HHS Public Access

Author manuscript

Stud Health Technol Inform. Author manuscript; available in PMC 2020 November 11.

Published in final edited form as:

Stud Health Technol Inform. 2019 August 21; 264: 1228–1232. doi:10.3233/SHTI190422.

Characterization of Behavioral Transitions Through Social Media Analysis: A Mixed-Methods Approach

Tavleen Singh[a], **Carlos A Perez**[a], **Kirk Roberts**[a], **Nathan Cobb**[b], **Amy Franklin**[a], **Sahiti Myneni**[a]

[a]University of Texas School of Biomedical Informatics, Houston, TX, USA

[b]Georgetown University Medical Center, Washington, DC, USA

Abstract

Unhealthy behaviors are a socioeconomic burden and lead to the development of chronic diseases. Relapse is a common issue that most individuals deal with as they adopt and sustain a positive healthy lifestyle. Proper identification of behavioral transitions can help design agile, adaptive, and just-in-time interventions. In this paper, we present a methodology that integrates qualitative coding, machine learning, and formal data analysis using stage transition probabilities and linguistics-based text analysis to track shifts in stages of behavior change as embedded in journal entries recorded by users in an online community for tobacco cessation. Results indicate that our semi-automated stage identification method has an accuracy of 90%. Further analysis revealed stage-specific language features and transition probabilities. Implications for targeted social interventions are discussed.

Keywords

Health behavior; social media; machine learning

Introduction

Poor lifestyle choices and unhealthy behaviors such as smoking, alcohol consumption, poor diet, and physical inactivity affects the development and management of chronic diseases, like obesity, Type 2 diabetes mellitus, hypertension, cardiovascular diseases, and several types of cancer [1]. Chronic diseases kill 38 million people every year and hence their prevention is important [1]. Efficient self-management of chronic illnesses and adoption of positive health behaviors has shown to be related to overall better physical and psychological health outcomes [2]. However, such self-management is not driven by individuals solely but rather in a social context which includes formal health care providers, informal social peers, and their physical surroundings [3]. Behavior change can be a daunting task, and oftentimes individuals relapse as they attempt to embrace and sustain a positive health change.

Address for Correspondence: Tavleen Kaur Ranjit Singh, MS. The University of Texas School of Biomedical Informatics at Houston, 7000 Fannin, Houston, TX 77030. tavleen.kaur.ranjit.singh@uth.tmc.edu.

Behavioral transitions are important to sustain lasting and improved health outcomes [4]. Several theories have been postulated to define and model these transitions and changes in behavior. Specifically, the Transtheoretical Model (TTM) of Change has been used to conceptualize the process of intentional behavior change by its two main components: stages and processes of change [5]. Stages explore the temporality of behavior change, while processes encompass cognitive and behavioral concepts such as decisional balance, self-efficacy, and rewards management [5]. It is important to identify these stages of change so that appropriate lifestyle interventions can be prescribed to individuals who are willing to quit an unhealthy habit and modify their behaviors. Many mobile interventions have been modeled after TTM. Adapting theory-driven interventions to the changing needs of individuals can help sustain positive health changes long-term [5].

Emerging research has shown that agile and adaptive interventions that respond to stage transitions can be effective tools of behavior change in real-time settings [6]. However, it is difficult to monitor stage transitions using traditional behavior modeling approaches at scale. Online communication forums, which form the dominant means of communication in the digital era, provide an opportunity to understand human behavior change at nuanced levels. According to the Global Digital Report 2018 the number of social media users worldwide is about 3.2 billion [7]. The electronic archival of digital interactions provide us with rich data sources that afford the opportunity to understand intricate processes and stages of behavior change. Emerging research suggests that online social media analysis can help us understand patterns of social factors underlying behavior change and help in the development of network interventions for health behavior changes [8]. Thus far, social media analysis has focused on (a) understanding the structure of social ties [9], (b) assigning peer-to-peer communication events to various social support categories [10]. However, prior research on health-related social networks has not merged theoretically-driven content-based approaches with machine learning methods to facilitate stage-based user behavior and transition classification. Such understanding can have important implications for the design of interventions that aim to enhance self-management behavior.

In this paper, we present a new methodology to understand an individual's behavior transitions using recorded journal entries in online communities. Our study has three major components: (a) qualitative analysis to manually label the user-generated data for different stages of behavior change, (b) automated categorization using machine learning models to scale the labeling of stages of behavior change to a large social media dataset, and (c) characterization of stage-specific language features and transition rates to enable personalized intervention refinement and technology development in scalable online settings. Our research described in this paper will help answer the following questions: 1) How can we adapt existing informatics approaches to automate stage identification in digital health platforms? 2) What are the relationship characteristics between user engagement on these platforms, behavior transitions, and language features?

Such an understanding can help in identifying the triggers for relapse and designing appropriate interventions for the users at risk for relapse or in need of more social support for long-term sustenance [11]. We apply our methodology to characterizing behavioral transitions among the users of an online community for tobacco cessation.

Methods

QuitNet is an online social communication forum that promotes smoking cessation amongst its members thereby leading to behavior change and health promotion [12]. It has been in existence from the past 16 years and has over 100,000 new registrants per year. The members of this platform are usually smokers who are willing to quit or ex-smokers who are willing to stay abstinent. Graham et al. [13] have shown a strong correlation of individual's participation in online community with abstinence compared to individuals who do not participate in such communities. The dataset used in this study consisted of de-identified journal text entries, spanning from 1999-2015 including 26,441 individuals, and 111,004 journal text entries. The journal text entries that were marked as public by the QuitNet users were used for this analysis. This study is exempted from human subjects review by the Institutional Review Board at the University of Texas Health Science Center at Houston.

Qualitative analysis

Firstly, we developed the annotation guidelines to inform the coding process and ensure objectivity in stage assignment. Table 1 shows the annotation guidelines that were adopted in our analysis. We used the TTM model to label every journal text entry with its associated stage of change. The TTM model defines Precontemplation, Contemplation, Preparation, Action, Maintenance, and Termination as the six stages of change, where each stage involves a process of progress (Figure 1). Fivehundred out of the 111,004 journal text entries were randomly selected and coded manually by two independent researchers by performing a line by-line analysis on the text entries to derive the stages of change codes from the data. The codes they assigned to the text entries had a Cohen's Kappa measure of 97%. Disagreement between the researchers was resolved through discussion.

Automated Text Analysis

Using the labeled dataset of 500 journal text entries, we first performed a feature inspection to get a better understanding of our textual data. The dataset was then cleaned using the following techniques - (a) tokenizing the text to split sentences into individual words, (b) removing the stop words to get rid of words that occur too frequently or infrequently as well as punctuation, (c) stemming the text to address word variants, and (d) vectorizing words since we needed to convert words into mathematical representations to feed them into various machine learning models. Latent Dirichlet Allocation (LDA) and word2vec models [14, 15] were trained using default parameters to create feature vectors and extract sensible feature data from the journal text entries. LDA is a probalistic topic model that creates probabilities on the word level, while word2vec is a deep learning method that creates feature vectors for the words in the text corpus [15]. After this we applied a few supervised classifiers to be able to predict what stage belongs to each journal text entry. The labeled dataset was split into a train and test set and a ten fold cross validation was performed. We used the features that were previously created to train three machine learning models: Logistic Regression (LR), Support Vector Machine (SVM) (linear), and Random Forest (RF). These models were chosen, given their performance on similar classification tasks [15]. We used recall, precision, and F-measure as our accuracy metrics to evaluate the machine learning algorithms used in this study. We selected the model with the highest

accuracy to make final predictions on the unlabeled dataset. Scikit-Learn package was used for analysis [16].

Data Analysis

User level analysis was performed to map transitions between stages of behavior change. Pearson Correlation coefficient was calculated to identify if there was any correlation between user engagement and stage transition rates (the numbers of messages posted and the number of transitions from one stage to another). A Markov model was used to estimate the stage membership probabilities and the stage transition probabilities of movement from one stage to another. Stage membership probabilities indicate the prevalence of each stage and transition probabilities indicate the probability of stage movements conditional on the stage membership at the previous time point [17]. Further, we applied The Linguistic Inquiry and Word Count (LIWC) [18] dictionary that comprises of psychologically meaningful word categories, and whose output includes the percentage of words within a given text that belong to each stage of change. Applying this technique allows for a direct comparison of text features across various stage-specific journal entries.

Results

On an average, a user posted four text entries. Table 2 below shows some of the textual characteristics of the unlabeled journal text entry corpus.

Qualitative Analysis

Table 3 shows examples of some of journal text entries and their associated stage of change as coded manually by the researchers. In the first example - as the individual was in the process of deciding to undertake the quit process and still thinking about it- the entry was coded as ‘Contemplation’. The second example entry was coded as ‘Preparation’ since the individual specifically mentioned that they plan to quit smoking in 5 days. The last example entry was coded as ‘Action’ since the individual mentioned being smoke free for a certain amount of time, with a past quit date.

Our dataset lacked examples of the ‘Pre-contemplation’, ‘Maintenance’ and ‘Termination’ stage and had limited examples of the ‘Contemplation’ stage (n=21). There were 210 text entries which were coded as ‘Preparation’ and 268 text entries which were coded as ‘Action’. There was one text entry which was coded as ‘NA’ (Not Applicable).

Automated Text Analysis

Table 4 shows the accuracy of various machine learning models used for making predictions on the unlabeled dataset using 10-fold cross validation. Since the Random Forest Model gave the highest accuracy, this model was chosen to make the final predictions on unlabeled dataset.

62% of the total entries were labeled as ‘Action’ (n=68,947) and 38% were labeled as ‘Preparation’ (n=42,041). Very few entries were labeled as ‘Contemplation’ (n=15) and

'NA' (n=1). Table 5 shows some example journal text entries that were labeled using the RF machine learning model.

Data Analysis

Because a limited number of samples were coded to be in the 'Contemplation' stage, we conducted further data analysis with only those journal text entries that were labeled as either 'Action' or 'Preparation' by the machine learning model. The highest number of journal text entries posted by an individual was 418 and the number of transitions made by this particular individual was 95 from one to stage of behavior change to another. The Pearson Correlation coefficient between the number of journal text entries posted by the users and their transitions between the stages of change calculated was 0.857 (Figure 2).

This value showed that the higher the frequency of journaling by an individual, the higher the number of behavioral transitions for that individual, which may occur in either a positive direction (Preparation to Action) or a negative direction (Action to Preparation). This engagement phenomena may indicate that the denser the digital footprint of an individual in a digital platform like QuitNet, the higher the probability to identify behavioral transitions.

Figure 3 shows a Markov chain model representing the movement pattern between the two stages of change – 'Preparation' and 'Action'. The membership probabilities for 'Preparation' is 0.37 and for 'Action' is 0.32. The transition probabilities from 'Preparation to Action' is 0.27 and from 'Action to Preparation' is 0.21.

As can be seen from the figure above, the probability of an individual staying in the 'Preparation' stage is higher compared to an individual staying in the 'Action' stage. The probability of transitioning from 'Preparation' to 'Action' stage is higher compared to transitioning from 'Action' to 'Preparation' stage.

Exploratory text analysis using LIWC revealed specific language features that were prominent within the 'Preparation' and 'Action' stages. As can be seen in Figure 4, interrogatives were prevalent in the 'Contemplation' stage (e.g seeking information), while numbers were highly used in the 'Action' stage, which is probably expected due to expressing quantities of both time of abstinence and cigarettes not smoked (e.g. demonstrating progress). Figure 5 indicates specific foci of journal entries within each stage of change. For example, in the 'Contemplation' stage language with individual drives and needs, achievement, power, and reward were shown to be common. Sense of achievement and work-induced stress (obstacles to quitting) were emphasized in the 'Action' stage.

Discussion

Technologies like online communities provide support for individuals to modify their behavior for better health outcomes. On the other hand, there are theory-driven principles and processes that work best at every stage of behavior change to reduce resistance, facilitate progress, and prevent relapse. It is important to develop advanced analytic tools to integrate theory and technology, which can result in enhanced just-in-time support infrastructure for sustained behavior change. Such an ecosystem will ultimately lead to a greater

understanding of the ways in which such social phenomena mediate behavior change at individual and community levels, thereby providing us with the opportunity to develop superior and scalable interventions. This to further assist the development of multilevel systems that harness behavior change mechanisms aimed at augmenting the support for individuals attempting to achieve their health goals [19].

To the best of our knowledge, the research reported in this paper is among the few studies that have attempted to apply various machine learning models in the context of online communities and social journaling to understand and characterize stages of behavior change in tobacco cessation. Our work provides new techniques to analyze user generated unstructured health data to understand an individual's behavior transition over a period of time. Our model was able to label the stages of behavior change with an accuracy of 90% in journal text entries. The word2vec based approach described in this paper allows for the extension of human-intensive qualitative analysis to large social media datasets.

We found that people who engaged more in journaling had higher transition flags, which may indicate (a) desired or undesired stage shifts, and (b) higher chances of being flagged for stage transitions. In either case, better support infrastructure through adaptive and personalized means aimed at increasing user engagement may ultimately lead to better health outcomes. Further, we have seen that there is a higher probability of an individual staying in the 'Preparation' stage and targeted interventions should be developed to encourage such individuals to move to the 'Action' stage and attain a positive behavior change. The individuals undergoing relapse can be identified based on their stage transitions and appropriate interventions can be designed for them – encouraging them to pick a new quit date or helping them form new relationships online so that they can stay on the path of quitting.

Specific language traits revealed using LIWC analysis, while preliminary in nature, still highlight the authenticity and individual drive embedded in journal entries highlighting the need to understand emotional tone, speech intentions, and cognitive focus within each stage of behavior change. Our study can provide new directions for developing network interventions [20] for tobacco cessation and health promotion by focusing on content-based, targeted behavior change strategies while addressing stage-specific constraints and associations simultaneously.

One of the limitations of our work was the limited number journal text entries that were coded manually to identify stages of change. To improve the generalizability of our results, it is important to have higher number of journal text entries in the qualitative sample so that higher training accuracy of machine learning models can be achieved. It is possible,given the low fraction of journal text entries coded for stages of behavior change, that the distribution of various stages may not have been accurately represented. It is important that a larger number of journal text entries are coded to reach stage saturation. Also, the content of the journal text entries plays an important role during the predictions made by machine learning models. For example, an individual who has just started the 'Action' but has provided lots of detail about the 'Preparation' could have led the classifier to make a wrong stage prediction. Individual-specific linguistic features and demographics should also be considered when

analyzing journal entries. There are also some limitations inherent to the TTM such as a lack of consideration of the social context, biological, and environmental issues related to changes in health behaviors [21], that should be considered when understanding behavior transitions. Another limitation of our study is focusing only on tobacco cessation. Future work should extend these methodologies to datasets in other areas such as diabetes, cancer prevention and survivorship, etc. A more formal social network analysis that utilizes peer interactions with journal entries needs to be performed so that content specific network-patterns can be identified as they have implications in the design of behavioral support systems to promote public health and wellness.

Conclusions

Risky health behaviors contribute to a large number of preventable deaths around the world. The ubiquity of online social platforms allows us to examine inter- and intra-personal processes and stages of health behaviors among individuals. In this paper, we described a mixed methods approach that combines qualitative coding and automated text analysis to provide deeper insight into the mechanisms underlying behavior change through the utilization of digital footprints in the form of online journaling. It is very important to develop scalable methods to help health researchers and professionals analyze large amounts of textual data generated from online communities in today's digital era. Such techniques can help design personalized and targeted interventions that persuade people to initiate or adhere to a positive behavior change. This can help in establishing novel digital and translational interventions in public health and behavioral sciences.

Acknowledgements

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R21CA220670. 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]. Noncommunicable Diseases And Their Risk Factors, World Health Organization. (n.d.) https://www.who.int/ncds/en/ (accessed April 1, 2019).

[2]. Gallant MP, The Influence of Social Support on Chronic Illness Self-Management: A Review and Directions for Research, Health Educ Behav. 30 (2003) 170–195. [PubMed: 12693522]

[3]. Audulv Å, Asplund K and Norbergh K-G, The integration of chronic illness self-management, Qual Health Res 22 (2012), 332–345. [PubMed: 22167155]

[4]. Abraham C and Michie S, A taxonomy of behavior change techniques used in interventions., Health Psychology. 27 (n.d.) 379–387.

[5]. Prochaska JO and Velicer WF, The Transtheoretical Model of Health Behavior Change, Am J Health Promot. 12 (1997) 38–48. [PubMed: 10170434]

[6]. Centola D, Social Media and the Science of Health Behavior, Circulation. 127 (2013) 2135–2144. [PubMed: 23716382]

[7]. Digital In 2018: World's Internet Users Pass The 4 Billion Mark - We Are Social, We Are Social. (2018). https://wearesocial.com/blog/2018/01/global-digital-report-2018 (accessed April 1, 2019).

[8]. Myneni S, Fujimoto K, Cobb N and Cohen T, Content-Driven Analysis of an Online Community for Smoking Cessation: Integration of Qualitative Techniques, Automated Text Analysis, and Affiliation Networks, Am J Public Health. 105 (2015) 1206–1212. [PubMed: 25880942]

[9]. Centola D, The Spread of Behavior in an Online Social Network Experiment, Science. 329 (2010) 1194–1197. [PubMed: 20813952]

[10]. Glanz K, Health Behavior and Health Education, John Wiley & Sons, 2008.

[11]. Myneni S, Cobb N and Cohen T, In Pursuit of Theoretical Ground in Behavior Change Support Systems: Analysis of Peer-to-Peer Communication in a Health-Related Online Community, J Med Internet Res. 18 (n.d.) e28. [PubMed: 26839162]

[12]. QuitNet, (n.d.). https://quitnet.meyouhealth.com/#/ (accessed April 1, 2019).

[13]. Graham AL, Cobb NK, Raymond L, Sill S and Young J, Effectiveness of an Internet-Based Worksite Smoking Cessation Intervention at 12 Months, Journal of Occupational and Environmental Medicine. 49 (2007) 821–828. [PubMed: 17693778]

[14]. Mikolov T, Chen K, Corrado G, and Dean J.J.a.p.a., Efficient estimation of word representations in vector space, (2013).

[15]. Classification Combining LDA And Word2Vec | Kaggle, (n.d.) https://www.kaggle.com/vukglisovic/classificationcombining-lda-and-word2vec (accessed March 31, 2019).

[16]. Model selection and evaluation, (n.d.) https://scikit-learn.org/stable/model_selection.html#model-selection (accessed April 1, 2019)

[17]. Brick LA, Redding CA, Paiva AL and Velicer WF, Intervention effects on stage transitions for adolescent smoking and alcohol use acquisition., Psychology of Addictive Behaviors. 31 (n.d.) 614–624.

[18]. Pennebaker JW, Francis ME, and Booth RJ. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates. 2001;71(2001):2001.

[19]. Myneni S, Fujimoto K and Cohen T, Leveraging Social Media for Health Promotion and Behavior Change: Methods of Analysis and Opportunities for Intervention, (2017) 315–345.

[20]. Valente TW, Network Interventions, Science. 337 (2012) 49–53. [PubMed: 22767921]

[21]. Armitage CJ, Is there utility in the transtheoretical model?, 14 (2009) 195–210.

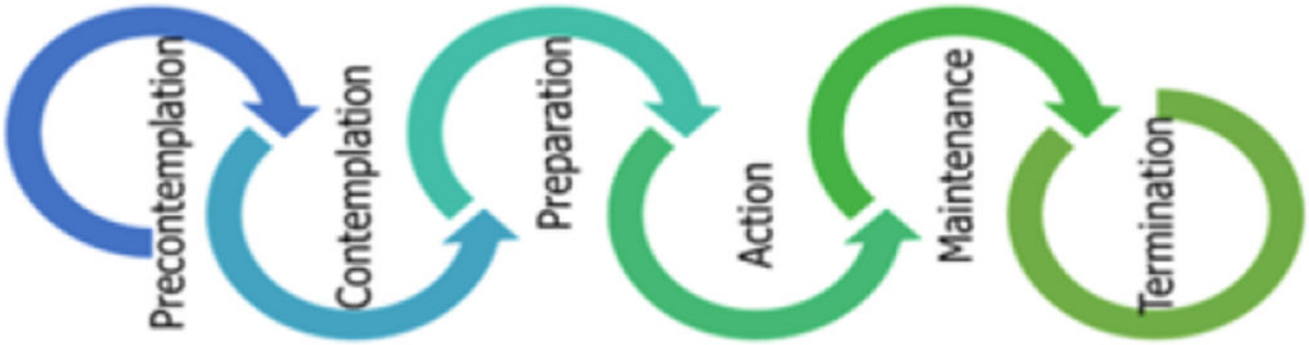

Figure 1–.
Stages of behavior change

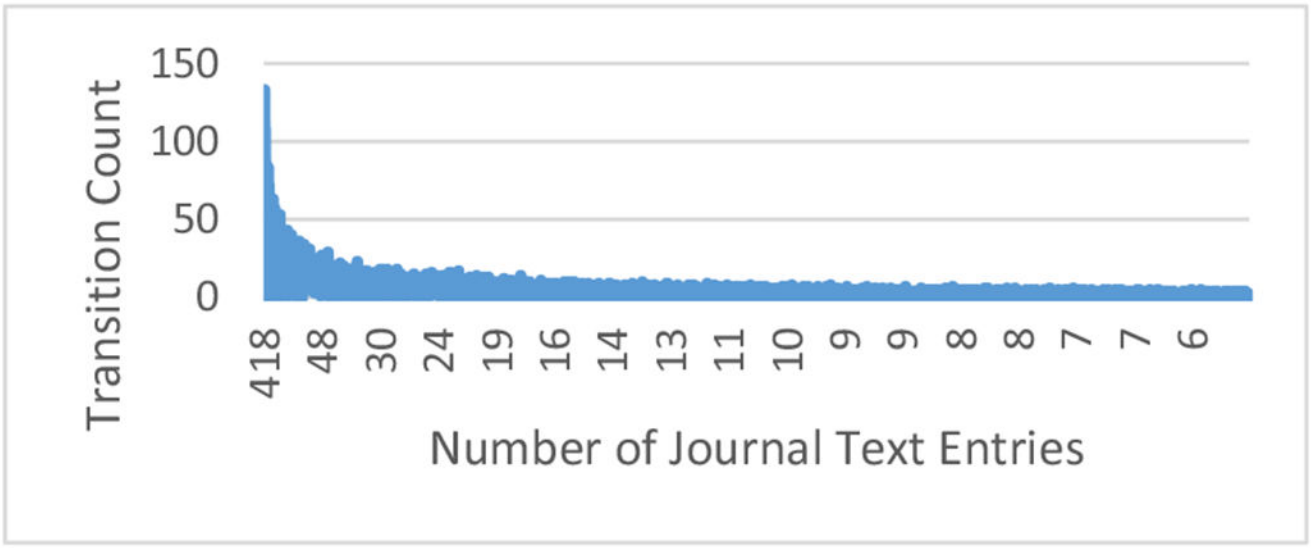

Figure 2–.
Line graph showing the relationship between the journal text entries posted and transitions between stages of change

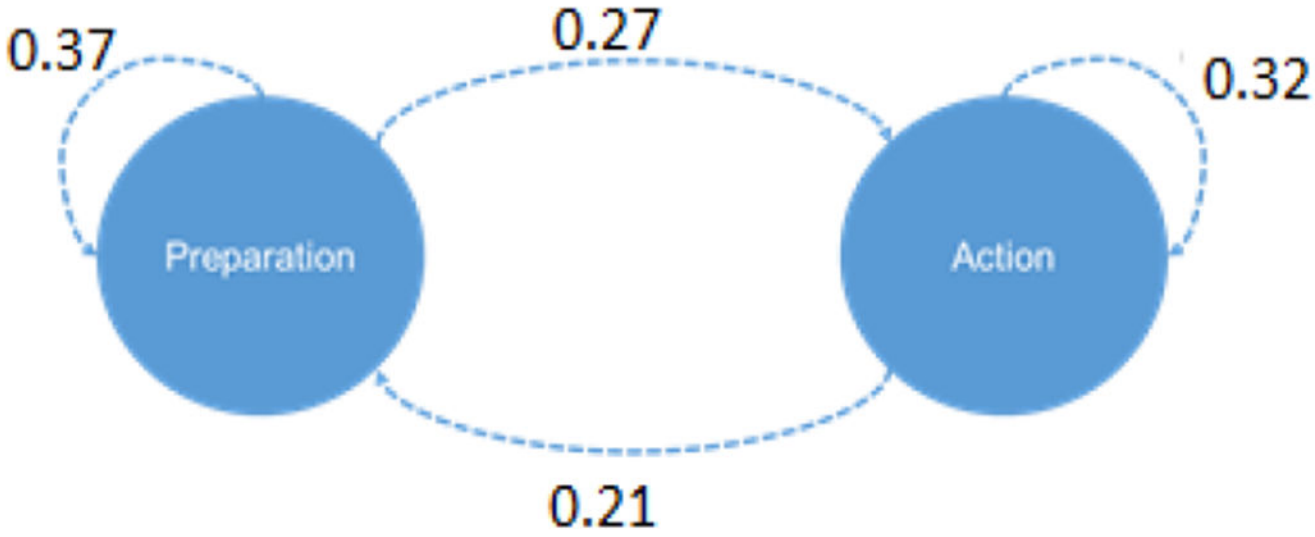

Figure 3–.
A two-state Markov chain representing the transition between stages of behavior change

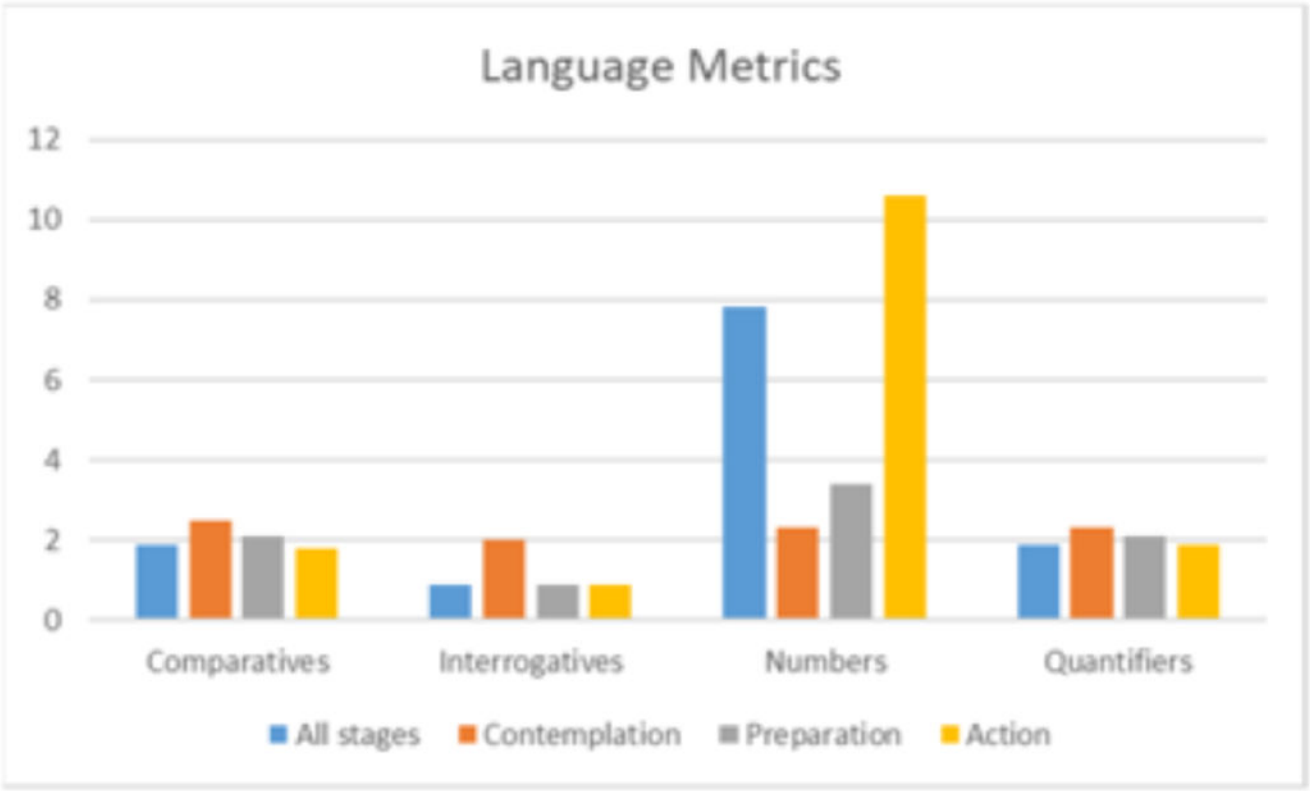

Figure 4-
Comparison of means for language metrics

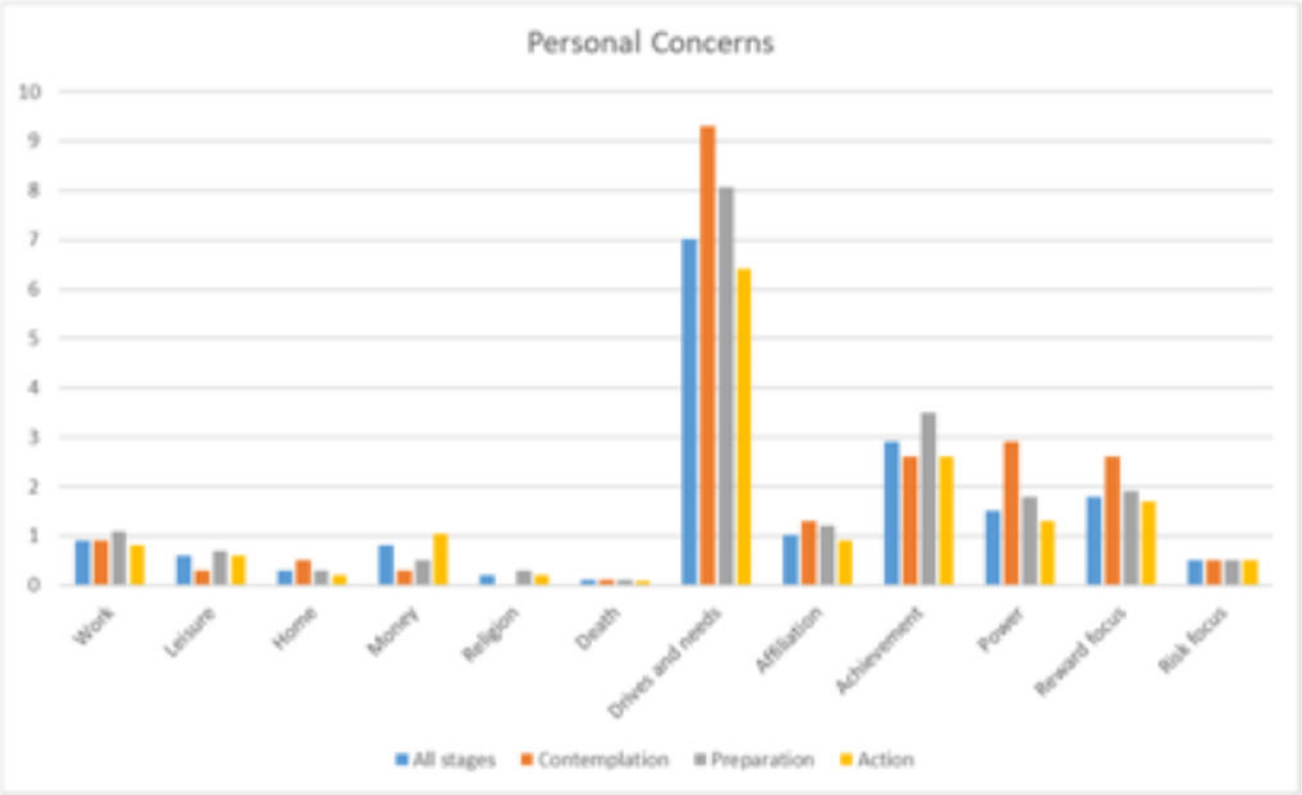

Figure 5-
Comparison of means for personal communication topics in different stages of behavior change

Table 1–

Annotation guidelines

Stage in TTM model	Characteristics of this stage as per TTM model	Patient's View / State of mind
Pre-contemplation (Not Ready)	People in this stage are not ready to change in the foreseeable future like for another 6 months.	Not thinking about change, Feeling of no control, Denial: does not believe it applies to self
Contemplation (Getting Ready)	People in this stage have the intention to change in the next 6 months.	Weighing benefits and costs of behavior, proposed change
Preparation (Ready)	People in this stage are ready to make a change immediately, they have a plan of action in place.	Experimenting with small changes
Action	In this stage people have made specific overt modifications in their lifestyles within the past six months.	Not all modifications count as Action in this model. Total abstinence is what counts as an action compared to switching to low nicotine cigarettes.
Maintenance	In this stage people have made specific overt modifications in their lifestyles and are working to prevent a relapse.	Maintaining a new behavior over time, has quit for over 6 months
Relapse	People in this stage have started smoking again following a quit attempt.	Experiencing a normal part of the process of change, usually feels demoralized

Table 2–

Characteristics of an unlabeled text corpus

Characteristics of the corpus	Frequency
Total number of journal text entries	111,004
Total number of unique users	26,441
Mean age of the users	40.3 years
Total number of females	18,614 (~71%)
Total number of males	7,614 (~29%)
Average length of the journal entry	150 words

Table 3–

Journal text entries labeled by manual coding

Stage of Change	Journal text entry example
Contemplation	I had originally set my quit date at Feb. 7 2003. I got to thinking about it and decided why wait?
Preparation	I was going to quit early but something upset me and I ran to the store. Ugh! Will I ever quit. I'm Quitting in 5 days!
Action	Tough day! Made it through! No slips today.4 days, 5 hours, 55 minutes and 17 seconds smoke free. 85 cigarettes not smoked. $33.60 and 15 hours of my life saved! My quit date: 8/15/2011 2:00:00 PM

Table 4–

Classification report for machine learning models

Machine Learning Models	Precision	Recall	F1-score
LR	0.69	0.70	0.69
SVM (Linear)	0.71	0.72	0.72
RF	0.91	0.90	0.90

Table 5–

Journal text entries labeled by the Random Forest model

Stage of Change	Journal text entry example
Contemplation	just hooked up on the quitnet today… checking the site out… seeing what's available… so far so good… let my younger brother know i'd joined up by sending him a q-card… maybe he'll join up too…
Preparation	Reasonable nights sleep. crazy day. headache, a few cravings. no exercise. just work until 6 p.m. 1 nic gum.
Action	last night was bad; no sleep; up every freakin hour…hope day 2 is better. 1 day, 14 hours, 6 minutes and 31 seconds smoke free. 24 cigarettes not smoked. $5.62 and 4 hours of my life saved! my quit date: 7/26/2010