheb E, and Yao L. Mining news media for understanding public health concerns. *Journal of Clinical and Translational Science* **5**: e1, 1–10. doi: 10.1017/cts.2019.434

Received: 3 June 2019
Revised: 14 October 2019
Accepted: 17 October 2019

Keywords:
News; Reuters; Sentiment analysis; Topic modeling; Public health issue

Address for correspondence: L. Yao, PhD, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. Email: lixia.cn.yao@gmail.com

*These authors contributed equally to this work.

Mining news media for understanding public health concerns

Maryam Zolnoori[1,*], Ming Huang[1,*], Christi A. Patten[2,3], Joyce E. Balls-Berry[1,4], Somaieh Goudarzvand[5], Tabetha A. Brockman[2,3], Elham Sagheb[1] and Lixia Yao[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; [2]Center for Clinical and Translational Science, Community Engagement Program, Mayo Clinic, Rochester, MN, USA; [3]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, USA; [4]Mayo Clinic College of Medicine and Science, Rochester, MN, USA and [5]School of Computing and Engineering, University of Missouri-Kansas, Kansas City, MO, USA

Abstract

Introduction: News media play an important role in raising public awareness, framing public opinions, affecting policy formulation, and acknowledgment of public health issues. Traditional qualitative content analysis for news sentiments and focuses are time-consuming and may not efficiently convey sentiments nor the focuses of news media. *Methods:* We used descriptive statistics and state-of-art text mining to conduct sentiment analysis and topic modeling, to efficiently analyze over 3 million Reuters news articles during 2007–2017 for identifying their coverage, sentiments, and focuses for public health issues. Based on the top keywords from public health scientific journals, we identified 10 major public health issues (i.e., “air pollution,” “alcohol drinking,” “asthma,” “depression,” “diet,” “exercise,” “obesity,” “pregnancy,” “sexual behavior,” and “smoking”). *Results:* The news coverage for seven public health issues, “Smoking,” “Exercise,” “Alcohol drinking,” “Diet,” “Obesity,” “Depression,” and “Asthma” decreased over time. The news coverage for “Sexual behavior,” “Pregnancy,” and “Air pollution” fluctuated during 2007–2017. The sentiments of the news articles for three of the public health issues, “exercise,” “alcohol drinking,” and “diet” were predominately positive and associated such as “energy.” Sentiments for the remaining seven public health issues were mainly negative, linked to negative terms, e.g., diseases. The results of topic modeling reflected the media’s focus on public health issues. *Conclusions:* Text mining methods may address the limitations of traditional qualitative approaches. Using big data to understand public health needs is a novel approach that could help clinical and translational science awards programs focus on community-engaged research efforts to address community priorities.

Introduction

Identifying sentiments and focuses of news media toward public health concerns is an emerging research topic of interest. News media play a substantial role in raising public awareness, framing public opinions, and affecting policy formulation and adoption of popular issues [1–3]. In the area of healthcare, news media use multiple channels to communicate evidence-based research findings to individuals and healthcare professionals and accelerate the translation of these research findings in healthcare to public health practice. For example, the news media have a drastic impact on changing the public’s perceptions, attitudes, and behaviors toward smoking, alcohol-impaired driving, and healthcare service utilization [4].

News media have a tendency to use language to influence the public’s opinions, behaviors, and perceptions related to specific health issues. For instance, antismoking articles emphasized the health risks of smoking with negative sentiment (e.g., fatal diseases such as lung cancer) and the benefits of quitting with positive sentiment (e.g., healthy life) using research findings and real patient cases [5]. The news media influence the understanding of public health concerns by selecting specific aspects of a topic and presenting the concerns as salient news articles [6].

Previous research assessing sentiments and the focus of news media toward public health were conducted using traditional qualitative content analysis. Glenn *et al.* analyzed the sentiments of national online news and the readers’ comments for weight loss surgery [7]. They found that the sentiments of the news articles were mostly positive and supportive, while the sentiments of readers’ comments were predominately negative and associated with some negative terms such as “piggy” and “fatty.” Patterson *et al.* analyzed the content of seven UK national newspapers to identify the style of presentation of news media for women’s and men’s drinking [8]. Their findings indicated a difference by participants’ gender. For instance, men’s drinking was mostly associated with the topics of “violence” and “disorderly,” while the women’s drinking was frequently linked to the topics “out of control,” “putting themselves in danger,” “harming their physical appearance,” and “burdening men.” Although providing useful insights into the sentiments and focuses of news media for a specific healthcare issue, qualitative content analytic approaches

https://doi.org/10.1017/cts.2019.434 Published online by Cambridge University Press

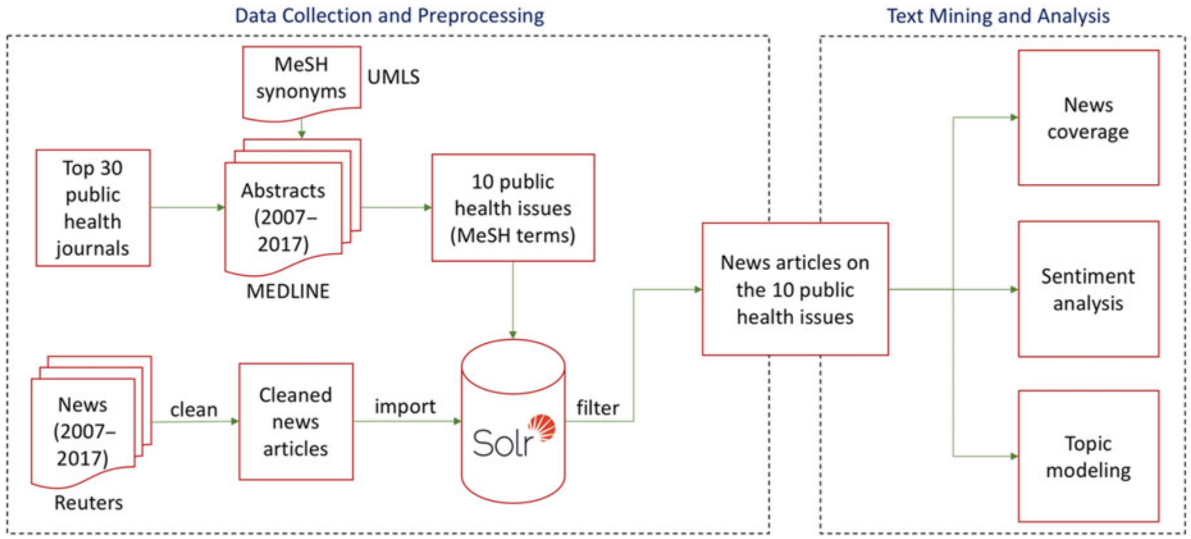

Fig. 1. A schematic view of methods for mining Reuters news. MeSH, Medical Subject Heading; UMLS, Unified Medical Language System.

are costly, time-consuming, and resource intensive. There is the subjective nature of traditional qualitative inquiry due to human perceptions and interpretations. This subjectivity does not avail itself to efficiently or systematically detecting sentiments and the focus of news media. Another key factor related to the sample sizes used in the qualitative content analysis is usually limited to a few hundred of news articles, which could limit the generalizability of the findings.

To address these challenges, we used state-of-art text mining methods including sentiment analysis and topic modeling, together with statistical analysis, to efficiently analyze more than 3 million Reuters news articles to identify news coverage, sentiments, and emphases toward public health issues from 2007 to 2017. We identified 10 major public health issues (i.e., "air pollution," "alcohol drinking," "asthma," "depression," "diet," "exercise," "obesity," "pregnancy," "sexual behavior," and "smoking") based on the top keywords from public health scientific journals. Sentiment analysis refers to the use of computerized algorithms for systemically evaluating opinions (e.g., negative or positive sentiments) and their intensities of the words and sentences in a large collection of text documents [9]. Topic modeling employs computerized algorithms to automatically discover the hidden topics in a large body of text documents related to a specific subject.

The analysis of news media data with advanced text mining techniques allows the discovery of sentiments and focuses of news media for public health issues. These discoveries could shed light on the understanding of the most pressing health concerns and provide insight for public policy. Moreover, to our knowledge, no previous work used sentiment analysis and topic modeling for identifying sentiments and focuses of news media for public health issues.

Methods

Fig. 1 shows a schematic view of the methods for mining Reuters news articles in this work. The methods consist of five main phases: (1) identifying the major public health issues from 30 top public health journals; (2) downloading, cleaning, and filtering news articles from Reuters news agency for the public health issues; (3) calculating the coverage of news articles over a decade linked to the public health issues and compare them with Google Trends searches; (4) analyzing sentiments of news articles related to the public health issues and their trends over time; and (5) identifying the focuses of the news articles associated with the public health issues. We briefly describe the five phases in the following subsections with detailed descriptions in Supplementary document 1. The Python scripts for mining news articles can be accessed via Github [10].

Identifying Major Public Health Issues

Public health research studies generally investigate major public health issues to provide new knowledge and insights to increase wellness of the general population. These studies are mainly published in public health journals that are indexed in the MEDLINE database [11]. Abstracts of the published articles including the keywords present their main research focuses. Thus, we selected 30 of the top public health journals (See Supplementary document 2) and downloaded 61,387 abstracts of articles published between January 1, 2007 and December 31, 2017 to identify the major publish health issues.

We mapped the article keywords to Medical Subject Headings (MeSH) terms. MeSH is a controlled terminology developed by National Library of Medicine for indexing articles in the MEDLINE database [12]. We identified the synonyms of the MeSH terms using Unified Medical Language System (UMLS) Metathesaurus [13,14]. The UMLS Metathesaurus is a collection of controlled terminologies and provides mapping structures between different medical vocabularies via concept unique identifier. Sequentially, we developed a Python script with regression expression [10] to identify the MeSH terms (and their synonyms) in the abstracts and calculated the frequencies of the MeSH terms. The frequency of a MeSH synonym is added to the frequency of the MeSH term.

After removing the MeSH terms which were not related to public health or whose relative frequencies are less than 1%, we found 90 popular MeSH terms on public health (see Supplementary document 3). We selected 10 major public health issues (Table 1) out of the 90 MeSH terms based on the frequencies.

https://doi.org/10.1017/cts.2019.434 Published online by Cambridge University Press

Table 1. *Frequencies of MeSH terms related to public health*

Public health issue (MeSH term)	Frequency
Smoking	4936
Pregnancy	3583
Obesity	1403
Air pollution	1009
Exercise	869
Diet	836
Sexual behavior	797
Alcohol drinking	788
Depression	755
Asthma	669

MeSH, Medical Subject Heading.

Collecting, Cleaning, and Filtering News Articles

Collecting News Articles From Reuters News Agency

Reuters News Agency is a leading global information media agency and the world's largest international text and television news provider [15]. We developed a web crawler in Python [10] to download news articles from an online archive of Reuters news agency [16]. We collected 3,763,737 articles between January 1, 2007 and December 31, 2017 to investigate major public health issues in the articles.

Cleaning News Articles

News articles contain noise and metadata that could affect the results of sentiment analysis and topic modeling toward public health issues. After reviewing a small sample of articles, we identified patterns that needed removing from the news articles. We removed repetitive special characters, such as "—", from the articles. We also removed the tags, such as "(Reuters)" and editorial information (i.e., "reporting by" or "editing by"). We replaced the sentence delimiters "*" and ">" using a period and replaced the hyperlinks, such as "http://topnews.session.com" with the word "link." In addition, readers' comments were deleted in the articles since the focus of the analysis is the content of the story reported.

Filtering News Articles on Public Health Issues

To filter articles related to the 10 public health issues (Table 1), we imported the articles into Apache Solr [17] for information indexing and searching. Apache Solr is an open source search platform built on Apache Lucene library. Apache Lucene provides rich features to handle document such as full-text search and real-time indexing for various applications [18,19]. The Reuters news data were filtered in Apache Solr to retrieve articles that mentioned the 10 public health issues.

News Coverage

Descriptive statistics were used to calculate the coverage of news media (i.e., numbers of news articles) for the 10 public health issues over time [20–23]. We compared the coverage trends of articles to Google Trends searches for the public health issues [20,21]. Google Trends analyzed the Internet search patterns of the individuals using Google search services over time [24]. The Internet search patterns reflect information-seeking behaviors of the individuals. Google Trends provides the majority of the Internet search services and makes the searching data publicly available.

Sentiment Analysis

Sentiment is a view of or attitude (e.g., positive or negative opinion) toward a situation or event. Sentiment analysis denotes systematic evaluation of opinions and their intensities of the words and sentences in a large collection of text documents by using computerized algorithms [9]. Sentiment analysis is widely used in the area of healthcare [9,15,25,26], particularly for identifying the attitudes and opinions of patients' posts in social media toward a specific healthcare issue. For example, Hopper and Uriyo used sentiment analysis to review patients' feedback for a selected group of gynecologists in Virginia [27]. In another study, Clark *et al.* applied sentiment analysis to quantify sentiments of patients toward breast cancer treatment experience [28]. In our previous work, we also used sentiment analysis to identify the sentiments of news articles toward hundreds of diseases and medical conditions [15]. Computerized algorithms were developed to automatize the process of sentiment analysis, enabling researchers to evaluate sentiments of a large volume of text documents.

In this work, we used a python module, Valence Aware Dictionary and sEntiment Reasoner (VADER) [29], to quantify the sentiments of the articles toward the 10 identified public health issues. VADER was specifically tuned to identify the sentiments of a wide range of social media data [15,30]. VADER reports a normalized and weighted sentiment score for a given sentence, according to predefined score of each word and embedded rules. The reported sentiment score is between −1.0 (the most negative) and 1.0 (the most positive), with 0.0 indicating neutral.

To improve the accuracy of sentiment analysis for each public health issue, we measured the sentiments of sentences containing MeSH terms and their synonyms for the public health issue. We calculated the average of sentiment scores of all sentences related to the public health issue in articles as a sentiment score for the public health issue. We computed the average sentiment scores of all sentences linked to a public health issue in all the news articles as a sentiment score of news for the public health issue in each year. We classified an article as positive, neutral, or negative, according to the threshold values suggested by VADER [30]. More specifically, if a sentiment score of a news article is equal to or larger than 0.05, the sentiment of the news article is positive; if a sentiment score of a news article is less than 0.05 and larger than −0.05, the news article has a neutral sentiment; otherwise, the news article is negative.

Topic Modeling

Due to the challenges (e.g., intensive human labor) of topic analysis with traditional manual qualitative methods, topic modeling methods were developed to automatically identify hidden topics in a massive collection of text documents [15,31,32]. The most frequently used topic modeling method is Latent Dirichlet Allocation (LDA) that was introduced by Blei *et al.* in 2003 [33]. LDA was extended and adopted in several domains for different purposes such as news themes on diseases [15], prognosis of human papillomavirus infection [34], and technology innovation in patents [31]. Although LDA is a powerful tool for discovering hidden topics in a large set of text documents, it is associated with some limitations. For example, LDA neglects the important word order in a text document. The text document is not treated as a sequence of words instead it is treated as a "bag" of words for topic

https://doi.org/10.1017/cts.2019.434 Published online by Cambridge University Press

modeling. LDA cannot automatically detect the number of topics in the text document and requires a predefined topic number as an input for topic modeling. Some posterior techniques such as perplexity and topic coherence [33,35] could help tune and find the appropriate number of topics in the text document, and they require extra computational time and involvement of domain experts to examine the generated topics for determining a good topic number. Given the number of topics, LDA infers topic distribution in each document (e.g., $0.5 * \text{topic}_1 + 0.3 * \text{topic}_2 + 0.2 * \text{topic}_3$ for the document_1) and word distributions over a topic (e.g., $0.4 * \text{word}_1 + 0.3 * \text{word}_2 + 0.2 * \text{word}_3 + 0.1 * \text{word}_4$ for the topic_1). During the inference of word distribution over a topic, LDA treats a common word (i.e., a word occurs equally across all topics) and a characteristic word (i.e., a word occurs dominantly in a few topics) equally when they have the same probability given a topic.

In response to these limitations, we used an advanced topic modeling method, Topic Keyword Model (TKM) [36,37], to identify the hidden topic structure of articles related to the 10 public health issues. TKM considers the word order in a text document for topic modeling. From a human perspective, it seems that multiple consecutive words in a text document have large probabilities to associate with the same topic. Similarly, TKM links a word to a topic if it or its adjacent words have a high association score with the topic. Thus, the topic that a word links to is heavily influenced by its nearby words. This way, the order of words was involved in topic modeling with TKM, compared to LDA. TKM measures the dissimilarity between topics and only keeps the topics that are significantly different from each other during topic modeling. TKM could potentially determine the appropriate number of distinct topics in the text document. In addition, TKM differentiates a common word and a characteristic word for a topic and adjusted the association probability (score) of a word with a topic according to its commonness and distinctiveness among topics whiling inferring word distributions over a topic.

After word stemming and lemmatization that reduce words into their base forms [38], we used the TKM package provided by its author in GitHub [37] to learn the topics of the articles linked to each of the 10 public health issues. Because topic modeling algorithms are unsupervised learning methods and they do not require prior annotation as gold standard to train the models, the standard metrics (e.g., recall, precision, and F-measure) for supervised learning methods are not suitable for evaluating the results of topic modeling. Therefore, we asked domain experts to evaluate the results of topic modeling [39].

Results

Findings of News Coverage

We calculated the coverage of articles associated with each of the 10 public health issues every year between 2007 and 2017. The results are compared with the numbers of Google Trends searches for the 10 public health issues as shown in Fig. 2. We rescaled these numbers relative to the highest number on each subfigure. Fig. 2 shows that the numbers of news articles for the seven public health issues "Smoking," "Exercise," "Alcohol drinking," "Diet," "Obesity," "Depression," and "Asthma" were constantly decreasing over years. The numbers of news articles for the remainder of the public health issues, "Sexual behavior," "Pregnancy," and "Air pollution" fluctuated in the study period. We found that the decreasing trends of Google searches for "Smoking" and "Obesity" are in line with the trends of relevant articles. In contrast, the number of Google searches for "Alcohol drinking" steadily increased over time, which had a negative correlation with the number of articles for "Alcohol drinking."

Findings of Sentiment Analysis

Fig. 3 shows the frequencies of positive, neutral, and negative sentiments of the articles toward the 10 public health issues. We found that the sentiments of the news articles on the three of the public health issues, "exercise," "alcohol drinking," and "diet" were predominately positive (i.e., 55.6%, 43.4%, and 45.6%, respectively), implying that the articles associated these issues with positive terms, such as "happiness," "energy," or terms showing overall healthy life. For "alcohol drinking," we found that there were more news articles with positive sentiment (43.4%) than news articles with negative sentiment (29.1%), which is surprising due to the public concern about alcohol misuse.

For the public health issues of "smoking," "obesity," "sexual behavior," "depression," "pregnancy," "asthma," and "air pollution," Reuters published 50.0%–89.1% articles with negative sentiment and 8.9%–31.6% articles with positive sentiment. This occurred more frequently because the articles were linked the topics with negative terms, such as diseases, symptoms, low quality of life, "pressure," "hopelessness," and "worrying." In addition, we found that among these public health issues, "smoking" is the mostly mentioned in the articles and "depression" had the largest coverage percentage (89.1%) of articles with negative sentiment.

Fig. 4 shows the sentiment scores of the media toward the 10 public health issues over 11 years (2007–2017). During 2007–2017, "depression" had the lowest sentiment score that fluctuated between −0.6 and −0.4, compared with other public health issues. On the other hand, "exercise" showed the highest sentiment score after 2007 and its sentiment score steadily increased between 2007 and 2017. This finding could imply that exercise was increasingly linked to terms for reducing disease risk and improving life quality in the news articles.

Findings of Topic Modeling

We used TKM to identify topics of articles associated with 10 public health issues. In this section, we presented the identified topics for two public health issues, "smoking" and "alcohol drinking" (Fig. 5). For the topics associated with the remaining eight public health issues, please see Supplementary document 4.

TKM identified 14 topics of the news articles related to "smoking." We selected five most meaningful topics and showed them in Fig. 5 (see Supplementary document 4 for the rest nine topics). By interpretation of the identified topic keywords, we could find that the five meaningful topics of the news articles on "smoking" was mostly related to "tobacco and cigarette," "industry," "adolescent smoking," "cancer," and "cardiovascular disease."

For articles on "alcohol drinking," TKM discovered 16 topics and we selected and illustrated five most meaningful topics in Fig. 5. The remainder of the 11 topics is in Supplementary document 4. After interpretation of the identified topic keywords, we found that the five meaningful topics on "alcohol drinking" are "health research," "driving," "wine industry," "culture and diet constraint," and "opioid."

Discussion

We used descriptive statistics and state-of-art text mining techniques including sentiment analysis and topic modeling to identify

https://doi.org/10.1017/cts.2019.434 Published online by Cambridge University Press

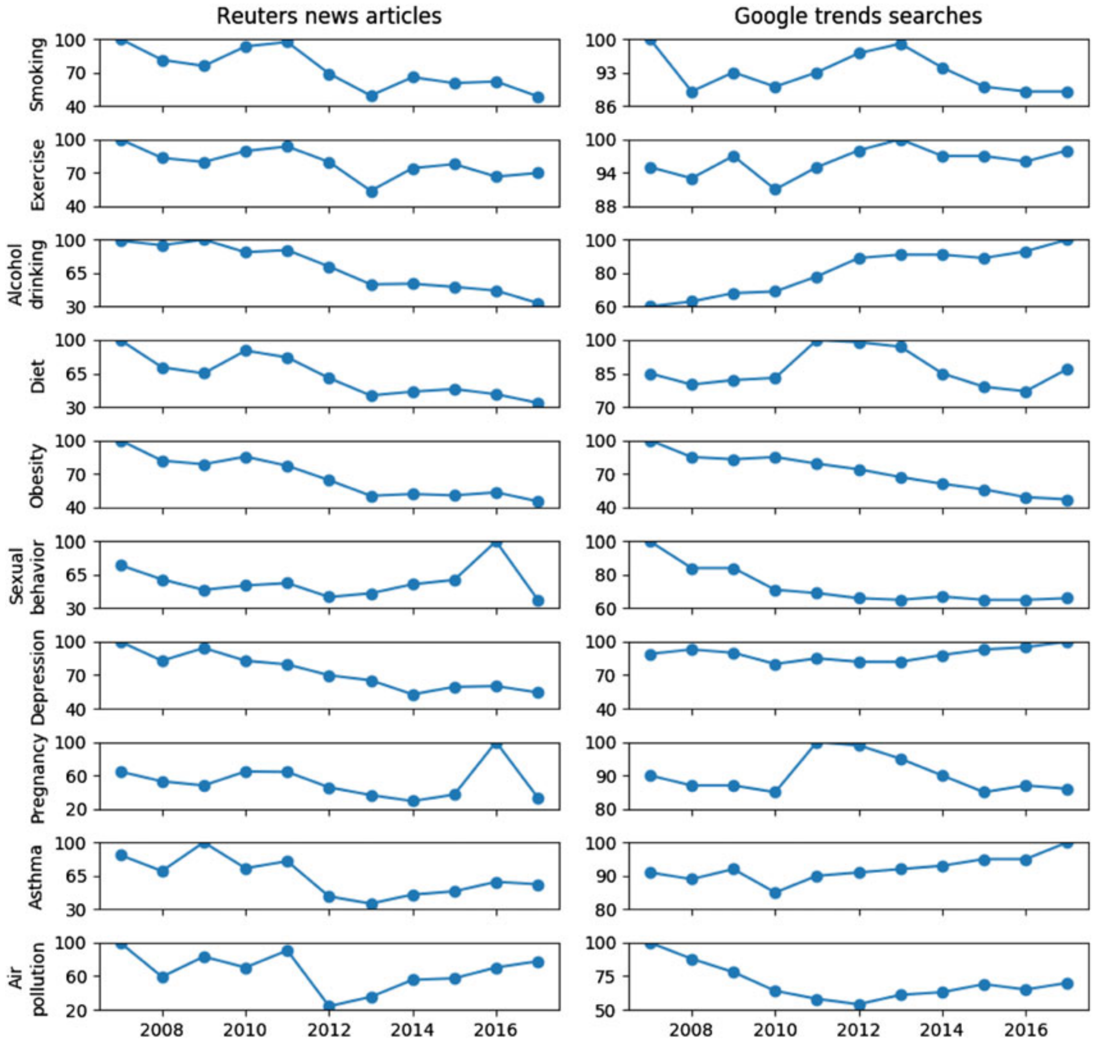

Fig. 2. Normalized numbers of articles and Google Trends searches for the 10 public health issues over time. The numbers are normalized to the highest point on each subfigure. A value of 100 represents the peak popularity for the public health issue.

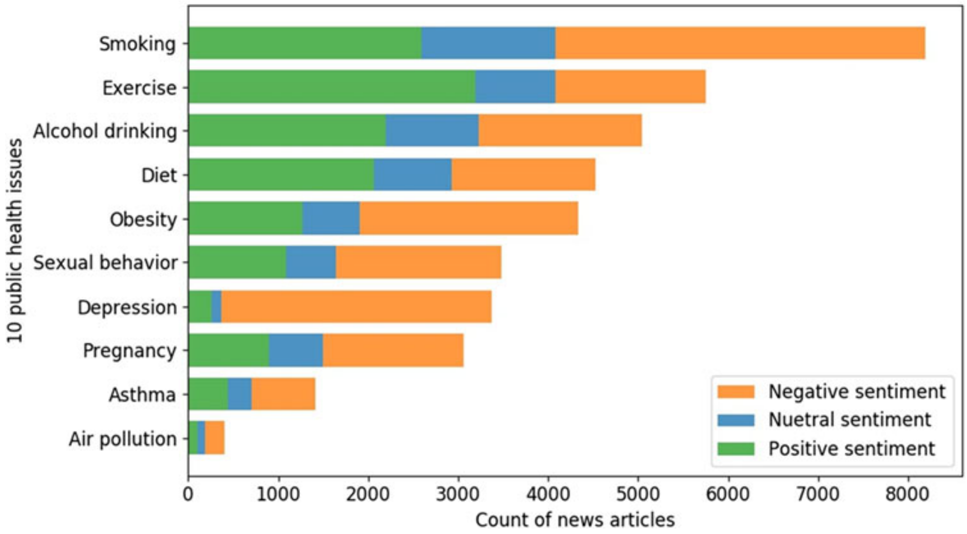

Fig. 3. Counts of news articles with positive, neutral, and negative sentiments toward 10 public health issues.

https://doi.org/10.1017/cts.2019.434 Published online by Cambridge University Press

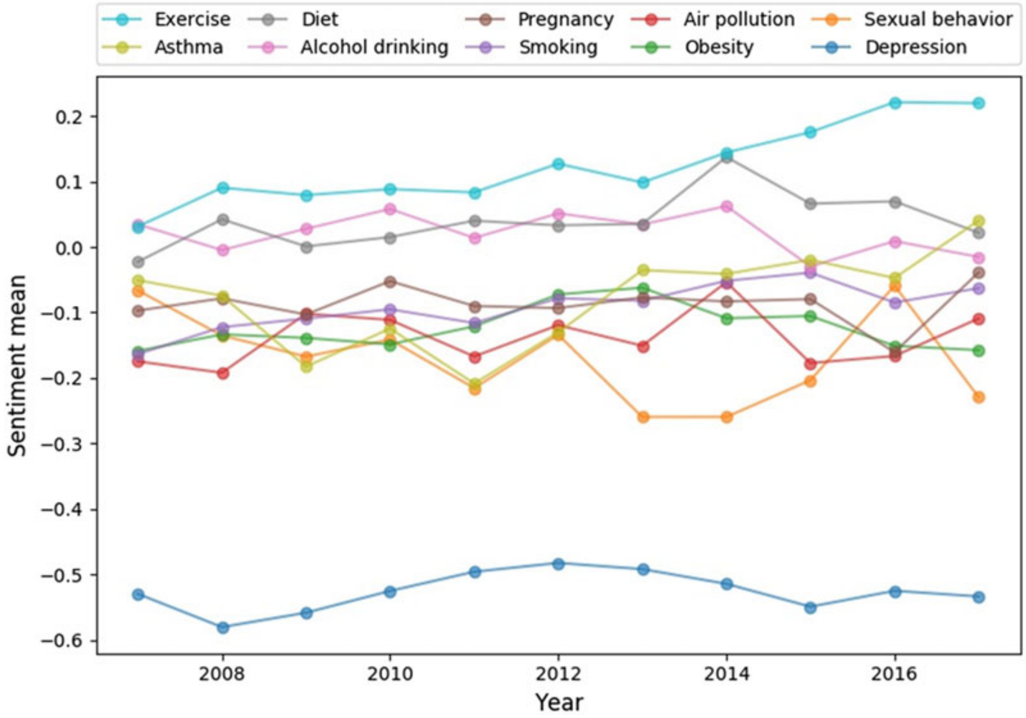

Fig. 4. Sentiment scores of news media toward 10 public health issues over 11 years (2007–2017).

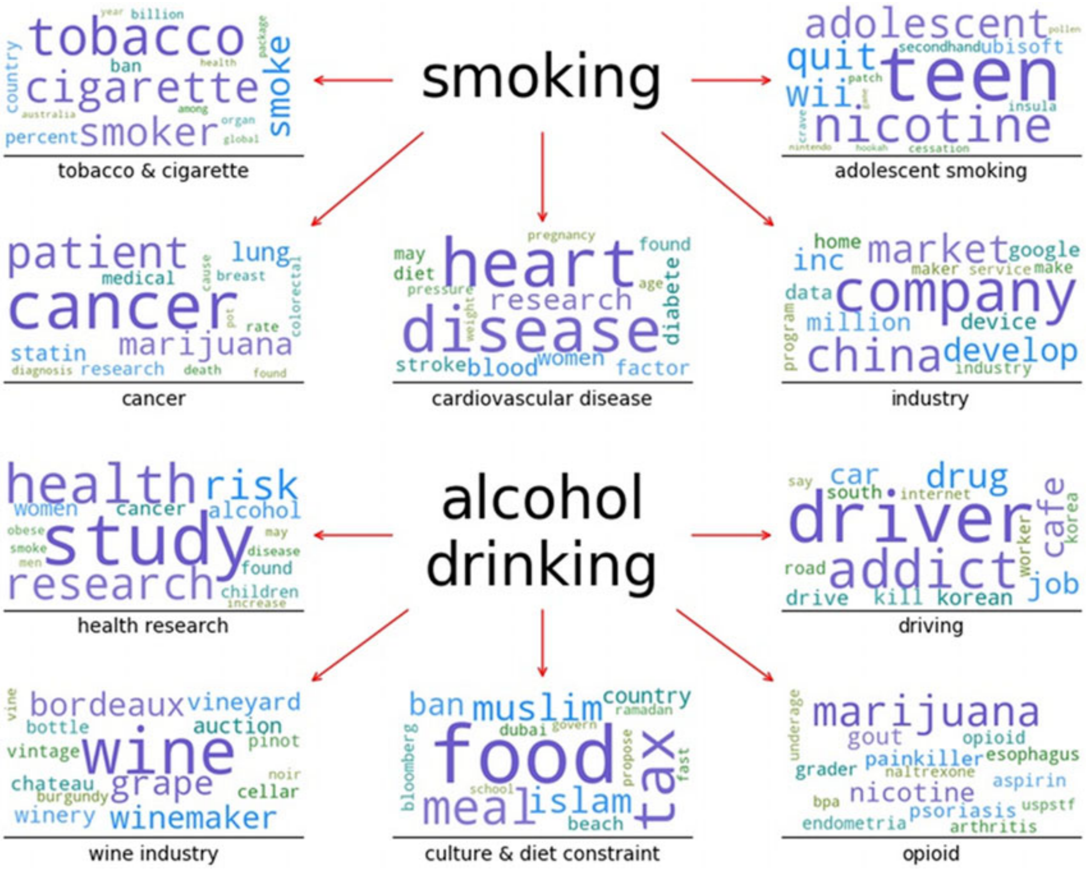
smoking
tobacco cigarette smoker smoke ban billion year health package country australia among organ global percent
tobacco & cigarette
adolescent teen quit nicotine wii secondhand ubisoft pollen patch insula crave nintendo hookah cessation
adolescent smoking
patient cancer lung medical breast cause colorectal pot rate statin marijuana diagnosis research death found
cancer
heart disease research pregnancy found may diet pressure weight age diabete stroke blood women factor
cardiovascular disease
home market google inc maker service make data company million device program china develop industry
industry
alcohol drinking
health risk women cancer alcohol obese smoke men study may disease found research children increase
health research
say car drug south internet driver worker cafe korea road addict drive kill korean job
driving
vine bordeaux vineyard bottle auction vintage wine pinot chateau grape noir burgundy cellar winery winemaker
wine industry
ban muslim country ramadan dubai govern food tax bloomberg propose fast school meal islam beach
culture & diet constraint
underage marijuana gout opioid grader painkiller esophagus naltrexone aspirin bpa nicotine psoriasis uspstf endometria arthritis
opioid

Fig. 5. Word clouds of five meaningful topics identified in news articles related to the public health issues, “smoking” and “alcohol drinking.”

https://doi.org/10.1017/cts.2019.434 Published online by Cambridge University Press

the sentiments of over 3 million news articles toward public health issues and discover the hidden topic structures of these articles that discussed the issues. We selected 10 popular public health issues from 30 top public health journals indexed in the MEDLINE database. The 10 major public health issues are "air pollution," "alcohol drinking," "asthma," "depression," "diet," "exercise," "obesity," "pregnancy," "sexual behavior," and "smoking."

The coverage of articles associated with each of the seven public health issues, "Smoking," "Exercise," "Alcohol drinking," "Diet," "Obesity," "Depression," and "Asthma" had a declining trend over years. The decreasing trend of articles on "Smoking" is correlated with the trends of Google Trends searches and adult smoking prevalence in the US [40]. The reduced concern of the public media and individuals on "Smoking" were probably resulted from a series of smoking regulations and actions taken by the US government such as tobacco control initiatives since the 1960s [41]. The same trends of news articles and Google Trends searches were observed for the public health issue, "Obesity."

The results of sentiment analysis showed that the sentiments of over 43% news articles toward "exercise," "diet," and "alcohol drinking" were positive. The dominantly positive sentiments "exercise" and "diet" could imply that the news articles mostly focused on the importance of healthy diet and regular exercise and their relationship with disease prevention and high quality of life. For example, articles such as "Study details how high fiber diets make for healthier lives" [42] and "Aerobic exercise eases depression, even in chronically ill" [43] linked the healthy diet and exercise to a healthy lifestyle and disease treatment. There is evidence indicating that "exercise" and/or "diet" (e.g., a Mediterranean diet) serve as a preventive or disease-modifying treatment of diseases such as dementia [44,45], Parkinson's disease [46–48], and cardiovascular disease [49–51]. These studies possibly explain the constant increase of sentiment score of news media toward "exercise" over time. It seems to be surprising that the majority of the articles on "alcohol drinking" has positive sentiment, due to the public concern about excessive alcohol use. However, news articles such as "Drinking alcohol may keep leg arteries healthy" [52] and "Moderate drinkers have a better health, study finds" [53] showed how articles had a positive sentiment by highlighting the positive aspects of "alcohol drinking." The attitude of articles toward "alcohol drinking" supported the findings by Mostofsky *et al.* [54]. According to their study, moderate alcohol drinking was associated with lower risk of cognitive decline and heart diseases. But moderate or higher alcohol intakes increase the risk of diseases such as breast cancer and bone fracture, particularly in women.

On the other hand, the sentiments of more than 50% articles were negative for "smoking," "obesity," "sexual behavior," "depression," "pregnancy," "asthma," and "air pollution." The predominantly negative sentiments for these public health issues could be justified in two different ways: (1) the potential link of the public health issues to diseases, low work productivity, or poor quality of life. For example, articles such as "Air pollution a leading cause of cancer" [55], "Poor mental health harming productivity, says OECD" [56], and "Sexual harassment, abuse tied to real health effects" [57] linked "air pollution," "mental disorder" (e.g., depression), and "inappropriate sexual behavior" to cancer, low productivity, and side effects on health, respectively. (2) The healthcare services and public health policies and programs are not effective enough to manage the issues. For example, articles such as "Kids with asthma often leave doctor's office with unanswered questions" [58], "Obesity medical bill could reach $1.2 trillion a year by 2025" [59], and "Pot-smoking on the rise among U.S. pregnant women" [60] indicate the ineffectiveness of public health and healthcare interventions for managing "asthma," "obesity," and "pregnancy," respectively. This is particularly the case for "depression" with the lowest sentiment score from the media, because it is currently one of the major public health concerns in our society and it is associated with several limitations in daily functioning and social participation [61].

We identified the main topics of the articles on the 10 public health issues using topic modeling. The highlighted five meaningful topics for two public health issues, "smoking" and "alcohol drinking." The five major topics emphasized in the articles related to "smoking" were "tobacco and cigarette," "industry," "cancer," "cardiovascular disease," and "adolescent smoking." The topics "cancer" and "cardiovascular disease" indicate that these fatal diseases were strongly associated with smoking in the media. It is well documented that smoking is linked to more than 12 cancers, particularly, lung cancer [62]. About 90% lung cancers are caused by tobacco smoking or secondhand smoke exposure [62]. Smoking is linked to cardiovascular health and cardiovascular disease with smoking attributing to about 140,000 premature deaths annually [63]. The topic "adolescent smoking" suggests that tobacco use in adolescence is another focus of the articles. In 2013, about 18% middle school students and 46% high school students were tobacco users [64]. Nearly 90% cigarette smokers first tried cigarette smoking by age 18, and the prevention of tobacco use in adolescence is critical to reducing tobacco epidemic in the US [41,65].

The five major topics uncovered in the articles on "alcohol drinking" are "health research," "driving," "wine industry," "culture and diet constraint," and "opioid." The topic "health research" may show that health-related research on "alcohol drinking" was mainly discussed in the articles, for example, the news articles "Cutting back on alcohol can prevent cancers: experts" [66] and "Moderate drinking helps heart, but don't binge" [67]. The topic "opioid" suggests the mixed use of opioid and alcohol and related risks. The opioids are effective painkiller but have the potential to be additive [68]. The combined use of opioid and alcohol increases the risk of overdose and injury [69]. The topic "driving" implies that the alcohol-impaired driving was another focus of many of the articles, possibly because driving under the influence of alcohol remains a public health problem [70].

The use of big data to understand public health issues is a novel way to provide clinical and translational science awards programs insight on community priorities that lend themselves to community-engaged research approaches [71]. Our other research indicates that meaningful community engagement offers the opportunity to promote bi-directional dialogue about health research with diverse communities [71–73]. The process used in this study offers a new way to identify topics for future dialogs with community-engaged stakeholders to set research priorities.

The discussion of this study cannot be considered fully without analyzing its limitations. The first limitation is related to the data source. We used articles published by Reuters between 2007 and 2017 for studying major public health issues because Reuters is a leading news media organization and the largest international text news provider. Although we are confident that the findings are not atypical for other media sources that cover public health issues, the results of coverage, sentiment analysis and topic modeling might not be generalized to other news media agencies (e.g., Associated Press) or other forms of media (e.g., television, radio or social media). In addition, the findings of news media may not reflect the concerns of the US population which is another indication that the findings lend themselves for starting a dialog about these topics as potential areas of focus for health-related research with the community.

The second limitation of the study is the synonyms generated using UMLS for each public health issue. Since UMLS is a compendium of standard medical terminologies, it might not include all synonyms that Reuters journalists use when writing about public health issues.

The third limitation relates to the sentiment analysis method (VADER) used for identifying sentiments of articles toward the public health issues. Although VADER has been evaluated using articles published by New York Times, it has not been tested or tuned on news articles from Reuters.

The fourth limitation is about the topic modeling method. We used TKM to identify topics in news articles for each public health issue. Although TKM addresses the limitations of LDA, topic modeling with TKM was performed in a completely unsupervised fashion. To evaluate the results of topic modeling, we relied on domain experts' judgments that might create bias in interpretation of these topics.

Conclusion

In this study, we identified 10 important public health issues after analyzing more than 60,000 abstracts of 30 top public health journals. We analyzed over 3 million Reuters articles during 2007–2017 identified their sentiments with linkages to the 10 public health issues, using state-of-art text mining methods including sentiment analysis and topic modeling.

Our results show that the coverage of news articles associated with each of the seven public health issues, "Smoking," "Exercise," "Alcohol drinking," "Diet," "Obesity," "Depression," and "Asthma" had a declining trend over years. The coverage of news articles for the rest three public health issues, "Sexual behavior," "Pregnancy," and "Air pollution" fluctuated over time. For sentiment analysis, the sentiments of the news articles for the three public health issues, "exercise," "alcohol drinking," and "diet," were predominately positive. It suggests that the articles associated these issues with positive terms such as energy. For the remainder of the seven public health issues including "smoking," "obesity," "sexual behavior," "depression," "pregnancy," "asthma," and "air pollution," most articles had negative sentiments. It may indicate that the articles mostly linked these issues to negative terms such as diseases or symptoms.

Our study showed that text mining methods may address the limitations associated with traditional qualitative approaches. Our analysis could provide valuable insights about the sentiments and topic structures of articles discussing public health issues. Our findings could offer valuable information for the healthcare professionals and policy makers.

Acknowledgments. Funding for this study was provided by Mayo Clinic Center for Clinical and Translational Science (UL1TR002377) from the National Institutes of Health/National Center for Advancing Translational Sciences (NIH/NCATS), and the National Library of Medicine (5K01LM012102).

Disclosures. All authors have no conflict of interest to declare.

Supplementary Material. To view supplementary material for this article, please visit https://doi.org/10.1017/cts.2019.434

References

1. **Vasudevan V**, ***et al.*** Effectiveness of media and enforcement campaigns in increasing seat belt usage rates in a state with a secondary seat belt law. *Traffic Injury Prevention* 2009; **10**(4): 330–339.
2. **Leurer MD**. Lessons in media advocacy: a look back at Saskatchewan's nursing education debate. *Policy, Politics, & Nursing Practice* 2013; **14**(2): 86–96.
3. **Gardner A**, ***et al.*** Clinic consortia media advocacy capacity: partnering with the media and increasing policymaker awareness. *Journal of Health Communication* 2010; **15**(3): 293–306.
4. **Bou-Karroum L**, ***et al.*** Using media to impact health policy-making: an integrative systematic review. *Implementation Science* 2017; **12**(1): 52.
5. **Wakefield MA, Loken B, Hornik RC**. Use of mass media campaigns to change health behaviour. *The Lancet* 2010; **376**(9748): 1261–1271.
6. **Weishaar H**, ***et al.*** Why media representations of corporations matter for public health policy: a scoping review. *BMC Public Health* 2016; **16**(1): 899.
7. **Glenn NM, Champion CC, Spence JC**. Qualitative content analysis of online news media coverage of weight loss surgery and related reader comments. *Clinical Obesity* 2012; **2**(5–6): 125–131.
8. **Patterson C**, ***et al.*** Content analysis of UK newspaper and online news representations of women's and men's 'binge' drinking: a challenge for communicating evidence-based messages about single-episodic drinking? *BMJ Open* 2016; **6**(12): e013124.
9. **Liu B, Zhang L**. A survey of opinion mining and sentiment analysis. In: Aggarwal CC, Zhai C-X, eds. *Mining Text Data*. New York: Springer, 2012, pp. 415–463.
10. **Huang M**, ***et al.*** News mining source codes. 2019; Retrieved from https://github.com/huangming6220/news_mining.
11. **National library of Medicine**. *MEDLINE®: description of the database.* 2019; Retrieved from https://www.nlm.nih.gov/bsd/medline.html.
12. **Lipscomb CE**. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 2000; **88**(3): 265.
13. **Schuyler PL**, ***et al.*** The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association* 1993; **81**(2): 217–222.
14. **Bodenreider O**. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 2004; **32**(suppl_1): D267–D270.
15. **Huang M**, ***et al.*** Public opinions toward diseases: infodemiological study on news media data. *Journal of Medical Internet Research* 2018; **20**(5): e10047.
16. **Reuters**. *Reuters site archive, United States*. Retrieved from: https://www.reuters.com/news/archive. Accessed February 5, 2018.
17. **Apache Software Foundation**. *Apache Solr*; 7.0.1; Retrieved from http://lucene.apache.org/solr/. Accessed October 5, 2017.
18. **Chen X**, ***et al.*** DataMed–an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association* 2018; **25**(3): 300–308.
19. **Singh G**, ***et al.*** QTLTableMiner++: semantic mining of QTL tables in scientific articles. *BMC Bioinformatics* 2018; **19**(1): 183.
20. **Hahn W**, ***et al.*** The effect of media attention on concern for and medical management of methicillin-resistant Staphylococcus aureus: a multimethod study. *Journal of Public Health Management and Practice* 2009; **15**(2): 150–159.
21. **Mahabir R**, ***et al.*** News coverage, digital activism, and geographical saliency: a case study of refugee camps and volunteered geographical information. *PLoS One* 2018; **13**(11): e0206825.
22. **Buckton CH**, ***et al.*** A discourse network analysis of UK newspaper coverage of the "sugar tax" debate before and after the announcement of the Soft Drinks Industry Levy. *BMC Public Health* 2019; **19**(1): 490.
23. **Motta M, Callaghan T, Sylvester S**. Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Social Science & Medicine* 2018; **211**: 274–281.
24. **Google Inc**. Google trends searches. 2019; Retrieved from https://trends.google.com/trends/?geo=US.
25. **Mäntylä MV, Graziotin D, Kuutila M**. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review* 2018; **27**: 16–32.

26. **Rastegar-Mojarad M**, ***et al.*** Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In: *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*; 2016.
27. **Hopper AM, Uriyo M**. Using sentiment analysis to review patient satisfaction data located on the internet. *Journal of Health Organization and Management* 2015; **29**(2): 221–233.
28. **Clark EM**, ***et al.*** *A Sentiment Analysis of Breast Cancer Treatment Experiences and Healthcare Perceptions Across Twitter*. arXiv preprint arXiv:1805.09959; 2018.
29. **Trilling D**. *Doing computational social science with python: an introduction*. Available at SSRN 2737682; 2018.
30. **Hutto CJ, Gilbert E**. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: *Eighth International AAAI Conference on Weblogs and Social Media*; 2014.
31. **Huang M**, ***et al.*** Technological innovations in disease management: text mining us patent data from 1995 to 2017. *Journal of Medical Internet Research* 2019; **21**(4): e13316.
32. **He K**, ***et al.*** Understanding the patient perspective of epilepsy treatment through text mining of online patient support groups. *Epilepsy & Behavior* 2019; **94**: 65–71.
33. **Blei DM, Ng AY, Jordan MI**. Latent dirichlet allocation. *Journal of Machine Learning Research* 2003; **3**(January): 993–1022.
34. **Kim YS**, ***et al.*** Clinical progress of human papillomavirus genotypes and their persistent infection in subjects with atypical squamous cells of undetermined significance cytology: statistical and latent Dirichlet allocation analysis. *Experimental and Therapeutic Medicine* 2017; **13**(6): 3032–3038.
35. **Stevens K**, ***et al.*** Exploring topic coherence over many models and many topics. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics; 2012.
36. **Schneider J, Vlachos M**. Topic modeling based on keywords and context. In: *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM; 2018.
37. **Johannes Schneider MV**. *TKM Package*. 2018; Retrieved from https://github.com/JohnTailor/tkm.
38. **Liu H**, ***et al.*** BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics* 2012; **3**(1): 3.
39. **Suaysom N, Gu W**. Expert opinion and coherence based topic modeling. *International Journal on Natural Language Computing* 2018; **7**(2): 1–14.
40. **Jamal A**. Current cigarette smoking among adults—United States, 2005–2015. *MMWR. Morbidity and Mortality Weekly Report* 2016; **65**(44): 1205–1211.
41. **US Department of Health and Human Services**. *The health consequences of smoking—50 years of progress: a report of the Surgeon General*. 2014; Atlanta, GA: US Department of Health and Human Services, Centers for Disease ....
42. **Kelland K**. *Study details how high fiber diets make for healthier lives*. 2019; Retrieved from https://www.reuters.com/article/us-health-fibre/study-details-how-high-fiber-diets-make-for-healthier-lives-idUSKCN1P42W5.
43. **Papaport L**. *Aerobic exercise eases depression, even in chronically ill*. 2019; Retrieved from https://www.reuters.com/article/us-health-depression-fitness/aerobic-exercise-eases-depression-even-in-chronically-ill-idUSKCN1QN2UI.
44. **Ahlskog JE**, ***et al.*** Physical exercise as a preventive or disease-modifying treatment of dementia and brain aging. In: *Mayo Clinic Proceedings*. Elsevier; 2011.
45. **Morris MC**, ***et al.*** MIND diet slows cognitive decline with aging. *Alzheimer's & Dementia* 2015; **11**(9): 1015–1022.
46. **Canning CG**, ***et al.*** Exercise for falls prevention in Parkinson disease: a randomized controlled trial. *Neurology* 2015; **84**(3): 304–312.
47. **Ahlskog JE**. Aerobic exercise: evidence for a direct brain effect to slow Parkinson disease progression. In: *Mayo Clinic Proceedings*. Elsevier; 2018.
48. **Maraki MI**, ***et al.*** Mediterranean diet adherence is related to reduced probability of prodromal Parkinson's disease. *Movement Disorders* 2019; **34**(1): 48–57.
49. **Gielen S**, ***et al.*** Exercise training in patients with heart disease: review of beneficial effects and clinical recommendations. *Progress in Cardiovascular Diseases* 2015; **57**(4): 347–355.
50. **Fiuza-Luces C**, ***et al.*** Exercise benefits in cardiovascular disease: beyond attenuation of traditional risk factors. *Nature Reviews Cardiology* 2018; **15**(12): 731–743.
51. **Estruch R**, ***et al.*** Primary prevention of cardiovascular disease with a Mediterranean diet supplemented with extra-virgin olive oil or nuts. *New England Journal of Medicine* 2018; **378**(25): e34.
52. **Hendry J**. *Drinking alcohol may keep leg arteries healthy*. 2007; Retrieved from https://www.reuters.com/article/us-alcohol-leg-arteries/drinking-alcohol-may-keep-leg-arteries-healthy-idUSSAT86642020071228.
53. **Kelland K**. *Moderate drinkers have a better health, study finds*. 2010; Retrieved from https://www.reuters.com/article/us-alcohol/moderate-drinkers-have-better-health-study-finds-idUSTRE64I2AK20100519.
54. **Mostofsky E**, ***et al.*** Key findings on alcohol consumption and a variety of health outcomes from the Nurses' Health Study. *American Journal of Public Health* 2016; **106**(9): 1586–1591.
55. **Nebehay KKS**. *Air pollution a leading cause of cancer - U.N. agency*. 2013; Retrieved from https://www.reuters.com/article/us-cancer-pollution/air-pollution-a-leading-cause-of-cancer-u-n-agency-idUSBRE99G0BB20131017.
56. **Russell KKR**. *Poor mental health harming productivity, says OECD*. 2011; Retrieved from https://www.reuters.com/article/us-mental-work/poor-mental-health-harming-productivity-says-oecd-idUSTRE7BB0MK20111212.
57. **Carroll L**. *Sexual harassment, abuse tied to real health effects*. 2018; Retrieved from https://www.reuters.com/article/us-health-assault-harassment/sexual-harassment-abuse-tied-to-real-health-effects-idUSKCN1MD26O.
58. **Crist C**. *Kids with asthma often leave doctor's office with unanswered questions*. 2019; Retrieved from https://www.reuters.com/article/us-health-asthma-kids/kids-with-asthma-often-leave-doctors-office-with-unanswered-questions-idUSKCN1QF2TP.
59. **World Health Organization**. *W.O.F. Factbox: Obesity medical bill could 'reach $1.2 trillion' a year by 2025*. 2017; Retrieved from https://www.reuters.com/article/us-health-obesity-problem/factbox-obesity-medical-bill-could-reach-1-2-trillion-a-year-by-2025-idUSKBN1CF34H.
60. **Rapaport L**. *Pot-smoking on the rise among U.S. pregnant women*. 2017; Retrieved from https://www.reuters.com/article/us-health-pregnancy-usa-marijuana/pot-smoking-on-the-rise-among-u-s-pregnant-women-idUSKBN1EK19Z.
61. **Gilbert P**. *Depression: The Evolution of Powerlessness*. New York: Routledge; 2016.
62. **US Department of Health and Human Services**. *Let's Make the Next Generation Tobacco-Free: Your Guide to the 50th Anniversary Surgeon General's Report on Smoking and Health*. Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention. National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2014.
63. **US Department of Health and Human Services**. *The Health Consequences of Smoking: A Report of the Surgeon General*. Vol. **62**. Atlanta, GA: Centers for Disease Control and Prevention; 2004.
64. **Arrazola RA**, ***et al.*** Tobacco use among middle and high school students—United States, 2013. *MMWR. Morbidity and Mortality Weekly Report* 2014; **63**(45): 1021.
65. **US Department of Health and Human Services and Center for Disease Control and Prevention**. Preventing Tobacco Use. Among Youth and Young Adults: A Report of the Surgeon General. 2012; Atlanta, GA.
66. **Weinstock CP**. *Cutting back on alcohol can prevent cancers: experts*. 2017; Retrieved from: https://www.reuters.com/article/us-health-cancer-alcohol/cutting-back-on-alcohol-can-prevent-cancers-experts-idUSKBN1D72ZX.
67. **Harding A**. *Moderate drinking helps heart, but don't binge*. 2010; Retrieved from: https://www.reuters.com/article/us-moderate-drinking/moderate-drinking-helps-heart-but-dont-binge-idUSTRE62M55F20100323.
68. **Højsted J, Sjøgren P**. Addiction to opioids in chronic pain patients: a literature review. *European Journal of Pain* 2007; **11**(5): 490–518.
69. **Rassool GH**. *Alcohol and Drug Misuse: A Guide for Health and Social Care Professionals*. New York: Routledge; 2017.

https://doi.org/10.1017/cts.2019.434 Published online by Cambridge University Press

70. **Fan AZ**, ***et al.*** Drinking and driving among adults in the United States: results from the 2012–2013 national epidemiologic survey on alcohol and related conditions-III. *Accident Analysis & Prevention* 2019; **125**: 49–55.
71. **Patten C**, ***et al.*** Addressing community health needs through community engagement research advisory boards. *Journal of Clinical and Translational Science* 2019; **3**(s1): 82–82.
72. **Khubchandani J**, ***et al.*** Community-engaged strategies to increase diversity of participants in health education research. *Health Promotion Practice* 2016; **17**(3): 323–327.
73. **Balls-Berry JE**, ***et al.*** Using Garden Cafés to engage community stakeholders in health research. *PLoS One* 2018; **13**(8): e0200483.

https://doi.org/10.1017/cts.2019.434 Published online by Cambridge University Press