Characterization of Temporal Semantic Shifts of Peer-to-Peer Communication in a Health-Related Online Community: Implications for Data-driven Health Promotion

Vishnupriya Sridharan, B.Tech[1], Trevor Cohen MBChB, PhD[1], Nathan Cobb, MD[2], Sahiti Myneni, PhD, MSE[1]

[1]The University of Texas School of Biomedical Informatics at Houston, TX, USA

[2]Georgetown University Medical Center, Washington, DC, United States

Abstract:

With online social platforms gaining popularity as venues of behavior change, it is important to understand the ways in which these platforms facilitate peer interactions. In this paper, we characterize temporal trends in user communication through mapping of theoretically-linked semantic content. We used qualitative coding and automated text analysis to assign theoretical techniques to peer interactions in an online community for smoking cessation, subsequently facilitating temporal visualization of the observed techniques. Results indicate manifestation of several behavior change techniques such as 'feedback and monitoring' and 'rewards'. Automated methods yielded reasonable results (F-measure=0.77). Temporal trends among relapsers revealed reduction in communication after a relapse event. This social withdrawal may be attributed to failure guilt after the relapse. Results indicate significant change in thematic categories such as 'social support', 'natural consequences', and 'comparison of outcomes' pre and post relapse. Implications for development of behavioral support technologies that promote long-term abstinence are discussed.

Introduction and Background:

According to CDC statistics of 2015, chronic diseases – such as hypertension, stroke, cancer and diabetes, are responsible for 7 of 10 deaths every year in the United States and treating people with chronic diseases accounts for 86% of all healthcare costs [1]. These chronic conditions are often caused by health-related behaviors such as tobacco smoking, which is "the leading cause of preventable disease and death in the United States, resulting in approximately 480,000 premature deaths and more than $300 billion in direct health care expenditures and productivity losses each year" [2]. Health promotion campaigns emphasize the need to create behavior change avenues through theoretically designed interventions to help people modify this risky health behavior and stay abstinent. Current trends show that online social communities are gaining popularity as behavior modification venues as users of these platforms reach out to their peers and experts for support and guidance irrespective of geographical and demographic boundaries [3]. End-users of these virtual support groups (patients or population at large) have expressed that they trust and rely on the information in social media, and even base their decisions on this information [4]. In addition, using social media peer interactions for behavior analysis has great potential because (a) unlike conventional survey-based or controlled laboratory studies, content on social media is spontaneous and unprompted [3], and (b) electronically captured communication is amenable to large scale text analysis thereby scaling up traditional socio-behavioral analytical methods to social media platforms in the digital era.

Previous research in this field has focused on analysis of the structural characteristics of the online communities [5], development of theory-guided interventions [6], and validation of social support in online communities for behavior change and chronic illness management [7]. Recent studies on social media analysis for smoking cessation have suggested new methodological advances for enhancing user engagement through content-mediated network modeling of peer interactions [8, 9]. Methods of distributional semantics, which learn the relatedness between terms from large electronic text collections, have been used in conjunction with social influence models to characterize content-specific communication patterns underlying behavior change in smoking cessation [10]. Such automated text analysis methods have been used for large scale analysis of content on social networking sites [11-13]. Latent Semantic Analysis (LSA) [14] has been used in conjunction with machine learning techniques to annotate large amount of health information on social media [15-17]. However, computational overhead may pose a problem while using LSA. More scalable distributional semantic methods such as Reflective Random Indexing (RRI) [18] can also learn semantic relatedness from text corpora, including meaningful relations between terms that do not occur together. Semantic space models have been developed to represent text characterizing a specific user's temporal transitions in online communities [19-21]. Further development of automated methods that enable understanding of individual and network level trends may

provide insight into social influence mechanisms in the context of sustained behavior change. Once analyzed, messages exchanged between users provide a *semantic context* for their health-related behavior. We posit that the relationship between this semantic context, behavior change events, and user engagement levels can be understood by modeling user communication over time.

Advances in text analytics can enable us to characterize temporal trends in semantic context underlying peer interactions [21-23]. While the overarching goals of our work are to develop informatics-driven methods to develop resource optimized analytics that account for granularity and scalability, and inform the design of consumer-facing digital health platforms for sustained user engagement and long term behavior change, the specific methodological objectives of this paper are three-fold: (a) to employ a behavior change taxonomy for annotation of peer interactions in QuitNet, an online community for smoking cessation, (b) to utilize automated text analysis to scale the results of qualitative analysis to a large dataset, and (c) to visualize peer interactions in QuitNet and characterize user semantics over time, and across smoking and abstinence behaviors.

Materials and Methods:

QuitNet is one of the first online social networks for health behavior change and has been in continuous existence for the past 16 years [24]. Forum interactions were the primary mode of communication and each forum message has a message id, a thread id (the thread in which the message was exchanged), a sender id and recipient id. For the purpose of this study, we considered two subsets of QuitNet data: 16,492 forum messages exchanged between March-April 2007 (we refer to this as "Dataset 1" from now on) and 65,910 forum messages exchanged from January -December 2014 (we refer to this as "Dataset 2" from now on). Overall, there are 82,402 messages and 2,354 unique users who have exchanged messages during these two time periods. The research reported in this manuscript has been reviewed and exempted by the Institutional Review Board at the University of Texas Health Science Center at Houston.

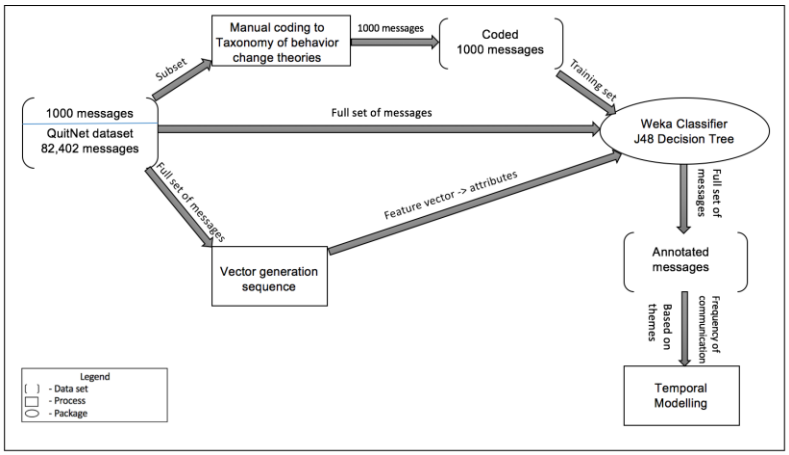

Figure 1. Methodological outline: Qualitative-Automated-Temporal Analysis

Qualitative coding, automated text analysis, and temporal visualization were integrated to conduct the study described in this paper. Figure1 represents the overall methodological details of the study. The first step in this study was the qualitative analysis, guided by an established behavior change taxonomy [25]. 1000 randomly selected messages from the QuitNet Dataset 1 were manually coded into 16 themes to characterize 93 theory-linked behavior change techniques outlined in the taxonomy. In order to validate the generalizability of thematic categorization, scalability of automated methods, and observed semantic shifts before and after relapse in temporal models, we extended our analysis to Dataset 2. The qualitative codes assigned to the subset of 1000 messages in Dataset 1 were then used to annotate the rest of messages in Dataset 1 and entire Dataset 2 using methods from distributional semantics as described in the next section. We then utilized temporal modeling to visualize changes in semantic context over these two time periods (2007 and 2014). On account of its clinical importance, temporal modeling was focused on relapse behavior (self-reported change from ex-smoker to active smoker) during the study period.

Qualitative methods:
We selected 1000 messages randomly from the Dataset 1 using a random number generator. Each of these messages was coded into thematic categories by two coders independently to ensure objectivity in the coding process. The themes to which each of the messages was assigned were obtained from the taxonomy of behavior change techniques, which was developed by a large panel of behavior change experts [25]. This taxonomy has 16 thematic categories drawn together from multiple behavior change theories [26] such as Social Change Theory, Social Cognitive Theory, the Health Belief Model, and the Integrative Model of Factors Influencing Smoking. As suggested by Michie at al., the messages were coded to appropriate themes that were explicitly linked to the target behavior (in our case smoking) and target population (QuitNet users). As evident from the sample messages and thematic definitions, a single message may be assigned to multiple taxonomy themes. The definition of each theme and the subcategories of each theme can be found in [25].

Automated Analysis:
Methods from distributional semantics in conjunction with a machine learning classifier were used as part of this approach. Incorporation of background semantic information facilitates derivation of meaningful interpretations of QuitNet vector representations, on account of their short and terse textual features [10, 27, 28]. To this end, we used the distributional information from the Touchstone Applied Science Associated (TASA) corpus [29] to provide sufficient semantic context. The TASA is a collection of 44,700 articles that contain 10 million words of unmarked high-school level English text on arts, health, home economics, industrial arts, science, social studies and business. We applied RRI, a variant of Random Indexing which was developed to recognize meaningful relationships between terms without requiring they co-occur directly. This method was applied using Semantic Vectors, an open source package for applying distributional semantics [30]. A stopword list was used in this process [31]. This list contains words that are frequently used in texts but offer little semantic context. We used a dimensionality of 500, with minimum term frequency of 10, maximum term frequency of 15000 and a 'logentropy' termweight. We applied RRI to the TASA corpus to obtain *TASA term vectors* - representation of terms in the TASA Corpus. We generated *TASA based QuitNet message vectors* by generating vector representations for each QuitNet message as the sum of the TASA term vectors of the terms it contains, with subsequent normalization. Similarly, we obtained *QuitNet term vectors* by adding the QuitNet message vectors for each term occurring in QuitNet. This step leads to meaningful vector representations for terms that occur in QuitNet, but not in TASA, such as neologisms developed by the community. Finally, we derived a second message vector for each QuitNet message by adding the QuitNet term vectors for the terms it contains, and normalizing the resultant vector to generate *QuitNet message vectors*. This procedure is illustrated in Figure 2.

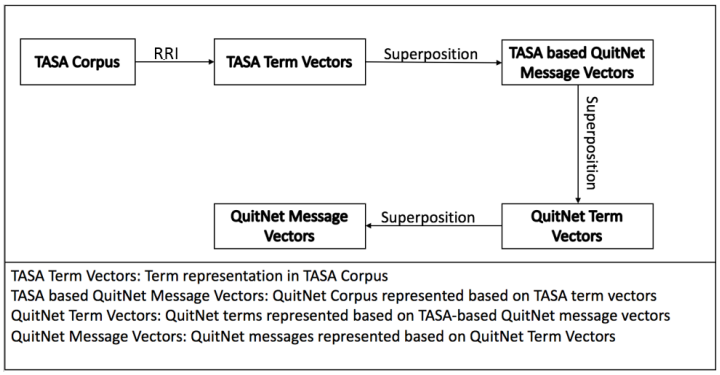

Figure 2: Vector generation sequence

From the QuitNet message vectors thus obtained, essential features were selected and extracted for machine learning techniques using the open source Weka package [32]. Each of these vector's 500 features was considered as individual attributes for the machine learning classification. Each of the themes was considered as a target for classification. Multi label classification was achieved by constructing individual binary classifiers for each theme. The classifier we used for this purpose was the J48 tree which is an improved version of the C4.5 tree [33, 34]. The classifier was trained using ten-fold cross-validation on the QuitNet message vectors representing the training set (manually coded 1000 messages). The trained J48 model was then used to classify QuitNet message vector representations of the entire sets

of messages Datasets 1 and 2. A random sample of 100 of these messages was used to test for reliability and accuracy of the machine coded messages. These messages were then coded manually by two independent coders and compared with the machine coding results to assess reliability measures. The inter-rater reliability for the machine versus the manual coding was calculated using Cohen's *Kappa* measure.

Temporal Modeling:
The communication pattern underlying peer interactions was modeled as a function of time by analyzing the proportion of messages that were categorized as belonging to each thematic category of the BCT taxonomy. The annotated messages from the automated analysis were used for this purpose. A specific category of QuitNet users, 'relapsers' (status change from ex-smoker to active smoker as self-reported by the users during study period) was considered for detailed temporal analysis, to characterize the frequency and thematic attributes of communication before and after relapse. The period of time modeled was the period of study (Mar-Apr 2007 & Jan-Dec 2014), i.e, the relapse was used an event marker and the entire set of messages exchanged by a given user before and after their relapse event in the study time period was taken into consideration.

Results and Discussion:

Over 72% users who exchanged messages during the study period were female, and the average age of the users at the time of registration was 45.

Qualitative analysis:
From the manual coding of the 1000 messages, the results indicated that 'feedback and monitoring' was the most commonly found theme. The second and third most frequently communicated themes were 'natural consequences' and 'social support' respectively. The Cohen's kappa measure between the two coders for the 1000 messages was 0.74. Most differences were related to coding of a single message into multiple themes and were all resolved by discussions. Messages which had users taking "pledges" to not smoke on that day were classified into 'goals and planning'. Messages where users mention how much time and money they have saved by quitting were classified into the categories 'feedback and monitoring' and 'comparison of outcomes'. QuitNet users also have traditions like "bonfire" where each user virtually throws unused cigarettes into the fire. They spell out the number of cigarettes they are throwing into the fire, thereby, monitoring their own progress. Table 1 shows sample messages for each thematic category. Since communication on social media is unprompted and uninhibited, QuitNet users discuss several extraneous aspects that are not specifically related to smoking. These messages were classified as 'miscellaneous' and not considered for further analysis since the taxonomy specifically focuses on target behavior, which in our case is smoking. However, it is important to note that these miscellaneous communications can aid in formation of social bonds, trust, and peer respect which are vital to long-term sustenance of user engagement in these online platforms. In addition, we observed that certain themes did not have ample representation. For instance, the theme 'associations', 'regulation' and 'antecedents' had only four samples out of the 1000 messages. In summary, the mapping of messages to the taxonomy of theoretically-linked behavior change techniques helped us understand if and how such techniques manifest in online platforms that promote health behavior changes. Describing the current landscape of behavioral techniques that manifest spontaneously in peer interactions could help us design better technology platforms that facilitate user interactions resembling other behavior change techniques, to enable users stay quit. For example, content recommendations to QuitNet users providing advice to peers about novel stress management strategies, would integrate 'shaping knowledge' techniques from the taxonomy of behavior change.

Automated Analysis:
The F-measures, precision and recall metrics for the cross validation of the machine learning technique J48 tree were 0.77, 0.79 and 0.77 respectively. The F-measures were calculated as an average of the individual binary classifiers using RRI vectors as attributes for the machine learning algorithm. The reliability measure between rater 1 and the automated classification system is 0.71, rater 2 and the system is 0.736. Therefore, the average system-rater agreement 0.72 approached inter-rater agreement of 0.74. A detailed characterization of thematic distribution over the years 2007 and 2014 as obtained using the automated classification system is shown in Figure 3. Although the users under consideration in these two time periods may not be the same, focusing on proportions of messages across themes over two different timelines gave us an overview of temporal patterns underlying thematic content of QuitNet user interactions.

$$\text{Percentage of messages in a specific theme in a given year} = \frac{(\text{Number of messages in that theme in that year}) * 100}{\text{Total number of messages in that given year}}$$

Table 1. Qualitative analysis of QuitNet messages using the taxonomy of behavior change techniques

Themes	Sample Message
Goals and planning	Good morning X and YYY.....**I pledge not to smoke today** and extend my hand to the next quitter who drops by....////ABC//day 4
Feedback and monitoring	You'll be fine. It takes some time for your body to heal !! Just hang in there and we will help you the best we can.////XXX////**36 days, 13 hours, 26 minutes and 28 seconds smoke free. 914 cigarettes not smoked**. $208.12 and 6 days, 23 hours of my life
Social Support	I want to pledge again today too! **Thanks for the support, I will not be smoking today, and I offer my hand to the next in line.**////I read YYY is also having a big storm with power outages. We may not see ZZZ today - no power, no computer. Everyone
Shaping Knowledge	**Read profiles and journals.//Learn as much as you can about what you are going to face**.//Don't walk into this thinking it is going to be easy or cute, this will be one of the toughest fights of and for your life.//Make the quit the most important thing in
Natural Consequences	good morning, **mine is my smell. I can't believe the other day when it was windy out i could smell my own shampoo and conditioner. And also that I am going to have more time on the earth with my family and friends**.////15 days, 10 hours, 3 seconds smoke fre
Comparison of behavior	That would be me. I know **I won't be smoking today**. Heck no. **Here is my hand for the next in line**...//XXX//Day 672
Associations	I am a nonsmoker**.//I became a nonsmoker the day I quit**, Feb 6th 2007.//
Repetition and substitution	((((((XXX))))))).////Nice warm fire tonite and it's just the ticket for these chilled bones!! **I'll be sending 33,657 unsmoked sick sticks to a blazing end**. I don't want or need them!////**I need a shot of apricot brandy and a relaxing hammock to settle**
Comparison of outcomes	Congratulations everyone//////XXX////39 days, 8 hours, 10 minutes and 22 seconds smoke free. 787 cigarettes not smoked. **$2,184.00 and 6 days of my life saved!** My quit date: 2/6/2007
Reward and threat	**Great job guys**! Congratulations!////XXX//
Scheduled Consequences	//HiXXX:////Hey I will gladly take those 571 unsmoke cigarettes for the freedom flames :)////Thank you so much for bringing them. **Now what can I get you to eat and drink while you relax and enjoy you**
Self-belief	"I'm one puff away from a pack a day" which, I chanted to myself every morning early in my quit - and it is so true, don't you think?//////I've been tempted to just have one "drag", or one "little taste**", but,I know,that I can't..... and won't!//Won't Ever**
Covert learning	Good morning ladies, i hope you are all well.//The smokefree zone is open, **drinks & Lemon Syrup Cake are ready for you.**////I pledge not to smoke today, offering my hand in friendship and support to the next.////XXX//D833
Regulation	**Go to the gym. Play a sport. Get adreneline rushing through your veins somehow. That's the only way I've found to release the anger and frustration**.////and BTW - you're boss probably NEEDS to hear that he's a moron!////Just as long as you're not smoking
Antecedents	**I've only had about 5 drinks since I quit.////I think if I had more than two drinks at one time I would definately smoke.////It just goes hand in hand with me.** //Maybe in time. I really don't miss drinking that much, and I used to drink quite a bit. I'm
Identity	I am a nonsmoker**.//I became a nonsmoker the day I quit**, Feb 6th 2007.//**A non smoker is defined as one who does not smoke, and I do not**.//I know if I ever smoked again I would be a smoker, and that is what I do not want, so I do not smoke.//I smoke in the p

The themes 'feedback and monitoring' and 'comparison of outcomes' were the most commonly used themes among QuitNet users in the years 2007 and 2014. The theme 'comparison of outcomes' has increased in proportion by 22% from 2007 to 2014. This could be attributed to the traditions within QuitNet community such as virtual bonfires where users account for unsmoked cigarettes and discuss consequent benefits with respect to quality of life. The increase in proportion of messages in this theme indicates that the users were aware of the outcomes of the behavior change (both positive and negative), which in turn could have motivated them to remain abstinent. The other theme which has increased in proportions of messages are the 'reward and threat', where the increase is 15%. Over the years, QuitNet

users began each of their messages by congratulating the recipient for their quit, which seems to be another new tradition, thus increasing content in this theme. The theme with most dip in percentage of messages were 'comparison of behavior' by 9.31%. Although the traditions on QuitNet were the same between the two years, huge reduction in the proportion indicates that number of users participating in the traditions may have reduced. The messages contributing to this theme are those in which users symbolically give their hand to others as a form of support and pledge not to smoke for the day. The messages in the themes 'feedback and monitoring', 'natural consequences' and 'goals and planning' have decreased by around 1%. Messages in which users account for unsmoked cigarettes and discuss the subsequent benefits in terms of quality of life were classified under "self-monitoring". The proportions of messages in the theme 'social support' has reduced by 7.72%. The proportion of messages in this theme was low overall in comparison to other themes across both years. This is not consistent with the results of the previous work done [10] and can be construed as a coding artifact of using the behavior change taxonomy.

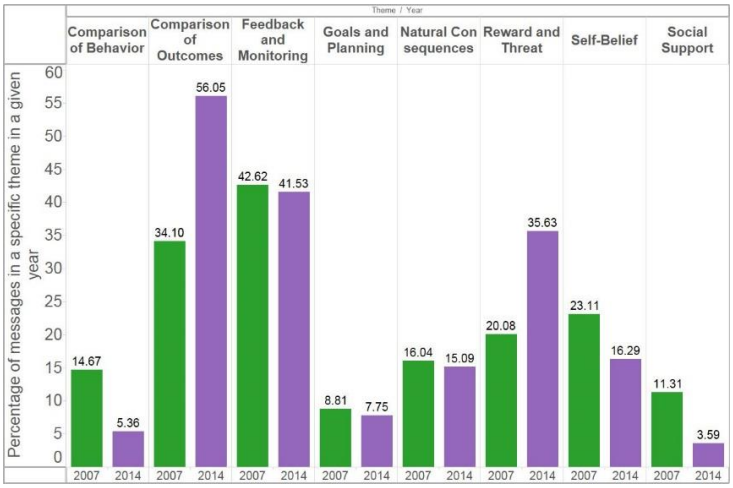

Figure 3. Distribution of messages across each theme after automated analysis over the years 2007 and 2014

Overall, this method has enabled visualization of the change in QuitNet users' communication across time. In summary, machine-learning techniques (to assign categories) in conjunction with distributional models (to provide additional semantic information not available in the content of short social media messages) have facilitated the extension of manually-coded thematic mapping to a large dataset with reasonable accuracy and reliability measures. This task would otherwise be prohibitively labor and resource intensive. Applying distributional semantic techniques such as RRI, to the content on social media has revealed implicit relationships between the messages without dramatically increasing computational overhead.

Temporal Modeling:
For the purpose of temporal modeling across active smoking and abstinence behavioral states, we chose to focus on QuitNet users who relapsed during the study period. Figure 4 portrays the change in percentage of messages before and after relapse.
For a specific theme:

$$\text{Percentage of messages before relapse} = \frac{(\text{Number of messages before relapse in a specific theme in a given year}) * 100}{\text{Total number of messages in the theme in that year}}$$

$$\text{Percentage of messages after relapse} = \frac{(\text{Number of messages after relapse in a specific theme in a given year}) * 100}{\text{Total number of messages in the theme in that year}}$$

The percentage difference between message frequency before and after relapse across all themes in the year 2007 was 80%, indicating an overall drop in frequency of communication. This pattern can also be seen in 2014 with a drop of 60% in proportion of messages before relapse and after relapse across all themes. The drop in percentage of messages before and after relapse in the year 2007 was 75% in the categories 'goals and planning' and 'feedback and monitoring', 66.6% for 'Social Support', 'comparison of behavior' and 'comparison of outcomes', 38.5% for 'natural consequences', 71% for 'miscellaneous' and 100% for 'reward and threat' and 'self-belief'. The pattern of communication among relapsers in the year 2014 was similar to the year 2007. The difference in proportion of

messages before and after relapse in the year 2014 is 70% in the category 'goals and planning', 80% in 'feedback and monitoring', 70% in 'social support', 60% for the categories 'comparison of behavior' and 'comparison of outcomes', 50% for the theme 'natural consequences', 95% for 'miscellaneous', 75% in 'rewards and threat'. Most frequently found themes embedded in communication of relapsing QuitNet users in 2007 belonged to the category 'feedback and monitoring' both before and after the relapse event. The theme that had the most significant reduction in number of messages before and after relapse, in the year 2007, was 'rewards and threat'. Most QuitNet messages which fell under this category are congratulatory in nature for the efforts of the users. Thus, it is coherent that number of messages in this thematic category have reduced. In 2014, the categories most prominent in users' messages before the event of their relapse, were 'feedback and monitoring' and messages that were in the 'miscellaneous' category (no relation to smoking cessation in particular). Whereas, the theme that had a huge fall in message count after relapse event in 2014, was 'miscellaneous' category. The messages in 'miscellaneous' category in QuitNet were related to postings on addictions other than smoking and everyday generic experiences of one's daily life. Examples of such messages were "my hd has gone out to get a dog", "Quit drinking without AA! From XXX on 4/9/2014 2:00:48 PM".

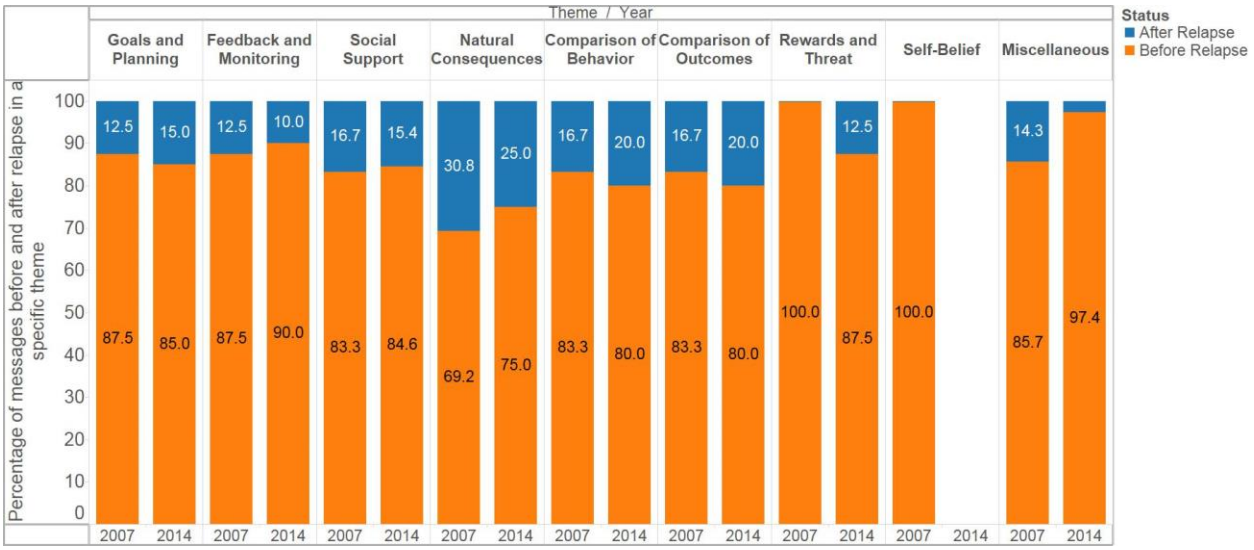

Figure 4. Pattern of communication of relapsing users before and after the event of their relapse

Both in 2007 and 2014, the message count in the thematic category 'self-belief' have remained low in comparison to count in other themes, before and after relapse. In 2014, specifically, message count in this theme had remained at 0% before and after relapse. Existing literature of behavior change suggest that users displayed lower self-belief when they were closer to a relapse episode and lower self-belief has been found to be an indicator for relapse [35]. It can also be seen that the number of messages in the category 'social support' have also reduced in both years, before and after relapse. The reduction in 'social support'-related messages after relapse may be attributed to the messages coded under other themes due to explicit discussions of ideas related to quit rebound and coping with failure. Although these messages offered moral support implicitly, they were categorized to 'shaping knowledge' and/or 'natural consequences' according to the taxonomy of behavior change theories. The overall drop in frequency of communication across both years, has indicated that users tend to communicate less after their relapse. Conversely, when a user quiets down it could indicate that they may relapse. Literature suggests that after relapse, users tend to be affected by guilt and shame, consequently tend to become reclusive [36], which can be observed from the pattern of communication of users over time. Again, the attitude of guilt and shame among relapsers after a relapse event is re-affirmed by the significant drop in messages in the 'feedback and monitoring' category. The users may not want others to revisit the event of their relapse unless they cope with their failure. The reduction in message count in the category 'comparison of behavior' may also be attributable to the shame that relapsers experience after quitting. Overall low presence of the theme 'goals and planning' after a relapse event could suggest that the relapsers may not be ready to set immediate quit goals [37].

When the post-relapse activity of users was represented as a percentage of the pre-relapse activity, the results revealed interesting insights into specific theoretical techniques embedded in the messages exchanged by the relapsers before

and after relapse. In the year 2007, techniques such as 'goals and planning' and 'feedback and monitoring' were retained in 14.3% messages exchanged post-relapse as compared to pre-relapse. The retention observed was highest in the themes 'natural consequences' at 44.4%, followed by 'comparison of behavior', 'comparison of outcomes' and 'social support' at 20%. Themes such as 'reward and threat' and 'self-belief' had the lowest retention in messages post-relapse. In the year 2014, themes that were retained in the messages post relapse were highest in 'natural consequences' at 33.3%, followed by 'comparison of behavior' and 'comparison of outcomes' at 25%. The messages in 'Miscellaneous' category were retained in 2.7% of the messages and was the lowest post-relapse retention in the year 2014.

In summary, understanding semantics and behavior change techniques before and after relapse has provided us with insights into the specific content that users might be interested in and can benefit from, at points in time corresponding to changes in behavioral states within a smoking cessation episode. Such information is essential to design targeted user-content and interactions to enhance the efficacy of existing health promotion platforms such as QuitNet. The information on the users' low self-efficacy and their attitude towards recovery is evident from analyzing the semantic content. The reduced communication indicating guilt and shame has been observed from the reduced frequency of interaction and also from the semantic content attributes.

Limitations and Future Work:

The BCT taxonomy used in this study specifically targets the identification of theoretically driven techniques that target a particular health behavior, in our case abstinence from smoking. However, applying the BCT taxonomy to analyze user-generated content and peer interactions in QuitNet like online platforms may result in omission of important social interactions that, though not recognized as techniques of behavior change, foster trust, bonding, and nurturing of the community. These aspects are ancillary to smoking cessation, yet are important mediators of behavior change in social platforms. Therefore, if it is to be used to annotate such data, the taxonomy should be extended to incorporate these interpersonal interactions – particularly as these may include community-initiated traditions that the users have developed over years of collective effort toward change in behavior. These are best captured through inductive coding techniques such as grounded theory. Our previous and ongoing work employs these techniques to capture emergent nature of user communication within QuitNet [26, 27, 38]. Certain themes did not have adequate representation during qualitative coding, so could not be considered for automated analysis on account of a lack of training cases. Other methods of representation learning (e.g. word embeddings [39]), machine learning involving deep neural networks [40]) may enhance the ability of our methods to generalize from small numbers of training examples. We have used data from two different years 2007 and 2014 to validate the generalizability of thematic categorization, scalability of automated methods, and observed semantic shifts before and after relapse in temporal models. In our future work, we will extend the study to incorporate additional datasets to examine continuous temporal trends. Such analysis will help us to identify semantic content that is predictive of behavior change, and vice versa. Relapse is an important behavioral phenomenon from a clinical perspective. Characterization of the temporal and social dimensions of relapse may permit proactive identification of relapsing users, with subsequent personalization of interventions to meet the associated user needs and behavioral targets. Understanding other user behavioral states beyond relapsers (e.g. active smokers, successful quitters) is equally important to help us distinguish specific communication characteristics underlying peer interactions based on smoking status.

Conclusion:

This study focuses on using peer-to-peer communication in an online community to understand the implications for data-driven health promotion. The major contributions of this study are as follows:

- This study uses a taxonomy of behavior change techniques to annotate peer communications in a health-related online community for smoking cessation. This offers an insight into the theoretical roots and related techniques embedded in QuitNet user communication.
- The automated analysis methods were used to extend the annotation from a set of 1000 manually coded messages to a large data set of about 84,204 messages. Use of distributional information that captures implicit meanings associated with peer interactions has provided us with better understanding of semantic context within terse social media interactions.
- The temporal modeling of QuitNet peer interactions has focused on understanding the semantic context surrounding a user's communication across multiple behavioral points surrounding an event of relapse. Such

semantic context offers insights into users' needs and triggers, which may be used for effective design of interventions appropriate to a particular stage of behavior change.

Annotating peer-to-peer communication to relate it to established theories of behavior change provides a bridge between these empirical data and scientific understanding of the mediators of behavior change. To our knowledge, this is the first effort to use the taxonomy of behavior change techniques (version 1 with 16 themes) to evaluate the construct of social media as a venue for behavior change. The application of the taxonomy for evaluation has revealed the theoretical manifestations underlying QuitNet user interactions that are primarily user-driven rather than expert-guided. Such findings can help us design better computer-mediated support technologies that nudge and prompt users to exchange theory-guided messages with embedded behavior change techniques. Distributional semantics in conjunction with machine learning, reduces human effort in dealing with large corpus of social media data making it amenable for large scale temporal modeling. Longitudinal trends in communication help us understand the traits of the user groups (classified based on their smoking behavior) that can help us customize the support infrastructure on social platforms for just-in-time guidance to sustain long term behavior change efforts. Indicators of relapse and understanding of user attitudes before and after relapse are essential to provide tailored assistance at the point in time when it is most relevant to the users of QuitNet like behavior change technologies in the digital era.

Acknowledgements:

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number 1R21LM012271-01. 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. Centers for Disease Control and Prevention. Chronic disease prevention and health promotion. [cited 2016 Mar 08]. Available from: http://www.cdc.gov/chronicdisease/
2. Centers for Disease Control and Prevention. The health consequences of smoking-50 years of progress: a report of the surgeon general. Atlanta, U.S. DHHS; 2014
3. Zhang M, Social media analytics of smoking cessation intervention: user behavior analysis, classification, and prediction. 2015; Drexel University, Pennsylvania
4. Chretien, Katherine C, and Kind T. Social media and clinical care ethical, professional, and social implications. Circulation. 2015;127(13): 1413–1421.
5. Cobb NL, Graham AL, Abrams DB. Initial evaluation of a real-world internet smoking cessation system. American Journal of Public Health. 2010; 100(7): 1282–1289, doi:10.2105/AJPH.2009.165449
6. Valente TW. Network interventions. Science. 2012;337(6090):49-53
7. Graham AL, Papandonatos GD, Kang H, Moreno JL, Abrams DB. Development and validation of the online social support for smokers scale. Journal of Medical Internet Research. 2011;13(3): e69. doi: 10.2196/jmir.1801.
8. Thrul J., Klein BA, Ramo DE. Smoking cessation intervention on facebook: which content generates the best engagement?. Journal of Medical Internet Research. 2015; 17(11). e244. doi:10.2196/jmir.4575
9. Rocheleau MS, Sadasivam RS, Baquis K, Stahl H, Kinney RL, Pagoto SL, et al. An observational study of social and emotional support in smoking cessation twitter accounts-content analysis of tweets. Journal of medical internet research.2015; 17(1).e18. doi: 10.2196/jmir.3768
10. Myneni S, Cobb NK, Cohen T. Finding meaning in social media: content-based social network analysis of quitnet to identify new opportunities for health promotion. Studies in Health Technology and Informatics. 2013;92:807–811.
11. Chen G, Warren J, Riddle P. Semantic Space models for classification of consumer webpages on metadata attributes. Journal of Biomedical Informatics. 2010;43(5):725-735.
12. Schwartz HA, Ungar LH. Data-driven content analysis of social media: a systematic review overview of automated methods: ANNALS, AAPSS.2015;659 DOI:10.1177/0002716215569197
13. Chouduhury C, Munmun, Gamon M, Counts S, Horvitza E. Predicting depression via social media. Proceedings of the seventh International AAAI 2013 Conference on Weblogs and Social Media.
14. Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review. 1997; 104:211-40.
15. MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. Journal of the American Medical Informatics Association. 2013;20(6):1120-1127.

16. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. Paper presented at: AMIA Annual Symposium Proceedings2011.
17. Wang YC, Kraut R, Levine JM. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. Paper presented at: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.
18. Cohen T, Schvaneveldt R, Widdows D. Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. Journal of Biomedical Informatics. 2010; 43(2):240-256
19. McArthur R, Bruza P, Warren J, Kralik D. Projecting computational sense of self: a study of transition in a chronic illness online community. System sciences. 2006;5:91.
20. Elhadad N, Zhang S, Driscoll P, Brody S. Characterizing the Sublanguage of Online Breast Cancer Forums for Medications, Symptoms and Emotions. Paper presented at: Proceedings of the AMIA Annual Symposium 2014, Washington, DC.
21. Zhang S, Grave E, Sklar E, Elhadad N. Longitudinal Analysis of Discussion Topics in an Online Breast Cancer Community using Convolutional Neural Networks. ArXiv Prepr ArXiv160308458 [Internet]. 2016
22. Qiu B, Zhao K, Mitra P, Wu D, Caragea C, Yen J, Greer GE, Portier K. Get online support, feel better--sentiment analysis and dynamics in an online cancer survivor community. InPrivacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on 2011 Oct 9 (pp. 274-281).
23. Durant KT, McCray AT, Safran C. Modeling the temporal evolution of an online cancer forum. InProceedings of the 1st ACM International Health Informatics Symposium 2010 Nov 11 (pp. 356-365).
24. QuitNet LLC. URL: https://quitnet.meyouhealth.com/ [accessed 2016-01-28] [WebCite Cache]
25. Michie S., Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting change interventions. Annals of Behavioral Medicine .2013; 46(1): 81-95
26. Glanz K, Rimer BK, Viswanath K, editors. Health behavior and health education: theory, research, and practice. John Wiley & Sons; 2008 Aug 28.
27. Myneni S, Fujimoto K, Cobb N, Cohen T. Content-Driven Analysis of an Online Community for Smoking Cessation: Integration of Qualitative Techniques, Automated Text Analysis, and Affiliation Networks. American journal of public health. 2015;105(6):1206-12.
28. Myneni, S., Cobb, N., & Cohen, T. (2016, accepted). Content-specific network analysis of peer-to-peer communication in an online community for smoking cessation. Paper to be presented at AMIA Annual Symposium Proceedings 2016, Chicago, IL
29. Landauer TK., Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes.1998; 25:259–284.
30. Widdows D. Cohen T. The semantic vectors package: new algorithms and public tools for distributional semantics. Fourth IEEE International Conference on Semantic Computing (IEEE ICSC); 2010: 9-15.
31. Salton G, McGill MJ. Introduction to modern information retrieval.2013. New York: McGraw-Hill
32. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update: SIGKDD Explorations, 2009, 11(1).
33. Patil TR, Sherekhar SS, Performance analysis of naive bayes and j48 classification algorithm for data classification. International journal of computer science and applications. 2013; 6(2) : 256-261
34. Yuan Lu, Application of random indexing to multi label classification problems: a case study with mesh term assignment and diagnosis code extraction. 2015. University of Kentucky, Kentucky
35. Bandura A. Self-efficacy: The exercise of control. New York: Freeman. 1997
36. Bowen S, Chawla N, Marlatt AG. Mindfulness-based relapse prevention for addictive behaviors: a clinician's guide. New York: The Guiford Press. 2010
37. O'Reilly, Christopher. Relapse and recovery: A crash course on the basics of addiction. 2016. Presentation.
38. Myneni S, Cobb NK. & Cohen T. In Pursuit of Theoretical Ground in Behavior Change Support Systems: Analysis of Peer-to-Peer Communication in a Health-Related Online Community. Journal of Medical Internet Research. 2016 ;18(2): e28
39. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pages 3111– 3119.
40. Zhang ML, Zhou ZH. Multilabel neural networks with applications to functional genomics and text categorization. IEEE Trans. Knowledge Data Engineering. 2006; 18(10) :1338-1351.