MEDINFO 2013
C.U. Lehmann et al. (Eds.)
© 2013 IMIA and IOS Press.

This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-289-9-807

Finding Meaning in Social Media: Content-based Social Network Analysis of QuitNet to Identify New Opportunities for Health Promotion

Sahiti Myneni[a], Nathan K. Cobb[b], Trevor Cohen[a]

[a] *National Center for Cognitive Informatics and Decision Making in Healthcare*
University of Texas School of Biomedical Informatics at Houston, TX, USA,
[b] *The Schroeder Institute for Tobacco Research and Policy Studies, Washington, DC, USA*

Abstract

Unhealthy behaviors increase individual health risks and are a socioeconomic burden. Harnessing social influence is perceived as fundamental for interventions to influence health-related behaviors. However, the mechanisms through which social influence occurs are poorly understood. Online social networks provide the opportunity to understand these mechanisms as they digitally archive communication between members. In this paper, we present a methodology for content-based social network analysis, combining qualitative coding, automated text analysis, and formal network analysis such that network structure is determined by the content of messages exchanged between members. We apply this approach to characterize the communication between members of QuitNet, an online social network for smoking cessation. Results indicate that the method identifies meaningful theme-based social sub-networks. Modeling social network data using this method can provide us with theme-specific insights such as the identities of opinion leaders and sub-community clusters. Implications for design of targeted social interventions are discussed.

Keywords:

Online social networks, content analysis, behavior change, smoking cessation.

Introduction

Smoking-related illnesses claim 440,000 lives a year and remain the leading preventable cause of death in the United States [1]. While most smokers wish to quit, success rates for cessation attempts remain low and tobacco use is considered a chronic, relapsing condition [2]. The impact of social relationships on health behaviors is well documented [3]. However, the mechanisms through which social connections might influence behavior change are currently poorly understood and further research is needed to inform the development of high-impact scalable intervention programs [4].

Background

With the ascendance of electronic and mobile health applications, the use of online social networks is increasing [5]. On account of their accessibility, these networks consist of users from all around the globe providing an international perspective on health-related issues. Also, online social networks provide a unique opportunity to understand the relationship between social interactions and behavior change, as communication among network members is electronically documented. To this end, several studies have investigated social network dynamics to deepen understanding of behavioral diffusion and the flow of ideas [6-8]. In the majority of cases, connections in a social network are characterized by the frequency of communication events between social actors, without considering the content of this communication. However, behavior modification techniques based on the existing behavior change theories emphasize the importance of communication content [3]. Therefore, we posit that content analysis of health-related online social networks can enable the development of tailored behavior support systems that better harness social influence.

Previous studies that considered online social network content examined the nature of social support using qualitative methods [9]. Recent advances in automated text analysis allow for large-scale analysis of the content of communication between members. Semantic space models have been utilized to classify health information on webpages using metadata attributes [10]. Methods of distributional semantics such as Latent Semantic Analysis (LSA) derive relatedness measures between terms from unannotated text by representing the terms in a high-dimensional vector space [11, 12]. Evidence suggests that the semantic relatedness measures derived using these techniques agrees with human estimates, and can be used to obtain human-like performance in a number of cognitive tasks [12].

In this paper, we present a new methodology to explore online social network data based on communication content that combines a distributional semantics approach with qualitative coding guided by grounded theory [13]. The method is applied to the analysis of messages exchanged by users of QuitNet, an online community for smoking cessation. QuitNet is one of the first online social networks for health behavior change and has been in existence for 14 years. It is widely used with over 100,000 new registrants per year. Previous studies of QuitNet indicated that participation was strongly correlated with abstinence [14]. Communication among QuitNet members can occur through private email, public forums, and chat rooms.

Materials and Methods

A database of 16,492 de-identified public messages between March 1, 2007 and April 30, 2007 was used in our study. These messages were analyzed using a hybrid method that enables the analysis of voluminous and context-rich social network data while providing a microscopic, scalable, and interpretable view of the data.

First, communication themes in QuitNet data were derived using grounded theory techniques. Using automated text analysis methods, the relatedness of each message to each of the themes was calculated. These measures were used to generate content-based social networks, by representing QuitNet users as nodes, and their communication as edges. Using statistically-determined theme-specific threshold values, meaning-based edges were retained such that QuitNet members were connected based on the thematic content of their exchanges. The resulting networks were then analyzed using traditional network analysis methods to understand content-specific network patterns. In the sections that follow, we discuss each of these steps in greater detail.

Qualitative analysis

100 out of the 15,050 messages were randomly selected and analyzed qualitatively using grounded theory techniques [13] to identify relevant socio-behavioral themes. The first step in the coding process involved *open-coding* where a line-by-line analysis was performed on the messages to derive the conceptual codes from the data. Consistency of code assignment was ascertained using *constant comparison* where instances of each code were compared in an iterative manner to make sure they reflected the same concept. Subsequently, *axial coding* was performed by re-organizing open codes into themes. Two researchers independently coded the data using the terminology developed above and the codes they assigned to the messages had a Cohen's Kappa measure of 81.6%. Disagreement was resolved through discussion and 12 of 16 discrepancies were attributable to messages related to more than one code.

Automated text analysis

Based on an initial analysis, including an attempt at automated theme discovery using unsupervised methods, we concluded that the distributional information in our QuitNet corpus was insufficient for the automated derivation of meaningful measures of semantic relatedness between terms. Therefore, we drew on distributional information from the Touchstone Applied Science Associated (TASA) corpus, a collection of 37,657 articles designed to approximate the average reading of an American college freshman. This corpus has been widely used in distributional semantics research, and when applied to this corpus LSA has been shown to approximate human performance on a number of cognitive tasks [12]. LSA was performed using the Semantic Vectors package [15], an open source package for distributional semantics. The log-entropy weighting metric was used, and terms occurring on the stopword list distributed with the General Text Parser software package [16] were ignored. This stopword list consists of frequently occurring terms that carry little semantic content. Figure 1 depicts the vector generation process that we utilized to ensure that terms present in the QuitNet corpus, but not in the TASA corpus, would obtain meaningful vector representations. Beginning with the term vectors produced by LSA (TASA term vectors), vector representations of all messages in the QuitNet corpus were derived by adding the vectors of the terms they contain (TASA based QuitNet message vectors). Representations for terms in the QuitNet corpus were then generated by adding the message vectors for each message they occurred in, and normalizing the resulting vector (QuitNet term vectors). Subsequently, a second set of message vectors was generated (QuitNet message vectors). This approach, which is similar in nature to the reflective approach [17] that we utilized previously to infer associations between terms that do not co-occur directly, resulted in intuitively interpretable term representations (see Table 1). A "pair vector" was constructed for each communicating pair of users, by adding the message vector for every message they exchanged, and normalizing the resulting vector. A "theme vector" was constructed for each identified theme by adding the vectors for terms representing this theme (Figure 1). The semantic relatedness between a pair of users and a given theme was measured as the cosine of the angle between their pair vector and the vector representing this theme.

Table 1- Nearest neighbors of terms "craving" and "depression" Underlined terms illustrate indirect inference. Association strength is measured with the cosine metric

Craving	Depression
0.901:cigarette	0.727:depressed
0.890:nicotine	0.696:horrific
0.862:crave	0.688:adjusts
0.856:craves	0.669:clinical
0.854:smoker	0.669:requiring
0.849:habit	0.656:emotions
0.847:chantix	0.645:stress
0.841:cig	0.630:emotional

Theme-based social network analysis

A network model of the QuitNet data was created by representing users as nodes, and their communication as edges. A statistical threshold for each theme was determined by considering the relatedness between selected representative terms and each of the 6,928 pair vectors. As the distribution of these relationship scores was normal, a threshold value of the mean + one standard deviation was used. Gephi, an open-source network analysis and visualization software package [18] was used to visualize and analyze network models for each theme. Differences in network structures across themes were examined using social network metrics [19], *degree* and *modularity*. Degree measures the number of links to and from a network member. Modularity is defined as the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random, and therefore can be considered as a measure of the cohesiveness of communities within the network.

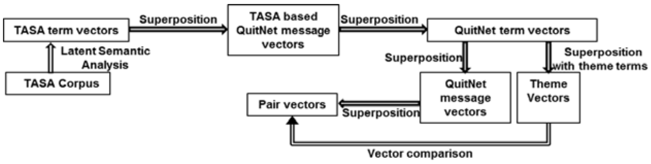

Figure 1- Vector generation sequence for automated analysis of QuitNet

Results

Qualitative analysis

A total of nine themes were discovered, of which six were predominant. An example message for each of the themes is presented in Table 2. Figure 2 shows the distribution of the coded messages across themes. Some message content revealed QuitNet-specific customs and behaviors. Consider the "Social Togetherness"-related message, through which a member affirms his/her decision about not smoking and "lends a hand" to the next member who in turn reciprocates and perpetuates the bond. This sort of serial affirmation is a tradition that has emerged within the QuitNet community. Similarly, the "Adherence"-related message was sent in the context of a virtual "bonfire" event conducted regularly as a QuitNet activity. Members who quit smoking keep count of the number of their unsmoked cigarettes and burn them all in "bonfire". Other QuitNet members attend this event as witness or to contribute their unsmoked cigarettes to the "bonfire". The virtual "bonfire" may serve multiple purposes: improving adherence, fostering self-efficacy, and community building.

Automated text analysis

Table 3 shows the thresholds and cue terms used to derive vector representation for each theme. For instance, the terms "sad", "emotion" were used to obtain similarity score for "support" theme. The theme vectors and statistical thresholds were evaluated using recall and precision metrics. For each theme a user was randomly chosen. Subsequently, all corresponding users that the selected user exchanged messages with irrespective of the theme were retrieved. A total of 82 messages from 27 unique users were evaluated. These 82 messages were coded to see how many of them belonged to that particular theme (above threshold). Thereby, the recall (messages retrieved/messages discussing theme) and precision (relevant messages retrieved/total messages retrieved) were estimated. On an average, the recall of the system was calculated to be 0.75 and the precision was 0.81.

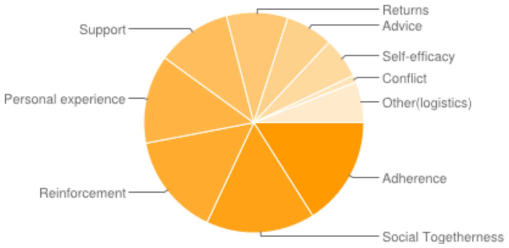

Figure 2- Distribution of QuitNet themes derived using qualitative coding across 100 coded messages

Table 3- Semantic similarity calculation

Theme	Representative terms	Threshold
Personal experience	experience, opinion	0.624
Adherence	pledge, bonfire, milestone	0.731
Advice	hang, help, advice	0.812
Reinforcement	congratulation, congratulations	0.787
Support	sad, emotion	0.820
Returns	cancer, exercise, weight, stats	0.685

Theme-based social network analysis

Figure 3 captures the network structure of the theme-based networks. The topology and structure of these three networks was quite different with respect to density and the high-degree nodes were different across the spectrum. The average degree of these three theme-based networks was 2.911, 2.34, and 3.162 with the average weighted degree being 2.355, 1.965, and 2.088 respectively. The modularity of the networks was 0.64, 0.766, and 0.607, which indicates there were sizable sub-community clusters within the theme-based network. The average path length was 4.475, 5.072, and 4.51 respectively.

Table 2- Illustration of message classification across nine themes

Themes	Definition	Example Message
Personal experience	Content deals with the personal experiences of the sender.	*I had a rough patch somewhere between 5 and 6 months.//I think I had a major turning point at about my 6 month milestone.*
Social Togetherness	Conduct activities to enhance the closeness of a network	*Thank you for your hand. I pledge not to smoke today, not one puff. In the spirit of friendship and support I take your hand in my left, and extend my right hand to the next to join us*
Advice	Information that helps a person in overcoming barriers	*Punch your pillow, yell, do whatever you want. It will help you vent. Close your eyes and take a deep breath.*
Reinforcement	Words of encouragement or aversion that would help a person to stay quit	*I am standing and applauding your accomplishments. What a great quit you have a wonderful attitude. Hoping to be like you one day*
Support	Express concern and/or share a person's emotional feelings	*That was a terrible experience. I think someone owes you an apology. You should not have been treated so badly in my opinion. Good for you for not smoking.*
Conflict	reflect a rift between two members	*For a lurker, you sure do post a lot though. So, one can ask, who is a perfect lurker.*
Adherence	promote compliance towards the decision to cease smoking	*Let's get rid of those 5,733 unsmoked nasties for you so you can start relaxing. Woosh! There they go, never to be seen again*
Self-efficacy	boost confidence levels so that a person can stay quit (or) attempt to quit	*It does not, and I repeat, does NOT have to be a lifetime battle. You CAN change how you think about smoking. You really can.*
Returns	Discussions about incentives of not smoking (quality of life, reduction in health risks, monetary savings).	*I'M SHOUTING AT THE TOP OF MY HEALTHY LUNGS!!* *33 days, 3 hours, 30 minutes smoke free. 663cigarettes not smoked. $280.50 and 5 days, 1 hour of my life saved!*

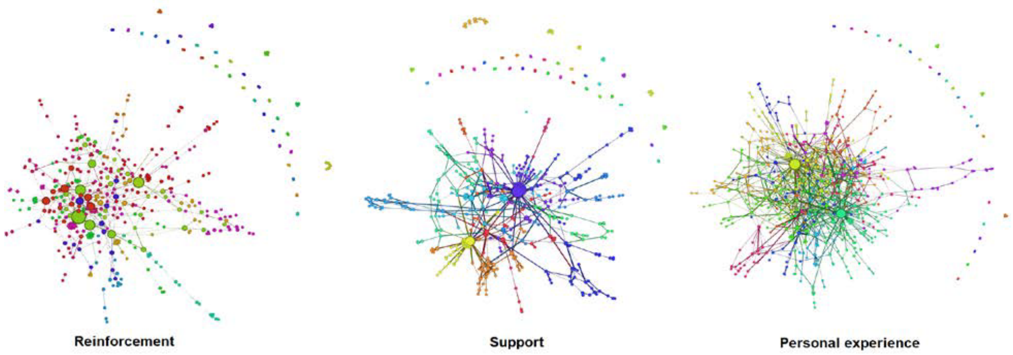

Figure 3- Visualization of "Reinforcement", "Support", and "Personal Experience" themes in QuitNet

Instead when theme-based thresholds were not applied and edges weighted based on communication frequency, the modularity was lower (~3.4) than that for the theme-based networks. Six large communities were found within the network with majority of the nodes concentrated in four modules with considerable crosstalk between them. These sub-communities were not readily intuitive for generating intervention strategies. In contrast, sub-community clusters identified in theme-based networks offered content-specific insights (explained in the next section) that were interpretable and actionable.

Discussion

To the best of our knowledge, this paper documents the first grounded theory-based empirical study of communication in a health-related online social network. Therefore, this analysis adds to our understanding of the mechanisms underlying behavior change in such contexts. Most themes generated using the qualitative coding are consistent with existing behavior change theories [3] such as Social Cognitive Theory, the Transtheoretical Model of Change, and the Health Belief Model. However, the "Conflict" theme is not considered by the predominant behavior change theories. Therefore, this study sheds light on the consistency between existing behavior change theories and emergent mechanisms of behavior change in online social networks.

Frequency-based edge weighting can provide insight into network topology, but this approach ignores content, which is considered to be important in theories of social influence. While using semantic relatedness to weight edges of a network is not without precedent [20], the application of this approach to the analysis of social networks related to health and wellness is novel. The distributional semantics based approach described in this paper has several advantages- (1) it allows for the extension of human-intensive qualitative analysis to larger data sets than would otherwise be possible; (2) it provides a convenient means to derive intuitive interpretations of QuitNet corpus; (3) its capability of indirect inference appears to be of value, given that human beings can express similar concepts in different ways; and (4) it allows for the representation of online social network content in a form that is amenable for adopting existing network analysis methods.

Network models obtained from formal network analysis using Gephi reveal differences in structure across themes, the most striking of which is the difference in the high-degree nodes, which indicate those users with the most connections with whom network members discuss content related to a particular theme. Consequently, these high-degree nodes represent the opinion leaders of the network with respect to specific content. Opinion leaders play an important role in social mobilization and social networks, acting as gatekeepers for interventions that aim to help change social norms, and accelerate behavior change [19, 21]. Previous work on QuitNet also emphasized the role played by opinion leaders as social integrators facilitating network formation [8]. Consequently, the method outlined in this paper has implications for the design of targeted interventions. Such insights into the distribution of key nodes can inform the design of targeted interventions that disseminate specific-theme related information and establish mentor-mentee relationships based on the mentee needs and mentor interests. In addition, large-scale personalized interventions can be delivered to a module of members in accordance with their community's interest in a theme.

Online social networks have attracted a great deal of attention in recent years on account of their potential to generate revenue through targeted advertising and related commercial ventures. However, as demonstrated by QuitNet, these venues can also provide a forum for a community of dedicated users to assist one another in the pursuit of better health, an activity that ultimately has societal benefit beyond the users of QuitNet itself. The development of better tools to analyze social network content of this nature allows us a greater understanding of the ways in which such social networks mediate behavior change, thereby providing us with the opportunity for empirically-grounded interventions to further assist these communities with the attainment of their laudable goals.

Limitations and Future Work

The QuitNet dataset considered in our analysis was recorded in 2007, and is limited in size. For future studies, we will attempt to obtain further data drawn from recent datasets. Given the small proportion of messages thematically coded, additional themes may have not been captured. Increasing the number of messages in the qualitative sample and examining the theoretical validity of the resulting themes is important to improve the generalizability of our results. The precision of our methods of automated text analysis is of particular importance, as low precision may lead to formation of erroneous links in the network, such that the derived theme-based network is not truly representative of the messages exchanged. The accuracy of the system may be further improved by more sophisticated choice of search terms [11]. In addition, the evaluation of the automated methods in this paper employs messages exchanged by a single user in each theme. This limitation will be addressed in our future work by developing an evaluation framework that builds on a large sample of messages drawn from multiple users. The formal social network analysis conducted in this paper was limited to two metrics. A more extensive approach toward network analysis needs to be adopted by taking into account the interplay between individual-level and network-level measures [19].The content-specific network patterns identified have implications for future work, as they suggest the design of behavioral support systems to promote public health and wellness. Application areas include a) identification of theme-specific opinion leaders, clusters, outliers, and b) "rewiring" networks to improve or reduce network cohesion to efficiently channel the spread of information and ideas. These findings can also inform individual lifestyle support systems that can be operationalized in the real-world using mobile health applications.

Conclusion

Health-related behaviors such as smoking contribute to the majority of deaths in United States and around the world. The development of scalable interventions to help people engage in healthy lifestyles is a high-priority task for health researchers and professionals. Online social networks allow us to examine the role of social relationships in health behavior at high granularity. The members of these networks such as QuitNet represent a global community of users, which brings an international perspective to the analysis. In this paper, we describe a method that combines qualitative coding, automated text analysis, and network analysis to provide further insight into the mechanisms underlying behavior change in social networks. Results obtained using this method can inform the design of personalized and targeted interventions that persuade people to initiate or adhere to a positive behavior change, thus setting the stage for a new generation of translational interventions in public health and behavioral science.

Acknowledgements

This study is supported in part by UTHealth Innovation for Cancer Prevention Research Pre-doctoral Fellowship, The University of Texas School of Public Health- Cancer Prevention and Research Institute of Texas grant # RP101503. *Disclaimer:* The content is solely the responsibility of the authors and does not necessarily represent the official views of the Cancer Prevention and Research Institute of Texas.

References

[1] The Health Consequences of Smoking: a report of the Surgeon General. Atlanta: Centers for Disease Control and Prevention, U.S. DHHS;2004.

[2] Chen PH, White HR, Pandina RJ. Predictors of smoking cessation from adolescence into young adulthood. Addictive behaviors. 2001;26(4):517–529.

[3] Heaney CA, Israel BA. Social networks and social support. Health behavior and health education: Theory, research, and practice. 2002;3:185-209.

[4] Smith KP, Christakis NA. Social Networks and Health. Annual Review of Sociology. 2008;34:405–29.

[5] Cobb NK, Graham AL, Byron, MJ, Niaura RS, Abrams DB. Online social networks and smoking cessation: a scientific research agenda. J Med Internet Res. 2011;13(4):e119.

[6] Centola D. The spread of behavior in an online social network experiment. Science. 2010;329(5996):1194

[7] Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. N Engl J Med. 2008; 358(21):2249-58.

[8] Cobb NK, Graham, AL, Abrams, DB. Social network structure of a large online community for smoking cessation. Am J Public Health. 2010;100(7): 1282-9.

[9] Chuang K., Yang C. A study of informational support exchanges in medhelp alcoholism community. Proceedings of SBP. 2012:9-17.

[10]Chen G, Warren J, Riddle P. Semantic Space models for classification of consumer webpages on metadata attributes. J Biomed Inform. 2010;43(5):725-35.

[11]Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. J Biomed Inform. 2009;42(2):390–405.

[12]Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev 1997;104:211-240.

[13]Strauss A, Corbin J. Basics of Qualitative Research: Grounded Theory Procedure and Techniques. Newbury-Park, London. Sage; 1990.

[14]Graham AL, Cobb NK, Raymond L, Sill S, Young J. Effectiveness of an internet-based worksite smoking cessation intervention at 12 months. JOEM. 2007;49(8):821.

[15]Widdows D, Ferraro K. Semantic Vectors: A scalable open source package and online technology Management application. Proceedings of LREC'08.2008:1183-90.

[16]Giles J, Wo L, Berry M. GTP (general text parser) software for text mining. Statistical data mining and knowledge discovery.2003:455–71.

[17]Cohen T, Schvaneveldt R, Widdows D. Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. J Biomed Inform. 2010;43(2):240-256.

[18]Bastian M., Heymann S., Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks; International AAAI Conference on Weblogs and Social Media; California: San Jose; 2009.

[19]Valente TW. Social Networks and Health: Models, Methods, and Applications: Models, Methods, and Applications: Oxford University Press, USA; 2010.

[20]Schvaneveldt RW. Pathfinder associative networks: Studies in knowledge organization: Ablex Publishing.1990.

[21]Rogers E. Diffusion of innovation. New York. Freepress;1983.

Address for correspondence

Sahiti Myneni, MSE
The University of Texas School of Biomedical Informatics at Houston, 7000 Fannin, UCT165, Houston, TX 77030
Sahiti.Myneni@uth.tmc.edu