HHS Public Access

Author manuscript

HT ACM Conf Hypertext Soc Media. Author manuscript; available in PMC 2015 December 02.

Published in final edited form as:

HT ACM Conf Hypertext Soc Media. 2015 September ; 2015: 139–148. doi:10.1145/2700171.2791247.

Characterizing Smoking and Drinking Abstinence from Social Media

Acar Tamersoy, **Munmun De Choudhury**, and **Duen Horng Chau**

College of Computing, Georgia Institute of Technology

Acar Tamersoy: tamersoy@gatech.edu; Munmun De Choudhury: munmund@gatech.edu; Duen Horng Chau: polo@gatech.edu

Abstract

Social media has been established to bear signals relating to health and well-being states. In this paper, we investigate the potential of social media in characterizing and understanding abstinence from tobacco or alcohol use. While the link between behavior and addiction has been explored in psychology literature, the lack of longitudinal self-reported data on long-term abstinence has challenged addiction research. We leverage the activity spanning almost eight years on two prominent communities on Reddit: StopSmoking and StopDrinking. We use the self-reported "badge" information of nearly a thousand users as gold standard information on their abstinence status to characterize long-term abstinence. We build supervised learning based statistical models that use the linguistic features of the content shared by the users as well as the network structure of their social interactions. Our findings indicate that long-term abstinence from smoking or drinking (~one year) can be distinguished from short-term abstinence (~40 days) with 85% accuracy. We further show that language and interaction on social media offer powerful cues towards characterizing these addiction-related health outcomes. We discuss the implications of our findings in social media and health research, and in the role of social media as a platform for positive behavior change and therapy.

Keywords

addiction; social media; smoking; drinking; abstinence; health; well-being; Reddit

1. INTRODUCTION

Health and well-being challenges such as smoking, alcoholism, and impulsive eating are known to be influenced by individuals' social environment [13], which are moving online, as social media sites become more popular. Indeed, the use of social media for health-related discourse have increased sharply in recent years [12]. Such use acts as a constantly available and conducive source of information, advice, and support, as well as known to foster positive behavior change [20]. Meanwhile, this new social interaction paradigm has begun

to provide us with an opportunity to observe individuals' psychological states and social milieu, often in a real-time, longitudinal fashion.

This paper focuses on the health challenge of addiction, specifically addiction to tobacco or alcohol. Alcohol and tobacco are among the top causes of preventable deaths in the United States [25]. In addition to contributing to traumatic death and injury, alcohol is associated with chronic liver disease, cancers, acute alcohol poisoning, and fetal alcohol syndrome. Similarly, smoking is associated with lung disease, cancers, and cardiovascular disease [17]. Achieving long-term abstinence of tobacco or alcohol is difficult [36]—most abstainers are known to relapse within one to three months of cessation. In fact, many individuals who want to quit have been observed to go through short phases of relapse and cessation [14]. While there is a rich body of research on identifying factors associated with such short-term relapse or cessation [35, 37, 28], limited research examines the cues associated with long-term abstinence. This is largely due to the difficulty in compiling high quality self-reported data on abstinence from suitable populations, spanning over long periods of time.

In this paper, we examine how social media language and interactions may be leveraged to *characterize long-term abstinence from tobacco or alcohol*. As of May 2013, 72% of online adults use social networking sites; the number is more than 80% for individuals under the age of 50.[1] Based on reports from the Centers for Disease Control and Prevention (CDC), this demographic aligns well with the age group in which heavy smoking and/or drinking are prevalent [34]. This suggests that social media may be a viable platform for mining cues associated with abstinence.

To this end, we focus on two prominent smoking and drinking abstinence communities on the social media site Reddit: StopSmoking[2] and StopDrinking[3]. These two communities together consist of more than 65,000 subscribed Reddit users as of April 2015, and as described on their public pages, serve as "a place for Reddit users to motivate each other to control or stop smoking/drinking". A participating user may request to have a "badge" (see Figure 1) that indicates self-reported information about the duration of their smoking/alcohol abstinence. The badges are dynamically updated in the system on a daily basis, unless the users request a change to their badges. The main contributions of this paper include:

- We collect and study a novel dataset from Reddit that describes 1,153 users' self-reported information on their duration of smoking or drinking abstinence via the badges. We use the badge information to identify short-term and long-term abstainers.
- We formulate and identify the key linguistic and interaction characteristics of short-term and long-term abstainers based on activity spanning eight years, from 2006 to 2014.

[1] http://www.pewinternet.org/data-trend/social-media/social-media-use-all-users/
[2] http://www.reddit.com/r/StopSmoking
[3] http://www.reddit.com/r/StopDrinking

- We build a supervised learning framework based on the characteristics above to distinguish long-term abstinence from short-term abstinence with over 85% accuracy, 88% precision, and 82% recall.
- Our findings present a number of significant discoveries that may help researchers better understand the role of social media language and interactions in assessing and determining tobacco or alcohol use. We find that:
  - the nature of affect manifested in Reddit posts and comments as well as the tenure of participation in Reddit communities are indicative of short-term or long-term abstinence;
  - the network properties of the users (e.g., indegree) based on their interaction patterns also bear significant explanatory power towards characterizing these addiction-related health outcomes.

We note here that our goal in this work is not to predict future success or failure in abstaining from tobacco or alcohol use. That is, we do not attempt to predict which individual will transition from being a short-term abstainer to long-term abstainer or will relapse while being a short-term or long-term abstainer. Rather, we study a set of successful abstainers and attempt to characterize the attributes of long-term smoking or drinking abstinence from social media. Through such characterization, we evoke the potential use of social media towards addressing public health challenges, in particular addiction to tobacco or alcohol.

2. BACKGROUND AND PRIOR WORK

2.1 Behavioral Science and Addiction

Clinical research on addiction shows that decreased psychosocial stress is associated with transitions from smoking to abstinence [30]. Smokers who fail to quit or relapse after a short period report high levels of stress prior to initial abstinence or at one, three, and six months after cessation [36]. Additionally, recent work analyzing the size and structure of individuals' social networks has found that their connections and interactions therein are related to health-related behaviors and goals [5]. Availability of a strong, trusting network of friends can provide practical and emotional support, which can reduce their smoking or drinking urges [22, 6].

The findings of this extensive body of research provide evidence on the relationship between behavior and addiction. However, they rely heavily on small, often homogeneous samples of individuals, not necessarily representative of the larger population. Furthermore, these studies are typically based on surveys, relying on retrospective self-reports about mood and observations regarding addiction episodes. This method limits temporal granularity as it involves recollection of historical facts. Some of these limitations are circumvented through the use of wearable sensors and other electronic equipment that capture behavioral and affective data in real time without explicit intervention [35]. However, these methods are often expensive and intrusive because they need participants to use the equipment over a period of time.

As such, most behavioral science research on substance abuse has focused on relapse [31, 24, 37]. In fact, few population-based cohort studies have examined long-term abstinence (a year or more) among former smokers or alcoholics. It is important to quantify the relationship between the duration of abstinence and the likelihood of continued abstinence for the evaluation of ongoing public health interventions and the design of smoking or drinking cessation programs. Additionally, understanding factors associated with long-term abstinence is critical due to the high rate of relapse—most individuals attempting to quit tobacco or alcohol abuse go through multiple short-term phases of abstinence and relapse [13].

Our research specifically tries to address this problem. We develop computational approaches that can characterize the attributes of long-term smoking or drinking abstinence from social media. We derive a promising non-intrusive way to examine psychosocial attributes associated with long-term health outcomes by analyzing longitudinal and fine-grained activity in online communities.

2.2 Social Media, Health, and Addiction

Social media research has indicated that individuals' psychological states and social support status relating to health and well-being may be gleaned via analysis of language and conversational patterns. These include utilizing social media, largely Twitter, to understand conditions and symptoms related to diseases [33], cyber-bullying and teenage distress [10], postpartum depression [8], mental health [9, 32, 19, 7], obesity and public health [1], exercise and mental health [11]. Broadly, this body of work investigated the role of linguistic attributes in describing or predicting health challenges.

We extend this body of research by examining the role of both language and social interactions gleaned from social media. Specifically, we build statistical language models that go beyond dictionary approaches. Additionally, we explore how network measures (e.g., indegree, neighborhood density, centrality, etc.) derived out of social interactions may bear explanatory power in the context of tobacco or alcohol addiction. Furthermore, we focus on Reddit, which remains underexplored in comparison to other social media platforms like Twitter.

There has been some research examining addiction behavior manifested on social media, however this body of work is limited. Relationship between displayed alcohol use on Facebook and self-reported information on alcohol abuse was examined in [26, 3, 27]. The authors in [4] explored sentiment manifested by individuals in Twitter by following a pro-marijuana profile. The structure of social circles of prescription drug abusers was investigated in [16]. Using Twitter, the authors in [29] examined perceptions of tobacco products. Another work conducted a study examining characteristics of individuals who express a desire to quit smoking on Twitter [28]. More recently, researchers have studied the prescription drug abuse recovery community Forum77 [23]. In a method similar to [28], they identified dictionary-based linguistic attributes of individuals in various phases of recovery, and were able to characterize recovery trajectory of these individuals.

With the exception of [27] and [23], none of the above pieces of research focuses on predicting health challenges related to addiction. Furthermore, it is important to note that, the ground truth labels on recovery in [28] and [23] were obtained via crowdsourcing. Simply looking at social media posts may not always allow third-party judges to reliably capture abstinence status. Additionally, reasons such as idiosyncratic or personal usage patterns of social media as well as differential social norms and stigma may motivate or preclude some individuals from explicitly reporting abstinence information in social media content. Hence, self-reported abstinence information is extremely valuable. In this paper, we leverage self-reported abstinence information on smoking and drinking.

3. DATA

We begin with a short overview of Reddit. Reddit is a highly popular social media platform, where the users are often referred to as “redditors”. They can submit content in the form of link posts or text posts. Posts are organized by areas of interest or subcommunities called “subreddits”. For instance, some popular subreddits are r/Politics, r/programming, and r/science.[4] Redditors can engage on a post via “upvotes” or “downvotes”; the post’s *score* is the difference between these two quantities. They can also post comments on a post and respond in a comment thread. Over time, redditors accrue reputation in two forms: *link karma* and *comment karma*. Link karma is proportional to the difference between the upvotes and downvotes in all the link posts users made. Comment karma refers to the same difference for all their comments. In 2014, Reddit had 71 billion page views, over 8,000 active communities, 55 million posts, and 535 million comments.[5]

In this paper, we focus on the following two self-improvement subreddits: StopSmoking and StopDrinking. We refer to them as SS and SD, respectively. Both subreddits host *public* content that can be viewed without a Reddit account. At the time of the writing of this paper, SS had 33,690 subscribed users, while SD had 25,542 subscribed users.

As we described before, both subreddits allow users to acquire “badges” to help track their abstinence progress (see Figure 1). Such badges are subreddit-specific, and are displayed next to the username whenever the user posts or comments on the subreddit (ref. Figure 1). Both SS and SD identify different stages of abstinence inside the badge icon (e.g., smiley face for “under one week”), although the actual number of days of abstinence is reported next to it as well.

Typically, a user makes a badge request to the moderators of the subreddit they are interested in, through the subreddit’s interface or by privately messaging the moderators. Badges are then awarded by the subreddit moderators either manually (SD) or automatically through an application known as “badgebot” (SS). Both subreddits are heavily moderated and follow a set of guidelines. For instance, SD cautions against providing medical advice on the forum, conducting surveys, or advertising links to recovery centers.

[4]Subreddits are typically referred to with the prefix “r/”. We omit the prefix when no ambiguity arises.
[5]http://www.redditblog.com/2014/12/reddit-in-2014.html

3.1 Data Collection

We used Reddit's official API[6] to collect posts, comments, and associated metadata from the subreddits. Our data collection proceeded in three phases.

Phase 1—We collected a sample of users in SS and SD. The Reddit API limits crawling historical posts on a subreddit to the past 1,000 posts, so we obtained the most recent 1,000 posts from each of the two subreddits. The crawl took place in November 2014. For each post, we collected the title of the post, body or textual content, ID, timestamp, author ID, author's comment and link karmas, and score of the post. We collected the same information for each comment on the post as well. We then used the API to obtain the badge value of the post author and each of the comment authors, if available.

Phase 2—We extracted the list of unique authors of the posts and comments who had a badge. This gave us 1,859 users for SS and 1,383 for SD (ref. Table 1). The distributions of the SS and SD users across the various abstinence stages displayed in the badges are shown in Figure 2. The badge values of these users were eventually used to construct ground truth data on smoking and drinking abstinence, which we will discuss below. We purposefully excluded the users for whom the API did not return any badge value. No badge information meant that we did not know about their smoking or drinking abstinence status at the time of the crawl.

Phase 3—For users with badges, we collected their posts, comments, and associated metadata, this time across Reddit. Note that these posts and comments could have been shared on any subreddit, outside of SS/SD. Like before, for every user, the Reddit API limits crawling to the most recent 1,000 posts or comments shared by the user. Using this method, we obtained 86,835 posts and 766,574 comments for the 1,859 SS users, and 59,201 posts and 492,573 comments for the 1,383 SD users.

We report the summary statistics of the crawled data in the "All data" columns for SS and SD in Table 1. Also important to note here that, per our crawl, each user in the dataset had a recent post or comment in SS/SD, therefore our dataset is likely to be free of any users who stopped being active in SS/SD and do not pay attention to their badges therein.

3.2 Ground Truth Creation

We constructed ground truth information on smoking and alcoholism abstinence from the crawled badges of the users. Since the badge information is self-reported, we consider it as a reliable, high-quality signal of a user's abstinence status. While characterizing the different abstinence statuses would be insightful, the skewness in the number of users among the different abstinence stages and the sparsity of users per stage (see Figure 2) debarred us from pursuing this direction. Instead, we examined whether we could utilize Reddit activity and interaction of users towards a binary classification task—determining whether a user is likely to belong to the short-term abstinence category or to the long-term abstinence category, given his or her historical data.

[6] http://www.reddit.com/dev/api

To identify the suitable durations to qualify for short-term or long-term abstinence, we leverage the cumulative distribution functions (CDFs) of abstinence duration obtained from the badges in SS and SD (Figure 3). The CDFs show stable patterns before the 30 percentile and after the 70 percentile. The 30 percentile mark for SS is 43 days while it is 44 days for SD; the 70 percentile mark is 350 days and 333 days for SS and SD, respectively. Prior research in addiction [36] indicates frequent relapse to happen at 1-2 months after quitting, which aligns with our 30 percentile mark. Furthermore, individuals who successfully abstain from smoking/alcohol for a year or more have been found to be less likely to relapse in the future [37]. Therefore, we consider the users within the 30 percentile mark to be the short-time abstainers and those beyond the 70 percentile mark to be the long-term abstainers.

This categorization gave us 635 users in SS (318 users/50.07% long-term abstainers) and 533 users in SD (268 users/50.28% longterm abstainers). In the rest of this paper, we use this user set for the task of characterizing long-term abstinence from tobacco or alcohol. Summary statistics on these users can be found in the “Ground truth data” columns for SS and SD in Table 1.

4. STATISTICAL METHOD

We now present the statistical method we employ to characterize long-term abstinence from tobacco or alcohol. For this goal, we introduce the variables outlined below and summarized in Table 2.

Response variable

Our binary response variable represents if a user is a *short-term* or a *long-term* abstainer of smoking/drinking.

Explanatory variables (Language)

Our first set of explanatory variables focuses on extracting linguistic attributes from a user’s posts and comments in SS/SD. Here, we converted the textual content of all the posts and comments in SS/SD to lowercase and extracted the top-100 most frequent unigrams, bigrams, and trigrams (three sets of 100 items each) following the conventional bag-of-words model.[7] These 300 n-grams do not include any phrase that is solely comprised of stopwords. We introduce a count variable for each n-gram, representing the total number of times that the corresponding n-gram appears in the user’s posts or comments.

As another dimension of language, we also consider the sentiment of the posts and comments with VADER [21]. VADER is a lexicon and rule-based sentiment analysis tool that is tailored to specifically detect sentiment expressed in social media. Using VADER, we introduce four variables that correspond to the mean and median of the positive sentiment (PA) and negative sentiment (NA) scores of a user’s posts and comments in SS/SD. Together, this set of explanatory variables contains 304 variables and we refer to them as the *language variables*.

[7]Our statistical models suffered from high dimensionality when we considered more than 300 n-grams.

Explanatory variables (Addiction)

Our second set of explanatory variables focuses on the content (posts or comments) shared by a user in subreddits other than SS/SD (we henceforth refer to this set of subreddits as OSR).[8] To examine if smoking or drinking related content in OSR could potentially help characterize long-term abstinence, we complied two addiction-related lexicons for smoking and drinking based on words in Urban Dictionary[9]. Urban Dictionary is a suitable choice due to the informal nature of online language. Specifically, we utilized a snowball approach in which we seeded the dictionary searches with "smok*" and "alcohol*". We followed the "related words" returned by the dictionary results on these two seed words. We recursively adopted this approach over three more iterations. The final two lexicons are shown in Table 3. Since a user is unlikely to use every word in the lexicon, we consider a single count variable that represents the total number of times that any of the words in the lexicon appears in the user's posts or comments. We also introduce four variables that correspond to the mean and median of the PA and NA scores of the users' posts and comments in OSR—we again use VADER for this purpose. This set of explanatory variables contains 5 variables and we refer to them as the *addiction variables*.

Explanatory variables (Interaction)

Our third set of explanatory variables focuses on the various aspects of interaction.

(1) Activity measures—We introduce variables for the number of posts and comments in SS/SD and OSR, the mean and median differences in hours (Δ) between consecutive contents in SS/SD and OSR, the mean and median content scores in SS/SD and OSR, the mean and median content lengths (in characters) in SS/SD and OSR, and the user's link and comment karmas. Also, we include variables that represent the number of days since the earliest and latest contents (tenure and recency, respectively) in SS/SD and OSR.

(2) Participation in related subreddits—Since abstainers might seek support from or contribute to other subreddits as well, we also extracted the list of the 100 most widely used subreddits, other than SS and SD themselves, based on the posts and comments of the users. Two researchers familiar with Reddit thereafter individually scanned the list to rate their relevance to our task. Researchers referred to prior addiction literature during this task to identify behavioral attributes associated with smoking/alcohol addiction [6]. Sub-reddits with the following characteristics were deemed relevant—emotional discourse subreddits (e.g., r/depression), religious discourse subreddits (e.g, r/Buddhism and r/atheist), fitness subreddits (e.g., r/Fitness), and subreddits on other types of addiction and recovery (e.g., r/cripplingalcoholism). Abstainers are known to engage to greater emotional expression, including personal and subjective topics like religion [30]. Fitness and exercise are also known to be a helpful characteristic of abstinence [6].

[8] SD (SS) becomes an OSR when we focus on smoking (drinking).
[9] www.urbandictionary.com

The final set of related subreddits considered here are shown in Table 4. For each of these subreddits, we introduce a count variable that represents the total number of posts and comments that the user made in the corresponding subreddit.

(3) Graph measures—To further quantify the interaction between the users in SS/SD, we construct a network based on the users' posting and commenting patterns in SS/SD. Specifically, if user A comments on user B's post or comment, we establish a directed edge with a weight of 1 from user A to user B in the network. The total weight of an edge denotes the number of "directed" interactions between the corresponding users. We introduce several graph-centric variables, representing a user's local and global relations with other users in SS/SD: the indegree, outdegree, and degree; reciprocity, the number of triangles to which the user participates (#triangles), and clustering coefficient; the betweenness, closeness and eigenvector centralities; and the number of users in the strongly (SCC) and weakly connected components (WCC) to which the user belongs. Note that for #triangles, clustering coefficient and the centrality measures, we consider an undirected network in which an edge exists only if it appears in both directions in the original network. We refer the reader to [2] for the details of these measures. This set of explanatory variables contains 48 variables and we refer to them as the *interaction variables*.

Statistical models

We employ Ridge regression [18] to classify our binary response variable (short-term or long-term smoking/drinking abstinence). Most of our explanatory variables correspond to English phrases, which posit the collinearity (i.e., excessive correlation between phrases) and sparsity (i.e., some phrases occurring infrequently) properties. Ridge regression guards against problems related to collinearity and sparsity by shifting the weights of the correlated and sparse variables to the more explanatory ones. We use 10-fold cross-validation to determine the best tuning constant that controls the strength of the ridge penalty and also to prevent overfitting to the dataset.

To understand the explanatory powers of our independent variables, we consider three statistical models: (i) the Language model, (ii) the Language + Addiction model, and (iii) the Language + Addiction + Interaction model, which consist of (i) the language, (ii) the language and addiction, and (iii) the language, addiction, and interaction variables, respectively. The first two models are motivated from prior work [28, 23], and through the third, we examine the additional role of interaction in characterizing abstinence. In these models, we represent each user as feature vectors that are standardized to zero mean and unit variance.

5. RESULTS

In this section, we present the results of our two tasks: characterizing long-term abstinence from tobacco and from alcohol.

5.1 Deviance Results

To evaluate the goodness of fits of our three models, namely Language, Language + Addiction, and Language + Addiction + Interaction, we use *deviance*. Briefly put, deviance

is a measure of the lack of fit to data, hence lower values are better. It is calculated by comparing a model with the saturated model—a model with a theoretically perfect fit, which we consider to be the intercept-only model and refer to as *Null*. Table 5 provides a summary of the different model fits. Due to the randomness introduced by cross-validation, we ran our models 10 times and here we report the results corresponding to the lowest deviances that we obtained in any of the runs.

Compared to the Null models, we observe that all three of our models provide considerable explanatory power with significant improvements in deviances in both SS and SD. The difference between the deviance of a Null model and the deviances of the other models approximately follows a χ^2 distribution, with degrees of freedom equal to the number of additional variables in the more comprehensive model. As an example, comparing the deviance of Language with that of Null in SS, we see that the information provided by the language variables has significant explanatory power: $\chi^2(304; N = 635) = 880.3 - 438{:}9 = 441{:}4$; $p < 10^{-6}$. This comparison with the Null model is statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.01}{3}$ since we consider three models). We observe similar deviance results for the Language + Addiction and Language + Addiction + Interaction models in both SS and SD, with the latter models possessing the best fits and highest explanatory powers.

From the fits of the Language + Addiction + Interaction models, Table 6 presents the top-30 positive and top-30 negative β values for the variables corresponding to the n-grams and the top-7 positive and top-7 negative β values for the other variables. The variables with negative and positive β values classify a user as short-term and long-term abstainer, respectively. Note that we standardize the feature vectors before regression, hence the β values correspond to standardized features. We do not report the statistical significance of the β values in the form of *p*-values because they are hard to interpret for strongly biased estimates such as those arise from Ridge regression [15].

The contribution of the different explanatory variables in the two characterization tasks is notable. In both, phrases are notable variables that distinguish short-term and long-term abstinence. In fact, the variables that have the highest explanatory power for short-term abstinence in SS/SD are the phrases *"i started"* and *"in the past"*, respectively. We conjecture that the short-term abstainers use these phrases to indicate new intentions: *"i started an attempt on monday…"* and *"it feels great to be sober and have my dark drinking days in the past"*, respectively. Furthermore, the phrases associated with short-term abstinence are related to current sensation, urge, or confession (*"i need to", "i feel"*), and appreciation and acknowledgement of support, perhaps because they are newcomers in the community (*"thanks for the", "thank you"*). E.g., notice the post excerpt below:

> i need to find more friends that don't drink so much

In contrast, the phrases associated with long-term abstinence are mostly about encouragement and boosting morale (*"keep it up", "hang in there"*) and advisory (*"worked for me', "was able to"*):

> for those of you behind me, keep it up! i believe in you!

Examining some of the non-phrase variables with negative β values, we observe that indegree is a strong indicator of short-term abstinence. This is likely because the short-term abstainers' contents are typically support-seeking in nature, which attract responses from a variety of users in the SS/SD communities. The negative sentiment of contents is also a significant indicator of short-term abstinence. We conjecture that this is likely due to the tendency of the short-term abstainers' disclosures about recent failures, challenges, and struggles related to quitting. Addiction literature also indicates that increased negative affect and stress are associated with early abstainers of smoking/drinking [35]:

> i […] struggle with depression and used alcohol to escape from my often difficult reality

Focusing on some of the non-phrase variables with positive β values, we observe that tenure in SS/SD and OSR are strong indicators of long-term abstinence. Prior work has indicated that long-term social engagement has a positive impact on the psychological states of individuals [8]. Hence, we conjecture that longer tenure on Reddit helps keep individuals intending to abstain from smoking/drinking more motivated and focused towards their respective self-improvement goals. Furthermore, users' comment karma characterizes long-term abstinence in SS, suggesting that social endorsement obtained from the greater Reddit community in the form of upvotes possibly motivated individuals to succeed in their abstinence goals.

We also see that the mean content score in SS and the mean positive sentiment of contents in SD are strong indicators of long-term abstinence from smoking and drinking, respectively, which are likely related to the supportive tone expressed in such content. Addiction literature indicates social support to act as a mediator of stress during smoking/drinking urges [35]. E.g., the following excerpt expresses positive sentiment:

> every time when i remember i quit smoking it makes me happy and a little proud

5.2 Classification Results

To evaluate how well our three statistical models distinguish the long-term and short-term abstinence categories, we randomly split the dataset into 90% training and 10% testing partitions. We trained our models only on the training partitions and measured their classification performance on the testing partitions. Due to the randomness introduced by cross-validation, we performed the aforementioned procedure 10 times to obtain accurate performance estimates. Assuming that long-term abstinence is our positive class, Table 7 presents the classification results with respect to the F1 score, accuracy, precision, recall, and specificity metrics. We report for each metric the mean and standard deviation of the 10 values that we obtained from the 10 iterations on the testing sets.

In general, we observe that the best performing model in both SS and SD is Language + Addiction + Interaction, which achieves the mean F1 scores of 0.86 and 0.85 in SS and SD, respectively. Considering the minimum of the values for SS and SD, this model also achieves a mean accuracy of 0.85, a mean precision of 0.88, a mean recall of 0.82, and a mean specificity of 0.88. This model is followed by Language + Addiction and then Language in terms of performance. Not only the mean values of the performance metrics for

Language + Addiction + Interaction are higher than those for the other two models, the ranges of the values are also narrower in Language + Addiction + Interaction (lower standard deviations).

The good performance of Language + Addiction + Interaction is also evident from the receiver operator characteristic (ROC) curves in Figure 4. To obtain the ROC curves, we first sorted the probabilities that the users are long-term abstainers as output by the models in ascending order. We then generated 250 threshold points equidistant in the range [0, 1] and applied them on the probabilities of the users in the testing partitions; for each threshold value, all users with probabilities above that value are labeled as long-term abstainers, or short-term abstainers otherwise. This process generated 250 pairs of true positive (TP) rate and false positive (FP) rate values for each testing partition, plotting the average of the 10 TP rate and FP rate values computed using the same threshold value across the 10 experiments on the testing partitions gave us the ROC curves in Figure 4. We observe from the figure that the performance of Language + Addiction + Interaction is superior to the other two models in both SS and SD in the whole spectrum of the average TP rate and FP rate values.

6. DISCUSSION

6.1 Clinical Relevance

Our findings indicate that linguistic and interaction cues gleaned from activity in SS and SD forums may be used to understand short-term or long-term abstinence tendencies among users. Such ability to proactively identify one's abstinence status may be used to create early warning systems or interventions that are integrated in social platforms. These early warning systems could analyze one's activity on the platform and engage appropriately if the probability of long-term abstinence drops below a certain level. Certainly, such systems could raise ethical and privacy concerns, and must therefore be carefully designed and developed. However, if successful, these systems may be used in clinically meaningful ways that provide great benefits. For instance, an individual may more easily keep track of his or her activities and interactions on a social media platform and share them with a therapist, which may subsequently lead to more effective treatment.

Broadly, tracking the patterns of changes in the explanatory variables we identified could help clinicians, medical professionals, and policy makers better understand people's experiences around long-term abstinence from tobacco or alcohol, and the strategies that may have worked for them. Since, traditionally, it has been challenging to understand and identify factors associated with long-term smoking or drinking abstinence [36], our research can also help identify previously underexplored variables that may contribute towards the success or failure of abstinence.

Finally, and importantly, through our statistical models that identify short-term and long-term abstainers, we can begin to determine the abstinence status of those individuals for whom badge or other self-reported information on abstinence is not available. This can be particularly valuable in bringing in-time help and support to individuals who intend to quit

smoking or drinking and use a social media platform, however have not adopted the practices of accruing badges, imbibed in the two online communities we study.

6.2 Implications for Social Media Research

Design Considerations—We believe our findings have strong design-related implications for social media research. Below, we describe several design ideas inspired by our research, which may help tailor social media platforms to cater to individuals aiming to abstain from smoking or drinking. Literature indicates that individuals desirous of quitting smoking or drinking often go through repetitive phases of cessation and relapse [14]. Hence, new users joining these abstinence communities, or those who have been short-term abstainers may benefit from content on the forum that discusses the challenges and struggles in this early phase. Mechanisms could be created to engage in a conversation with other long-term members on what to expect during this phase, how to combat desires of smoking or drinking urges, or for general positive reinforcement of their abstinence goal.

Post excerpts containing phrases and other linguistic constructs associated with long-term abstinence may also be promoted to users intending to quit smoking or drinking. They may also be directed to connect with other users in the community who have had success in tobacco or alcohol abstinence over a period of time—social support and higher levels of social capital have been known to help individuals fight addiction urges [13]. Moderators of these recovery communities may also direct requests for advice or help to appropriate users in the community who are actively engaged and have had experiences of long-term abstinence. Since we also found that posting activity or commentary in certain other subreddits were associated with long-term abstinence, users may also be recommended to participate in those other communities or forums where they might additionally obtain support for beating addiction urges or gather general positive reinforcement of their desire to abstain from smoking or drinking.

In addition, our work showed that network features derived out of the social interaction offered considerable explanatory power. That is, the presence of a strong support network on the forum is likely to play an important role in encouraging long-term abstinence. As a design idea, newcomers' posts could be promoted to prominent positions in the forums' timelines to attract more attention, increasing their likelihood of receiving responses. In turn, this would broaden engagement of the whole community, decrease user churn, and thereby increase member retention. This could lead to a self-reinforcing positive cycle that attracts and helps increasingly more people.

Furthermore, in these Reddit communities, reputation is associated with "badges" that indicate the duration of abstinence of a user from smoking/drinking. In a way, making such badge information accessible to visitors and users of the forum not only is likely to boost self-esteem because of improved reputation in the community, but also in general, is likely to induce positive feelings towards abstinence, and encourage and inspire others to do so as well.

Uniqueness of Reddit—We also discuss the effectiveness of addiction recovery communities like SS or SD in general. Although many online communities exist to help

individuals in addiction recovery, SS and SD are unique because they encourage long-term abstinence. This is indicated by the fact that almost 50% of the users in our dataset were abstainers for three or more months. We thus believe that participation in these Reddit forums are likely to help individuals adopt a positive attitude and approach towards addiction recovery. Moreover, the ability to be anonymous or pseudonymous can be an additional facilitating element of abstinence—Reddit accounts do not need any personally identifiable information. Users can thus engage in candid and honest discourse, without worrying about the social stigma that often comes with being a victim of addiction. In fact, a considerable fraction (10%) of users in our dataset explicitly only posted on these two subreddits, perhaps indicating that either they are on Reddit simply to participate in these abstinence forums, or have alternate account(s) on Reddit for non-addiction recovery related discourse. Also, even though some of the explanatory variables that we consider in our statistical models are Reddit-specific, our statistical models can be generalized to other social media platforms, especially to those that possess similar attributes implicitly or explicitly (e.g., link karma on Reddit vs. number of retweets on Twitter as a manifestation of a user's reputation on the online platform).

6.3 Limitations and Future Directions

Our work is of course not free from limitations. We acknowledge that generalizations of our work might not be easily applied across large populations or on arbitrary addiction contexts. As we pointed out, SS and SD are specialized self-improvement communities; most likely, individuals who choose to join them are already motivated to quit addiction. Moreover, since these are largely communities of abstainers, it is possible that individuals new to quitting may feel uncomfortable joining the communities or can feel uncomfortable to be participating. Further biases inherent to Reddit exist as well—the average redditor is a 20-something male[10], perhaps more "tech-savvy", and therefore more likely to resort to online platforms to obtain abstinence support compared to the general population. Additionally, since we did not have information on whether the long-term abstainers sought support through offline means, we are limited in the way we evaluate the effectiveness of the particular forums for addiction recovery. We also note that we focused on smoking and drinking addiction recovery, obviously extending our findings to other kinds of addiction (e.g., prescription or recreational drugs) would need additional investigation.

As we also pointed out earlier, an important point to note about this work is that we do not *predict* abstinence of individuals in SS/SD. That is, based on our findings, we are not able to make (causal) claims as to whether someone will continue to abstain smoking or drinking in the future, or will relapse. This requires tracking an individual's activity and their abstinence reports, i.e., the badge values, over time, which construes an important future research direction we intend to pursue. In fact, in prior literature on clinical studies of addiction behavior, use of survival analysis methods have been found to be particularly helpful in forecasting the likelihood of experiencing a relapse. We intend to leverage these statistical approaches in the future to predict smoking or drinking relapse based on social media activities.

[10] http://www.pewinternet.org/files/old-media/Files/Reports/2013/PIP_reddit_usage_2013.pdf

We also note that a known concern with many recovery communities is member retention—failure to recover often demotivates individuals and leads them to leave the platform. While it is challenging to measure the overall retention rate for SS and SD based on our data, the focus on both self-reported abstinence information through badges and the users who had a recent post or comment in SS/SD ensures that we consider a population of individuals who are attempting to abstain from smoking/drinking and continuing to use Reddit. Also, as mentioned earlier, in our ground truth dataset, we had nearly 50% users who are short-term abstainers. However, per our current data, we cannot be sure of the nature of such short-term abstinence—i.e., whether individuals were attempting to quit smoking/drinking for the first time, or it followed a recent relapse experience. This is because Reddit's API allows our program to access only the *current* badge of a user. Hence, we were not able to determine the nature of short-term abstinence of users in our dataset. For instance, we do not know if they had relapsed shortly before, or if they are attempting to quit for the first time. Finally, as Reddit also imposes that only the most recent thousand posts and comments of every user may be retrieved, we were limited in how far back we could go to examine redditors' historical activity.

7. CONCLUSION

We presented a computational framework to understand smoking and drinking abstinence of individuals from social media. We compiled and studied a previously unexplored source of data—activity on the Reddit communities StopSmoking and StopDrinking. We leveraged the badge feature in these forums to construct self-reported ground truth information on the abstinence status of users to characterize long-term abstinence. Our statistical models incorporated a variety of language and interaction attributes to distinguish long-term abstinence from smoking or drinking from short-term abstinence with 85% accuracy. We found that linguistic cues like affect, activity cues like tenure, and network features like indegree to be indicative of short-term or long-term abstinence. Through our findings, we provided insights into how social media may be leveraged to tackle addiction-related health challenges.

References

1. Abbar S, Mejova Y, Weber I. You tweet what you eat: Studying food consumption through twitter. Proc CHI. 2015
2. Aggarwal, CC. Social Network Data Analytics. 1. Springer; 2011.
3. Beullens K, Schepers A. Display of alcohol use on facebook: a content analysis. Cyberpsychology, Behavior, and Social Networking. 2013; 16(7):497–503.
4. Cavazos-Rehg P, Krauss M, Grucza R, Bierut L. Characterizing the followers and tweets of a marijuana-focused twitter handle. Journal of Medical Internet Research. 2014; 16(6):e157. [PubMed: 24974893]
5. Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. New England Journal of Medicine. 2008; 358(21):2249–2258. [PubMed: 18499567]
6. Cook SH, Bauermeister JA, Gordon-Messer D, Zimmerman MA. Online network influences on emerging adults' alcohol and drug use. Journal of Youth and Adolescence. 2013; 42(11):1674–1686. [PubMed: 23212348]
7. Coppersmith G, Harman C, Dredze M. Measuring post traumatic stress disorder in twitter. Proc ICWSM. 2014

8. De Choudhury M, Counts S, Horvitz E, Hoff A. Characterizing and predicting postpartum depression from facebook data. Proc CSCW. 2014
9. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Proc ICWSM. 2013
10. Dinakar K, Jones B, Lieberman H, Picard RW, Rosé CP, Thoman M, Reichart R. You too?! mixed initiative lda story-matching to help teens in distress. Proc ICWSM. 2012
11. Dos Reis VL, Culotta A. Using matched samples to estimate the effects of exercise on mental health from twitter. Proc AAAI. 2015
12. Fox S, Jones S. The social life of health information. Pew Internet & American Life Project. 2009
13. Galea S, Nandi A, Vlahov D. The social epidemiology of substance use. Epidemiologic Reviews. 2004; 26(1):36–52. [PubMed: 15234946]
14. Gilpin EA, Pierce JP, Farkas AJ. Duration of smoking abstinence and success in quitting. Journal of the National Cancer Institute. 1997; 89(8):572–576. [PubMed: 9106646]
15. Goeman JJ. L1 penalized estimation in the cox proportional hazards model. Biometrical Journal. 2010; 52(1):70–84. [PubMed: 19937997]
16. Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An exploration of social circles and prescription drug abuse through twitter. Journal of Medical Internet Research. 2013; 15(9):e189. [PubMed: 24014109]
17. Harwood HJ. Updating estimates of the economic costs of alcohol abuse in the United States: Estimates, update methods, and data. NIH Publication No 98-4327. 2000
18. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1):55–67.
19. Homan CM, Lu N, Tu X, Lytle MC, Silenzio V. Social structure and depression in trevorspace. Proc CSCW. 2014
20. Huh J, Ackerman MS. Collaborative help in chronic disease management: Supporting individualized problems. Proc CSCW. 2012
21. Hutto CJ, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proc ICWSM. 2014
22. Kaskutas LA, Bond J, Humphreys K. Social networks as mediators of the effect of alcoholics anonymous. Addiction. 2002; 97(7):891–900. [PubMed: 12133128]
23. MacLean D, Gupta S, Lembke A, Manning C, Heer J. Forum77: An analysis of an online health forum dedicated to addiction recovery. Proc CSCW. 2015
24. Marlatt, G.; Donovan, D. Relapse prevention: Maintenance strategies in treatment of addictive behaviors. 2. Guilford Publications; 2005.
25. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. Journal of the American Medical Association. 2004; 291(10):1238–1245. [PubMed: 15010446]
26. Moreno MA, Christakis DA, Egan KG, Brockman LN, Becker T. Associations between displayed alcohol references on facebook and problem drinking among college students. Archives of Pediatrics & Adolescent Medicine. 2012; 166(2):157–163. [PubMed: 21969360]
27. Moreno MA, D'Angelo J, Kacvinsky LE, Kerr B, Zhang C, Eickhoff J. Emergence and predictors of alcohol reference displays on facebook during the first year of college. Computers in Human Behavior. 2014; 30:87–94.
28. Murnane EL, Counts S. Unraveling abstinence and relapse: Smoking cessation reflected in social media. Proc CHI. 2014
29. Myslín M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. Journal of Medical Internet Research. 2013; 15(8):e174. [PubMed: 23989137]
30. Niaura RS, Rohsenow DJ, Binkoff JA, Monti PM, Pedraza M, Abrams DB. Relevance of cue reactivity to understanding alcohol and smoking relapse. Journal of Abnormal Psychology. 1988; 97(2):133–152. [PubMed: 3290304]

31. Pagano ME, Friend KB, Tonigan JS, Stout RL. Helping other alcoholics in alcoholics anonymous and drinking outcomes: Findings from project match. Journal of Studies on Alcohol. 2004; 65(6): 766–773. [PubMed: 15700515]

32. Park M, McDonald DW, Cha M. Perception differences between the depressed and non-depressed users in twitter. Proc ICWSM. 2013

33. Paul MJ, Dredze M. You are what you tweet: Analyzing twitter for public health. Proc ICWSM. 2011

34. Schoenborn C, Adams P, Peregoy J. Health behaviors of adults: United States, 2008-2010. Vital and Health Statistics. 2013; 10(257)

35. Shiffman S. Relapse following smoking cessation: A situational analysis. Journal of Consulting and Clinical Psychology. 1982; 50(1):71–86. [PubMed: 7056922]

36. Whitworth AB, Fischer F, Lesch OM, Nimmerrichter A, Oberbauer H, Platz T, Potgieter A, Walter H, Fleischhacker WW. Comparison of acamprosate and placebo in long-term treatment of alcohol dependence. Lancet. 1996; 347(9013):1438–1442. [PubMed: 8676626]

37. Zhou X, Nonnemaker J, Sherrill B, Gilsenan AW, Coste F, West R. Attempts to quit smoking and relapse: Factors associated with success or failure from the attempt cohort study. Addictive Behaviors. 2009; 34(4):365–373. [PubMed: 19097706]

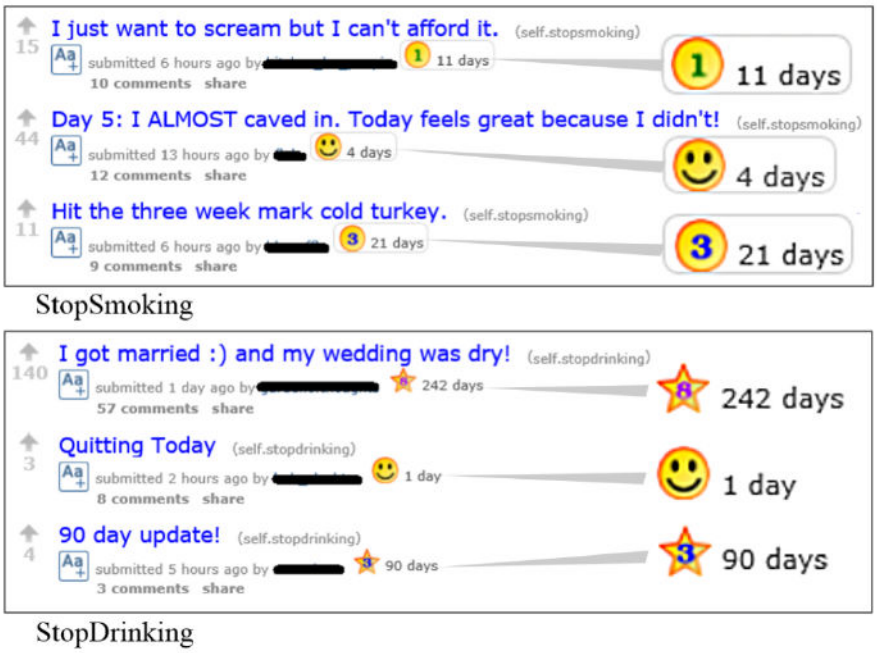

Figure 1.

Examples of the users' abstinence badges on the StopSmoking and StopDrinking subreddits. The abstinence stage is displayed inside the badge icon (e.g., smiley face for "under one week") and the actual number of days of abstinence is reported next to it (e.g., 4 days).

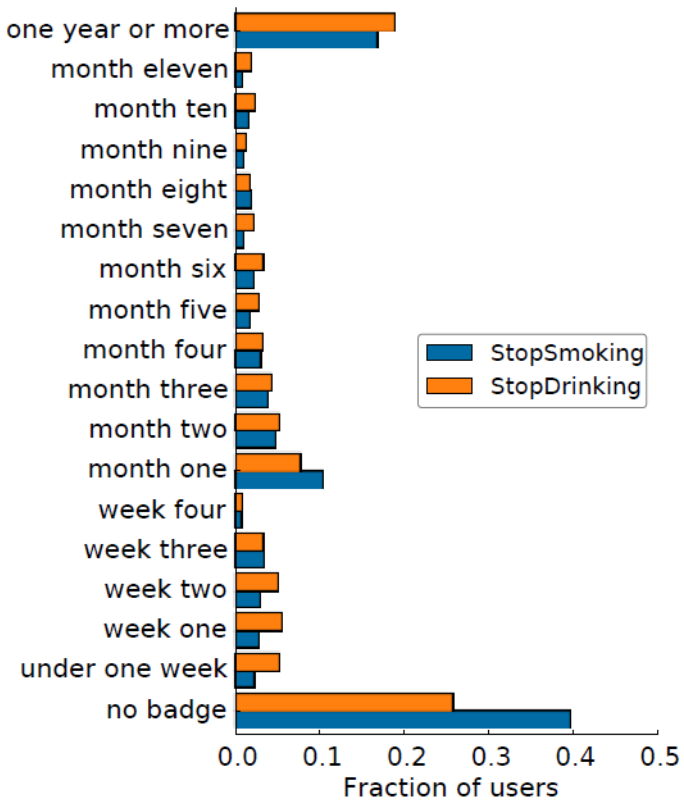

Figure 2.

Distributions of the users in StopSmoking (SS) and Stop- Drinking (SD) across the various smoking and drinking abstinence stages, displayed in the subreddit-specific badges.

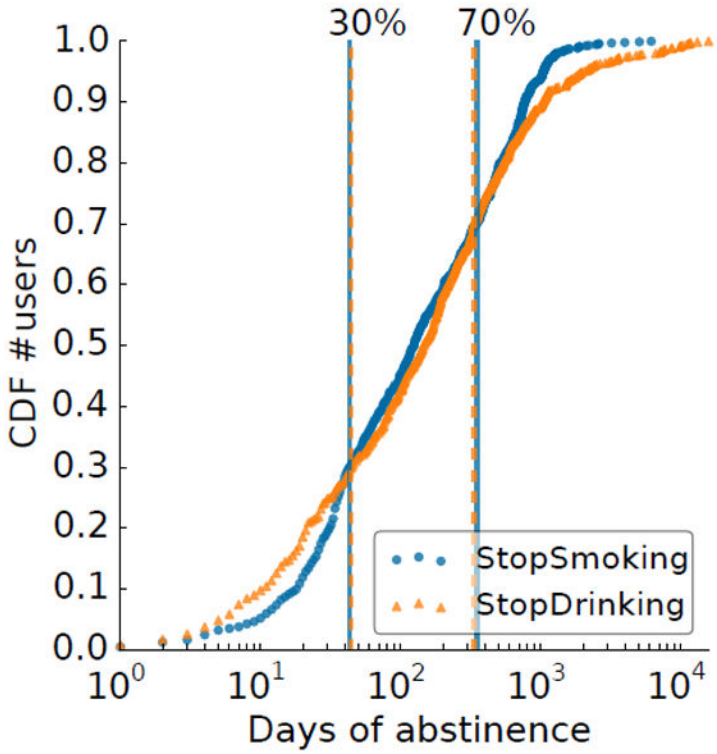

Figure 3.
Cumulative distribution functions (CDFs) of the number of users over the abstinence duration (in days) in StopSmoking (SS) and StopDrinking (SD).

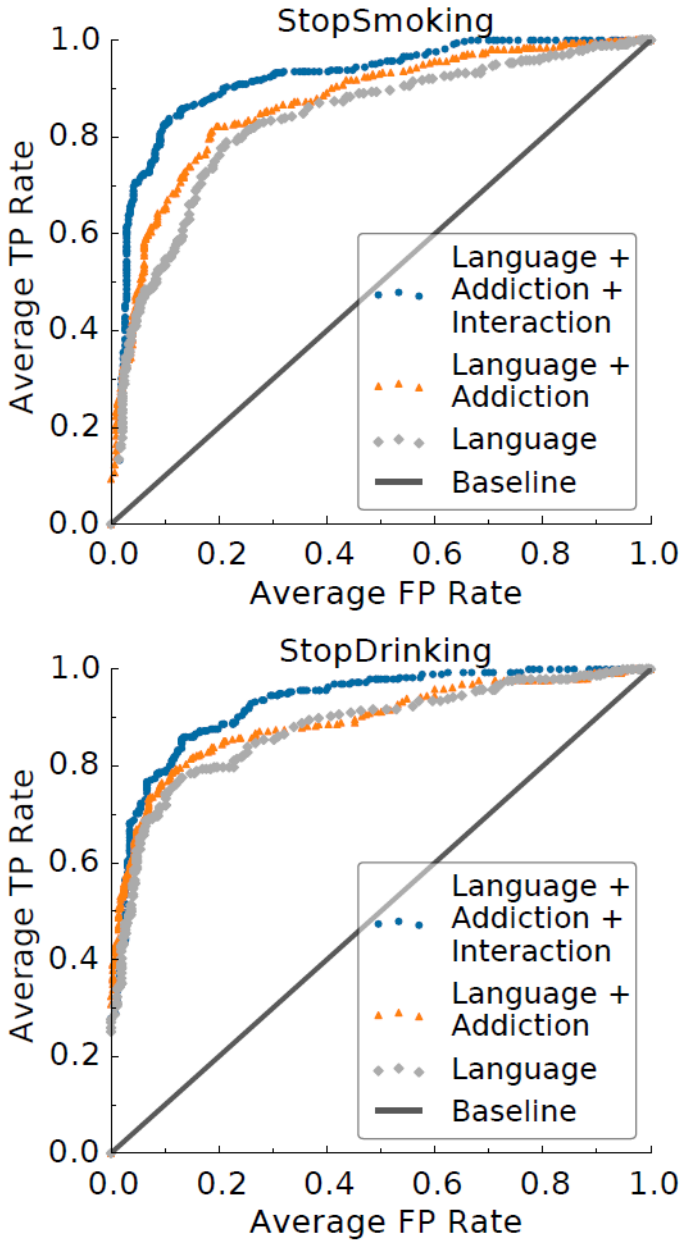

Figure 4.

Receiver operating characteristic (ROC) curves corresponding to the three statistical models for StopSmoking (SS) and StopDrinking (SD). Long-term abstinence is the positive class.

Table 1

Summary statistics of the crawled dataset.

	StopSmoking (SS)		StopDrinking (SD)	
	All data	**Ground truth data**	**All data**	**Ground truth data**
Users	1,859	635	1,383	533
Total posts from users	86,835	36,713	59,201	30,178
Total comments from users	766,574	306,560	492,573	229,656
Date of earliest post	Dec. 09, 2006	Dec. 09, 2006	Feb. 18, 2006	Feb. 18, 2006
Date of earliest comment	Aug. 29, 2006	Aug. 29, 2006	Aug. 02, 2007	Aug. 02, 2007
Date of latest post	Nov. 23, 2014	Nov. 23, 2014	Nov. 23, 2014	Nov. 23, 2014
Date of latest comment	Nov. 23, 2014	Nov. 23, 2014	Nov. 23, 2014	Nov. 23, 2014
Mean / Median comment karma	4,390.2 / 846	5,065.4 / 1,391	3,808.6 / 406	4,610.2 / 745
Mean / Median link karma	1,312.7 / 88	1,626.2 / 201	1,184.7 / 7	1,794.9 / 38
Mean / Median comments per post	6.8 / 5	7.1 / 5	12.6 / 9	13.2 / 9
Mean / Median post score	37.5 / 4	36.9 / 4	34.3 / 5	34.1 / 5
Mean / Median comment score	5.5 / 1	5.5 / 2	5.2 / 2	5.0 / 2
Mean / Median post length in words	55.2 / 15	55.3 / 14	67.5 / 17	62.7 / 15
Mean / Median comment length in words	31.9 / 16	32.6 / 17	36.7 / 18	39.2 / 19

Table 2

List of the explanatory variables used in the statistical models for StopSmoking (SS) and StopDrinking (SD).

Explanatory variables
Language variables:
counts for the 300 uni/bi/trigrams
mean, median PA, NA SS/SD
Addiction variables:
addiction words count
mean, median PA, NA OSR
Interaction variables:
#posts, #comments SS/SD
#posts, #comments OSR
mean, median Δ between contents SS/SD
mean, median Δ between contents OSR
mean, median content scores SS/SD
mean, median content scores OSR
mean, median content lengths SS/SD
mean, median content lengths OSR
link, comment karma
tenure, recency SS/SD
tenure, recency OSR
#contents in each of the 15 related subreddits
indegree, outdegree, degree
reciprocity, #triangles, clustering coefficient
betweenness, closeness, eigenvector centralities
SCC size, WCC size

Table 3

Addiction-related lexicons for smoking and drinking.

Smoking:	acid, alcohol, baked, blaze, blazed, blunt, blunts, bong, bongs, bowl, bowling, bowls, bud, cannabis, chew, chronic, cig, cigar, cigarette, cigarettes, cocaine, coke, crack, dank, dip, doobie, dope, drug, drugs, drunk, ecstasy, fag, ganja, grass, grizzly, herb, heroin, high, hit, hookah, joint, joints, lsd, marijuana, meth, nicotine, party, piece, pills, pipe, pipes, pot, reefer, ripped, roach, school, sex, shit, skoal, smoke, smokes, smoking, snuff, spliff, stone, stoned, stoner, stones, tobacco, toilet, toke, toking, wasted, weed, fucked up, mary jane
Drinking:	acid, alcohol, alcoholic, alcoholism, awesome, bar, beer, beers, beverage, booze, boozing, brew, cocaine, cocktail, coke, college, crack, crazy, crunk, dance, dope, drink, drinking, drinks, drug, drugs, drunk, ecstasy, friends, fucked, fun, girls, hammered, hangover, heroin, high, intoxicated, liquor, lsd, marijuana, meth, parties, party, partying, pills, pissed, pong, pot, rave, rum, sex, shitfaced, shot, shots, smashed, smoke, sober, stoned, trashed, up, vodka, wasted, weed, whiskey, wine

Table 4

Related subreddits—subreddits other than StopSmoking (SS) and StopDrinking(SD) where users post/comment.

Smoking:	StopDrinking, electronic_cigarette, BabyBumps, Fitness, relationships, Christianity, personalfinance, atheism, IAmA, MakeupAddiction, SkincareAddiction, loseit, Frugal, Showerthoughts, Buddhism
Drinking:	REDDITORSINRECOVERY, alcoholism, StopSmoking, relationships, cripplingalcoholism, depression, Christianity, Drugs, CasualConversation, IAmA, atheism, Fitness, MakeupAddiction, electronic_cigarette, DebateReligion

Table 5

Summary of different model fits. Null is the intercept-only model. Deviance measures the goodness of fit. All comparisons with the Null models are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.01}{3}$).

	StopSmoking (SS)				StopDrinking (SD)			
Model	**Deviance**	**df**	**χ^2**	**p-value**	**Deviance**	**df**	**χ^2**	**p-value**
Null	880.3	0			738.9	0		
Language	438.9	304	441.4	$< 10^{-6}$	353.5	304	385.4	10^{-3}
Language + Addiction	418.5	309	461.8	$< 10^{-7}$	340.8	309	398.1	$< 10^{-3}$
Language + Addiction + Interaction	326.9	357	553.4	$< 10^{-9}$	273.2	357	465.7	$< 10^{-4}$

Table 6

β values corresponding to the 74 features with the highest explanatory power for StopSmoking (SS) and StopDrinking (SD). “OSR” stands for subreddits other than SS/SD. The prefix “r/” indicates a related subreddit. “aa” stands for Alcoholics Anonymous.

StopSmoking (SS)				StopDrinking (SD)			
feature	β	**feature**	β	**feature**	β	**feature**	β
indegree	-0.28	tenure SS	0.75	indegree	-0.26	tenure SD	0.83
median content length SS	-0.24	#comments OSR	0.35	closeness centrality	-0.20	#comments OSR	0.25
degree	-0.23	tenure OSR	0.24	median NA SD	-0.16	r/REDDITORSIN RECOVERY	0.24
r/Buddhism	-0.18	mean content score SS	0.20	mean NA OSR	-0.16	mean PA SD	0.18
recency SS	-0.17	comment karma	0.18	r/Fitness	-0.15	tenure OSR	0.16
median NA SS	-0.16	addiction words count	0.18	link karma	-0.15	#posts OSR	0.14
outdegree	-0.16	r/electronic_cigarette	0.14	SCC size	-0.15	r/relationships	0.13
feature (n-gram)	β	feature (n-gram)	β	feature (n-gram)	β	feature (n-gram)	β
i started	-0.31	year	0.32	in the past	-0.33	year	0.33
i need to	-0.26	keep it up	0.27	i’m going to	-0.31	i got sober	0.27
this time	-0.23	think about it	0.21	week	-0.24	months	0.25
i’m going to	-0.23	pack a day	0.20	i know i	-0.18	i quit drinking	0.23
i want to	-0.22	i still	0.19	i need to	-0.17	i don’t drink	0.23
as much as	-0.19	keep it	0.18	day	-0.17	a drink	0.19
trying to quit	-0.19	never	0.18	i need	-0.17	meetings	0.19
thanks for the	-0.19	since i quit	0.18	i feel	-0.16	find	0.19
if you don’t	-0.18	if you want	0.18	i don’t know	-0.14	was able to	0.17
in the morning	-0.18	a year	0.17	to quit	-0.14	years	0.16
feel like	-0.17	worked for me	0.16	and i don’t	-0.14	as much as	0.16
i don’t want	-0.16	you want	0.16	last	-0.13	keep up the	0.15
started	-0.16	going to be	0.16	want to be	-0.13	stay	0.14
the last	-0.13	i would	0.15	the first time	-0.12	stay sober	0.14
try to	-0.13	i smoked	0.15	have a problem	-0.12	in the first	0.13
feeling	-0.13	hang in there	0.15	so much	-0.12	sobriety	0.13
last	-0.13	a non smoker	0.14	back to	-0.12	at a time	0.13

HT ACM Conf Hypertext Soc Media. Author manuscript; available in PMC 2015 December 02.

StopSmoking (SS)				StopDrinking (SD)			
feature	***β***	**feature**	***β***	**feature**	***β***	**feature**	***β***
i want	-0.13	you'll	0.14	don't know	-0.11	still	0.13
thanks for	-0.13	get a	0.14	i'm	-0.11	part of	0.13
you don't have	-0.13	you're	0.14	i can't	-0.11	one day at	0.13
i've	-0.13	so much	0.13	i think i	-0.11	people	0.12
right now	-0.12	keep	0.12	i'm not	-0.11	a time	0.12
2	-0.12	you don't need	0.12	i know that	-0.11	i was drinking	0.12
in the past	-0.12	helped me	0.12	i don't want	-0.11	congrats on	0.12
in my life	-0.11	you quit	0.12	not drinking	-0.11	i really	0.12
to quit smoking	-0.11	it gets	0.12	drinking i	-0.10	i got	0.12
i quit smoking	-0.11	like a	0.12	i've been	-0.09	aa	0.12
able to	-0.11	years	0.12	thank you	-0.09	life	0.12
i got	-0.11	you want to	0.12	i feel like	-0.09	if you don't	0.12
as well	-0.10	a pack a	0.12	i want to	-0.09	you don't want	0.11

HT ACM Conf Hypertext Soc Media. Author manuscript; available in PMC 2015 December 02.

Table 7

Performance metrics corresponding to the three statistical models for StopSmoking (SS) and StopDrinking (SD).

	Language		Language + Addiction		Language + Addiction + Interaction	
Measure	**StopSmoking**	**StopDrinking**	**StopSmoking**	**StopDrinking**	**StopSmoking**	**StopDrinking**
F1 score	0.70 ± 0.06	0.78 ± 0.04	0.78 ± 0.05	0.80 ± 0.05	0.86 ± 0.03	0.85 ± 0.05
Accuracy	0.74 ± 0.05	0.81 ± 0.04	0.80 ± 0.04	0.81 ± 0.04	0.86 ± 0.03	0.85 ± 0.05
Precision	0.81 ± 0.06	0.91 ± 0.07	0.83 ± 0.05	0.91 ± 0.06	0.90 ± 0.04	0.88 ± 0.06
Recall	0.62 ± 0.09	0.69 ± 0.06	0.74 ± 0.06	0.71 ± 0.07	0.82 ± 0.04	0.83 ± 0.06
Specificity	0.86 ± 0.04	0.93 ± 0.04	0.86 ± 0.05	0.93 ± 0.05	0.91 ± 0.04	0.88 ± 0.05