OJPHI

Understanding Discussions of Health Issues on Twitter: A Visual Analytic Study

Oluwakemi Ola[1*], Kamran Sedig[2]

[1]University of British Columbia, [2]Western University

ABSTRACT

Social media allows for the exploration of online discussions of health issues outside of traditional health spaces. Twitter is one of the largest social media platforms that allows users to post short comments (i.e., tweets). The unrestricted access to opinions and a large user base makes Twitter a major source for collection and quick dissemination of some health information. Health organizations, individuals, news organizations, businesses, and a host of other entities discuss health issues on Twitter. However, the enormous number of tweets presents challenges to those who seek to improve their knowledge of health issues. For instance, it is difficult to understand the overall sentiment on a health issue or the central message of the discourse. For Twitter to be an effective tool for health promotion, stakeholders need to be able to understand, analyze, and appraise health information and discussions on this platform. The purpose of this paper is to examine how a visual analytic study can provide insight into a variety of health issues on Twitter. Visual analytics enhances the understanding of data by combining computational models with interactive visualizations. Our study demonstrates how machine learning techniques and visualizations can be used to analyze and understand discussions of health issues on Twitter. In this paper, we report on the process of data collection, analysis of data, and representation of results. We present our findings and discuss the implications of this work to support the use of Twitter for health promotion.

Keywords: Twitter; Social Media; Online Discussion Analysis; Health Issues; Visual Analytics; Interactive Visualizations; Machine Learning; Health Sentiments

Correspondence: *kemiola@cs.ubc.ca

DOI: 10.5210/ojphi.v12i1.10321

1 Introduction

Health information can be gathered from diverse media, including social media. Using social media allows stakeholders to explore online discussions occurring outside of traditional health spaces in a rapid fashion [1], [2]. Twitter is one of the largest social media platforms, with over 320 million active accounts [3]. This platform allows users to post short comments (i.e., tweets) that contain 280 characters or less. Tweets may also contain pictures, videos, or links to webpages. Users can like, retweet (i.e., repost a tweet), and reply to tweets. Unregistered users can only read

tweets. The unrestricted access to opinions and a large user base has made Twitter a source for the collection and dissemination of information for various domains including health [4-7].

Currently, health organizations are using Twitter to promote healthy lifestyle choices, identify disease outbreaks, explore human behavior, and assess the public's perception of health issues [2], [8-11]. These organizations also use Twitter for health promotion. The Department of Health and Human Services in the United States is one such organization that uses Twitter to provide the public with actionable health information [12]. In addition to health organizations, individuals, news organizations, businesses, interest groups, and a host of other entities discuss health issues on Twitter.

On any given day, over 500 million tweets are posted [3]. The voluminous number of tweets presents many challenges to those who seek to use Twitter to improve their knowledge of a wide variety of health issues and understand ongoing discussions. Observational studies on specific health issues on Twitter show many formal and informal conversations taking place [13]. While following a health organization's Twitter account may be beneficial for learning about a specific health hazard, for stakeholders who want to obtain a high-level understanding of the social discourse on a wide variety of health issues, challenges abound. Currently, it is difficult for stakeholders to understand the overall sentiment on a health issue, the types of users involved in the discourse, and the content of their tweets. The platform's open participatory nature and the brevity of a given tweet message can result in the distortion of information [14], [15]. In addition, the quality of the information varies and the identity of the individual tweeting, which is helpful in evaluating the tweet's credibility, is not always known [13], [15]. For Twitter to be an effective tool for health promotion, stakeholders need to be equipped to understand and appraise health information on the platform [16]. A high-level understanding can help address misinformation and equip individuals with a better conceptual model to assess how health issues are discussed. In addition to supporting the information-seeking tasks of the public, an analysis of the health discourse on Twitter can benefit health professionals and social scientists by providing them with a lens through which they can better understand the public's perception of these issues and effectively utilize Twitter for health promotion [17], [18].

Manual content annotation and computational models have been used to analyze the discourse of health on Twitter. Studies that utilize manual content analysis have looked at health issues such as swine flu, dental pain, concussions, breast cancer, and marijuana use [19-23]. These studies typically involve content analysis of a small set of tweets (e.g., 1,000 to 10,000). Manual content analysis studies are typically time-consuming because they require the manual coding of tweets by individuals. On the other hand, computational models have been employed to analyze large samples of Twitter data promptly. Some of the work has focused on sentiment analysis. Sentiment analysis involves using natural language processing and computational linguistics to characterize sentiment, opinion, attitudes, and emotion from written language [24]. Salathé and Khandelwal [10] applied sentiment analysis to understand the perception of the H1N1 vaccine on Twitter. Myslin et al. [25] used machine learning classifiers to deduce sentiment for tweets related to tobacco usage. In addition to sentiment, Cole-Lewis et al. [1] used machine learning techniques to classify tweets based on user description, genre, theme, and relevance to the topic of e-cigarettes. Existing research has focused predominantly on understanding one or two health topics on Twitter.

The goal of this paper is to build on this research and provide insight into a variety of health issues through a visual analytic study. Visual analytics enhances the understanding of data by combining computational models with interactive visualizations [26-28]. A recent survey on visual analytics highlights the need for more research in supporting the use of social media data in public health practice [29]. Our study is meant to demonstrate how machine learning techniques and visualizations can be used to analyze and understand discussions of health issues on Twitter. To this end, we retrieved over half a million health-related tweets, and randomly selected a sample of 3000 on which we conducted manual content analysis. We used the sample to create models that classified tweets based on their content and user category. These models were then applied to the larger tweet dataset. Finally, we created a visualization that supports the exploration of the discourse of health issues in the tweet corpus. In this paper, we report our findings and discuss their implications.

The rest of the paper is organized as follows. Section 2 presents the research method. Section 3 discusses the results. Section 4 highlights the limitations of the work. Section 5 covers the implications of this work. The final section, Section 6, presents general conclusions.

2 Method

In this study, supervised machine learning was used to build classification models that assess themes of tweets and categories of Twitter users. For our analysis, we are more concerned about what is being said about certain health issues as opposed to their frequency or popularity. In addition to the tweet text, Twitter allows developers to access relevant metadata about the user who posted the tweet. User information includes username, description of the account, the number of followers, the number of people the user is following, and the number of tweets the user has posted [30]. In this section, we describe how the data was collected and processed.

2.1 Data Collection

In the past, hashtags and search terms have been used to search for health-related tweets [31-34]. We opted to use search terms. Our initial list of search terms comprised causes of death that have been identified by the Institute for Health Metrics and Evaluation (IHME) [35]. We utilized these causes as search terms primarily because this work is part of a larger research plan to facilitate sensemaking of health data and we wanted to have a consistent terminology. IHME classifies causes into 21 cause-clusters, which are aggregated into three main groups: 1) non-communicable, 2) injury-based, and 3) communicable, maternal, neonatal, and nutritional.

To get a better understanding of the ability of these terms to provide relevant tweets, we collected a sample of over 50,000 tweets. We utilized Tweepy [36]—a Twitter application programming interface—to search for and retrieve the tweets. Iteratively, for each search term, we retrieved up to 200 recent tweets to determine whether the search terms predominantly retrieved health-related tweets. In certain situations, search terms were modified to improve results. For instance, the forces of nature search term was expanded to include earthquake death, tsunami death, flood death, and hurricane death. Appendix A includes the final list of the 117 search terms used. Over a one-month period, we retrieved tweets using the search terms. The total number of English language tweets retrieved during this period was 547,921. The tweets were stored in a MongoDB database.

2.2 Analysis

2.2.1 Sentiment Analysis

Similar to existing research practices, we measured sentiment as being either negative, positive, or neutral [1], [10], [32]. For our study, we utilized AlchemyAPI's sentiment analysis tool to assign polarity and sentiment value to our tweets. AlchemyAPI, a company acquired by IBM, was a text mining platform that extracted metadata such as keywords and sentiment from text-based documents [37]. We selected AlchemyAPI because, at the time of this study, it was one of the leading free sentiment analysis tools with a high accuracy rate [38-40]. For a text fragment, AlchemyAPI returns a sentiment category and score. The sentiment score is in the range -1 to +1 and expresses the strength of the sentiment. The category is based on the score value. For a score less than 0, the category is negative, for a score over 0, the category is positive, and for a score of 0, the category is neutral. Table 1 includes some of the tweets and the corresponding sentiment score and category it was assigned.

Table 1. Sample of AlchemyAPI sentiment analysis of health tweets

Tweet	Score	Category
Involved lymph nodes in HPV positive oropharyngeal cancer Regional control is preserved after dose de excavated	0.0000	neutral
Ambulance came in hospital with atrial flutter on like this	-0.2296	negative
Share the love via CandyGram amp support to feed people affected by HIV AIDS valentinesday	0.4615	positive

2.2.2 Manual Annotation

To obtain a better understanding of who was tweeting and the content of each tweet, we analyzed 500 tweets that were randomly selected from the corpus. Based on previous research [1] and our analysis, five content themes and six categories of users were established. The five identified content themes are as follows:

- Educational: post about relevant health-related news, factoid, resource, research, or public health announcement. Tweet that contains general health information, research, or information to raise awareness on a health issue. For example,
    - "Brain cancer two essential amino acids might hold key to better outcome cancer News"
    - "Preparation and Characterization of Irinotecan Loaded Cross Linked Bovine Serum Albumin Heads for Liver Cancer"
- Fundraising: post that seeks to raise funds or solicit money or services for a health organization, cause, or individual needing medical treatment. For example,
    - "That dollar goes to the Measles and Rubella Initiative to buy a vaccine for a child against Measles and Rubella"
    - "LETS SAVE A LIFE Baron has suffered with Throat cancer for 5 years and lung cancer for eyes Your contribution matters"

- Personal: post in which the user is giving an opinion on a health issue, reporting on their own personal health status, or asking health-related questions. For example,
  - "His bronchitis has my chest feeling heavyyyyyy"
  - "I am wheeling like an old man with asthma after a joy Thank you of"
- Promotional: post promoting or advertising a for-profit health event or product. For example,
  - "Find out how you can prevent and reverse diabetes won The At Real Good Health Summit"
  - "Or Lane Vishnubala will be teaching our coming Of Obesity and Diabetes Specialist Instructor course"
- Unrelated: post that contains search terms but is unrelated to health. For example,
  - "I feel like I am drowning without your looooooveeeeeeeeee"
  - "Nationalism is an infantile disease It is the measles of mankind"

The user categories are as follows:

- Businesses: for-profit organizations, e.g., retailers, pharmaceutical companies, fitness companies.
- Celebrities: famous people in pop culture, politics, sports and news media.
- Interest Groups: unofficial organizations for specific health interests, e.g., school groups, health food groups, anti-vaccination groups.
- Media: reputable news source such as New York Times, Washington Post, Wall Street Journal, Associated Press and reputable journals that publish health research.
- Official Agencies: government agencies and large non-government health agencies, e.g., National Institutes of Health, Centers for Disease Control and Prevention, American Heart Association.
- Public: general public that does not fall into one of the aforementioned categories.

After establishing the categories, three thousand tweets were coded. Table 2 shows the categorization of these tweets. Overall, 74.3% of the tweets were found to be health-related tweets. The predominant user category is the general public which accounts for 75.5% of the tweets. For the content category, the predominant theme is education with 45.7%. These tweets served as the test and training data for our classification models. In the next section, we describe how the classification models were constructed.

Table 2. Categorization of tweets by user and content (n = 3000)

User		**Content**	
Category	**Frequency**	**Theme**	**Frequency**
Public	2264 (75.5%)	Educational	1370 (45.7%)
Interest Groups	227 (7.6%)	Personal	770 (25.7%)
Media	227 (7.6%)	Unrelated	761 (25.3%)

Businesses	215 (7.2%)	Promotional	66 (2.2%)
Celebrities	40 (1.3%)	Fundraising	33 (1.1%)
Official Agencies	27 (0.9%)		

2.2.3 Model Construction

Our models were constructed with the Scikit Learn library [41] for Python. We used the Bag-of-Words approach, which is a 3-step process that involves transforming the text into numerical features, which are then analyzed. The first step is tokenization, which involves splitting each document (i.e., tweet or text) into words based on whitespace and punctuation. Next, the occurrences of each word are counted and stored in a matrix. The last step involves normalizing and weighting the occurrences. Normalization is important because when dealing with a large corpus, common words like 'a' and 'the', which frequently appear, typically convey little meaningful information about the content of the document. Re-weighting was done with the frequency-inverse document frequency (tf-idf) transform, which helps to measure how important a word is to a document in a collection by taking into consideration the number of times a word appears in a document and the frequency of the word across the entire corpus [42]. In the following subsections, we discuss how models were constructed for user categories and content themes.

User Category

We utilized Support Vector Machine (SVM) models for classification as previous research points to the benefits of using SVM for short text (e.g., tweets) [1], [25]. Models were created based on the following attributes:

- **User description:** a user-provided string that describes their account (e.g., "United Nations Development Programme helps empower lives & build resilient nations. To learn more, follow @ASteiner & visit: http://www.undp.org").
- **User verification status:** indicates whether the account has been deemed authentic by Twitter. Twitter authenticates an account so that the public is aware that the account holder's identity has been verified. This is typically done for individuals in the entertainment, government, religious, media, business, or sports spheres.
- **User screen name:** unique user name or handle name that is used to identify the tweeter, typically preceded by the @ symbol in tweets (e.g., @UNDP, @WHO, @UNICEF).
- **Influence score:** this attribute helps determine how influential an account is on Twitter. Past research notes that influence is not solely based on the number of people that follow an account on Twitter but is also affected by the number of people the account follows [43]. The score is calculated by dividing the number of followers by the number of people that the account follows. For instance, for @UNDP the number of followers is 1.13 million while the following is 4656. The influence score is 242.70.

Table 3 shows the average accuracy rate for 100 runs for four different models. Accuracy rate is defined as the percentage of observations that were correctly classified in the test dataset. For each run, 80% of the coded tweets (i.e., 2400 tweets) were used to train the model, while the remaining 20% (i.e., 600 tweets) were used to test the model. The experiment was run 100 times for each of the models created. The model with the highest accuracy rate was Model A1, which used the user description alone. Subsequent models that incorporate the username, influence score, and user verification status of the account, resulted in lower accuracy rates.

Table 3. Accuracy rate for user category model construction (n = 600)

Model	Average Accuracy Rate (%)
A1: description	86.86
B1: description and screen name	79.83
C1: description, screen name, and influence score	79.84
D1: description, screen name, influence score, and user verification status	79.75

Tweet Theme

Machine learning models were built for the tweet theme. We used a Bag-of-Words approach and Support Vector Machine technique for our models. The first model uses the tweet, the second model uses the tweet text as well as the number of reserved news words (e.g., newspaper, news, official), the third model uses the tweet and the verification status of the tweeter's account and the last model uses the tweet, the verification status, and the number of reserved news keywords. Table 4 shows the average accuracy rate for the tweet themes for the four models. The experiment was run 100 times for each of the models created. For each run, 2400 tweets (i.e., 80% of the coded tweets) were used to train the model, while the remaining 600 tweets (i.e., 20%) were used to test the model.

Table 4. Accuracy rate for tweet theme model construction (n = 600)

Model	Average Accuracy Rate (%)
A2: tweet	80.99
B2: tweet and number of reserved keywords	81.09
C2: tweet and user verification status	81.14
D2: tweet, number of reserved keywords and user verification status	81.44

Based on the experimental analysis of model construction, we used Model A1 to classify the user categories and Model D2 to classify the tweet themes. 24% of the tweets were classified as unrelated and were removed. In the next section, we discuss the results of the remaining tweets.

3 Results

A total of 416,900 tweets remained in our corpus after unrelated tweets were removed; these tweets represent over 100 different causes that contribute to mortality. Each tweet has a sentiment score and type, category for the user who sent the tweet, and content theme. In this section, we first present a brief overview of the results, describe the design of a visualization we created to facilitate understanding discussions of health on Twitter, and then highlight results for certain cause-clusters.

Table 5 shows the frequency of tweets categorized by sentiment, theme, and user group. 73% of the tweets were deemed negative, while 27% of the tweets were either positive or neutral. Similar to the manually coded data, the majority of tweets in our corpus were tweeted by the general public (84 %). The tweets by the media and official agencies made up less than 5% of the corpus. This is important to note because individuals may assume that a significant portion of health-related tweets are from reputable sources, which is not the case. In terms of the content, 66% of the tweets were educational tweets, while personal themed tweets made up 34% of the corpus. Combined, fundraising and promotional tweets were less than 1 percent.

Table 5. Frequency of tweets by sentiment, theme, and user categories (n = 416,900)

Sentiment	Percent (%)	Theme	Percent (%)	User	Percent (%)
negative	72.85	educational	65.99	businesses	4.98
neutral	14.47	fundraising	0.16	celebrities	0.01
positive	12.68	personal	33.62	interest group	6.71
		promotional	0.23	media	4.73
				official agency	0.04
				public	83.52

The visualization described in this section includes prevalent words (i.e., non-search terms that frequently appear in the corpus) and the net sentiment rate for causes as well as clusters of causes. In the context of tweets, net sentiment rate is defined as the subtraction of the number of negative tweets from the number of positive tweets, divided by the total number of tweets.

$$\text{Net Sentiment Rate} = \frac{\text{number of positive tweets} - \text{number of negative tweets}}{\text{total number of tweets}}$$

3.1 Description of Sentiment Visualization

Using JavaScript and the d3.js visualization library [44], we created a visualization to facilitate the exploration of the results. Figure 1 shows the default configuration of the visualization. The visualization has three main parts. The first part is comprised of circular arcs that frame the rest of the visualization. These arcs represent the top 50 words across the entire corpus. The size and location of each arc depict its prevalence. The larger the arc, the more times the word appeared in

the corpus. By hovering over the arc (i.e., a word), the number of occurrences appears. The arcs are arranged from left to right in descending order based on prevalence. As shown, the words get, health, like, women, may, type, and new are frequent words in the corpus.

Some of the screenshots used in the figures only include partial representations of the entire visualization; this is done to aid in the reading of the textual content in the visualizations. The central portion of the visualization (see Figure 1), is a variation of a visualization developed by Bremer [45]. It depicts the breakdown of tweets by cause-clusters, user category, and tweet theme. In the center of the visualization is a list of the 21 cause-clusters arranged in descending order according to the number of tweets. The diabetes, urogenital, blood/endocrine cluster has the highest number of tweets in the corpus, while the transport injuries cluster has the least.

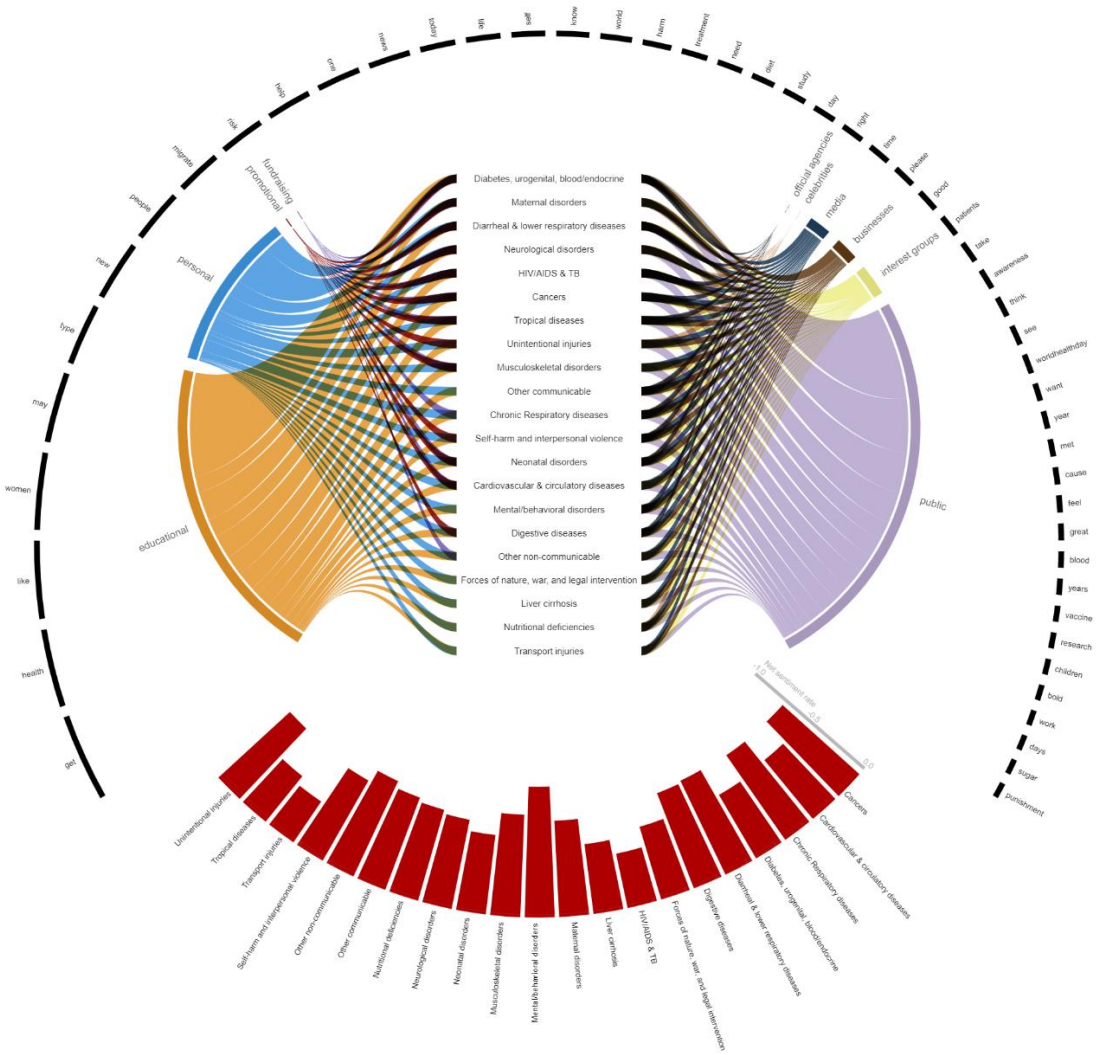

Figure 1: Default configuration of the sentiment visualization

On the left side of the cluster list is a sub-visualization of the tweets by content themes. The links that branch out of each theme represent the presence of tweets for a cause-cluster. For instance, Figure 2 shows a partial screenshot of the visualization when the promotional theme is selected.

There are 13 links for the promotional theme because there are no promotional tweets for the other eight cause-clusters. The clusters that do not have promotional-themed tweets are greyed out.

Figure 2: Screenshot of sentiment visualization with the promotional theme selected

The right sub-visualization shows the breakdown of user categories, and is encoded in a similar fashion as the left sub-visualization. For instance, Figure 3 shows the state of the visualization when the celebrities user category is selected.

Figure 3: Screenshot of sentiment visualization with the celebrities user category selected

It is worth mentioning that the content themes and user categories are arranged based on the number of tweets. We use both size and location to encode quantity so that it is easier for users to determine which group is bigger. For example, for the user categories (see Figure 1), the media (4.73%) and businesses (4.98%) arcs appear to be the same size but because the arcs are ordered, users of the tool can deduce that the businesses category has more tweets. The lower portion of the visualization has two alternating views. The first view is shown in Figure 1 and it depicts the net sentiment rate for cause-clusters. The second sub-visualization (see Figure 4) depicts sentiment for the causes that make up a specific cluster. This sub-visualization contains curved heatmaps and is divided into two parts. The first part shows the breakdown of sentiment by the user categories and the second by the theme of the tweets. The sections of the heatmap are encoded with color, where red is used to indicate negative sentiment, green for positive, and grey is used to depict the absence of data. For instance, as shown in Figure 4 when the cardiovascular & circulatory diseases cluster is selected, the visualization shows that there are no tweets from official agencies or celebrities for all the causes that make up the cluster. Also, the atrial flutter, hemorrhagic stroke, cardiomyopathy, and peripheral arterial disease causes have a net sentiment score that is positive for certain themes and user categories. With this visualization, users can explore the sentiment for different causes and cause-clusters, learn about the different user groups, and understand the general themes of the tweets.

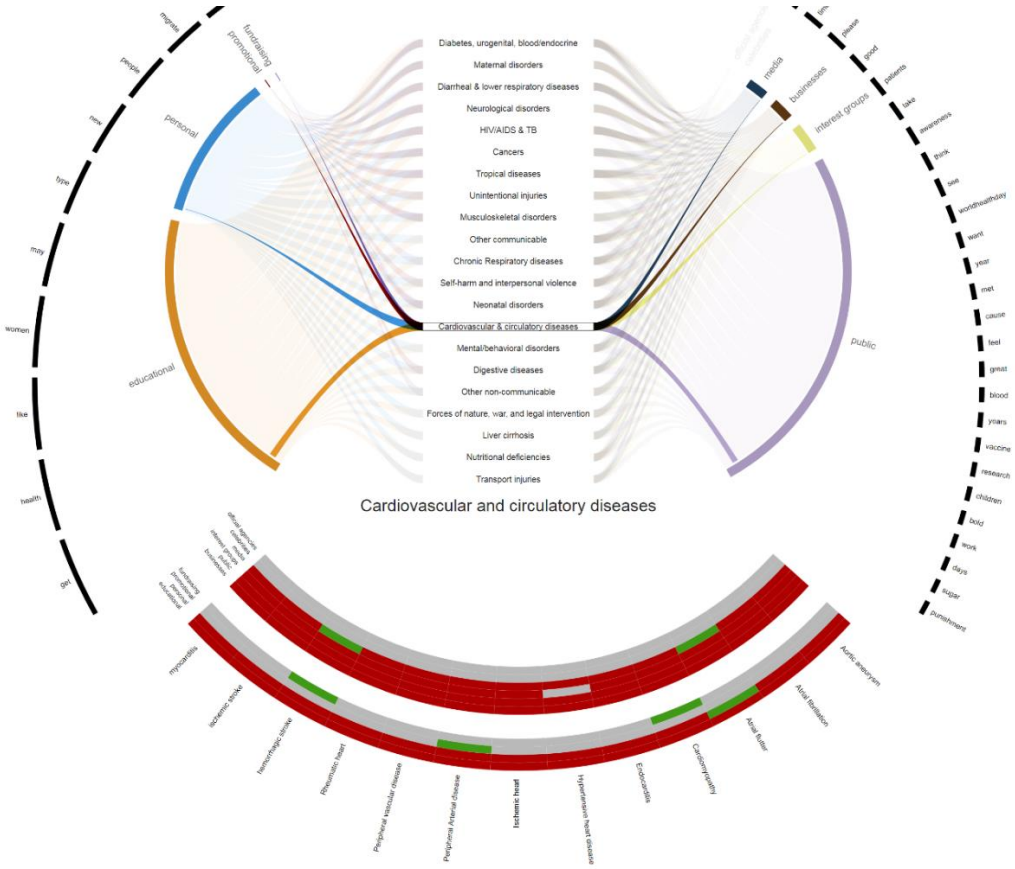

Figure 4: Screenshot of sentiment visualization with the cardiovascular & circulatory diseases cluster selected

3.2 Exploration of Tweet Corpus with Visualization

Now that the visualization has been described, let us take a close look at how it aids in the understanding of the discussions on HIV/AIDS&TB, mental and behavioral disorders, and neglected tropical diseases. Figure 5a depicts the breakdown of tweets for the HIV/AIDS&TB cluster by user category and content theme. This cluster is one of the clusters in which tweets on all four content themes are present in the corpus. In addition, all user categories are tweeting on at least one cause in this cluster. Figure 5b depicts the sentiment across the various categories. With this sub-visualization, one can notice that the tweet corpus does not include any tweets from celebrities on tuberculosis, but the discussion on HIV/AIDS includes all user groups. Another observation is that for promotional and fundraising tweets, the sentiment is positive for both HIV/AIDS and tuberculosis. It may seem intuitive that promotional and fundraising tweets are more positive than other themes, but the same pattern is not observed for other cause-clusters.

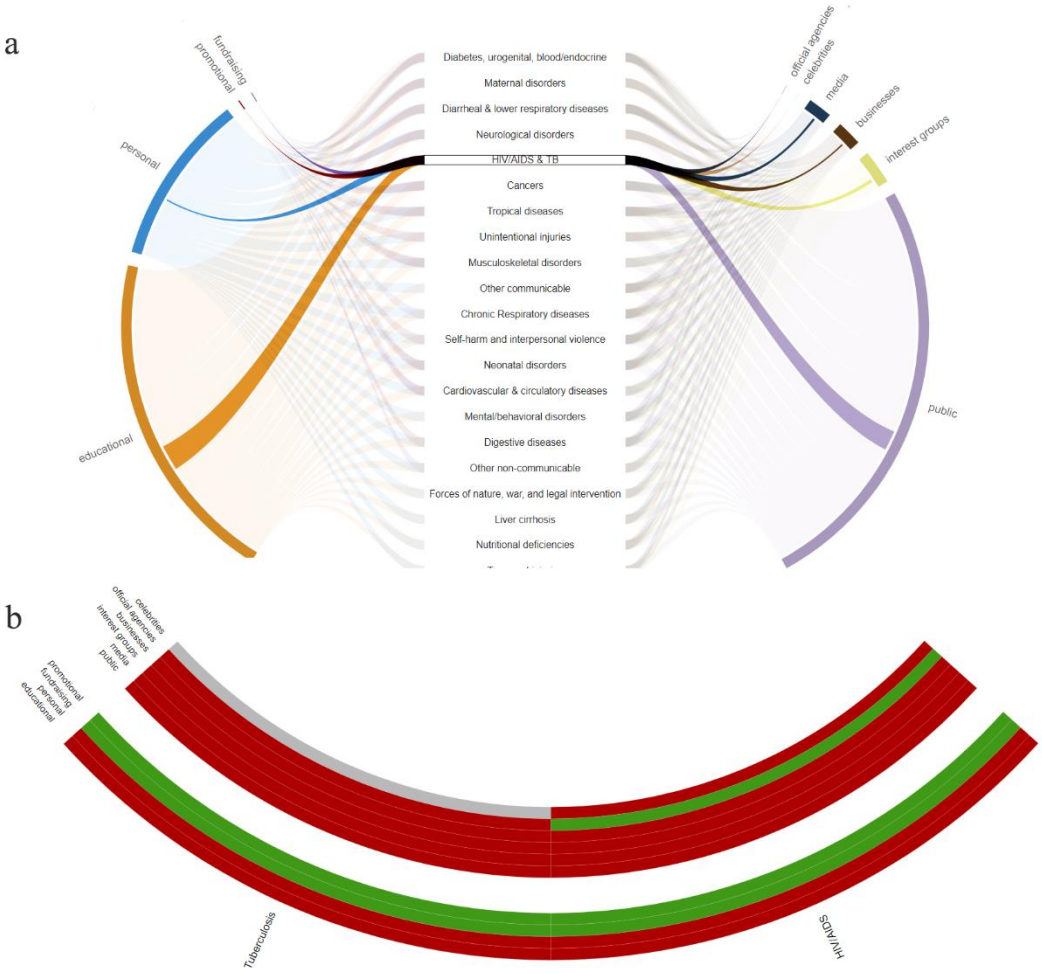

Figure 5: (a-b) Screenshots of sentiment visualization with the HIV/AIDS & TB cluster selected

Figure 6a shows the lower portion of the visualization when the mental and behavioral cause-cluster is selected. For the mental and behavioral cause-cluster, the tweets in the corpus do not include fundraising and promotional-themed tweets. Furthermore, official agencies are not

tweeting on alcohol use disorders, but they are tweeting on drug use disorders. Another observation worth highlighting is the positive net sentiment of alcohol use tweets and the negative sentiment of drug use tweets by personal accounts. The discussion on tropical diseases such as malaria, dengue, ebola, and chikungunya is highly varied. Figure 6b depicts the net sentiment rate for tropical diseases. The sentiment for the discussion of Ebola is mostly positive. This may seem erroneous, given that the 2014-15 outbreak resulted in thousands of deaths. Our data collection coincided with the release of a statement by the World Health Organization in which they discussed the successful containment of the disease [46]. This observation emphasizes the importance of context when using visualizations for data exploration.

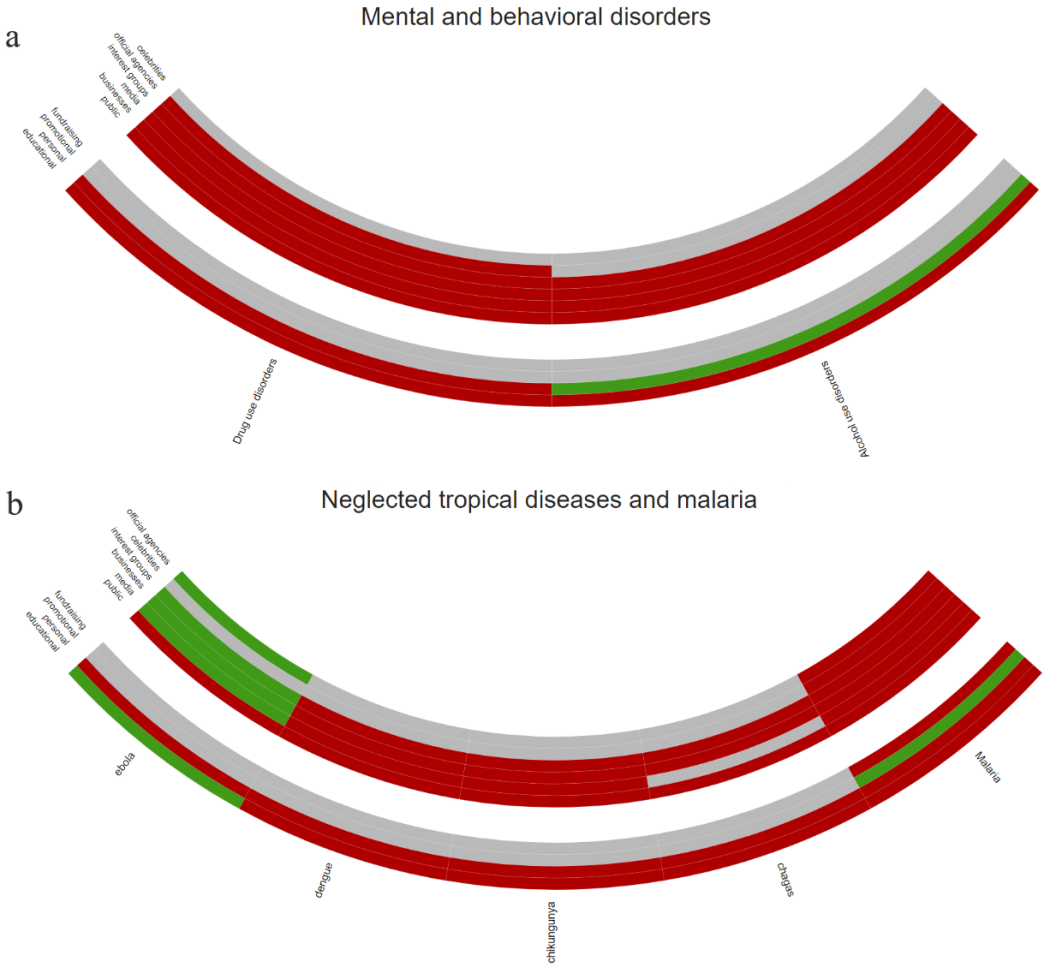

Figure 6: (a-b) Screenshots of sentiment visualization with the mental & behavioral cluster and the neglected tropical diseases & malaria cluster selected

This work provides a cross-sectional analysis of the discussion on Twitter for a broad range of health issues for a limited time frame. Subsequent steps would be to provide real-time analysis that includes historical data so that users can better understand the discussion of health issues and how it changes over time.

4 Limitations

This paper has presented a visual analytic study that contributes to the growing body of literature on understanding how health issues are portrayed on social media platforms. One of the contributions of the study is a demonstration of how supervised machine learning methods can be combined with interactive visualizations to help with an understanding of health issues on Twitter. In this study, we analyzed over half a million tweets that were retrieved from Twitter. Although we tried to apply as much rigor as possible, certain limitations exist. First, our data collection occurred over one month, which may have resulted in certain health issues being oversampled and others being under sampled. Future studies can examine the discourse for more extended periods or explore real-time analysis of tweets. Secondly, we only retrieved English-language tweets. As a result, our findings cannot be generalized to other languages. Despite this limitation, we did not specify a geographical location, and consequently, our analysis may be relevant in countries in which English is widely used.

Our use of search terms to retrieve tweets and machine learning models to classify data resulted in some non-health related tweets being included in the analysis. In addition, while AlchemyAPI was a sentiment analysis tool, its veracity at categorizing health tweets remains largely untested. Furthermore, our categorization of Twitter accounts did not include the verification of whether individual accounts are managed by Twitter bots or trolls. In addition, our analysis is of the discussions on Twitter, and as Twitter is not widely used across all demographics, our study cannot be generalized to the entire public discourse on health issues. Lastly, our constructed classification models are based on manual content analysis, which may be subject to bias.

5 Discussion

Despite the limitations mentioned above, valuable findings emerge from this study. On Twitter, the discussions of health topics are primarily mediated by the general public. Though 66% of the tweet corpus is educational, most of these tweets come from the general public, and not reputable health organizations. The discussions mainly revolve around topics such as treatment options and news reports on health ailments. For public health stakeholders, the fact that the public plays a significant role, and that the majority of the content is educational, presents both an opportunity and challenge for health promotion efforts. The use of Twitter to spread misinformation on Zika and Ebola outbreaks across the globe highlights one challenge [47]. While the burden of stemming misinformation may rest on social media organizations, an awareness of these issues can help inform policy on proper social media engagement. More research is needed to determine the influence (i.e., reach of tweets) for different types of users (e.g., see [7]). While efforts exist to use social media platforms for health education, our research highlights that there is still more work to be done. Official health and news agencies, which typically provide reputable information, are largely underrepresented in the discussion. These findings corroborate research that suggests that public health organizations may not yet effectively use Twitter to educate or engage in dialogue with the general public [4].

While our paper has primarily focused on the development of analytic models and how interactive visualizations can be used to synthesize the results, there is a need to discuss the implications beyond the methods and results. One area of future research is user categories. In this work, we broadly categorized users into six groups. While this categorization is beneficial for a high-level

understanding of participants in the discussion, there is a need to explore the interplay and communication between subsets that may exist within a category. Network analysis can support the identification of such communities and help professionals better understand the interaction between them. Recent work highlights that on Twitter, the communication between pro-vaccine and anti-vaccine communities was minimal when compared to the communication within each of the communities [48]. For public health stakeholders, understanding the structure of the communities that exist within a category is important, especially if the goal is to educate the populace on polarizing issues. Furthermore, more work is needed to understand how bots and trolls are used to spread health misinformation on Twitter. For example, one study on the discussion of vaccines on Twitter, observed that Russian trolls posted messages on both sides of the discussion that were divisive and political [49].

This work suggests that visual analytic tools may be beneficial in supporting public health stakeholders charged with developing targeted and effective health campaigns that debunk misinformation. However, before such tools can be adopted, user studies must be conducted to understand how the tools will fit into public health practice. Such studies are critical for the successful deployment of visual analytic tools because they will help developers understand the workflow processes of public health stakeholders, as well as their expectations. A survey of visualization and analytics tools highlights barriers that exist to the successful adoption of such tools in public health practice [50]. One such barrier is the risk of misinterpreting the encoded information [50], [51]. In addition, when analytic models are employed, there is an additional risk of not understanding how the data was analyzed. To address these challenges, there is a need for exploring approaches in which computer scientists and public health stakeholders work together to design visual analytic solutions [29].

Our work highlights how interactive visualizations that allow for the rapid exploration of data can support hypothesis generation. For instance, with our sentiment visualization, one is able to observe that the overwhelming categorization of health issues on Twitter is negative. Some may postulate that the very nature of health issues and the challenges they present may influence the sentiment of tweets and seek to explore whether there is a difference between the sentiment of health and non-health issues. Other researchers may choose to explore why the overall sentiment for one health issue is more positive than another. Research in this area may avoid using pre-packaged sentiment analysis tools, in order to better understand how sentiment is calculated. Unlike surveys, which have been crafted for a specific objective, the analysis of the public's discourse on Twitter is not mediated. As a result, the findings of analyzing the discourse tend to serve as a useful starting point. For instance, let us consider a situation in which every adult in a local government uses Twitter and that the data has been analyzed and visualized. Knowing that the sentiment is overwhelmingly positive for a specific health issue does not answer the question, "Why is the sentiment positive?" Therefore, for practitioners, there is a need for policies that offer guidance on how the results of studies, such as ours, can be used to inform health practice [52]. As social media becomes more embedded in society, and the data it generates increasingly valued, we need to not only develop tools to facilitate the quick analysis and exploration of data, but also create guidelines on how to effectively use both the tools and social media for health promotion.

6 Conclusions

This work demonstrates how combining machine learning methods with interactive visualizations can help with an understanding of health discussions on Twitter. Findings from this work highlight the need for studies to understand the reach of content by the various user categories and how visual analytics tools can be incorporated into public health practice. Furthermore, it provides a foundation on which further research that involves the real-time analysis of Twitter data can be built upon. It also provides a way to understand which topics are being discussed and by whom, which has implications for health literacy. This research provides a reference point for public health officials engaged in using social media to promote health policies. While our focus has been on the discussion of health matters on Twitter, the approach presented in this paper can be adapted to make sense of the discussion of other issues on social media platforms.

References

1. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, et al. 2015. Assessing Electronic Cigarette-Related Tweets for Sentiment and Content Using Supervised Machine Learning. *J Med Internet Res*. 17(8), e208. PubMed https://doi.org/10.2196/jmir.4392

2. Park H, Rodgers S, Stemmle J. 2013. Analyzing Health Organizations' Use of Twitter for Promoting Health Literacy. *J Health Commun*. 18(4), 410-25. PubMed https://doi.org/10.1080/10810730.2012.727956

3. Aslam S. Twitter by the Numbers: Stats, Demographics & Fun Facts [Internet]. 2019 [cited 2019 Oct 21]. Available from: https://www.omnicoreagency.com/twitter-statistics/

4. Gurman TA, Clark T. 2016. #ec: Findings and implications from a quantitative content analysis of tweets about emergency contraception. *Digit Health*. 2, 2055207615625035. PubMed https://doi.org/10.1177/2055207615625035

5. Hughes E. 2016. Can Twitter improve your health? An analysis of alcohol consumption guidelines on Twitter. *Health Info Libr J*. 33(1), 77-81. PubMed https://doi.org/10.1111/hir.12133

6. Logghe HJ, Selby LV, Boeck MA, Stamp NL, Chuen J, et al. 2018. The academic tweet: Twitter as a tool to advance academic surgery [Internet]. *J Surg Res*. 226, viii-xii. http://www.ncbi.nlm.nih.gov/pubmed/29622401. PubMed https://doi.org/10.1016/j.jss.2018.03.049

7. Parwani P, Choi AD, Lopez-Mattei J, Raza S, Chen T, et al. 2019. Understanding Social Media: Opportunities for Cardiovascular Medicine [Internet]. *J Am Coll Cardiol*. 73(9), 1089-93. https://www.sciencedirect.com/science/article/pii/S0735109719301081?via%3Dihub. PubMed https://doi.org/10.1016/j.jacc.2018.12.044

8. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, et al. 2015. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic

Literature Review. *PLoS One*. 10(10), e0139701. PubMed https://doi.org/10.1371/journal.pone.0139701

9. Finfgeld-Connett D. 2015. Twitter and Health Science Research. *West J Nurs Res*. 37(10), 1269-83. PubMed https://doi.org/10.1177/0193945914565056

10. Salathé M, Khandelwal S. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. Meyers LA, editor. PLoS Comput Biol. 2011 Oct 13;7(10):e1002199.

11. Weeg C, Schwartz HA, Hill S, Merchant RM, Arango C, et al. 2015. Using Twitter to Measure Public Discussion of Diseases: A Case Study. *JMIR Public Health Surveill*. 1(1), e6. PubMed https://doi.org/10.2196/publichealth.3953

12. Osborne H. Using Twitter and Other Social Media to Communicate About Health Literacy (HLOL #80) [Internet]. 2012 [cited 2019 Oct 21]. Available from: http://healthliteracy.com/2012/07/10/using-twitter-and-other-social-media-to-communicate-about-health-literacy-hlol-80/

13. Schein R, Wilson K, Keelan J. Literature review on effectiveness of the use of social media: A report for Peel Public Health [Internet]. 2010. Available from: https://www.peelregion.ca/health/resources/pdf/socialmedia.pdf

14. Chou WY, Hunt YM, Beckjord EB, Moser RP, Hesse BW. 2009. Social media use in the United States: implications for health communication. *J Med Internet Res*. 11(4), e48. PubMed https://doi.org/10.2196/jmir.1249

15. Kamel Boulos MN. Social media and mobile health. In: Kickbusch I, Pelikan JM, Apfel F, Tsouros AD, editors. Health literacy: the solid facts. Copenhagen: WHO Regional Office for Europe; 2013. p. 63–7.

16. Sørensen K. 2017. The Need for "Health Twitteracy" in a Postfactual World. *Health Lit Res Pract.* 1(2), e86-89. PubMed https://doi.org/10.3928/24748307-20170502-01

17. Ghosh DD. 2013. (Debs), Guha R. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartogr Geogr Inf Sci*. 40(2), 90-102. PubMed https://doi.org/10.1080/15230406.2013.776210

18. Korda H, Itani Z. 2013. Harnessing Social Media for Health Promotion and Behavior Change. *Health Promot Pract*. 14(1), 15-23. PubMed https://doi.org/10.1177/1524839911405850

19. Chew C, Eysenbach G. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. Sampson M, editor. PLoS One. 2010 Nov 29;5(11):e14118.

20. Heaivilin N, Gerbert B, Page JE, Gibbs JL. 2011. Public health surveillance of dental pain via Twitter. *J Dent Res*. 90(9), 1047-51. PubMed https://doi.org/10.1177/0022034511415273

21. Sullivan SJ, Schneiders AG, Cheang C-W, Kitto E, Lee H, et al. 2012. "What's happening?" A content analysis of concussion-related traffic on Twitter. *Br J Sports Med*. 46(4), 258-63. PubMed https://doi.org/10.1136/bjsm.2010.080341

22. Thackeray R, Burton SH, Giraud-Carrier C, Rollins S, Draper CR. 2013. Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. *BMC Cancer*. 13(1), 508. PubMed https://doi.org/10.1186/1471-2407-13-508

23. Krauss MJ, Grucza RA, Bierut LJ, Cavazos-Rehg PA. "Get drunk. Smoke weed. Have fun." Am J Heal Promot. 2016;31(3).

24. Liu B. Sentiment Analysis and Opinion Mining [Internet]. Morgan & Claypool Publishers; 2012. 180 p. Available from: https://dl.acm.org/citation.cfm?id=3019323

25. Myslín M, Zhu S-H, Chapman W, Conway M. 2013. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *J Med Internet Res*. 15(8), e174. PubMed https://doi.org/10.2196/jmir.2534

26. May R, Hanrahan P, Keim DA, Shneiderman B, Card SK. The state of visual analytics: Views on what visual analytics is and where it is going. In: 2010 IEEE Symposium on Visual Analytics Science and Technology. IEEE; 2010. p. 257–9.

27. Ola O, Sedig K. 2014. The challenge of big data in public health: an opportunity for visual analytics. *Online J Public Health Inform*. 5(3), 1-21. PubMed

28. Parsons P, Sedig K, Mercer RE, Khordad M, Knoll J, et al. Visual analytics for supporting evidence-based interpretation of molecular cytogenomic findings. In: Proceedings of the 2015 Workshop on Visual Analytics in Healthcare - VAHC '15 [Internet]. New York, New York, USA: ACM Press; 2015. p. 1–8. Available from: http://dl.acm.org/citation.cfm?doid=2836034.2836036

29. Preim B, Lawonn K. A Survey of Visual Analytics for Public Health. Comput Graph Forum [Internet]. 2019 Nov 28 [cited 2020 Jan 28];cgf.13891. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13891

30. Twitter. Twitter Developer Documentation [Internet]. 2019 [cited 2019 Oct 21]. Available from: https://dev.twitter.com/rest/public

31. Attai DJ, Cowher MS, Al-Hamadani M, Schoger JM, Staley AC, et al. 2015. Twitter Social Media is an Effective Tool for Breast Cancer Patient Education and Support: Patient-Reported Outcomes by Survey [Internet]. *J Med Internet Res*. 17(7), e188. http://www.ncbi.nlm.nih.gov/pubmed/26228234. PubMed https://doi.org/10.2196/jmir.4721

32. Palomino M, Taylor T, Göker A, Isaacs J, Warber S. 2016. The Online Dissemination of Nature-Health Concepts: Lessons from Sentiment Analysis of Social Media Relating to "Nature-Deficit Disorder". *Int J Environ Res Public Health*. 13(1), 142. PubMed https://doi.org/10.3390/ijerph13010142

33. Paul MJ, Dredze M. Discovering Health Topics in Social Media Using Topic Models. Lambiotte R, editor. PLoS One. 2014 Aug 1;9(8):e103408.

34. Symplur. Healthcare Hashtag Project [Internet]. 2019 [cited 2019 Oct 21]. Available from: https://www.symplur.com/healthcare-hashtags/

35. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, et al. 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 380(9859), 2095-128. PubMed https://doi.org/10.1016/S0140-6736(12)61728-0

36. Tweepy. Tweepy [Internet]. 2019. Available from: http://www.tweepy.org/

37. IBM. AlchemyLanguage Overview [Internet]. [cited 2020 Apr 13]. Available from: https://mediacenter.ibm.com/media/1_v8ulaavw?mhsrc=ibmsearch_a&mhq=AlchemyLanguage

38. Meehan K, Lunney T, Curran K, McCaughey A. Context-aware intelligent recommendation system for tourism. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). IEEE; 2013. p. 328–31.

39. Saif H, He Y, Alani H. Semantic Sentiment Analysis of Twitter. In: International Semantic Web Conference. Springer, Berlin, Heidelberg; 2012. p. 508–24.

40. Serrano-Guerrero J, Olivas JA, Romero FP, Herrera-Viedma E. 2015. Sentiment analysis: A review and comparative analysis of web services. *Inf Sci (Ny)*. 311, 18-38. https://doi.org/10.1016/j.ins.2015.03.040

41. Scikit. Scikit [Internet]. 2019. Available from: https://scikit-learn.org/stable/

42. Silge J, Robinson D. Term Frequency and Inverse Document Frequency (tf-idf) Using Tidy Data Principles [Internet]. 2017 [cited 2019 Oct 21]. Available from: https://cran.r-project.org/web/packages/tidytext/vignettes/tf_idf.html

43. Anger I, Kittl C. Measuring influence on Twitter. In: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '11. New York, New York, USA: ACM Press; 2011. p. 1–4.

44. D3 [Internet]. 2019. Available from: https://d3js.org/

45. Bremer N. The words in the LotR [Internet]. 2016. Available from: https://www.visualcinnamon.com/portfolio/words-lord-of-the-rings

46. WHO. Statement on the 9th meeting of the IHR Emergency Committee regarding the Ebola outbreak in West Africa [Internet]. 2016 [cited 2019 Oct 20]. Available from: https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html

47. Wang Y, McKee M, Torbica A, Stuckler D. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. Soc Sci Med [Internet]. 2019 Nov 1 [cited 2020 Feb 5];240:112552. Available from: https://www.sciencedirect.com/science/article/pii/S0277953619305465

48. Gunaratne K, Coomes EA, Haghbayan H. 2019. Temporal trends in anti-vaccine discourse on Twitter [Internet]. *Vaccine*. 37(35), 4867-71. https://www.sciencedirect.com/science/article/pii/S0264410X1930876X. PubMed https://doi.org/10.1016/j.vaccine.2019.06.086

49. Broniatowski DA, Jamison AM, Qi SH, AlKulaib L, Chen T, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health [Internet]. 2018 [cited 2020 Jan 28];108(10):1378–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30138075

50. Carroll LN, Au AP, Detwiler LT, Fu T-C, Painter IS, et al. 2014. Visualization and analytics tools for infectious disease epidemiology: A systematic review. *J Biomed Inform*. 51, 287-98. PubMed https://doi.org/10.1016/j.jbi.2014.04.006

51. Zakkar M, Sedig K. Interactive visualization of public health indicators to support policymaking: An exploratory study. Online J Public Health Inform [Internet]. 2017 [cited 2020 Jan 28];9(2):e190. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29026455

52. Colditz JB, Chu K-H, Emery SL, Larkin CR, James AE, et al. 2018. Toward Real-Time Infoveillance of Twitter Health Messages [Internet]. *Am J Public Health*. 108(8), 1009-14.https://ajph.aphapublications.org/doi/10.2105/AJPH.2018.304497 PubMed https://doi.org/10.2105/AJPH.2018.304497

Appendix A: List of search terms

abortion	hemorrhagic stroke	pancreatic cancer
alcohol use disorders	hepatitis	pancreatitis
alzheimer	hiv/aids	paralytic ileus
aortic aneurysm	hurricane death	parkinsons disease
asthma	Hypertensive heart disease	peptic ulcer
atrial fibrillation	influenza	peripheral arterial disease
atrial flutter	interpersonal violence	peripheral vascular disease
bile duct disease	intestinal ischemic syndrome	pharyngeal cancer
biliary tract cancer	intestinal obstruction	pneumoconiosis
bladder cancer	iron-deficiency anemia	pneumonia

brain cancer	ischemic heart	poisonings
breast cancer	ischemic stroke	pregnancy hypertensive
bronchitis	kidney cancer	preterm birth complications
cardiomyopathy	kidney disease	prostate cancer
cervical cancer	laryngeal cancer	protein-energy malnutrition
chagas	leukemia	pulmonary sarcoidosis
chikungunya	liver cancer	rheumatic heart
chronic obstructive pulmonary disease	liver cirrhosis	rheumatoid arthritis
colon cancer	low back pain	road injury
congenital anomalies	lung cancer	self-harm
dengue	malaria	sepsis
diabetes	male infertility	skin disease
diarrhea diseases	maternal hemorrhage	skin melanoma
diffuse parenchymal lung disease	measles	stds
drowning	medical treatment adverse effect	stomach cancer
drug overdose	meningitis	subcutaneous disease
earthquake death	migraine	syphilis
ebola	mouth cancer	tetanus
encephalitis	multiple myeloma	tornado death
endocarditis	multiple sclerosis	trachea cancer
epilepsy	myocarditis	transport injury
esophageal cancer	nasopharyngeal cancer	tsunami death
falls	neck pain	tuberculosis
fire death	neonatal encephalopathy	typhoid
gall bladder	nervous system cancer	typhoon death
gallbladder cancer	non-hodgkin lymphoma	urinary disease
glomerulonephritis	oropharyngeal cancer	urinary organ cancer

gout	osteoarthritis	uterine cancer
heat death	ovarian cancer	whooping cough