Original Paper

Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection

Didi Surian[1], PhD; Dat Quoc Nguyen[2], MSc; Georgina Kennedy[1], MEng; Mark Johnson[2], BSc, MA, PhD; Enrico Coiera[1], MBBS, PhD; Adam G Dunn[1], PhD

[1]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, North Ryde, New South Wales, Australia

[2]Department of Computing, Faculty of Science and Engineering, Macquarie University, North Ryde, New South Wales, Australia

Corresponding Author:
Didi Surian, PhD
Centre for Health Informatics
Australian Institute of Health Innovation
Macquarie University
Level 6, 75 Talavera Road
North Ryde, New South Wales, 2109
Australia
Phone: 61 +61298502455
Fax: 61 +61298502499
Email: didi.surian@mq.edu.au

Abstract

Background: In public health surveillance, measuring how information enters and spreads through online communities may help us understand geographical variation in decision making associated with poor health outcomes.

Objective: Our aim was to evaluate the use of community structure and topic modeling methods as a process for characterizing the clustering of opinions about human papillomavirus (HPV) vaccines on Twitter.

Methods: The study examined Twitter posts (tweets) collected between October 2013 and October 2015 about HPV vaccines. We tested Latent Dirichlet Allocation and Dirichlet Multinomial Mixture (DMM) models for inferring topics associated with tweets, and community agglomeration (Louvain) and the encoding of random walks (Infomap) methods to detect community structure of the users from their social connections. We examined the alignment between community structure and topics using several common clustering alignment measures and introduced a statistical measure of alignment based on the concentration of specific topics within a small number of communities. Visualizations of the topics and the alignment between topics and communities are presented to support the interpretation of the results in context of public health communication and identification of communities at risk of rejecting the safety and efficacy of HPV vaccines.

Results: We analyzed 285,417 Twitter posts (tweets) about HPV vaccines from 101,519 users connected by 4,387,524 social connections. Examining the alignment between the community structure and the topics of tweets, the results indicated that the Louvain community detection algorithm together with DMM produced consistently higher alignment values and that alignments were generally higher when the number of topics was lower. After applying the Louvain method and DMM with 30 topics and grouping semantically similar topics in a hierarchy, we characterized 163,148 (57.16%) tweets as evidence and advocacy, and 6244 (2.19%) tweets describing personal experiences. Among the 4548 users who posted experiential tweets, 3449 users (75.84%) were found in communities where the majority of tweets were about evidence and advocacy.

Conclusions: The use of community detection in concert with topic modeling appears to be a useful way to characterize Twitter communities for the purpose of opinion surveillance in public health applications. Our approach may help identify online communities at risk of being influenced by negative opinions about public health interventions such as HPV vaccines.

(J Med Internet Res 2016;18(8):e232) doi: 10.2196/jmir.6045

KEYWORDS

topic modelling; graph algorithms analysis; social media; public health surveillance

Introduction

The human papillomavirus (HPV) vaccine was first introduced to reduce the incidence of HPV and the majority of cervical cancers [1]. Despite evidence of its safety and efficacy [2-5], the quality of information about the vaccine on the Web is varied [6,7], and coverage of the vaccine is low in some countries, including the United States [8]. For vaccines generally, there is evidence to suggest that negative information about vaccines from celebrities, health practitioners, and news media can increase vaccine hesitancy and refusal [9-11]. Although the HPV vaccine is still a recent addition to the armament of public health, it is important to perform surveillance on social media to understand various opinions about vaccination.

The use of social media information in public health applications has previously centered on forecasting clinical outcomes that have traditionally been measured using surveys and registries. Applications of data mining using Twitter have included influenza surveillance [12-14] and measuring spatial differences in language or mood [15,16]. Sentiment and language analyses on Twitter have been used as indicators for the geographical variation in heart disease mortality [17]. Examples that are relevant to our research include the use of topic modeling to extract tobacco-related tweets in the United States [18] and the surveillance of information about the misunderstanding and misuse of antibiotics from online media [19]. There is a growing area of research considering the spread of information, news, and opinions about vaccines [20-24], and research in this area focuses on measuring associations between misinformation, beliefs, and decision making across a range of community health practices [25].

People interact and form relationships to each other on social media. With these relationships, communities are formed. The structure of online communities influences—and can be influenced by—the information that enters and is diffused through them. Studies examining the spread of information through online communities, social media, and news media have shown that the heterogeneous social structure of the network and external factors can play a role in how far and fast new information spreads [26,27]. That competition between memes and natural attrition may affect how far and fast they can spread and how quickly they decay [28-30], and that topics of interest in a community can also influence the formation and persistence of social structure [31].

Differences in the content of the information posted in online news and social media influence the variation in opinions and beliefs in online communities. While the opinions held by community members and the content they post are not equivalent, we assume that the content provides a reasonable proxy for the opinions that might influence decision making. Topic modeling methods are appropriate for identifying thematic structure (topic) within a set of tweets because they can be applied to an unstructured corpus of documents and there is no requirement that the topic be defined in advance [32]. The primary challenge associated with applying topic modeling to tweets comes from the short length of tweets (140 characters). Despite this challenge, topic modeling has been used to examine topics across a range of subjects on Twitter—by pooling tweets to produce longer documents to analyze [33-35], or applying extensions or alternatives to existing models that work better on shorter documents [36-40].

We expected to find that homophily and contagion would lead to a clustering of opinions in online communities, but to date there has been little research done to measure this important information for vaccines. Our aim was to evaluate the combination of community structure and topic modeling methods in measuring the distribution of topics about HPV vaccines from the tweets posted by users within communities on Twitter, with the broader goal of evaluating a new process for characterizing online communities by the public health information expressed by the community members.

Methods

Study Data

Using repeated searches via the Twitter Search Application Programming Interface, we collected tweets about HPV vaccines between October 1, 2013, and October 29, 2015, using the keywords shown in Table 1. For each tweet labeled as English language by Twitter, we stored the text of the tweet and the related metadata. Each time a new user tweeted about HPV vaccines for the first time in the period, we additionally collected the lists of users they followed and who followed them. This information on relationships was used later to construct the network of users for our analysis. At the conclusion of the data collection period, there were 302,856 tweets (including retweets) and 112,944 users. Following the data collection, we removed users who were suspended, protected, or deleted, which left 101,519 users and 285,417 tweets for the analysis.

Table 1. Search keywords used to collect tweets about HPV vaccines for our analysis.

No.	Keywords
1.	“HPV” and “vaccine”
2.	“HPV” and “vaccination”
3.	“gardasil”
4.	“cervical” and “vaccination”
5.	“cervical” and “vaccine”
6.	“cervarix”

We pre-processed the tweets prior to topic modeling. For words that were hashtags (beginning with "#") or usernames (beginning with "@"), we made no further modifications. The remaining words in the text of the tweet were converted to lowercase, and we removed stop words, the word "RT" (which represents a retweet), and any numerical values. We then applied the Porter stemmer [41]. We excluded URLs (uniform resource locator) generated from generic URL shortening services (eg, "http://bit.ly") and included the domain of any full URLs identified from the list of expanded URLs. Document size plays an important role in topic modeling methods [42], so we chose to assign all of the tweets with fewer than three words to a single extra topic (1114 tweets), leaving 284,303 tweets.

We ran Latent Dirichlet Allocation (LDA) and the Dirichlet Mixture Model (DMM) to infer the topics of the 125,003 unique tweets that were identified within the set of 284,303 tweets after pre-processing the text. After inferring the topics using LDA and DMM, we mapped those topics back onto the full set, so that each tweet was associated with a single topic.

We constructed the network from the 4,387,524 follower connections among the 101,519 users using an undirected graph. A node represents a user, and an edge between two users is established if one was found to be following the other. The network included 100,826 (99.32%) users comprising the single largest connected component, 500 (0.49%) users who formed smaller islands disconnected from the largest connected component, and 193 (0.19%) disconnected users with no connections to the core. Within the largest connected component, the average number of social connections was 86.98 and the largest number of connections was 18,635. We measured the alignment between topics and communities for the users who were part of the largest connected component. More details on network construction are provided in Multimedia Appendix 1.

Community Detection

Community detection algorithms aim to find sets of nodes in a graph that have a greater density of connections within their set compared to across sets. Traditionally, community detection algorithms produced a hard clustering—where each node belongs to only one community [43-47]. Some of the more recent methods have considered overlapping communities [48]. In this work, we chose two algorithms that assign each node to a single community, are known to produce reliable results, and work efficiently in large networks.

The Infomap algorithm was developed to extract community structure in large complex networks [49]. Using random walks as a proxy for the way that information flows through a system, the method first determines the probability of visiting each node in the network and then characterizes the community and node structure of the network as a Huffman code. By progressively modifying the community affiliation, the aim is to compress the code describing the network to its smallest size. We used the implementation of Infomap from igraph [50].

The Louvain algorithm is a relatively fast community detection algorithm to compute and can therefore scale to large networks [51]. The algorithm is agglomerative—nodes are initialized to belong to a community of size one and sequentially aggregated with the neighboring community that produces the greatest gain in modularity (if a positive gain exists). Communities detected in this first phase become nodes in a new network with edge weights determined by the number of connections between the communities from the first phase. The algorithm therefore constructs a hierarchical representation of the network and proceeds until no more modularity gains can be identified. The final clustering that results from this procedure is used to define the community structure. We used the code released by the author of Louvain from MapEquation [52].

Topic Inference

Topic modeling is used to find natural clusters based on the co-occurrence of words. We used the Latent Dirichlet Allocation (LDA) model [53] and the Dirichlet Multinomial Mixture (DMM) model [54]. The LDA model is a standard method for topic modeling, and the DMM model is a variant especially developed for short documents such as tweets. When applying the DMM model, only one topic is assigned to each document, so we labeled each tweet according to the topic inferred by the DMM model. For LDA—where a probability topic distribution is produced for each document—tweets were labeled using the topic with the largest probability [33,40,55]. We used the implementation of LDA from gensim [56] and the jLDADMM implementation of DMM [57]. For both methods, we used standard settings for each model [39,55] and did not attempt to optimize the parameters further. Details of the formal specification and notations for the methods are provided in Multimedia Appendix 1.

Alignment Measures

The aim of measuring the alignment between topics and communities is to determine if the topics appear more frequently within some communities relative to all others. Since each tweet was associated with a single topic, we represented communities by the distribution of topics in the tweets posted by the users in that community. We adapted measures of alignment that are typically used to quantify the quality of an estimated clustering against an observed clustering to compare between the clustering methods that use the observed structure (social connections) and the clustering methods that use the observed content (topics in tweets). There are several appropriate metrics for assessing cluster quality in this scenario, including purity, normalized mutual information (NMI), and the adjusted Rand index (ARI) [39,58] (see Multimedia Appendix 1 for definitions).

While these typical metrics provide a general measure of the alignment between community structure and the topics of the tweets posted by users in those communities, they were not useful for summarizing how topics may be disproportionately represented within a small subset of the communities. We therefore additionally considered a measure of topic concentration (*TC*). We defined a *TC* value by the smallest number of communities required to cover a specified percentage of the tweets about a given topic, so TC_{95} is the number of communities required to cover 95% of the tweets in that topic, and TC_{100} is the number of communities that covers every tweet labeled with that topic. A lower TC_{95} value therefore implies a

higher concentration of topics within a small number of communities.

When comparing measures of topic concentration across multiple networks to determine alignment, the differences in the number of tweets associated with each topic can influence the measures independently of the alignment, so we used permutation tests to produce a fair comparison. The permutation tests create a baseline distribution of *TC* values that may occur in the absence of any real alignment, which can then be used to establish the level of alignment relative to the levels of alignment that could be produced by chance [59]. To do this, we randomly permuted the topics associated with each tweet such that the distributions of tweets per topic and tweets per community remained the same as the observed network. We then compared the observed TC_{95} values against the distribution of TC_{95} values produced in the permutation tests. Typical permutation tests report the percentile of a single observed value within the distribution of values produced after permutation. In the permutation tests we applied, distributions of TC_{95} values (one for each topic) were produced rather than single values, so we used a two-sample Kolmogorov Smirnov test to compare the distributions. The Kolmogorov-Smirnov test statistic varies between 0 and 1, and a higher test statistic means that the topics were more concentrated within individual communities than would be expected if the same number of tweets per topic were randomly distributed across the communities.

Manual Intrusion Tests

We performed intrusion tests on the topics from the tweets. One investigator, blinded to the results of the topic modeling, was presented with sets of five test cases per pairwise combination of topics. Each test case included the text of five tweets chosen at random from one topic and one tweet chosen at random from a different topic. The investigator was tasked with identifying the tweet that did not belong to the topic. The results of these intrusion tests indicated how well the topic modeling was able to capture semantic differences in the tweets. We additionally used the results of the intrusion tests to construct a hierarchy of topics based on their semantic dissimilarity by applying multidimensional scaling [60-62]. The method produces a distance between every pair of topics, which is then used to merge the closest topics to construct the hierarchy.

Results

Community Detection and Topic Modeling

The two community detection algorithms were applied to the largest connected component of 100,826 users. Applying the Louvain algorithm, we identified 38 distinct communities of sizes between 3 and 21,733 users. The Infomap algorithm identified 1334 distinct communities, ranging in size from 2 to 18,974 users.

We constructed a series of LDA and DMM models by varying the number of topics between 5 and 200. From the purity, NMI, and ARI scores, we found that the alignment between the community structure and the topics was higher across all measures for DMM compared to LDA. The highest purity score (0.495) and the highest ARI scores (0.166) were found when applying the DMM model with the Louvain algorithm. The highest NMI score (0.185) was found when applying the DMM model with the Infomap algorithm. The results of these experiments suggest that the DMM topic model may have produced a more realistic clustering of the tweets by topic.

The TC_{95} scores were consistently higher when using the DMM model compared to the LDA model (see Multimedia Appendix 1 for detailed results). In combination with the Infomap algorithm, TC_{95} scores were highest between 10 and 25 topics, and in combination with the Louvain algorithm, TC_{95} scores were highest between 20 to 30 topics. Considering these results, we used the DMM model (with 30 topics) and the Louvain algorithm to demonstrate the characterization of the communities by topic in what follows.

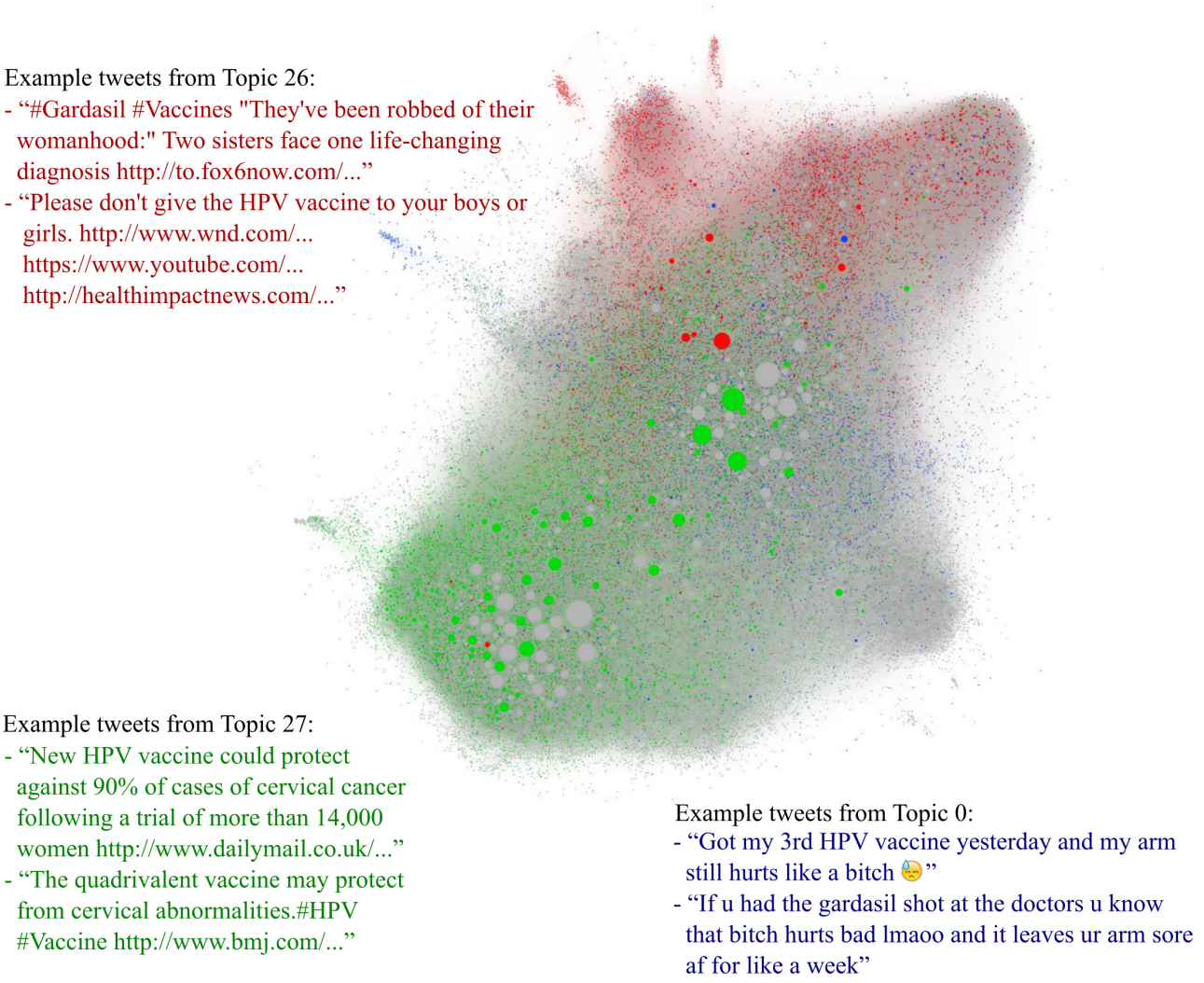
To illustrate how the topics tend to cluster within communities, we selected three representative topics and visualized them in the network constructed from the set of followers among the 100,826 users (Figure 1). The topics include one of the topics that captured clinical and scientific evidence (Topic 27), the topic comprising experiential tweets (Topic 0), and one of the topics describing side effects and harms (Topic 26).

Topic 27 includes words that are common to published studies about the efficacy of the vaccine such as “prevent,” “protect,” “study,” “news,” and “research.” Links to news media alongside other published articles and related media tended to be grouped within this topic, and the topic is broadly represented throughout the majority of the core network, including among the users with the greatest number of connections (typically news organizations in the center, and news organizations, health-related magazines, and scientific journals to one side).

Topic 0 captures a large number of tweets from users describing their own experiences with the vaccine, including temporal words such as “today,” “get,” “got,” and “go.” Tweets including phrases such as “my arm hurts like a...” were commonly assigned to this topic, and these users appeared to share fewer connections with other users posting about HPV vaccines.

In Topic 26, emotive words like “kill,” “victim,” and “death” are common. Tweets that include links to specific antivaccine websites were commonly assigned to this topic, and users posting tweets in Topic 26 appeared to cluster with different densities in three distinct groups that were separated from the groups of users posting tweets labeled as Topic 27.

Figure 1. A network of 100,826 users (nodes) who posted tweets about HPV vaccines in the period. The sizes of the nodes are proportional to the number of social connections they have in this network. Nodes are colored if they posted tweets labeled as Topic 0 (blue), Topic 26 (red), or Topic 27 (green). Node position was determined by a heuristic that attempts to locate connected nodes closer together, partially revealing the community structure.



Topic Grouping

We measured the quality of the topic modeling using the manual intrusion tests. Overall, the correct intruder was identified in 63.7% of the 4650 tests, which is a clear departure from the 16.7% that would be expected by chance. The hierarchy constructed from the manual intrusion tests revealed the semantically similar topics (Figure 2). The topic groups were (1) media debates, (2) politics and policy debates, (3) scandals and conspiracies, (4) side effects and harms, (5) public health advocacy, (6) clinical evidence, and (7) experiences. When measured across the groups of topics, the intrusion test accuracy was 76% and when measured within the groups of topics, the intrusion test accuracy was 49%. These results suggest that the separation among topic groups is clear (high score for intergroup accuracy and low score for intragroup accuracy).

Using the topics groups, we were then able to characterize the communities by the distribution of topics among the set of tweets posted by the users in those communities. Figure 3 details the topic distributions for three selected communities, notable because they illustrate the concentration of vaccine harms/conspiracies, evidence/advocacy, and experiential themes within different communities. Note also that the number of tweets per user is highest for users in the community that posts tweets labeled mostly among the vaccine harms/conspiracies theme and lowest for users in the community posting mostly about their experiences with the HPV vaccine.

Across the set of all communities, we found that users posting about their own experiences with the HPV vaccine belonged to communities for which the majority of tweets were related to evidence and advocacy. Of the 4548 users who posted tweets labeled as experiential, 3449 (75.84%) belonged to communities for which the majority of tweets were related to evidence/advocacy, 674 (14.8%) belonged to communities for which the majority of tweets were related to harms/conspiracies, 196 (4.3%) belonged to communities for which the majority of tweets were experiential, and 229 (5.0%) belonged to the group of users who were not connected to the core of the network. Figures 4 and 5 detail the distribution of themes within the communities.

Figure 2. A dendrogram of 30 topics (Topic 0-29) from the Dirichlet Mixture Model and one separate topic (Topic 30) for the tweets with fewer than 3 words. The groups were identified post-hoc and the colors represent themes—harms/conspiracies (red), evidence/advocacy (green), and experiential (blue).

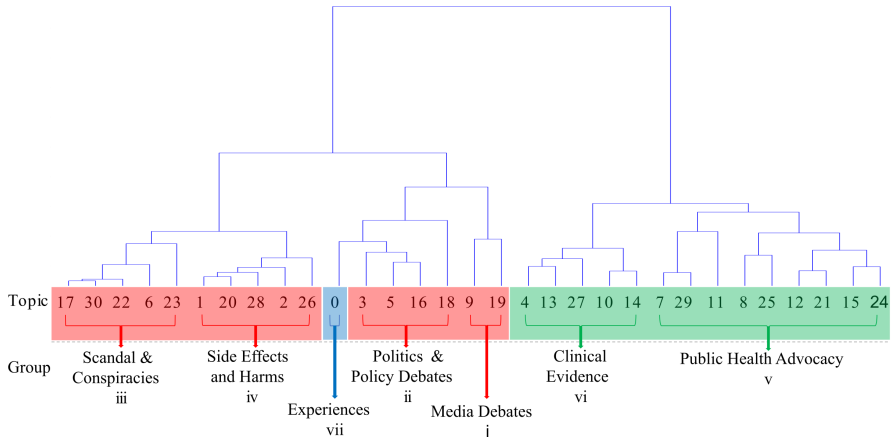


Figure 3. Topic distributions for 3 selected communities ordered by group and theme: (1) community 24 included 5275 users and an average of 2.46 tweets per user; (2) community 20 included 11,047 users and an average of 1.96 tweets per user; and (3) community 34 included 187 users and an average of 1.16 tweets per user.

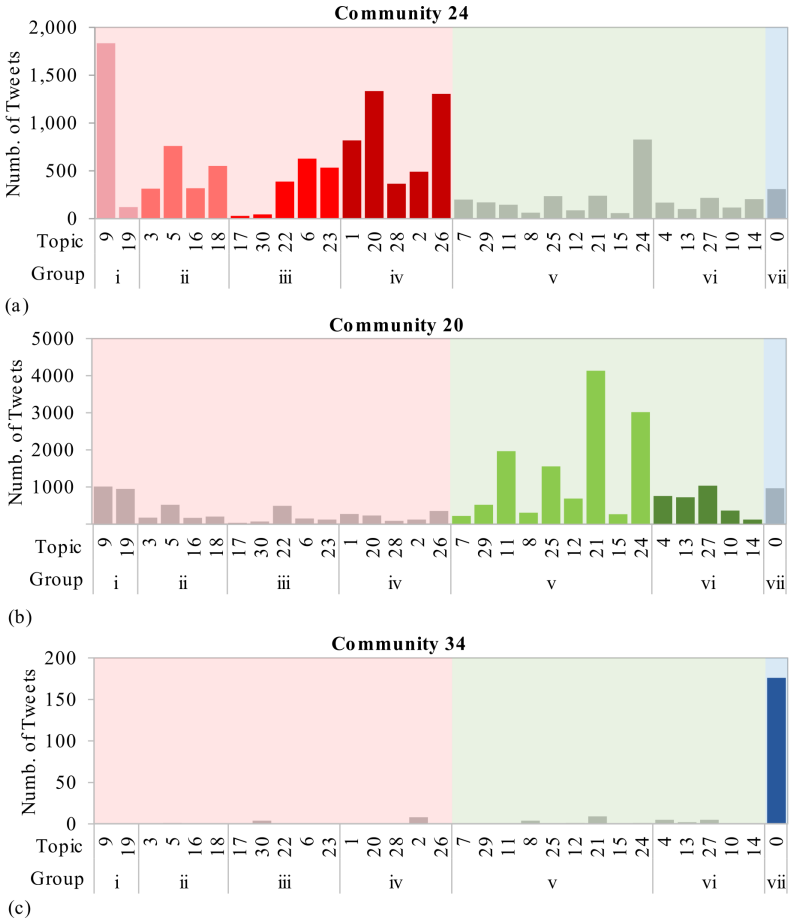


Figure 4. The proportion of evidence/advocacy and harms/conspiracies themes for the identified 39 communities. Each circle represents a community and the size is proportional to the number of tweets in the respective community. Communities further from the diagonal include greater proportions of experiential theme tweets.

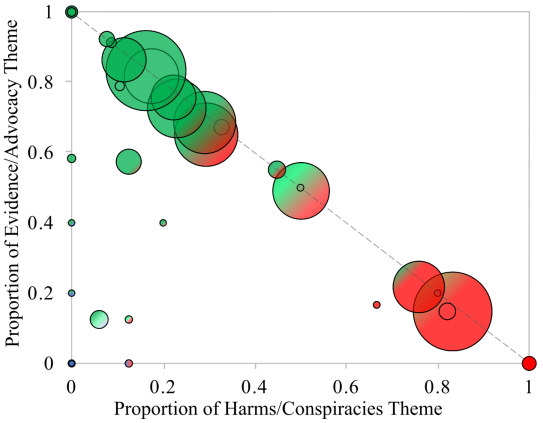
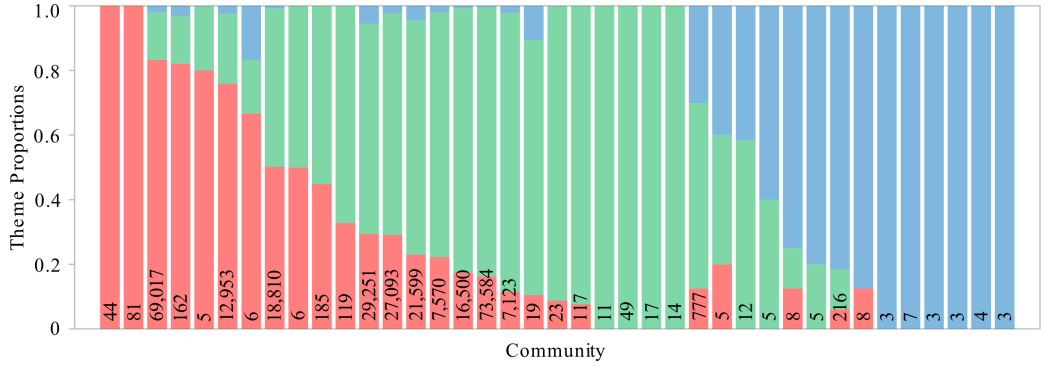


Figure 5. Theme proportions across the 39 communities, representing the proportion of tweets that were assigned to the themes of harms/conspiracies (red), evidence/advocacy (green), and experiential (blue). Values along the horizontal axis are the total number of tweets in the community.



Discussion

Principal Results

In this study, we sought to measure the alignment between the community structure implicit among the follower network of Twitter users posting tweets about HPV vaccines and the topics about which they posted. Given what is already known about the variable quality of information about HPV vaccines on Twitter [22,23], we expected to find that some communities would more often perpetuate negative opinions about HPV vaccines and that these communities would be distinct from the communities describing the favorable evidence or advocating for its uptake. Using a statistical measure quantifying the strength of the topic concentration, we found that some topics were heavily concentrated within a small number of communities, which was consistent with our expectations. Compared to our previous work in this application domain [22,23], the process described here provides a more nuanced view of the specific concerns about HPV vaccines expressed on social media and the ability to identify communities in which these concerns were the predominant topics. Using the combination of topic modeling and community structure to characterize communities, we were able to identify communities in which specific concerns about safety or politics were predominant, as well as identify younger Twitter users who posted experiential tweets and were at risk of greater exposure to safety concerns than to evidence and advocacy, which may occur between the first and subsequent doses of the vaccine. Analysis of tweets for public health where opinions and experiences are mixed have been investigated previously for influenza, where some tweets may help identify influenza incidence and others represent evidence dissemination or opinion [21].

Comparison With Prior Work

A growing set of methods has been developed using either structural information to improve inferences about the content of a corpus, or the information characterizing the nodes in a network to improve the analysis of the structure. Those aiming to understand the content of tweets have used social connections to improve tweet classification [22,63-65]. These studies have considered mentions, retweets, and other forms of interaction that are available on Twitter, but the use of information about followers generally produced the highest levels of performance. Other researchers have proposed methods for incorporating network structure into topic modeling approaches in networks other than Twitter [66,67]. Conversely, some studies have considered the use of content associated with nodes in networks to improve the quality of community detection [68,69]. Among the studies examining documents and the structure between them—such as emails [70], co-authorship [71], and Wikipedia [72]—one study produced topic profiles for communities in a similar fashion to the way we have done in Figure 3 [73]. The approach we presented here differs from these studies because we applied community detection and topic modeling independently, rather than attempting to leverage the information available about social connections to improve the quality of the

topic modeling process, or to use content information to improve community structure or predict new connections.

Limitations

A limitation of this work is that we considered a single application domain. While the uptake of vaccines is of critical importance to public health, further testing on other application domains would be required to determine the generalizability of assessing topic concentrations as a way of characterizing Twitter communities. A further limitation in our work is that we did not consider the temporal dynamics of the topics or the community structure in any detail. Given that the topics related to HPV vaccines are likely to produce similar temporal patterns to those observed by Leskovec et al [30], future work in this area may benefit from further analysis of the relationship between the temporal dynamics of the topics and the economy of attention within communities, which has been explored elsewhere [28,29]. Finally, we considered the follower network as an undirected network and did not incorporate weights or directionality based on mutual followers, or the presence of retweets and mentions, which would have provided a more nuanced representation of the social connections and may have produced a different community structure.

Future Directions

Our work here has potential implications for public health practices. Applying topic modeling and community detection methods in concert to a corpus of tweets about HPV vaccines, we found that it was possible to characterize online communities by the topics that are most heavily concentrated among their tweets. One way of translating these methods into public health practice would be to use these methods in combination with new spatial and demographic estimation methods [74-76], to produce spatiotemporal indicators that determine where and when the growth of specific concerns may lead to increased vaccine hesitancy or refusal. We think that these indicators may have a future role in helping public health organizations design interventions and communication strategies that are better targeted and thus more efficient.

Conclusions

In this work, we demonstrated a novel process for characterizing the concentration of certain opinions in online communities by independently applying existing community detection and topic modeling methods, and quantifying the differences in the topic distributions across communities. Among tweets about HPV vaccines, we found that there were clear differences in the distribution of topics across communities defined by the follower network. In practice, public health organizations may wish to consider identifying the locations and demographics of communities that are at risk of exposure to antivaccine information in order to intervene with positive messages targeting the specific concerns identified through topic modeling. The value of this work in public health includes a more nuanced representation of the variety of concerns expressed about HPV vaccines online and some practical steps towards the development of an automated system for the surveillance of public opinions with the purpose of understanding localized differences in decision making and health behaviors.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Network construction, formal specification/notations, and detailed experiment results.

[PDF File (Adobe PDF File), 981KB-Multimedia Appendix 1]

References

1. Forman D, de Martel C, Lacey CJ, Soerjomataram I, Lortet-Tieulent J, Bruni L, et al. Global burden of human papillomavirus and related diseases. Vaccine 2012 Nov 20;30 Suppl 5:F12-F23. [doi: 10.1016/j.vaccine.2012.07.055] [Medline: 23199955]
2. Crowe E, Pandeya N, Brotherton JML, Dobson AJ, Kisely S, Lambert SB, et al. Effectiveness of quadrivalent human papillomavirus vaccine for the prevention of cervical abnormalities: case-control study nested within a population based screening programme in Australia. BMJ 2014;348:g1458 [FREE Full text] [Medline: 24594809]
3. Skinner SR, Szarewski A, Romanowski B, Garland SM, Lazcano-Ponce E, Salmerón J, et al. Efficacy, safety, and immunogenicity of the human papillomavirus 16/18 AS04-adjuvanted vaccine in women older than 25 years: 4-year interim follow-up of the phase 3, double-blind, randomised controlled VIVIANE study. Lancet 2014 Dec 20;384(9961):2213-2227. [doi: 10.1016/S0140-6736(14)60920-X] [Medline: 25189358]
4. Tabrizi SN, Brotherton JML, Kaldor JM, Skinner SR, Liu B, Bateson D, et al. Assessment of herd immunity and cross-protection after a human papillomavirus vaccination programme in Australia: a repeat cross-sectional study. Lancet Infect Dis 2014 Oct;14(10):958-966. [doi: 10.1016/S1473-3099(14)70841-2] [Medline: 25107680]
5. Brotherton JML, Fridman M, May CL, Chappell G, Saville AM, Gertig DM. Early effect of the HPV vaccination programme on cervical abnormalities in Victoria, Australia: an ecological study. Lancet 2011 Jun 18;377(9783):2085-2092. [doi: 10.1016/S0140-6736(11)60551-5] [Medline: 21684381]
6. Madden K, Nan X, Briones R, Waks L. Sorting through search results: a content analysis of HPV vaccine information online. Vaccine 2012 May 28;30(25):3741-3746. [doi: 10.1016/j.vaccine.2011.10.025] [Medline: 22019758]

7. Robbins SCC, Pang C, Leask J. Australian newspaper coverage of human papillomavirus vaccination, October 2006-December 2009. J Health Commun 2012;17(2):149-159. [doi: 10.1080/10810730.2011.585700] [Medline: 22136302]
8. Printz C. HPV vaccination rates remain low: cancer prevention community needs to continue promoting the vaccine's safety and efficacy, say experts. Cancer 2015 May 1;121(9):1341-1343 [FREE Full text] [doi: 10.1002/cncr.28993] [Medline: 25902756]
9. Mason BW, Donnelly PD. Impact of a local newspaper campaign on the uptake of the measles mumps and rubella vaccine. J Epidemiol Community Health 2000 Jun;54(6):473-474 [FREE Full text] [Medline: 10818125]
10. Hoffman SJ, Tan C. Following celebrities' medical advice: meta-narrative analysis. BMJ 2013 Dec 17;347(dec17 14):f7151. [doi: 10.1136/bmj.f7151]
11. Betsch C, Renkewitz F, Betsch T, Ulshöfer C. The influence of vaccine-critical websites on perceiving vaccination risks. J Health Psychol 2010 Apr;15(3):446-455. [doi: 10.1177/1359105309353647] [Medline: 20348365]
12. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 2011;6(5):e19467 [FREE Full text] [doi: 10.1371/journal.pone.0019467] [Medline: 21573238]
13. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr 2014;6:- [FREE Full text] [doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117] [Medline: 25642377]
14. Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. In: 1st Workshop on Social Media Analytics. USA: ACM Press; 2010 Presented at: The 1st Workshop on Social Media Analytics; July 25, 2010; Washington, DC p. 115-122 URL: http://snap.stanford.edu/soma2010/papers/soma2010_16.pdf[WebCite Cache ID 6jyz7IJSE] [doi: 10.1145/1964858.1964874]
15. Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q, Vespignani A. The Twitter of Babel: mapping world languages through microblogging platforms. PLoS One 2013;8(4):e61981 [FREE Full text] [doi: 10.1371/journal.pone.0061981] [Medline: 23637940]
16. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS One 2011;6(12):e26752 [FREE Full text] [doi: 10.1371/journal.pone.0026752] [Medline: 22163266]
17. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci 2015 Feb;26(2):159-169 [FREE Full text] [doi: 10.1177/0956797614557867] [Medline: 25605707]
18. Prier KW, Smith MS, Giraud-Carrier C, Hanson CL. Identifying health-related topics on Twitter: an exploration of tobacco-related tweets as a test topic. In: Proceedings of the 4th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction. Heidelberg: Springer-Verlag; 2011 Presented at: The 4th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction; March 29-31, 2011; College Park, Maryland p. 18-25 URL: http://mail.smithworx.com/publications/SBP11.pdf[WebCite Cache ID 6jzLj57N1]
19. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: Twitter and antibiotics. Am J Infect Control 2010 Apr;38(3):182-188 [FREE Full text] [doi: 10.1016/j.ajic.2009.11.004] [Medline: 20347636]
20. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011 Oct;7(10):e1002199 [FREE Full text] [doi: 10.1371/journal.pcbi.1002199] [Medline: 22022249]
21. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010;5(11):e14118 [FREE Full text] [doi: 10.1371/journal.pone.0014118] [Medline: 21124761]
22. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong M, Dunn AG. Using social connection information to improve opinion mining: identifying negative sentiment about HPV vaccines on Twitter. Stud Health Technol Inform 2015;216:761-765. [Medline: 26262154]
23. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations between exposure to and expression of negative opinions about Human Papillomavirus vaccines on social media: an observational study. J Med Internet Res 2015;17(6):e144 [FREE Full text] [doi: 10.2196/jmir.4343] [Medline: 26063290]
24. Mahoney LM, Tang T, Ji K, Ulrich-Schad J. The digital distribution of public health news surrounding the Human Papillomavirus Vaccination: a longitudinal infodemiology study. JMIR Public Health Surveill 2015;1(1):e2. [doi: 10.2196/publichealth.3310] [Medline: 27227125]
25. Oliver JE, Wood T. Medical conspiracy theories and health behaviors in the United States. JAMA Intern Med 2014 May;174(5):817-818. [doi: 10.1001/jamainternmed.2014.190] [Medline: 24638266]
26. Romero DM, Tan C, Ugander J. On the interplay between social and topical structure. In: 7th International Conference on Weblogs and Social Media. Palo Alto, California: AAAI Press; 2013 Presented at: The 7th International Conference on Weblogs and Social Media; July 8-11, 2013; Cambridge, Massachusetts p. 516-525 URL: http://www.cs.cornell.edu/~chenhao/pub/social-topical-structure.pdf[WebCite Cache ID 6jyzAnVnX]
27. Weng L, Menczer F, Ahn Y. Virality prediction and community structure in social networks. Sci Rep 2013;3:2522 [FREE Full text] [doi: 10.1038/srep02522] [Medline: 23982106]

28. Weng L, Flammini A, Vespignani A, Menczer F. Competition among memes in a world with limited attention. Sci Rep 2012;2:335 [FREE Full text] [doi: 10.1038/srep00335] [Medline: 22461971]
29. Lehmann J, Gonc¸alves B, Ramasco JJ, Cattuto C. Dynamical classes of collective attention in Twitter. In: 21st International Conference on World Wide Web. USA: ACM; 2012 Presented at: The 21st International Conference on World Wide Web; April 16-20, 2012; Lyon, France p. 251-260. [doi: 10.1145/2187836.2187871]
30. Leskovec J, Backstrom L, Kleinberg J. Meme-tracking and the dynamics of the news cycle. In: 15th ACM International Conference on Knowledge Discovery and Data Mining. USA: ACM; 2009 Presented at: 15th ACM International Conference on Knowledge Discovery and Data Mining; June 28-July 1, 2009; Paris, France p. 497-506 URL: https://cs.stanford.edu/people/jure/pubs/quotes-kdd09.pdf[WebCite Cache ID 6jz03c62k] [doi: 10.1145/1557019.1557077]
31. Backstrom L, Huttenlocher D, Kleinberg J, Lan X. Group formation in large social networks: Membership, growth, and evolution. In: 12th ACM International Conference on Knowledge Discovery and Data Mining.: ACM; 2006 Presented at: The 12th ACM International Conference on Knowledge Discovery and Data Mining; August 20-23, 2006; Philadelphia, USA p. 44-54 URL: http://www.cs.cornell.edu/~lars/kdd06-comm.pdf[WebCite Cache ID 6jz066dOT] [doi: 10.1145/1150402.1150412]
32. Blei DM. Probabilistic topic models. Commun. ACM 2012 Apr 01;55(4):77. [doi: 10.1145/2133806.2133826]
33. Mehrotra R, Sanner S, Buntine W, Xie L. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: 36th International Conference on Research and development in Information Retrieval. USA: ACM; 2013 Presented at: 36th International Conference on Research and development in Information Retrieval; July 28-August 1, 2013; Dublin, Ireland p. 889-892 URL: http://users.cecs.anu.edu.au/~ssanner/Papers/sigir13.pdf[WebCite Cache ID 6jz08zvPi] [doi: 10.1145/2484028.2484166]
34. Weng J, Lim EP, Jiang J, He Q. Twitterrank: finding topic-sensitive influential twitterers. In: 3rd ACM International Conference on Web Search and Data Mining. USA: ACM; 2010 Presented at: 3rd ACM International Conference on Web Search and Data Mining; 2010; New York, USA p. 261-270 URL: http://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1503&context=sis_research[WebCite Cache ID 6jz0CEI6Q] [doi: 10.1145/1718487.1718520]
35. Balasubramanyan R, Kolcz A. “w00t! feeling great today!” chatter in Twitter: identification and prevalence. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. USA: ACM; 2013 Presented at: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining; August 25-29, 2013; Niagara, Canada p. 312-316. [doi: 10.1145/2492517.2492611]
36. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, et al. Comparing Twitter and traditional media using topic models. In: 33rd European Conference on Advances in Information Retrieval. Heidelberg: Springer-Verlag; 2011 Presented at: 33rd European Conference on Advances in Information Retrieval; April 18-21, 2011; Dublin, Ireland p. 338-349 URL: http://www.mysmu.edu/faculty/jingjiang/papers/ECIR'11.pdf[WebCite Cache ID 6jz0FUxJj] [doi: 10.1007/978-3-642-20161-5_34]
37. Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models. In: 4th International AAAI Conference on Weblogs and Social Media. USA: AAAI Press; 2010 Presented at: 4th International AAAI Conference on Weblogs and Social Media; May 23-26, 2010; Washington, DC p. 130-137 URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/download/1528/1846[WebCite Cache ID 6jz0IYjJQ]
38. Yang S, Kolcz A, Schlaikjer A, Gupta P. Large-scale high-precision topic modeling on Twitter. In: 20th ACM International Conference on Knowledge Discovery and Data Mining. USA: ACM; 2014 Presented at: 20th ACM International Conference on Knowledge Discovery and Data Mining; August 24-27, 2014; New York, USA p. 1907-1916 URL: http://cobweb.cs.uga.edu/~squinn/mmd_s15/papers/p1907-yang.pdf[WebCite Cache ID 6jzMST8PR] [doi: 10.1145/2623330.2623336]
39. Yin J, Wang J. A Dirichlet Multinomial Mixture model-based approach for short text clustering. In: 20th ACM International Conference on Knowledge Discovery and Data Mining. USA: ACM; 2014 Presented at: 20th ACM International Conference on Knowledge Discovery and Data Mining; August 24-27, 2014; New York, USA p. 233-242 URL: http://dbgroup.cs.tsinghua.edu.cn/wangjy/papers/KDD14-GSDMM.pdf[WebCite Cache ID 6jz0NhILQ] [doi: 10.1145/2623330.2623715]
40. Nguyen DQ, Billingsley R, Du L, Johnson M. Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics. 2015. p. 299-313 URL: https://transacl.org/ojs/index.php/tacl/article/viewFile/582/132[WebCite Cache ID 6jz0QYY58]
41. Porter MF. An algorithm for suffix stripping. In: Readings in Information Retrieval (Morgan Kaufmann Series in Multimedia Information and Systems). San Francisco, USA: Morgan Kaufmann; 1997:313-316.
42. Jian T, Zhaoshi M, Xuanlong N, Qiaozhu Mei MZ. Understanding the limiting factors of topic modeling via posterior contraction analysis. In: 31st International Conference on Machine Learning.: International Machine Learning Society (IMLS); 2014 Presented at: 31st International Conference on Machine Learning; June 21-26, 2014; Beijing, China p. 190-198 URL: http://jmlr.org/proceedings/papers/v32/tang14.pdf[WebCite Cache ID 6jz0T9bdl]
43. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci U S A 2002 Jun 11;99(12):7821-7826 [FREE Full text] [doi: 10.1073/pnas.122653799] [Medline: 12060727]
44. Newman MEJ. Detecting community structure in networks. The European Physical Journal B - Condensed Matter 2004 Mar 1;38(2):321-330. [doi: 10.1140/epjb/e2004-00124-y]
45. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A 2006 Jun 6;103(23):8577-8582 [FREE Full text] [doi: 10.1073/pnas.0601602103] [Medline: 16723398]

46. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW. Statistical properties of community structure in large social and information networks. In: 17th International Conference on World Wide Web. USA: ACM; 2008 Presented at: 17th International Conference on World Wide Web; April 21-25, 2008; Beijing, China p. 695-704 URL: https://cs.stanford.edu/people/jure/pubs/ncp-www08.pdf[WebCite Cache ID 6jz18eFhf] [doi: 10.1145/1367497.1367591]
47. Fortunato S. Community detection in graphs. Physics Reports 2010 Feb 17;486(3-5):75-174. [doi: 10.1016/j.physrep.2009.11.002]
48. Ahn Y, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature 2010 Aug 5;466(7307):761-764. [doi: 10.1038/nature09182] [Medline: 20562860]
49. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci U S A 2008 Jan 29;105(4):1118-1123 [FREE Full text] [doi: 10.1073/pnas.0706851105] [Medline: 18216267]
50. Python igraph (version 0.7.0). URL: http://igraph.org/python/ [accessed 2016-05-26] [WebCite Cache ID 6hmWgjDT3]
51. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech 2008 Oct 09;2008(10):P10008. [doi: 10.1088/1742-5468/2008/10/P10008]
52. Edler D, Rosvall M. Source code for multilevel community detection with Infomap (version Sep 07). 2015. URL: http://www.mapequation.org/code.html [accessed 2016-05-26] [WebCite Cache ID 6hmXBQpp5]
53. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003;3:993-1022.
54. Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine Learning - Special Issue on Information Retrieval 2000;39(2-3):103-134 [FREE Full text] [doi: 10.1023/A:1007692713085]
55. Lu Y, Mei Q, Zhai C. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Inf Retrieval 2010 Aug 5;14(2):178-203. [doi: 10.1007/s10791-010-9141-9]
56. Řehůřek R. Software framework for topic modelling with large corpora. In: LREC 2010 Workshop on New Challenges for NLP Frameworks. Malta: ELRA; 2010 Presented at: LREC 2010 Workshop on New Challenges for NLP Frameworks; May 22, 2010; Valletta, Malta p. 45-50 URL: https://radimrehurek.com/gensim/lrec2010_final.pdf[WebCite Cache ID 6jz2QxAvb]
57. Nguyen DQ. jLDADMM: A Java package for the LDA and DMM topic models (version 1.0, 2015-07-06). 2015. URL: http://jldadmm.sourceforge.net/ [accessed 2016-05-26] [WebCite Cache ID 6hmXKnvie]
58. Yan X, Guo J, Lan Y, Cheng X. A biterm topic model for short texts. In: 22nd International Conference on World Wide Web. USA: ACM; 2013 Presented at: 22nd International Conference on World Wide Web; May 13-17, 2013; Rio de Janeiro, Brazil p. 1445-1456. [doi: 10.1145/2488388.2488514]
59. Dunn AG, Westbrook JI. Interpreting social network metrics in healthcare organisations: a review and guide to validating small networks. Soc Sci Med 2011 Apr;72(7):1064-1068. [doi: 10.1016/j.socscimed.2011.01.029] [Medline: 21371798]
60. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964 Mar;29(1):1-27. [doi: 10.1007/BF02289565]
61. Cox TF, Cox MAA. Multidimensional scaling. In: Handbook of Data Visualization. Heidelberg: Springer-Verlag; 2008:315-347.
62. Borg I, Groenen PJF. Modern multidimensional scaling: Theory and applications. New York: Springer-Verlag; 2005.
63. Tan C, Lee L, Tang J, Jiang L, Zhou M, Li P. User-level sentiment analysis incorporating social networks. In: 17th ACM International Conference on Knowledge Discovery and Data Mining. USA: ACM; 2011 Presented at: 17th ACM International Conference on Knowledge Discovery and Data Mining; August 21-24, 2011; San Diego, CA p. 1397-1405. [doi: 10.1145/2020408.2020614]
64. Hu X, Tang L, Tang J, Liu H. Exploiting social relations for sentiment analysis in microblogging. In: 6th ACM International Conference on Web Search and Data Mining. USA: ACM; 2013 Presented at: 6th ACM International Conference on Web Search and Data Mining; February 4-8, 2013; Rome, Italy p. 537-546 URL: http://faculty.cs.tamu.edu/xiahu/papers/wsdm13Hu.pdf[WebCite Cache ID 6jz1POZoe] [doi: 10.1145/2433396.2433465]
65. Speriosu M, Sudan N, Upadhyay S, Baldridge J. Twitter polarity classification with label propagation over lexical links and the follower graph. In: Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics; 2011 Presented at: Conference on Empirical Methods in Natural Language Processing; July 27-31, 2011; Edinburgh, Scotland p. 53-63 URL: http://anthology.aclweb.org/W/W11/W11-2207.pdf[WebCite Cache ID 6jz1RI5q5]
66. Mei Q, Cai D, Zhang D, Zhai CX. Topic modeling with network regularization. In: 17th International Conference on World Wide Web. USA: ACM; 2008 Presented at: 17th International Conference on World Wide Web; April 21-25, 2008; Beijing, China p. 101-110 URL: http://www-personal.umich.edu/~qmei/pub/www08-netplsa.pdf[WebCite Cache ID 6jz1e33Vf] [doi: 10.1145/1367497.1367512]
67. Sun YZ, Han JW, Gao J, Yu YT. iTopicModel: Information network-integrated topic modeling. In: 9th IEEE International Conference on Data Mining.: IEEE; 2009 Presented at: 9th IEEE International Conference on Data Mining; December 6-9, 2009; Miami, FL p. 493-502. [doi: 10.1109/ICDM.2009.43]
68. McAuley J, Leskovec J. Learning to discover social circles in ego networks. In: Advances in Neural Information Processing Systems 25. 2013 Presented at: 26th Annual Conference on Neural Information Processing Systems 2012; December 3-8,

2012; Nevada, USA p. 548-556 URL: http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2012_0272.pdf[WebCite Cache ID 6jz1n4wak]
69. Yang J, McAuley J, Leskovec J. Community detection in networks with node attributes. In: 13th IEEE International Conference on Data Mining. 2013 Presented at: 13th IEEE International Conference on Data Mining; December 7-10, 2013; Dallas, TX p. 1151-1156 URL: https://cs.stanford.edu/people/jure/pubs/cesna-icdm13.pdf[WebCite Cache ID 6jz1pWpAH] [doi: 10.1109/ICDM.2013.167]
70. Zhou D, Manavoglu E, Li J, Giles CL, Zha H. Probabilistic models for discovering e-communities. In: 15th International World Wide Web Conference. USA: ACM; 2006 Presented at: 15th International World Wide Web Conference; May 22-26, 2006; Edinburgh, Scotland p. 173-182. [doi: 10.1145/1135777.1135807]
71. Chang J, Blei DM. Relational topic models for document networks. In: 12th International Conference on Artificial Intelligence and Statistics. 2009 Presented at: 12th International Conference on Artificial Intelligence and Statistics; April 16-18, 2009; Florida, USA p. 81-88.
72. Chang J, Blei DM. Connections between the lines: Augmenting social networks with text. In: 15th International Conference on Knowledge Discovery and Data Mining. USA: ACM; 2009 Presented at: 15th International Conference on Knowledge Discovery and Data Mining; June 28-July 1, 2009; Paris, France p. 169-178. [doi: 10.1145/1557019.1557044]
73. Pathak N, DeLong C, Erickson K, Banerjee A. Social topic models for community extraction. In: 2nd SNA-KDD Workshop. 2008 Presented at: 2nd SNA-KDD Workshop; August 24, 2008; Las Vegas, NV URL: http://www-users.cs.umn.edu/~banerjee/papers/08/snakdd08.pdf[WebCite Cache ID 6jz1tbTTy]
74. Burger JD, Henderson J, Kim G, Zarrella G. Discriminating gender on Twitter. In: Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics; 2011 Presented at: Conference on Empirical Methods in Natural Language Processing; July 27-31, 2011; Edinburgh, Scotland p. 1301-1309.
75. Jurgens D, Finnethy T, McCorriston J, Xu YT, Ruths D. Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In: 9th International AAAI Conference on Web and Social Media. California: The AAAI Press; 2015 Presented at: 9th International AAAI Conference on Web and Social Media; May 26-29, 2015; Palo Alto, CA p. 188-197 URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10584/10502[WebCite Cache ID 6jz1xM3Dd]
76. Jurgens D, Allen D. Geotagging one hundred million Twitter accounts with total variation minimization. In: IEEE International Conference on Big Data. USA: IEEE; 2014 Presented at: IEEE International Conference on Big Data; October 27-30, 2014; Washington, DC p. 393-401.

Abbreviations

ARI: adjusted Rand index
DMM: Dirichlet Multinomial Mixture
HPV: human papillomavirus
LDA: Latent Dirichlet Allocation
NMI: normalized mutual information

Edited by G Eysenbach; submitted 30.05.16; peer-reviewed by D Arachi, A MacKinlay; comments to author 20.07.16; accepted 03.08.16; published 29.08.16

Please cite as:
Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG
Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection
J Med Internet Res 2016;18(8):e232
URL: http://www.jmir.org/2016/8/e232/
doi: 10.2196/jmir.6045
PMID: 27573910