Tracking Health Related Discussions on Reddit for Public Health Applications

Albert Park, PhD[1], Mike Conway, PhD[1]
[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

We use Reddit to demonstrate social media's potential for public health applications. First, we employ a lexicon-based approach to track the prevalence of keywords indicating public interest in Ebola, electronic cigarette, influenza, and marijuana. Second, to better understand the public reactions, we use the Latent Dirichlet Allocation algorithm, to identify either the general themes or motivations for extreme changes in the volume of discussion over time. We observe that discussions related to Ebola and influenza, infectious diseases of public health interests, surged when the first case of Ebola was diagnosed and a new strain of H1N1 influenza virus was confirmed in the United States. We also observed that discussions of a controversial health topic like marijuana increased with the announcement of a major change in United States federal policy. Discussions of electronic cigarette highlighted opportunities for better health education. Lastly, we discuss the implications of our findings for utilizing Reddit data for public health applications.

Introduction

Nearly two-thirds of American adults (65%) use social media: a nearly a tenfold increase in the past 10 years[1]. Social media provides a platform for users to freely express their thoughts and provides an opportunity to interact with geographically dispersed likeminded individuals. These social media users discuss a wide variety of topics ranging from ordinary details of their daily life to information about infectious diseases of public health interest like Ebola[2]. Due to the popularity and ubiquitous nature of social media, researchers advocate for utilizing social media for public health applications[3–5]. Public health agencies are in an early adoption stage of using social media for information distribution[6]. In addition to the substantial potential for using social media as a disease surveillance tool[3–5] and means of information distribution[6], social media also has the potential to provide other opportunities to improve public-health practice.

Studying the reactions or opinions of a population has traditionally involved nationally distributed data collection, such as surveys from government agencies. However, these methods are expensive, and perhaps more importantly, time consuming. Some researchers suggest that mining social media data can provide opportunities to reduce time and expense when understanding the reactions or opinions of a population on health issues[7–10]. For example, social media allows for accessing first person accounts of experiences[7,8], public sentiments[9], public knowledge[10], and public attitudes[10] that may help public health agencies and researchers to develop policies that improve public health outcomes. Moreover, social media can provide the contextual information and prevalence of public interests more efficiently than traditional public health methods. Tracking the prevalence of public interests and understanding the general public reactions and opinions on various health issues have the potential to expand the scope of public-health practice.

In this paper, we report on findings derived from social media data gathered from Reddit for the purpose of tracking the prevalence of public interests and understanding public reactions towards infectious diseases of public health interests like Ebola and influenza as well as controversial health issues, such as electronic cigarettes and marijuana. In fact, although Reddit is one of the most popular public social media platforms, it has been underutilized for public health applications. Reddit's size and range of topics make it difficult to make use of the data without any knowledge of how the platform is used in practice. Thus, we aim to fill this gap in the literature with the current study and answer the following two research questions (RQ):

- (RQ1) Is Reddit an effective source for tracking the prevalence of public interests on infectious diseases (i.e., Ebola and influenza) and controversial health related issues (i.e., electronic cigarette and marijuana) over time?
- (RQ2) What do Reddit members discuss regarding these health issues (a) in times of elevated discussion volume or (b) in general, if the issues have a steady level of discussions?

The work described in this paper was exempted from review by the University of Utah's Institutional Review Board (IRB) [ethics committee] (IRB 00076188).

Background

A growing body of research has demonstrated the successful use of social media for public health applications[11–13]. Often referred to as *digital disease detection*[3], *Infoveillance*[4], and *digital epidemiology*[5], many studies have used Twitter data for applications in public health, primarily due to the real-time nature of the data. For example, Twitter data have been used to monitor or estimate influenza[12,14], seasonal allergies[12], alcohol sales and consumption[15], cholera outbreaks[16], earthquake[17], and smoking behavior[18], as well as to examine sentiment towards marijuana use[19]. Although Twitter is highly popular and tweet analysis has performed well with the aforementioned topics, tweets provides relatively limited context due to a length limitation of 140 characters.

Other social media data, such as Facebook and online health community data, have also been mined to, for example, characterize and predict postpartum depression[20], classify opioid addiction phrases[21] and predict adverse drug reactions[22]. Google search queries allowed researchers to provide timely estimation of influenza rates[23]. However, a previous study suggested that Facebook users are reluctant to discuss certain negative topics on Facebook, due to users' desire to convey positive images of themselves[24]. Online health communities can provide rich details of first person accounts of experiences[25], however, online health communities typically are single topic focused groups, often with a small number of members and attracting a substantial number of "lurkers"[26] (i.e., individuals who participate without posting) and dropouts[27]. Google search queries can be useful and timely, however, search queries are relatively limited in providing context and have been shown to overestimate disease rates, due (in part) to heightened media coverage[28].

Recently, Reddit, due to the availability of a public Application Programming Interface (API)[29], the capability of providing contextual information, and the support for throwaway accounts, has become a widely studied social media platform for controversial discussions. For example, using Reddit data, researchers have found empirical evidence that Reddit members openly discuss and exchange information support for potentially stigmatized issues like mental health illnesses[30], detected increases in suicidal content following reports of several celebrity suicides[31], identified distinct markers of shifts to suicidal ideation from mental illnesses[32], explored the relationship between social feedback and community participation[33], identified distinctive linguistic characteristics that are associated with mental illnesses[34], characterized smoking and drinking problems[35], and examined user experiences with different tobacco products[36]. Thus, in this study, we explore Reddit's utility as a data source for public health applications for tracking and understanding public opinions and reactions to health issues.

Data: Social Media Site

The data for this study is hosted in the popular social media platform, Reddit (http://www.reddit.com). We use Reddit to track and understand discussions of Ebola, influenza, electronic cigarettes, and marijuana for the following three reasons. First, Reddit is a highly active social media platform that had 83 billion page views from over 88,000 active sub-communities (subreddits) in 2015. Members of Reddit made over 73 million individual posts with over 725 million associated comments in the same year[37]. Second, Reddit allows for throwaway and unidentifiable accounts that are suitable for controversial discussions, such as thoughts and feelings on electronic cigarette and marijuana as well as epidemic concerns like Ebola and influenza that may be inappropriate or sensitive for identifiable accounts. Third, Reddit content is publicly available, in contrast to other health focused social media platforms like Facebook Groups or specifically health-focused online communities like PatientsLikeMe, where the content is typically not available on the open web.

Reddit members converse via a forum like platform. Reddit discussion consists of posts (i.e., a submission that starts a conversation) and associated comments (i.e., a submission that replies to posts or other comments) in various topically focused subreddits. Members who have achieved a certain status within the community are able to create new subreddits. For this study, we used a dataset[38] released by a Reddit member. The dataset has been used in previous studies[34,39,40]. The dataset for the current study is comprised of 239,772 (including both active and inactive) subreddits, 13,213,173 unique member IDs, 114,320,798 posts, and 1,659,361,605 associated comments that were made from October 2007 to May 2015.

Methods

RQ1. Is Reddit an effective source for tracking the prevalence of public interests on infectious diseases and controversial health related issues over time?

We used a lexicon-based approach to track discussions on Ebola, electronic cigarettes, influenza, and marijuana from all subreddits available in Reddit. First, we identified key terms associated with the topics of our interests. A

summary of key terms for each issue is shown in Table 1. Second, we preprocessed the entire dataset, which included converting text to lower case and removing punctuation. Third, to extract submissions (i.e., posts and comments) containing key terms from all available 239,772 subreddits, we employed a lexicon-based approach and extracted timestamps, comment or post IDs, member IDs, and subreddit IDs of the submissions. We extracted and included any partial matches in this process to cover a wide variation of terms. For example, a partial match of 'cig' can cover a variation of 'cig', 'cigs', 'cigarette', and 'cigarettes' for electronic cigarette. Fourth, we counted unique member IDs, subreddits, posts, and comments containing key terms. Fifth, we normalized the frequencies over time by dividing the frequency counts by the total number of the respective variables from all available subreddits for that period. Since the total number of submissions in Reddit generally increases over time, we report normalized frequencies over time counts.

Table 1. Key terms used in the lexicon-based approach

Issues	Key terms
Ebola	ebola
Electronic cigarette	e cig, elec cig, electronic cig
Influenza	flu, influenza, H1N1
Marijuana	weed, marijuana, ganja, cannabis, bong, spliff, Mary Jane

RQ 2. What do Reddit members discuss on these health issues (a) in times of elevated activities or (b) in general, if the issues have a steady level of discussions?

Based on results of RQ 1, we created two scenarios deciding which time periods to further investigate for understanding the discussions on Ebola, electronic cigarette, influenza, and marijuana. (a) If the issue has a sudden elevated level of discussion, we investigated the time period in which the elevation occurs along with prior discussions of the same temporal length to understand the underlying causes for these sudden changes in public interest. Similar methods that contrast to prior time periods have been used to detect emerging topics[41,42]. (b) If the issue has a steady level of discussions, we investigated the entire discussions on the issue to understand the main themes.

We used natural language processing (NLP) and language modeling for this research question. Due to the size of the dataset and range of topics discussed on Reddit, we used automated methods. Similar automated methods have been used in the health care domain to extract information and analyze data, and to enhance the personal health care experiences[43–45]. First, we preprocessed the entire dataset as we did in RQ1. Second, to improve the language modeling results, we removed the URLs and comments and posts with less than 5 words, and then extracted nouns using Python Natural Language Toolkit (NLTK) package[46]. The extracted nouns were used to create language models—a set of topics generated from document-level word co-occurrences for a given set of documents—using Latent Dirichlet Allocation[47] (LDA) for the time period of our interests. We elected to use LDA, an unsupervised algorithm, due to the lack of a ground truth dataset. We considered each post and its associated comments as a single document.

One advantage of using LDA as opposed to other unsupervised clustering techniques is that the algorithm considers each document with multiple topics. A previous study of online health discussions suggested that discussions could have multiple topics due to topic drift[48]. Thus, we employed LDA for this study. One disadvantage of using LDA is, however, it requires a pre-determined number of topics. After experimenting with varying numbers of topics, we generated 50 topics to understand Ebola, electronic cigarette, influenza, and marijuana related issues. We used the Python package genism[49] to conduct LDA analysis. We then present the main topics and their top 50 associated words as the word cloud overview using the Python package wordcloud[50]. Despite its simplicity, word cloud overview remains one of the more preferred and user-friendly visualizations that can also scale to different data sizes[51]. We then manually investigated the identified topics and their associated words to thoroughly examine the LDA results.

Lastly, we performed two types of validity checks. First, for health issues with a sudden elevated level of discussion, we verified the LDA results via a systematic analysis of news at the time of the change. LDA results reflect motivations for the extreme changes, thus news can be an effective source for a validity check. Second, we extracted URLs using regular expressions and categorized the results. A previous study concerning electronic cigarettes—a product with few marketing restrictions in the US until recently—suggested that up to 90 percent of social media (in this case, Twitter) content could be related to product marketing[52]. Thus, because marketing content can skew our result, we used URLs as a proxy to marketing content and reported the percentage of posts with URLs. We also manually examined several extracted URLs to ensure the quality of the validation process.

Results

RQ1. Is Reddit an effective source for tracking the prevalence of public interests on infectious diseases and controversial health related issues over time?

The lexicon-based approach identified Reddit posts, comments, and members discussing Ebola, electronic cigarette, influenza, and marijuana from October 2007 to May 2015 (Table 2). The most discussed matter was influenza, followed by marijuana, electronic cigarettes, and then Ebola. The raw counts of discussions and members who mentioned each topic generally increased with time.

Table 2. The total number and average normalized count of posts, comments, members, and subreddits identified using the lexicon-based approach

Issues	Total posts and comments (n)	Average normalized count of posts and comments (%)	Total members (n)	Average normalized count of members (%)	Number of subreddits containing the key terms
Ebola	252,243	7.18E-05	113,546	9.68E-04	6,039
Electronic cigarette	355,839	2.17E-04	176,252	3.75E-03	4,454
Influenza	6,876,684	4.48E-03	1,443,223	0.06	30,856
Marijuana	4,809,337	3.31E-03	968,892	3.75E-02	18,236

We identified one notable increase in discussion each for Ebloa, influenza, and marijuana using the normalized frequencies over time (Figure 1). First, the normalized count on marijuana almost doubled from the previous month in February 2009. The heightened level of discussions continued for two months then slowly dropped back to the previous level. Second, in April of 2009, the normalized count on influenza almost doubled from the previous month. Third, October 2014 accounts for the Ebola discussions. The discussions on Ebola showed the most increase, jumping more than five times from the previous month. The number of members discussing each issue increased in a similar manner (Figure 1). The Discussions on electronic cigarette was relatively steady from October 2007 to May 2015.

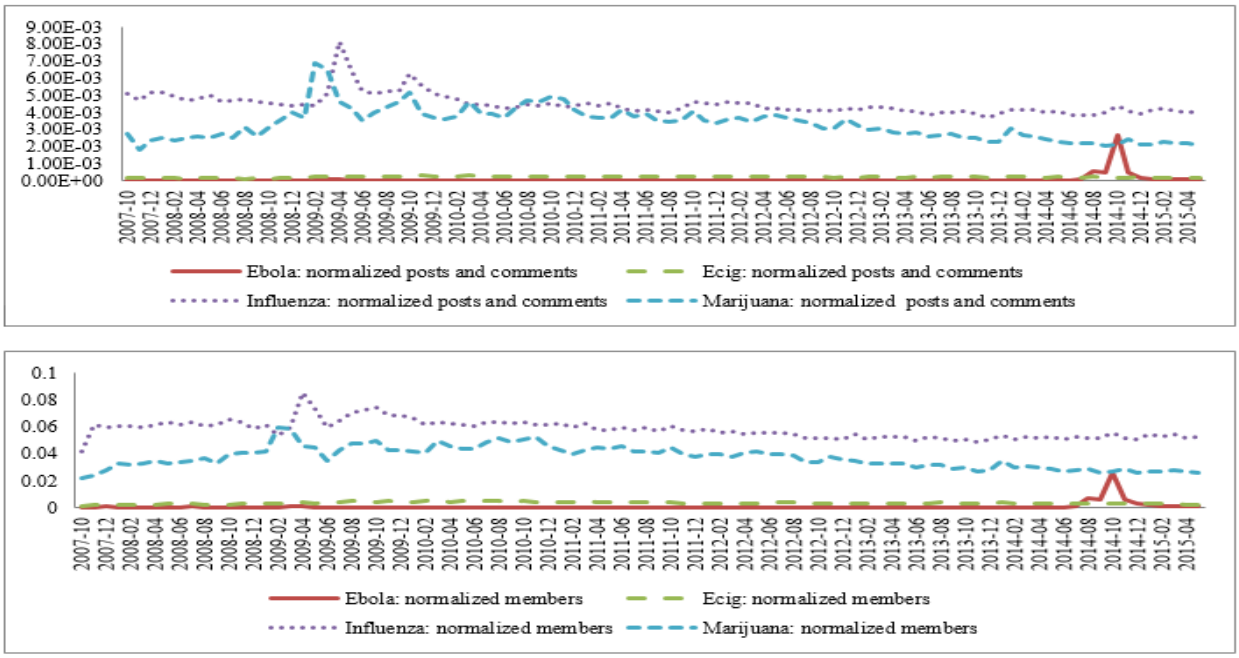

Figure 1. The Line Graphs of normalized frequencies over time for posts and comments with key terms and members who used the key terms

Discussions on Ebola, electronic cigarettes, influenza, and marijuana, however, only accumulated to a fraction of the overall discussions on Reddit (Table 2). Although the community as a whole did not frequently talk about these health-related issues, this still amounted to more than 3,000 members for Ebloa, the least discussed issue, and more than 137,000 members for influenza, the most discussed issue in a month with a normal level of discussion, May 2015.

New subreddits to discuss Ebloa, Electronic Cigarette, Influenza, and Marijuana

Members of Reddit created a number of subreddits specifically focusing on Ebola, electronic cigarettes, influenza, and marijuana, although they also discussed the issues in many different subreddits (Table 3). Using the key terms (Table 1), we detected a total of 450 topically dedicated subreddits that were created between October 2007 and May 2015. For example, marijuana was casually discussed in 18,236 subreddits (i.e., subreddits with key terms in posts or comments), while members created at least 244 subreddits (i.e., key terms in names of subreddits) to talk about marijuana.

Table 3. Newly created communities dedicated to focus on Ebola, Electronic Cigarette, influenza, and Marijuana

Issues	Subreddits, n		Example Subreddits
	Names	Posts, comments	
Ebola	37	4,454	EbolaOutbreak, batebola, Ebolaworld, ebola2014, ebolaUS, ILoveEbola, EbolaEpidemic, EbolaWestAfrica, ebolanews
Electronic cigarette	3	6,039	Ecigclassifieds, ecigclassifiedseu, ecigclassifiedsuk
Influenza	166	30,856	Influenza, Birdflu, flu,
Marijuana	244	18,236	Marijuana, ganja, cannabiscultur, ganjaoutlaw, cannabis_marijuana

RQ 2. What do Reddit members discuss on these health issues (a) in times of elevated activities or (b) in general, if the issues have a steady level of discussions?

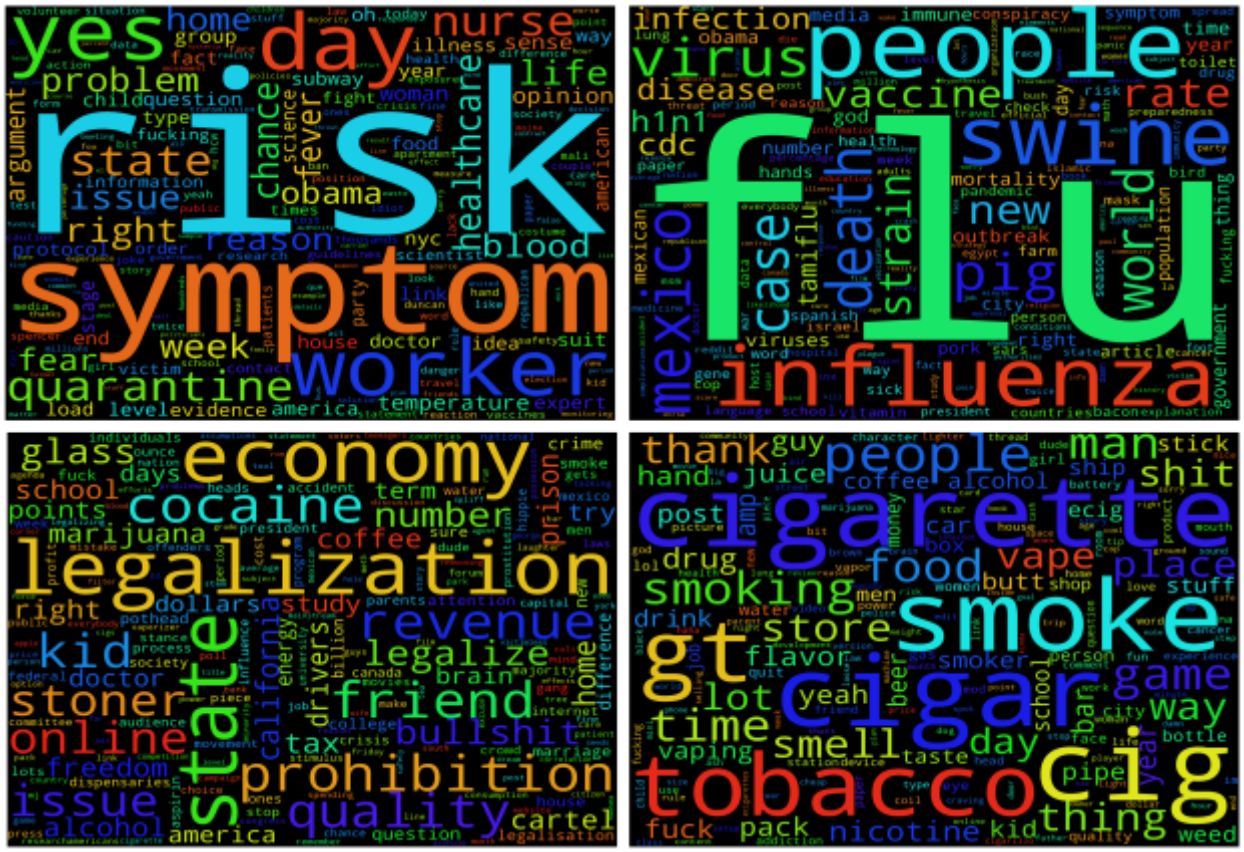

Figure 2. Word cloud overviews of emerging topics for Ebola (top left), influenza (top right), and marijuana (bottom left) as well as general word cloud overviews for electronic cigarette (bottom right)

From RQ1, we learned that discussions focusing concerning Ebola, influenza, and marijuana, each had one sudden increase of activities. Thus, we created word cloud overviews of emerging topics for Ebola, influenza, and marijuana, while creating a general word cloud overviews for electronic cigarette (Figure 2).

According to the word cloud overview generated by the LDA topic modeling algorithm, we can infer that Reddit members are most concerned about 'risk' and 'symptoms' regarding Ebola. For influenza, members used terms like 'Mexico', 'Obama', 'CDC', and 'conspiracy', along with H1N1 influenza related terms (e.g., 'H1N1', 'Swine') as well as H5N1 related terms like 'Egypt' and 'pig'. Topics regarding 'legalization', 'prohibition', 'economy', and 'state' appeared in discussions regarding marijuana. The general word cloud overview for electronic cigarettes has more commercially related terms such as 'quality', 'prices', 'shop', and 'store' than the other three discussions, however substantially more terms related to tobacco (e.g., 'tobacco', 'cigarette', 'cigar') are shown in Figure 2. Other notable topics for electronic cigarette that were identified via the LDA were 'quitting smoking', 'fun experience', and 'health information'.

The LDA algorithm identified 'quit', 'addiction', 'habit', 'cravings', 'gum' and 'turkey' for 'quitting smoking', associated 'fun', 'experience', 'safe', and 'pleasure' with 'fun experience', and linked 'cancer', 'risk', 'study', 'evidence', 'research', 'article', 'data', and 'science' with 'health information'. These topics highlighted a great opportunity for better health education (See Discussion).

Table 4. Posts and comments containing URLs

Issues	URL, n	Percentages of posts/comments with URLs to the total number of posts/comments
Ebola	32,863	13.03%
Electronic cigarette	22,839	6.42%
Influenza	783,350	11.39%
Marijuana	390,675	8.12%

To check the validity of the results, we extracted and investigated the URLs to ensure that frequencies are not inflated by marketing content. The types of URLs shared by members were similar in nature for all four issues. Members shared websites that are concerning information (e.g., Wikipedia, CDC), news (e.g., NY Times), personal stories (e.g., blogs), other social media platforms, (e.g., Youtube), different Reddit posts, and commercial resources (e.g., amazon). Although the proportion of each type of URLs is different, members shared a relatively small number of posts and comments with URLs compared to the overall posts and comments focusing on all four issues.

Discussion

Principal Findings

We examined four different infectious disease related or potentially stigmatized health related issues discussed on Reddit. We discovered three periods with higher levels of activities on Reddit. We observed that there were almost twice as many marijuana related discussions in February 2009 compared to the previous month, due – we suspect – to the announcement of a major shift in federal policy. Attorney General Eric Holder confirmed that Drug Enforcement Administration would halt medical marijuana raids and give states the power to regulate medical marijuana usage for pain control in February of 2009[53]. In April of 2009, discussions about influenza almost doubled from the previous month. This is likely due to the fact that a novel strain of H1N1 influenza virus was discovered in North America in the spring of 2009[54] and the Centers for Disease Control and Prevention (CDC) confirming the first two cases of human infection with H1N1 influenza virus in the United States in April of 2009[55]. On September 30, 2014, the United States had its first diagnosed case of Ebola in Texas, and the first Ebola related death on October 8, 2014[56]. We observed that discussions on Ebola, a potentially fatal infectious disease, surged more than five times from the previous month in October of 2014. The news related to Ebola, influenza, and Marijuana align well with the results from topic model analyses (RQ2). On the basis of these changes of activities, Reddit may be a valuable source of data for tracking the prevalence of public interests on infectious diseases (i.e., Ebola and influenza) and controversial health related issues (i.e., electronic cigarette and marijuana) over time (RQ1).

The result of our analysis on electronic cigarette discussions suggests that Reddit contains more than just commercial content despite the fact there are at least three subreddits focusing on classified content (Table 4). For instance, a subreddit called 'Ecigclassifieds' consists mainly of commercial content, thus the content of these subreddits deserves further investigation to better utilize the data. From electronic cigarette discussion, we identified three topics, 'quitting smoking', 'fun experience', and 'health information' that highlighted opportunities for better health education. From their associated terms (see Results), we can infer that Reddit members are seeking

information on these three topics. Information seeking behavior on Reddit suggests Reddit's utility as another social media platform for information distribution and as a data source for understanding user groups (e.g., electronic cigarette smokers) and identifying better health education. Why members are seeking health information on Reddit is an unanswered research question, although a recent study suggests that electronic cigarette related health information from public health agencies may be too difficult for the general public to comprehend[57].

Reddit members also created at least 450 relevant new subreddits specifically focusing on these four issues. How the content from these subreddits contrast with the content from multiple subreddits on the same issue is an unanswered question. Previous studies[30,31,33,34,39] analyzed content from a handful of especially dedicated subreddits for their studies. However, our finding suggests that at least for discussions of Ebola, influenza, electronic cigarettes, and marijuana, members mentioned these issues on thousands of subreddits (Table 3). For instance, a common issue like influenza was discussed in over 30,000 subreddits, and even a focused topic like Ebloa were discussed in over 4,400 subreddits. Thus, we believe analyzing a wider number of subreddits can improve recall of the relevant content.

Limitation, Future Directions, and User Privacy

Reddit offers substantial potential for understanding the public reactions to health-related topics, however, not without a number of limitations. Although Reddit is a widely-used platform, it is more frequently used by young males[58,59] and may be subjective to self-selection bias. Reddit members are not necessarily representative of the general public, however, the levels of activity on Reddit aligned with the United States news and deserve a further investigation, especially with respect to location of postings and the overall reactions in Reddit. To better understand the reaction of the general public, studying different platforms and avenues, Facebook and Twitter for example, is warranted. Our analysis suggests that given the increasing popularity and use of Reddit, as well as the increasing frequency of discussions concerning our topic of interests, Reddit provides a productive starting point for investigating infectious disease related or controversial health issues.

Another limitation lies in the methodology. In RQ1, we used a relatively rudimentary lexicon-based approach to extract posts and comments explicit mentioning variations of pre-specified key terms. One major shortcoming of such approach is the selection of key terms. For example, utilizing a large set of key terms will undoubtedly create more false-positives, whereas too limited a set of key terms will surely result in more false-negatives. Moreover, partial matches can produce false-positive matches. We believe the figures for influenza were inflated because 'flu' can be a part of a longer word such as 'fluorine' or 'flute'. In future studies, we suggest that precision rather than recall should be emphasized in order to eliminate irrelevant discussions. Other difficulties in mining social media data include the fact that social media text is frequently characterized by extensive of acronyms, abbreviations, and slang terms[60]. Although we included the most frequently found abbreviations and slang terms, lexicon-based approaches are to omit unknown forms of abbreviations and slang. More sophisticated methods utilizing knowledge-based[61] or corpus-based[62] approaches could produce different results. Furthermore, a smaller timeframe can better measure the timeliness of the observed reactions as oppose to the one month timeframe used in RQ1. In RQ2, we relied on a systematic analysis of the news to verify the result of our investigation. However, data driven qualitative analysis[63] can further bolster our findings and provide the contextual information on the discussions of our interests. Sentiment analysis on the extracted discussion can also provide further clues about general public reactions on various health related topics[9].

Research and applications using social media data should be highly sensitive to user privacy, especially for potentially stigmatized topics. Although at least some social media data are publicly available, researchers should consider ethical implications when processing data even for population-level social media research using public data[64–66]. For this reason, we have refrained from using direct quotations from Reddit users in this paper.

Conclusion

As evident by the frequencies over time of discussions, inflated discussions after major news, as well as newly created subreddits specifically focusing on these health-related issues, Reddit could be a useful platform for understanding the concerns and opinions of the general public, especially for issues focusing on controversial topics, such as abuse and addiction as well as infectious diseases of public health interest. By utilizing the content, we also identified opportunities for better health education that could improve public health outcomes. We created topic models using LDA and generated topically associated words and created word cloud visualizations to show (1) emerging topics by contrasting to the prior topic models or (2) main themes of the discussions. We believe our insights and analyses can be generalized to other similar health related issues in the Reddit platform. Understanding public reactions to these issues has the potential to expand the scope of public-health practice.

Acknowledgments

We restricted our analysis to publicly available discussion content. The study was exempted from review by the University of Utah's Institutional Review Board (Ethics Committee) [IRB 00076188].

Author AP was funded by National Library of Medicine of the National Institutes of Health under award number T15 LM007124. Author MC's contribution to this research was supported by National Library of Medicine of the National Institutes of Health under award numbers R00LM011393 & K99LM011393.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Perrin A. Social Media Usage : 2005-2015. 2015.
2. Fung IC-H, Duke CH, Finch KC, Snook KR, Tseng P-L, Hernandez AC, et al. Ebola virus disease and social media: A systematic review. Am J Infect Control. 2016;
3. Brownstein JS, Freifeld CC, Madoff LC. Digital Disease Detection — Harnessing the Web for Public Health Surveillance. N Engl J Med. 2009 May 21;360(21):2153–7.
4. Eysenbach G. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. J Med Internet Res. 2009 Mar 27;11(1):e11.
5. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital Epidemiology. Bourne PE, editor. PLoS Comput Biol. 2012 Jul 26;8(7):e1002616.
6. Thackeray R, Neiger BL, Smith AK, Van Wagenen SB. Adoption and use of social media among public health departments. BMC Public Health. 2012 Dec 26;12(1):242.
7. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter. Drug Saf. 2014 May 29;37(5):343–50.
8. Alvaro N, Conway M, Doan S, Lofi C, Overington J, Collier N. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. J Biomed Inform. 2015 Dec;58:280–7.
9. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. PLoS Comput Biol. 2011;7(10).
10. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control. 2015 Jun;43(6):563–71.
11. Conway M, O'Connor D. Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications. Curr Opin Psychol. 2016 Jun;9:77–82.
12. Paul MJ, Dredze M. You are what you Tweet: Analyzing Twitter for public health. Proc Fifth Int AAAI Conf Weblogs Soc Media. 2011;265–72.
13. Dredze M. How Social Media Will Change Public Health. IEEE Intell Syst. 2012 Jul;27(4):81–4.
14. Culotta A. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. Lang Resour Eval. 2013 Mar 13;47(1):217–38.
15. Culotta A. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. Lang Resour Eval. 2013 Mar;47(1):217–38.
16. Chunara R, Andrews JR, Brownstein JS. Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. Am J Trop Med Hyg. 2012 Jan 1;86(1):39–45.
17. Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-Time Event Detection by Social Sensors. In: Proceedings of the 19th international conference on World wide web - WWW '10. New York, New York, USA: ACM Press; 2010. p. 851.
18. Myslín M, Zhu S-H, Chapman W, Conway M. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res. 2013 Aug 29;15(8):e174.
19. Cavazos-Rehg PA, Krauss M, Fisher SL, Salyer P, Grucza RA, Bierut LJ. Twitter chatter about marijuana. J Adolesc Heal. 2015;56(2):139–45.
20. De Choudhury M, Counts S, Horvitz EJ, Hoff A. Characterizing and predicting postpartum depression from shared facebook data. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14. New York, New York, USA: ACM Press; 2014. p. 626–38.
21. Maclean D, Gupta S, Lembke A, Manning C, Heer J. Forum77 : an analysis of an online health forum dedicated to addiction recovery. In: Proceedings of the companion publication of the 18th ACM conference on Computer supported cooperative work & social computing (CSCW). 2015.

22. Chee BW, Berlin R, Schatz B. Predicting Adverse Drug Events from Personal Health Messages. In: AMIA Annu Symp Proc. American Medical Informatics Association; 2011. p. 217–26.
23. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009 Feb 19;457(7232):1012–4.
24. Newman MW, Lauterbach D, Munson SA, Resnick P, Morris ME. It's not that i don't have problems, i'm just not putting them on facebook. In: Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11. New York, New York, USA: ACM Press; 2011. p. 341.
25. Hartzler A, Pratt W. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. J Med Internet Res. 2011 Jan;13(3):e62.
26. Sun N, Rau PP-L, Ma L. Understanding lurkers in online communities: A literature review. Comput Human Behav. 2014 Sep;38:110–7.
27. Park A, Hartzler AL, Huh J, McDonald DW, Pratt W. Homophily of Vocabulary Usage: Beneficial Effects of Vocabulary Similarity on Online Health Communities Participation. AMIA . Annu Symp proceedings AMIA Symp. 2015;2015:1024–33.
28. Copeland P, Romano R, Zhang T, Hecht G, Zigmond D, Stefansen C. Google Disease Trends: An update. In: International Society of Neglected Tropical Diseases 2013. 2013.
29. Reddit. Reddit API Documentation [Internet]. 2015. Available from: https://www.reddit.com/dev/api; Archived at: http://www.webcitation.org/6dGU2ksOW
30. De Choudhury M, De S. Mental health discourse on Reddit: self-disclosure, social support, and anonymity. In: Proceedings of ICWSM, AAAI. Ann Arbor, Michigan, USA; 2014. p. 71–80.
31. Kumar M, Dredze M, Coppersmith G, De Choudhury M. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15. New York, New York, USA: ACM Press; 2015. p. 85–94.
32. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16. New York, New York, USA: ACM Press; 2016. p. 2098–110.
33. Cunha TO, Weber I, Haddadi H, Pappa GL. The Effect of Social Feedback in a Reddit Weight Loss Community. In: Proceedings of the 6th International Conference on Digital Health Conference - DH '16. New York, New York, USA: ACM Press; 2016. p. 99–103.
34. Gkotsis G, Oellrich A, Hubbard TJP, Dobson RJB, Liakata M, Velupillai S, et al. The language of mental health problems in social media. In: Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, California: Association for Computational Linguistics; 2016. p. 63–73.
35. Tamersoy A, Choudhury M De, Chau DH. Characterizing smoking and drinking abstinence from social media. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media. ACM; 2015. p. 139–48.
36. Chen AT, Zhu S-H, Conway M. What Online Communities Can Tell Us About Electronic Cigarettes and Hookah Use: A Study Using Text Mining and Visualization Techniques. J Med Internet Res. 2015 Sep 29;17(9):e220.
37. Reddit. Reddit in 2015 [Internet]. 2015. Available from: http://www.redditblog.com/2015/12/reddit-in-2015.html; Archived at: http://www.webcitation.org/6eTHN0TFD
38. Reddit_Member. I have every publicly available Reddit comment for research. ~ 1.7 billion comments @ 250 GB compressed. Any interest in this? [Internet]. 2015. Available from: https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/; Archived at: http://www.webcitation.org/6kgAuNxDE
39. Park A, Conway M. Longitudinal Changes in Psychological States in Online Health Community Members: Understanding the Long-Term Effects of Participating in an Online Depression Community. J Med Internet Res. 2017 Mar 20;19(3):e71.
40. Park A, Conway M. Towards Tracking Opium Related Discussions in Social Media. Online J Public Health Inform. 2017 May 2;9(1):e73.
41. Cheng X, Yan X, Lan Y, Guo J. BTM: Topic Modeling over Short Texts. IEEE Trans Knowl Data Eng. 2014 Dec 1;26(12):2928–41.
42. Kalyanam J, Velupillai S, Conway M, Lanckriet G. From event detection to storytelling on microblogs. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). San Francisco, CA: IEEE; 2016. p. 437–42.
43. Friedman C, Elhadad N. Natural Language Processing in Health Care and Biomedicine. In: Biomedical Informatics. London: Springer London; 2014. p. 255–84.

44. Hartzler AL, McDonald DW, Park A, Huh J, Weaver C, Pratt W. Evaluating health interest profiles extracted from patient-generated data. In: AMIA . Annual Symposium proceedings AMIA Symposium. 2014. p. 626–35.
45. Hartzler AL, Taylor MN, Park A, Griffiths T, Backonja U, McDonald DW, et al. Leveraging cues from person-generated health data for peer matching in online communities. J Am Med Informatics Assoc. 2016 Feb 5;ocv175.
46. Bird S. NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions.)Association for Computational Linguistics.; 2006. p. 69–72.
47. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res. 2003 Mar;3:993–1022.
48. Park A, Hartzler AL, Huh J, Hsieh G, McDonald DW, Pratt W. "How Did We Get Here?": Topic Drift in Online Health Discussions. J Med Internet Res. 2016 Nov 2;18(11):e284.
49. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010. p. 45–50.
50. Mueller A. wordcloud [Internet]. MIT. 2013. Available from: https://pypi.python.org/pypi/wordcloud; Archived at: http://www.webcitation.org/6oD8n8hqa
51. Chen H, Kelliher A. Conversational Lives: Visualizing Interpersonal Online Social Interactions. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2011. p. 241–50.
52. Huang J, Kornfield R, Szczypka G, Emery SL. A cross-sectional examination of marketing of electronic cigarettes on Twitter. Tob Control. 2014 Jul;23(suppl 3):iii26-iii30.
53. Johnson MA. DEA to halt medical marijuana radis: Holder confirms states to have final say on use of drug for pain control. nbcnews.com. 2009;
54. The New England journal of medicine. Clinical Aspects of Pandemic 2009 Influenza A (H1N1) Virus Infection. N Engl J Med. 2010 May 6;362(18):1708–19.
55. Jain S, Kamimoto L, Bramley AM, Schmitz AM, Benoit SR, Louie J, et al. Hospitalized Patients with 2009 H1N1 Influenza in the United States, April–June 2009. N Engl J Med. 2009 Nov 12;361(20):1935–44.
56. Centers for Disease Control and Prevention (CDC). Questions and Answers: 2014 Ebola Outbreak [Internet]. Available from: https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/qa.html; Archived at: http://www.webcitation.org/6lhkiWPCI
57. Park A, Zhu S-H, Conway M. The Readability of Electronic Cigarette Health Information and Advice: A Quantitative Analysis of Web-Based Information. JMIR public Heal Surveill. 2017 Jan 6;3(1):e1.
58. Duggan M, Smith A. 6% of online adults are Reddit users. Pew Internet Am Life Proj. 2013;3.
59. Bogers T, Wernersen R. How `Social' are Social News Sites? Exploring the Motivations for Using Reddit.com. In: iConference 2014 Proceedings. Berlin, Germany: iSchools; 2014. p. 329–44.
60. Park A, Hartzler AL, Huh J, McDonald DW, Pratt W. Automatically Detecting Failures in Natural Language Processing Tools for Online Community Text. J Med Internet Res. 2015 Aug 31;17(8):e212.
61. Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: WordNet: An electronic lexical database. The MIT Press; 1998. p. 265–83.
62. Turney PD. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001). 2001.
63. Strauss AL, Corbin JM. Basics of qualitative research. Newbury Park, CA: Sage Publications; 1990.
64. Conway M. Ethical Issues in Using Twitter for Public Health Surveillance and Research: Developing a Taxonomy of Ethical Concepts From the Research Literature. J Med Internet Res. 2014 Dec 22;16(12):e290.
65. Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical Challenges of Big Data in Public Health. PLoS Comput Biol. 2015;11(2):1–7.
66. Mikal J, Hurst S, Conway M. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. BMC Med Ethics. 2016;17(1):1–11.