Study of Twitter communications on cardiovascular disease by state health departments

Aibek Musaev, Rebecca K. Britt, Jameson Hayes, Brian C. Britt, Jessica Maddox, and Pezhman Sheinidashtegol

The University of Alabama, Tuscaloosa, AL 35487

Abstract. The present study examines Twitter conversations around cardiovascular health in order to assess the topical foci of these conversations as well as the role of various state departments of health. After scraping tweets containing relevant keywords, Latent Dirichlet Allocation (LDA) was used to identify the most important topics discussed around the issue, while PageRank was used to determine the relative prominence of different users. The results indicate that a small number of state departments of health play an especially significant role in these conversations, and that irregular events like ebola outbreaks also exert a strong influence over the volume of tweets made in general by state departments of health.

Keywords: Twitter, cardiovascular disease, LDA, PageRank

1 Introduction

Cardiovascular disease (CVD) is the leading cause of death in the United States [6], with contributing factors including poor health and risk factors such as obesity and diabetes, among others [9]. While prevention is the optimal approach towards reducing CVD, the potential applicability of social media communication remains understudied. Health care professionals have increased their use of social media to engage with the public, to increase health care education, patient compliance, and organizational promotion [21]. To date, social media based health communication research has prioritized studies of theory, message effects, or disseminating interventions to end users [16]. However, state health departments' social media communication can engage with patients to improve their care [22].

Concurrently, research using social media has advanced considerably in recent years. API scrapers [1] enable the rapid collection of data from social media, while data analysis approaches such as time series analysis of user behaviors [15] and topic modeling through Latent Dirichlet Allocation (LDA) [14] allow large textual datasets to be rapidly analyzed in order to draw insights for basic research as well as in applied contexts, such as those related to public health care and associated campaigns [3].

In this study, we analyzed social media activity of state health departments related to cardiovascular disease. The objectives of this study were 1) to determine the most active state health departments on Twitter with respect to cardiovascular disease, and 2) to determine the most important topics that were discussed and the most important terms used in those discussions.

See the next section for an overview of the proposed methodology and related work. Section 3 presents the experimental results using real data and section 4 concludes the paper.

2 Overview and related work

In this study, we analyzed both the tweets posted by state departments of health and the users that posted them.

In analyzing tweets, we first propose to determine the peaks of public activity by aggregating the tweets posted by all users in the collected dataset for each month under study. To understand the key drivers of those peaks, we then perform a detailed analysis by identifying the most popular topics discussed during those peaks. See Section 2.1 for a description of the topic modeling approach used in this study. Finally, we determine and visualize the most important terms used by the public based on a word cloud approach [10].

For the user analysis, we first examine the total number of messages posted by state health departments since opening Twitter accounts. Then we analyze their communications with respect to cardiovascular disease using an extension of PageRank algorithm. See Section 2.2 for a description of the algorithm used to identify the most important users in the collected dataset.

2.1 Towards understanding the most important topics discussed by the public

Topic modeling in machine learning and natural language processing is a popular approach to uncover hidden topics in a collection of documents. Intuitively, given that a document, such as a tweet, is about a particular topic, one would expect certain words to appear more or less frequently than others. For example, words such as 'cardiovascular', 'heart', and 'stroke' will appear more frequently in tweets on cardiovascular disease, 'congress', 'vote', and 'policy' in documents about politics, and 'the', 'a', and 'is' may appear equally in both. Furthermore, a document typically discusses multiple topics in different proportions, e.g., 60% about politics and 40% about cardiovascular disease in a news article about passing a bill on CVD.

Popular topic modeling algorithms include Latent Semantic Analysis (LSA) [23, 5], Hierarchical Dirichlet process (HDP) [4, 11], and Latent Dirichlet Allocation (LDA) [13, 14]. In this project, we use LDA to uncover the most important topics discussed by the public posted by or mentioning state departments of health. The study of alternative topic modeling algorithms will be explored in our future work.

In LDA, each document can be described by a distribution of topics and each topic can be described by a distribution of words [8]. Here topics are introduced as a hidden (i.e., latent) layer connecting documents to words. Note, that the number of topics is a fixed number that can be chosen either as an informed estimate based on a previous analysis or via a simple trial-and-error approach. See Section 3 for an application of LDA approach to determine the topics discussed during the peaks of public activity.

2.2 Ranking state departments of health by social media influence

With the rise of social media platforms, such as Twitter, identification of the most influential users garnered a huge amount of interest [18]. Different Twitter influence measures have been proposed; some are based on simple metrics provided by the Twitter API [7], while others are based on complex mathematical models [12].

In this study, we wanted to rank state departments of health by social media influence. Specifically, we wanted to identify the Twitter accounts whose posts on cardiovascular disease attracted the most amount of attention.

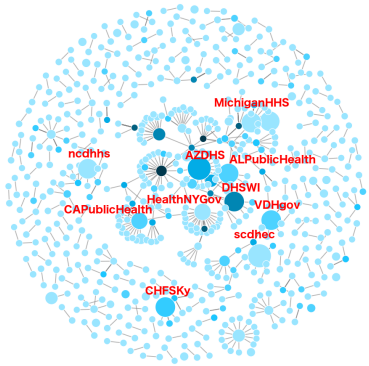

Fig. 1: User ranking based on reply relationships

Given the dataset of Twitter users and their tweets, we built a graph where the users serve as nodes. The links between nodes are represented by reply relationships, such that the direction of a link is a directed edge from the author of the reply to the author of the original tweet that was replied to.

Given the directed graph, we can now apply a PageRank algorithm, which is a way of measuring the importance of nodes in a directed graph such as website pages [19]. PageRank works by counting the number and quality of the links pointing to a page to determine a rough estimate of their importance.

See Figure 1 and the corresponding discussion in Section 3 for a visualization of the proposed approach.

3 Evaluation using real data

In this section, we present a pilot study of Twitter communications involving state health departments. We begin by describing the data collection process used in this study in Section 3.1. Then we design two sets of experiments as follows. The first set of experiments (Section 3.2) analyzes the Twitter activity of

all communication involving those departments while the second set (Section 3.3) focuses on the communication related to cardiovascular disease.

3.1 Data collection

In this pilot study, we performed a set of experiments based on real-world data collected from Twitter. For data collection, we used a two-step process. In step 1, we downloaded basic data about tweets containing the keywords that we are interested in. Specifically, we used a JavaScript module called *scrape-twitter*[1]. It allows for querying Twitter for matching tweets based on keywords. In step 2, we used Twitter's statuses/lookup[2] feature that returns a complete set of data for up to 100 tweets at a time.

For keywords, we used the Twitter handles of state departments of health for each state, such as *@ALPublicHealth* for Alabama or *@HealthNYGov* for New York. This allowed us to collect 319k tweets ranging from November 2, 2007 to December 26, 2018 posted by 52.5k users including the 50 state departments of health. This represents a full dataset of all tweets either posted by or mentioning the state health departments.

For the dataset on cardiovascular disease, we filtered the full dataset based on the keywords related to CVD as follows. We included the keywords directly related to CVD, such as 'cardiovascular' and 'CVD.' We also added the keywords related to CVD symptoms and risk factors, including 'heart failure', 'heart disease', 'heart stroke', 'heart failure', 'blood pressure', 'atherosclerosis', 'arrhythmia', 'cardiac', and 'obesity.'

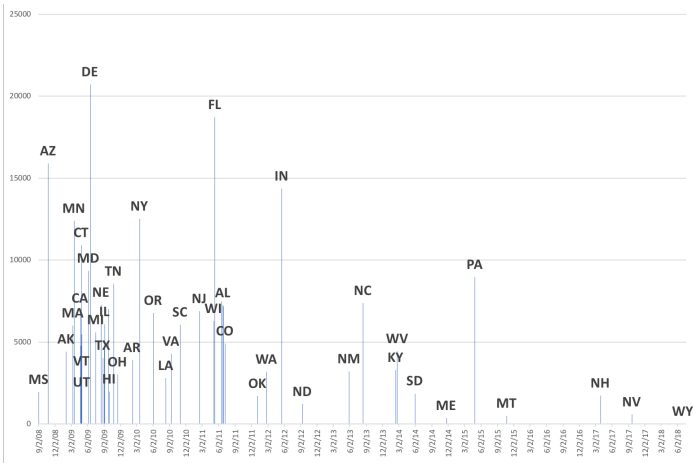

Fig. 2: Total number of tweets since opening accounts by state departments of health

[1] https://www.npmjs.com/package/scrape-twitter

[2] https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup

We released both datasets as a contribution to the research community[3]. These are the first published datasets that contain Twitter activity related to all 50 state departments of health covering an 11 year period. The datasets are provided as listings of tweet IDs in accordance with the Twitter policy[4].

3.2 Analysis of Twitter activity involving state health departments

In this experiment, we analyzed the overall Twitter activity involving state departments of health. Specifically, we aggregated the total number of tweets posted by each department since they opened their Twitter accounts and plotted the results in Figure 2.

As expected, we observed that the departments that opened their Twitter accounts earlier have a larger number of tweets compared to the departments that joined Twitter at a later date. However, there are recent accounts that managed to generate an unusually high amount of activity despite joining Twitter later, such as the Pennsylvania Department of Health account (*@PAHealthDept*) which joined Twitter on April 28th, 2015.

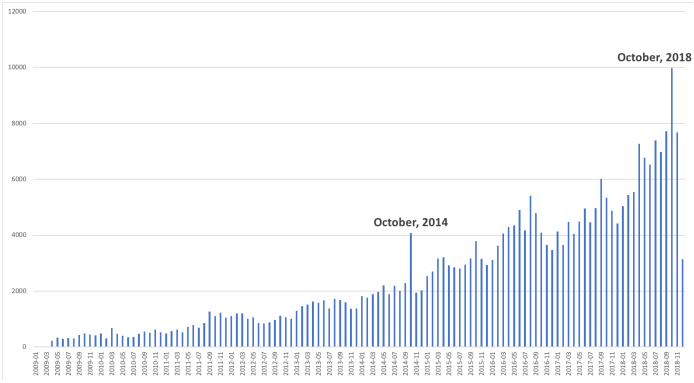

Fig. 3: Monthly Twitter activity involving state health departments, 2009-2018

Next, we analyzed the monthly Twitter activity across all state health departments and plotted the results in a time series graph in Figure 3. Although we observed an overall increase of the total number of tweets over time, there are several peaks with an unusually high amount of activity compared to other months, such as October, 2014 and October, 2018.

To understand the key drivers of the peak in October, 2014, we analyzed the topics discussed during that month. Using sklearn [17], we trained an LDA model based on an arbitrarily chosen number of topics $n\_topics = 10$. For illustration

[3] http://aibek.cs.ua.edu/files/sdh_all_ids.txt, http://aibek.cs.ua.edu/files/sdh_cvd_ids.txt
[4] https://developer.twitter.com/en/developer-terms/agreement-and-policy

purposes, we visualized the computed topics and the most important words in those topics using t-SNE, or t-distributed stochastic neighbor embedding [20].

Based on the visualization shown in Figure 4, we observed that ebola was an important topic of discussion during October 2014, which is when ebola spread outside of Africa[5]. For example, one topic contains keywords, such as 'says', 'dallas', 'test', 'health', and 'negative' as discussed in the following tweet: *JUST IN: Texas Health Presbyterian says test results for a Dallas Co Sheriff's deputy came back negative for ebola.*

Fig. 4: Visualization of topics discussed on Twitter involving state health departments in October, 2014 using LDA model

3.3 Analysis of Twitter activity on cardiovascular disease

In this experiment, we analyzed Twitter communication on cardiovascular disease across all state health departments and plotted the results in a time series graph in Figure 5. We observed that the largest peak of activity occurred in February 2018. Using the tweets posted during that month, we generated a word cloud to visualize the most important terms: see the results in Figure 6. As expected, the words 'heart' and 'disease' are the most important words. Similarly, 'blood' and 'pressure' are also significant according to this visualization.

Finally, we plotted a diagram of user rankings based on reply relationships in Figure 1. The diagram was implemented using a JavaScript visualization library

[5] https://en.wikipedia.org/w/index.php?title=Ebola_virus_disease&oldid=876053081

D^3 [2]. Each node represents a Twitter user in the dataset, directed edges are based on reply relationships, the color intensity is used to represent the number of tweets containing the search keywords each user has. Lastly, the size of a node is computed based on the PageRank algorithm as described in Section 2.2.

For visualization purposes, we added labels to the users with the highest ranking scores. Based on these results, Arizona Department of Health (*@AZDHS*) and South Carolina Department of Public Health (*@scdhec*) are among the most important social media users on cardiovascular disease.

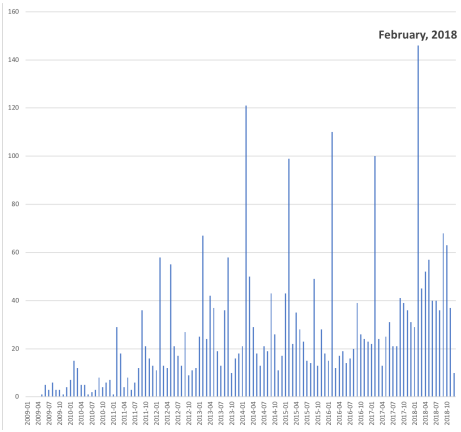

Fig. 5: Monthly Twitter activity on CVD, 2009-2018

Fig. 6: Word cloud using Twitter activity on CVD in February, 2018

4 Conclusion and future work

This study demonstrates the increasing Twitter usage of state departments of health and the role of cardiovascular disease and other health-related issues in their communication with the public. It also demonstrates the wide disparity in the Twitter presence of different state departments of health, with a minority of institutions taking an especially central role in conversations about health.

More broadly, this paper demonstrates an approach to mine and analyze social media data in order to draw conclusions about the behaviors and relative importance of various human and institutional actors, as well as the topics that drive conversation and the points in time at which those conversations shift.

Future research should address how individual users' activities affect their influence over time and how prominent actors, in turn, shape the social media conversation. Researchers should also study finer aspects of discourse about cardiovascular disease and other health issues, such as the social structure of conversations among underserved communities, as well as how health-related organizations can influence those conversations to promote healthy behaviors.

References

[1] Bogdan Batrinca and Philip C. Treleaven. "Social media analytics: a survey of techniques, tools and platforms". In: *AI and Society* 30.1 (2015), pp. 89–116.

[2] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. "D^3 data-driven documents". In: *TVCG* 17.12 (2011), pp. 2301–2309.

[3] Brian C. Britt et al. "Finding the invisible leader: When a priori opinion leader identification is impossible". In: *NCA*. 2017.

[4] Sophie Burkhardt and Stefan Kramer. "Multi-label classification using stacked hierarchical Dirichlet processes with reduced sampling complexity". In: *Knowl. Inf. Syst.* 59.1 (2019), pp. 93–115.

[5] Zhiqiang Cai et al. "Impact of Corpus Size and Dimensionality of LSA Spaces from Wikipedia Articles on AutoTutor Answer Evaluation". In: *Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018, Buffalo, NY, USA, July 15-18, 2018*. 2018.

[6] Centers for Disease Control and Prevention. *Heart disease in the United States*. `https://www.cdc.gov/heartdisease/facts.htm/`. Accessed on 1/14/2019.

[7] Meeyoung Cha et al. "Measuring user influence in Twitter: The million follower fallacy." In: *ICWSM* 10.10-17 (2010), p. 30.

[8] Stefan Debortoli et al. "Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial". In: *CAIS* 39 (2016), p. 7.

[9] Luc F Van Gaal, Isle L Mertens, and Christophe E De Block. "Mechanisms linking obesity with cardiovascular disease". In: *Nature* 444 (2006), pp. 875–880.

[10] Florian Heimerl et al. "Word cloud explorer: Text analytics based on word clouds". In: *HICSS*. 2014, pp. 1833–1842.

[11] Vagia Kaltsa et al. "Multiple Hierarchical Dirichlet Processes for anomaly detection in traffic". In: *Computer Vision and Image Understanding* 169 (2018), pp. 28–39.

[12] Georgios Katsimpras, Dimitrios Vogiatzis, and Georgios Paliouras. "Determining influential users with supervised random walks". In: *WWW*. ACM. 2015, pp. 787–792.

[13] DongHwa Kim et al. "Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec". In: *Inf. Sci.* 477 (2019), pp. 15–29.

[14] Chao Li et al. "Mining Dynamics of Research Topics Based on the Combined LDA and WordNet". In: *IEEE Access* 7 (2019), pp. 6386–6399.

[15] Sorin Adam Matei and Brian C. Britt. *Structural differentiation in social media: Adhocracy, entropy and the "1% effect"*. Springer, 2017.

[16] S Anne Moorehead et al. "A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication". In: *JMIR* 15.4 (2013), e85.

[17] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *JMLR* 12 (2011), pp. 2825–2830.

[18] Fabián Riquelme and Pablo González-Cantergiani. "Measuring user influence on Twitter: A survey". In: *IPM* 52.5 (2016), pp. 949–975.

[19] Keita Sugihara. "Using Complex Numbers in Website Ranking Calculations: A Non-ad hoc Alternative to Google's PageRank". In: *JSW* 14.2 (2019), pp. 58–64.

[20] Laurens Van Der Maaten. "Accelerating t-SNE using tree-based algorithms". In: *JMLR* 15.1 (2014), pp. 3221–3245.

[21] C L Ventola. "Social media and health care professionals: Benefits, risks, and best practices". In: *P&T* 39 (2014), pp. 491–499.

[22] R J Widmer et al. "Social media platforms and heart failure". In: *Journal of Cardiology Failure* 23.11 (2017), pp. 809–812.

[23] Chandra Shakhar Yadav and Aditi Sharan. "A New LSA and Entropy-Based Approach for Automatic Text Document Summarization". In: *Int. J. Semantic Web Inf. Syst.* 14.4 (2018), pp. 1–32.