Decision Support Systems 116 (2019) 26–34

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach

Xi Wang[a], Kang Zhao[b,*], Sarah Cha[c], Michael S. Amato[c], Amy M. Cohn[c,d], Jennifer L. Pearson[c], George D. Papandonatos[f], Amanda L. Graham[c,d,e]

[a] School of Information, Central University of Finance and Economics, Beijing, China
[b] Tippie College of Business, The University of Iowa, Iowa City, IA, United States of America
[c] Schroeder Institute, Truth Initiative, Washington, DC, United States of America
[d] Department of Oncology, Georgetown University Medical Center, Washington, DC, United States of America
[e] Cancer Prevention and Control Program, Lombardi Comprehensive Cancer Center, Washington, DC, United States of America
[f] Center for Statistical Sciences, Brown University, Providence, RI, United States of America

ARTICLE INFO

Keywords:
Machine learning
Text mining
Smoking cessation
Online community
Social network

ABSTRACT

Online smoking cessation communities help hundreds of thousands of smokers quit smoking and stay abstinent each year. Content shared by users of such communities may contain important information that could enable more effective and personally tailored cessation treatment recommendations. This study demonstrates a novel approach to determine individuals' smoking status by applying machine learning techniques to classify user-generated content in an online cessation community. Study data were from BecomeAnEX.org, a large, online smoking cessation community. We extracted three types of novel features from a post: domain-specific features, author-based features, and thread-based features. These features helped to improve the smoking status identification (quit vs. not) performance by 9.7% compared to using only text features of a post's content. In other words, knowledge from domain experts, data regarding the post author's patterns of online engagement, and other community member reactions to the post can help to determine the focal post author's smoking status, over and above the actual content of a focal post. We demonstrated that machine learning methods can be applied to user-generated data from online cessation communities to validly and reliably discern important user characteristics, which could aid decision support on intervention tailoring.

1. Introduction

Smoking causes about 20% of all deaths in the United States [1]. The majority of smokers want to quit [2], and over 12 million turned to the internet for information about quitting smoking in 2017 [3]. For health promotion programs, it is important to provide tailored interventions, as they exert positive effects on health behavior change and program participation [4–6]. Users are more likely to attend to content they perceive as being personally relevant, and more likely to remain engaged with interventions that they find satisfying or helpful in achieving their goals. Tailored content is thought to elicit a greater degree of cognitive processing [7,8] as it is more likely to be read, understood, recalled, rated highly, and perceived as credible compared to one-size-fits-all intervention content [8]. Tailored content may also lead to longer and more robust engagement with an intervention [9–12]. Specifically for smoking cessation, tailored information – delivered via print [13] and Internet interventions [14] – has been shown to be effective in helping people quit. Development of automated decision support tools that can accurately identify an individual's smoking status will help designers of Internet cessation interventions better deliver tailored support.

Key constructs involved in tailoring are typically assessed at program initiation or at coarse intervals tied to follow-up assessments [8,15–17]. However, more fine-grained, dynamic tailoring may have important treatment advantages [18]. Knowing when a smoker is planning to quit could enable the timely presentation of skills training content and additional support around a quit date; real-time response to a slip could preclude a full-blown relapse; fluctuating levels of cravings and confidence throughout the cessation process may be important to acknowledge and respond to with varying intervention strategies and

* Corresponding author at: Tippie College of Business, The University of Iowa, S224 Pappajohn Business Building (PBB), Iowa City, IA 52242-1994, United States of America.
E-mail address: kang-zhao@uiowa.edu (K. Zhao).

https://doi.org/10.1016/j.dss.2018.10.005
Received 20 June 2018; Received in revised form 11 September 2018; Accepted 10 October 2018
Available online 15 October 2018

recommendations. Despite the intervention opportunities created by the availability of such metrics, gathering such data from users through traditional survey methods can create an unacceptable response burden, be ignored by users, and/or cause them to abandon an intervention.

An advantage of interventions delivered via the Internet, or Web-based intervention support systems [19], is their ability to unobtrusively gather real-time data [20] that can be used to support this kind of dynamic tailoring with minimal burden on the user. In particular, the proliferation of social networks and online communities represents exciting opportunities for large-scale data analysis and intervention design. Researchers have leveraged data from these online platforms to study user engagement [21,22], predict population-level health status (e.g., influenza) [23,24], and analyze individuals' offline health status (e.g., depression and drug usage) [25,26].

Online communities for smoking cessation are used by thousands of current and former smokers each year who post about their quitting journey, their struggles in staying abstinent, and their achievements and celebrations. Dynamic interventions that change over time in response to a user's pattern of engagement and progress toward their goals have been noted as a promising target for novel systems for behavior change [27–29]. To support decisions for such interventions, important insights into a user's current "state" or potential treatment needs may be discerned through user-generated content throughout their engagement with an online community.

To date, there have been relatively few studies of user-generated content in online smoking cessation communities [30–37]. Previous studies have primarily focused on the prevalence of specific topics of discussion, rather than making person-level inferences about the author. Selby et al. [32] found that the most common theme of first posts was from recent quitters who were struggling with quitting and seeking support or advice. In a forum specifically for recent quitters, Burri et al. [38] found that the most prevalent message type was "giving emotional support" (i.e., messages of solidarity and encouragement) which were replies to posts signaling emotional distress or relapse to smoking. These studies provide an important foundation for understanding the nature of online smoking cessation communities and the ways that members interact and support each other, but many have relied on manual coding of a fraction of the available content. Such an approach cannot capture the complex, dynamic nature of online social ties, or address the multi-faceted nature of large-scale social interactions in online communities. Additionally, manual coding studies have often relied on cross-sectional snapshots of network ties, sentiment, and content that may obscure important changes over time.

Powerful computational methods have been used to mine large volumes of user-generated content in other areas of health [39–45] and have been used to discern smoking status from free-text clinical data in electronic medical records [46,47] and medical discharge summaries [48–50]. However, machine learning methods have only begun to be applied to the rich, real-time user-generated data in online health communities. Whereas classification of free-text from medical records is useful for post-hoc observational studies, the ability to automatically identify users' smoking status at a large-scale and in real-time would help to guide treatment decisions via a post recommender system. For example, post recommenders can push content about success stories and coping with the urge to smoke to users who are still smoking, yet prioritize posts about relapse prevention to those who recently quit. Such a recommender system could also go beyond a user's immediate status and consider their past trajectories. For example, the tailored support provided to a user starting their first quit attempt should be different from the tailored support provided to a user starting their 10th quit attempt. In addition, machine-learning-based identifications of smoking status, especially when used longitudinally, can help community managers decide if the design, management, or content of the community need to be adjusted to achieve better outcomes for community members.

Mining user-generated content from an online smoking cessation community presents several methodological challenges, but also some unique opportunities. Compared to clinical data, the content of online discussions may be quite "noisy" for a variety of reasons. Community members may talk about a wider range of different topics than a clinician recording a discharge summary, many of which may not relate specifically to smoking or abstinence [32,38]. Indeed, off-topic posts are quite common among mature online communities and are often a hallmark of when an online community has evolved beyond an inception phase [43,51]. Noise may also stem from the fact that a post about smoking status could refer to the author of the post or someone else in the community (e.g., congratulating another user's abstinence). In addition, user-generated content often contains informal expressions (e.g., "Day 14 for me!") or community vernacular (e.g., "DOF 145" meaning 145 Days of Freedom from smoking), which may be challenging to interpret and code [52]. Despite these challenges, user-generated content, along with the clickstream data and meta-data that accompany a user's involvement in an online community, are easy to track and readily available. To the extent that user-generated content typically represents interactions among users, additional information about the original post may be gleaned from the ways in which other users react to it. For example, a blog post with many congratulatory replies/comments is likely more indicative of the author's abstinence than a post with few such comments. Using such novel features as classifier inputs, in addition to the text of users' posts, may improve performance over what could be achieved by relying on text alone.

This study aimed to demonstrate the feasibility of using machine learning classifiers to predict smoking status from user-generated content in an online cessation community by extracting novel features. This study focused on classifying a user's smoking status, a key concept that is often used to tailor smoking cessation treatment. We illustrate the effectiveness of combining texts of a focal post with domain-specific features, author features, and thread features in smoking status detection, and discuss other potential applications of this methodology.

Previous machine-learning-based studies on user-generated content from online cessation communities aimed at finding social-support categories [53] or distinguishing users' short-term quits from long-term quits [54,55]. By contrast, in this study, we built predictive models that sought to distinguish smokers who have quit from those who have not. In addition, we identified quit status for all users who have posted. While previous studies have been limited to investigating only users who explicitly declared quit dates in user profiles [54,55], our approach allows a larger proportion of users to be studied and reduces sampling bias. The approach can be applied to other online cessation communities where explicit declarations of quit dates are unavailable, rarely used, or unreliable (e.g., such a date may have been changed multiple times by a user with more than one attempt). Overall, leveraging the largely untapped wealth of information available in online communities to identify users' smoking status in terms of "quit vs not" could inform the development of more powerful interventions tailored in real-time [56]. Also, doing so on all users who have posted, rather than a subsample can have important clinical difference that substantially expands the potential impact of interventions based on this work.

2. Data and methods

2.1. Source of data

The study involved data from BecomeAnEX.org, a publicly available web-based smoking cessation program.[1] BecomeAnEX was developed in collaboration with the Mayo Clinic Nicotine Dependence Center [57] and has had over 800,000 users register on the site since it was

[1] The study protocol for these analyses was reviewed and approved by Chesapeake Institutional Review Board (Pro00010302).

launched in 2008. To register on BecomeAnEX, individuals must agree to the site's Terms of Use and Privacy Policy. The Privacy Policy states that 1) BecomeAnEX collects information about users and their use of the site; 2) Information is used for research and quality improvement purposes only; and 3) Personal information is kept confidential. Thus, de-identified data from all registered users was available for analysis. BecomeAnEX provides problem-solving and coping skills to quit smoking, educates users about cessation medications, and facilitates social support through a large online community. The online community is comprised of thousands of current and former smokers who interact via several asynchronous communication channels [58] (e.g., private messages, blogs, blog comments). Our analyses focused on blogs and blog comments since they are the most popular communication channels and typically comprise of longer and more elaborate posts from users. Thus, our dataset includes 38,156 blog posts and 316,886 blog comments published by 5435 users in the BecomeAnEX community between January 2012 to May 2015.

2.2. Domain expert annotations

Training and evaluating a machine learning classifier requires labeled data so that algorithms can learn differences between instances from different classes. Specifically for this research, we need some posts whose authors' smoking status is labeled. Thus we recruited five long-standing members of the BecomeAnEX community as domain experts with in-depth familiarity with community norms[2] on how users "talk about" smoking status. We randomly sampled 2120 community posts (750 blog posts, 1370 blog comments) which were manually coded by domain experts in accordance with an annotation guide. The random sample was created by first sampling 120 posts for intensive reviews and discussions during creation of the annotation guide, and then four batches of 500 posts each (resulting in the full training set of 2120 posts).

The annotation process has been previously described [59]. Briefly, each post was coded by two domain experts; a study team member served as a tiebreaker for any posts where the two original coders disagreed. Posts were coded for the author's smoking status at the time the post was written. Available codes were "Clearly smoking," "Clearly not smoking," or "Unclear." Coders were instructed to use inference and make their best guess based on the text and subtext of each post, but to use the "Unclear" code whenever they did not feel confident that a reliable judgment could be made. Table 1 shows example posts with each label. Like many other online cessation communities, BecomeAnEX does not offer a structured way for users to declare their own smoking status in their online profiles. However, the validity of the manual annotations, as assessed by comparison with self-report data, was high in previous work (initial Kappa = 0.82, with disagreements resolved by a third tie-breaking coder) [59].

In the current study, those manual annotations were recoded into a binary scheme that emphasized accurate classification of abstinence. "Clearly not smoking" posts were the positive class; "Clearly smoking" and "Unclear" posts were combined into a single category serving as the negative class, so that we have a binary classification problem. The two classes were relatively balanced: the positive class constituting 48% (n = 1015) of all the annotated posts, including 44.4% among blogs, and 55.6% among blog comments. The decision to select "Clearly not smoking" as the positive class was based on treatment implications. Specifically, accurately identifying when a smoker has begun a quit attempt (i.e., their first instance of "Clearly not smoking") would allow a tailored intervention to provide them with relapse prevention support, which qualitatively differs from skills training support that is most appropriate for smokers before they begin a quit attempt. Other applications might be better served by focusing on identification of "Clearly smoking" which could be achieved by adapting the methods applied here.

Table 1
Example posts with each label.

Post content	Label
"Thank u all for ur support and info, the links I have checked out so far already seem very helpful, so I can't wait to finish them all. THE DECISION HAS BEEN MADE! I can beat this.;)"	Clearly smoking
"As of today, I haven't smoked for 32 days!"	Clearly not smoking
"Thank you so much for the information."	Unclear

2.3. Features of user-generated data for machine learning

Finding informative features to differentiate between positive and negative classes is critical for the success of classifications. Our analyses leveraged five sets of features to identify the smoking status for the author of a focal post (Table 2). The first two sets are the standard text feature sets derived from the text of a focal post. One is the unigram features, in which we used the standard bag-of-words model and extracted 5600 unigrams after stop-word removal (e.g., removing "the" and "at") and stemming (e.g., converting "smoking" to "smoke"). These are the most common features for most text mining problems [60] and have been used to classify posts by users' short and long-term quit status [54,55].[3] The other one is the Doc2Vec feature set, a document embedding technique that generates vector representations for documents [61–63]. After varying the size of vectors from 100 to 400, we selected 200 as the best-performing vector size, so that each post was represented with a vector of length 200. These two feature sets serve as the baseline in our study.

While these standard text features are intuitive, they are generic for most text classification tasks and do not take advantage of the specific domain of smoking cessation. Therefore, our second set is made up of domain-specific text features (i.e., selected phrases or n-grams that community members use to report their own smoking status) compiled from the study team. One such feature measures the use of various first-person words, such as "I", "me", "my" and "mine", which was included to ensure a post referred to the author's own smoking status (e.g., "Today is my first day to a new and improved me") instead of someone else's smoking status (e.g., "Congratulations on 200 days of freedom!"). In addition, posts often mention how long the author has been abstinent (e.g., "14 Days of Freedom!", "I quit 3 months ago."). Therefore, we also included the mention of timespan by creating a list of words for time units, including "hour", "day", "week", "month" and their possible variations, such as "hrs" and "days". This list allowed us to check the usage of timespans by matching phrases in the formats of {numeric values + time unit}, such as "5 days", "first month", and "8 hrs", or {time unit + numeric values}, such as "week 3" and "month 2".

In addition to what is expressed within a focal post's own content, characteristics of the author may also be important. On one hand, higher levels of engagement in an online cessation community are often associated with abstinence [58,64,65]. On the other hand, it may be rare for a new member to report abstinence upon joining a community. Therefore, our third feature set is comprised of author-based features that capture the focal post authors' past community activities. For the author of each post, we extracted their length of tenure as a community member, the total number of posts published, and the total number of visits to the community (based on clickstream logs). All three features were calculated from the date a user joined the community until the

[2] These domain experts who annotated posts are not co-authors of this paper, and are different from co-authors who designed domain-specific features.

[3] We also tried to add bigrams of focal posts, but the F-measure and AUC of our classifiers did not improve compared to using unigrams only.

Table 2
Five feature sets in this study.

Feature set	Description
1	Standard text features (unigrams) of focal post content
2	Doc2Vec features from focal post content
3	Domain-specific features from focal post content
4	Author-based features of focal posts
5	Thread-based features of a focal post's thread

time the post was published, and they could vary for the same author if s/he published more than one post over time.

Whereas these first three feature sets are about a focal post and its author, the fourth set of thread-based features goes beyond the focal post and examines the whole "thread" that the post belongs to. Previous studies have shown that thread-based features can improve thread-level classification [66], but have not leveraged them for post-level classification. For each focal post, we first extracted basic features such as type of the post (i.e., the original blog post or comments), number of words in the post, how long the thread was active (i.e., the time difference between the original blog post and the last comment), total number of comments in the thread, and the number of unique users who contributed to the thread.

In addition, the content of neighboring posts may also help to classify the focal post. All other posts and comments in the same thread with a focal post were considered its neighboring posts. Compared to posts in the negative class (i.e., no evidence of abstinence), a focal post with a success story may be preceded and followed by different discussions. For instance, a focal post indicating the author's abstinence status could be surrounded by other posts with similar discussions on smoking status (e.g., "It has been 41 days since my last smoke… I have been smoke-free because of the friends I've made here"), and/or congratulatory posts (e.g., "Congrats to Mary for 50 days!"). However, if we include every word or phrase that appears in neighboring posts, the dimensionality of the feature set will increase greatly. Therefore, we conducted feature selection for text features of neighboring posts. We represented each neighboring post with frequency counts of unigrams. Then, we created two groups of neighboring posts: neighbors of positive-class posts, and neighbors of negative-class posts. Comparing the two groups of neighboring posts, we calculated information gain [67], a popular feature selection method, and picked the top 100 unigrams. To further highlight the discriminating power of each of the top-100 unigrams, we also proposed a class-discriminating term weighting scheme inspired by prototypical words scoring [68]. Eq. (1) defines class-based term frequency $TF_{C,\,i}$, which represents term *i*'s frequency among all neighboring posts of class *C* (positive or negative). It is the fraction between the total number of appearances of term *i* among all neighboring posts of class *C* ($T_{C,\,i}$) and the total number of neighboring posts of class *C* (P_C).

$$TF_{C,i} = \frac{T_{C,i}}{P_C},\ C \in \{+,-\} \tag{1}$$

$$R_i = \frac{TF_{+,i}}{TF_{-,i}} \tag{2}$$

If term *i* appears frequently in neighboring posts of the positive class, but rarely in neighboring posts of the negative class, it will have high $TF_{+,\,i}$ and low $TF_{-,\,i}$, and thus high values in the between-class term ratio R_i defined in Eq. (2). For example, "congrat", the root form of "congratulate" and so on, features a between-class term ratio of 3.11, which suggests that a post containing the unigram "congratulate" and its variants is approximately three times more likely to appear among neighboring posts of a positive-class post than a negative-class post. Meanwhile, the unigram "smoke" has a ratio of 1.19, which means the term's frequency of appearance is nearly equivalent in the neighboring posts of both classes. The between-class term ratio was calculated and used as the weight for each of the top-100 unigrams from neighboring posts in the feature set. We also would like to note that such feature selection via between-class term ratio was only based on the training set.

2.4. Evaluation of classifier performance

We evaluated the contributions of the feature sets listed in Table 2 to the classification of smoking status using 5 different models. Model 1 includes only standard unigram features of focal posts (Feature set 1) and Model 2 is based only on Doc2Vec (Feature set 2). On top of the better performing baseline model, Model 3 adds Feature set 3; Model 4 adds Feature sets 3 and 4; and Model 5 adds all the three new Feature sets (3, 4, and 5). This approach allows us to quantify how much new classification power is gained after adding one additional feature set. Feature sets were added in the order of increasing levels of abstraction from the focal post.

We used weighted F1-score and weighted AUC to evaluate the performance of classification models with 10-fold cross validation. F1-score is a harmonic mean of precision (the number of correct predictions divided by the number of all predictions in that class) and recall (the number of correct predictions divided by the number of actual instances in that category). AUC (area under the ROC curve) measures the probability that a positive sample is ranked higher than a negative sample and provides robust measurements of classification performance even in datasets with unbalanced class distributions. We selected six different classification algorithms to represent distinct, classic approaches to machine learning and examined their performance: Naïve Bayes, Logistic Regression, J48 decision tree, SVM (with polynomial kernel), and AdaBoost with two weak leaners (DecisionStump and J48).

3. Results

3.1. Classifier performance

The performance of binary classification with various algorithms on each of the four models is shown in Table 3 (F1 scores) and Table 4 (AUC values). Fig. 1 provides a visual comparison. Between two baseline models, Model 1 outperforms Model 2 in 5 out of the 6 algorithms we tried. Therefore, Models 3, 4, and 5 are based on adding new feature sets to Model 1.

Overall, Feature sets 3, 4, and 5 that we proposed contribute to better performance of the classifier compared to using only standard Feature sets 1 or 2. The best overall performance across all the models and algorithms is achieved by Model 5 with AdaBoost, using J48 as the weak learner. That combination yields the best F1-score of 0.759, which is 10.2% higher than the same algorithm's performance using only the features in Model 1 (0.689). The difference is statistically significant

Table 3
Weighted F1-scores (10-fold CV) for different models with different algorithms (the highest values are in bold for each algorithm, and standard deviations are in parentheses).

Algorithms	Model 1	Model 2	Model 3	Model 4	Model 5
Naïve Bayes	0.670	0.578	0.672	**0.678**	0.621
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Logistic regression	0.612	**0.699**	0.618	0.608	0.645
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)
J48 decision tree	0.669	0.557	0.692	**0.707**	0.705
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
SVM (Poly-Kernel)	0.704	0.692	0.720	0.730	**0.744**
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
AdaBoost (DecisionStump)	0.705	0.551	0.717	0.744	**0.755**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
AdaBoost (J48)	0.689	0.608	0.716	0.731	**0.759**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)

Table 4
Weighted AUC (10-fold CV) for different models with different algorithms (the highest values are in bold for each algorithm, and standard deviations are in parentheses).

Algorithms	Model 1	Model 2	Model 3	Model 4	Model 5
Naïve Bayes	0.718	0.640	0.723	0.720	**0.724**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Logistic regression	0.631	**0.737**	0.639	0.628	0.700
	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)
J48 decision tree	0.667	0.559	0.686	**0.699**	0.695
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
SVM (Poly-Kernel)	0.703	0.693	0.719	0.729	**0.743**
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
AdaBoost	0.774	0.613	0.780	0.827	**0.837**
(DecisionStump)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
AdaBoost (J48)	0.755	0.643	0.779	0.809	**0.828**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)

with p-value < 0.001 in a paired *t*-test. Model 5 with AdaBoost (J48) also has the best precision (0.759), recall (0.759) and accuracy (75.9%), all approximately 7% higher than those from Model 1, which has a precision of 0.690, a recall of 0.690, and an accuracy of 69.0% (p-value < 0.001 in a paired *t*-test). With DecisionStump as the weak learner, Model 5 with AdaBoost has the best AUC of 0.837, 8.1% higher than that from Model 1 (0.774) (p-value < 0.001 in a paired *t*-test). These comparisons indicate that the novel features we proposed (Feature sets 3, 4, and 5) provide valuable signals that improve the detection of smoking status.

In addition to Models 3, 4 and 5, where feature sets were added accumulatively, we also compared the predictive power of the 3 new feature sets when each is used alone for smoking status classification (Fig. 2). For comparison, we also included in Fig. 2 the performance of the baseline Feature set 1 (i.e., Model 1), along with the best-performing Model 5, which incorporates all the 5 feature sets. We did not include Model 2 because its performance is almost dominated by Model 1. Compared to the standard Feature set 1 (Model 1), Feature set 3 has similar performance: it has better F1-score with 3 of 6 algorithms and better AUC with 2 algorithms, even though this feature set only examines the appearance of a few domain-specific phrases, instead of all unigrams. Not surprisingly, Feature set 4, which only contains information about the author of each focal post, has the lowest F1-score across the 6 algorithms, and only outperforms the AUC of Feature set 1 with 2 algorithms. Feature set 5, which does not consider the content of a focal post at all, still has a better F1-score and AUC than Feature set 1 with 1 algorithm.

3.2. Abstinence status among all community members

We also applied the best-performing classification model (AdaBoost with DecisionStump on Model 5) to the remaining 352,922 blog posts

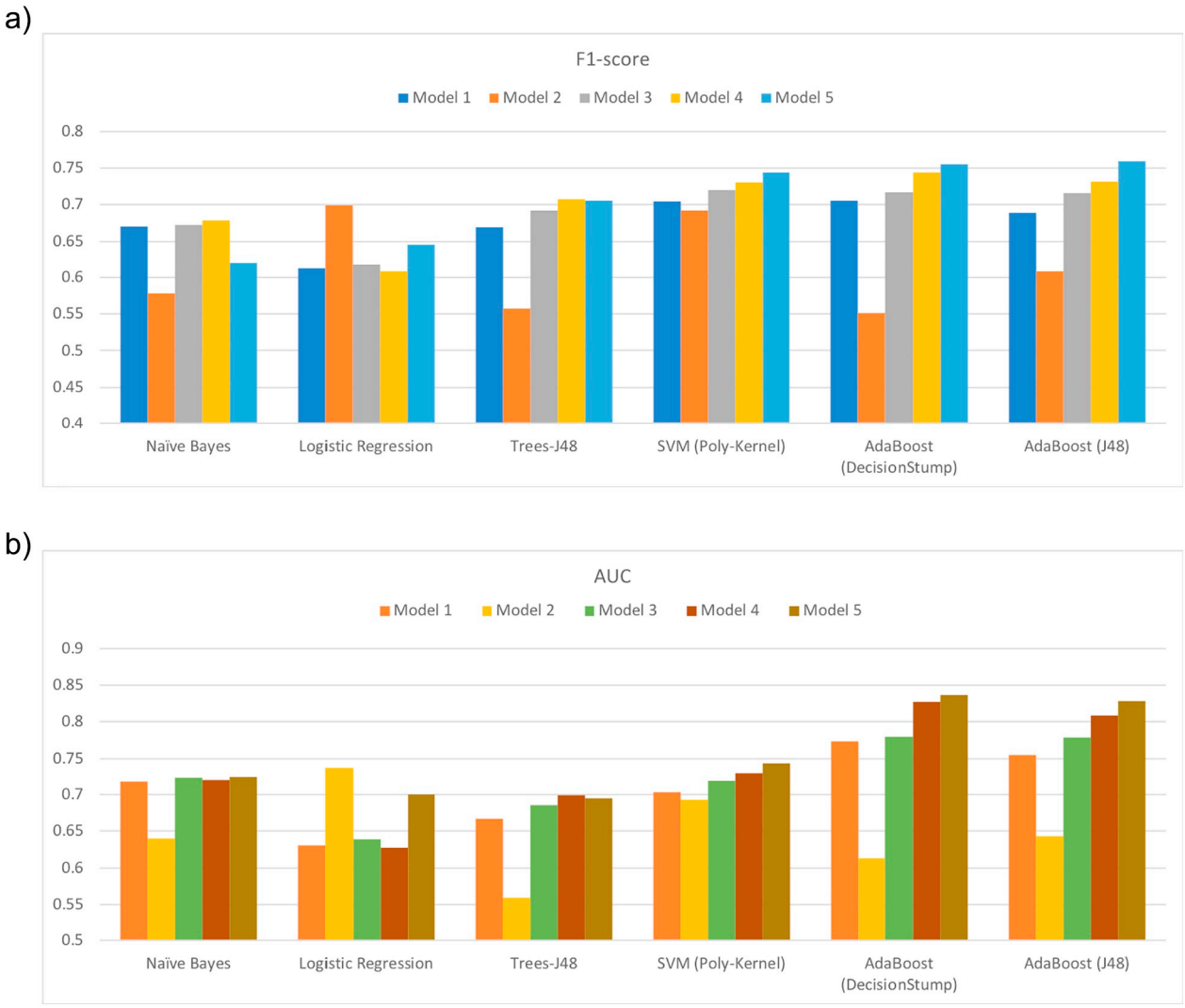

Fig. 1. Comparing the performance of different models with different algorithms using F1 scores (a), and AUC (b).

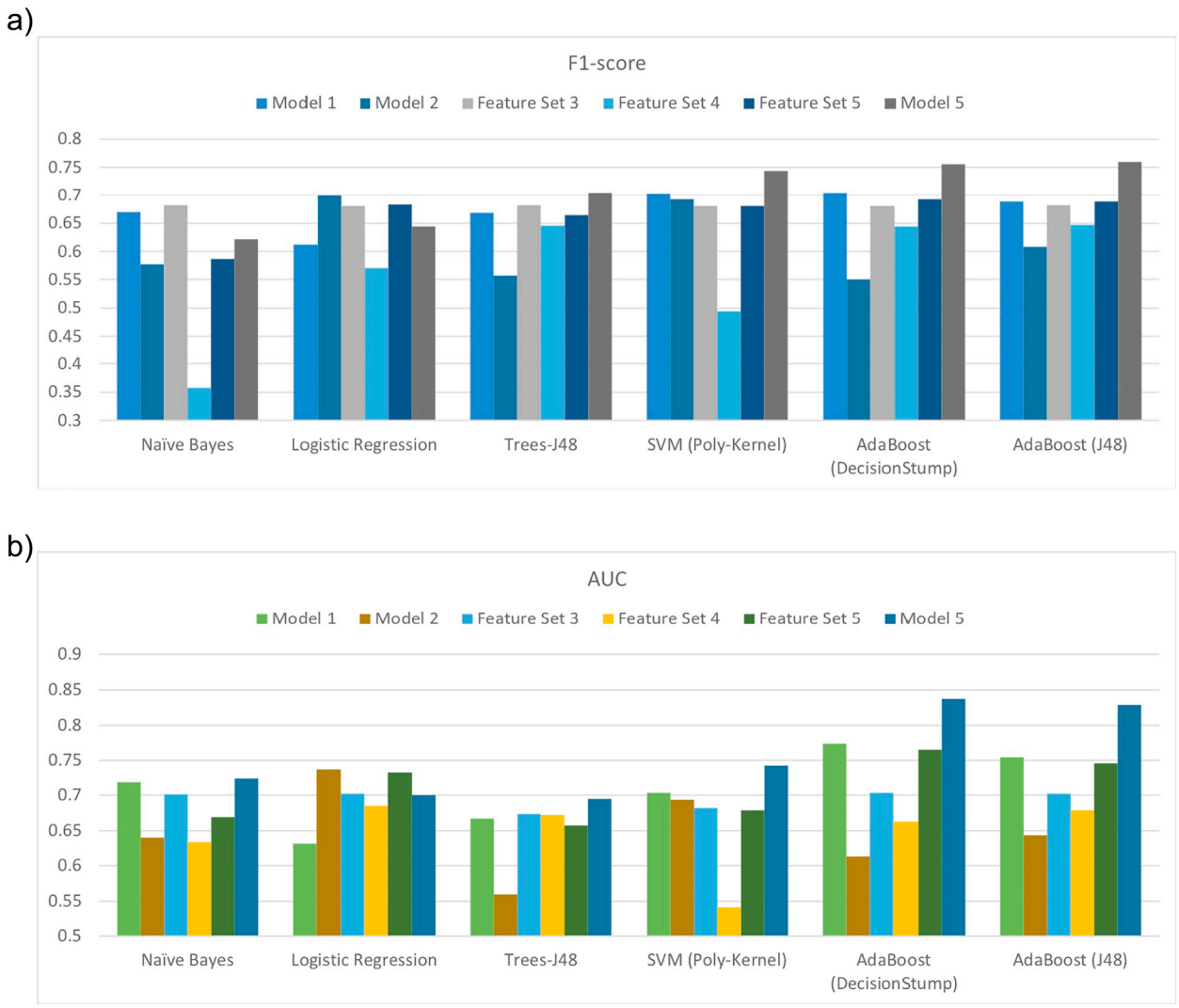

Fig. 2. Comparing the performance of using one of the 3 new feature sets in smoking status classification using F1-score (a) and AUC (b).

and comments. A total of 181,321 posts written by 3240 users indicated that the author was not smoking at the time of the post. This analysis suggests that 60% of users (3240/5435) who authored a blog or blog comment wrote at least one post indicating a period of abstinence.

4. Discussions and conclusions

This proof-of-concept study demonstrated the effectiveness of a new approach to automatically detect individuals' smoking status from large-scale data in an online smoking cessation community. Our approach went beyond the traditional approach that only examines the content of a user's post. Instead, we incorporated into the machine-learning-based classifier domain-specific features related to an online smoking cessation community, author-specific features related to patterns of user online engagement, and thread-specific features that signaled abstinence. Adding these novel features improved the classifier's performance by approximately 10% and pointed to the importance of incorporating domain knowledge, considering characteristics of the author along with preceding and subsequent posts in detecting the smoking status of a focal post's author.

To our knowledge, this is the first study to use a machine learning approach to detect the smoking status of users as “quit vs not” from individual posts in an online smoking cessation community. Compared to previous studies that identified individual health status by mining UGCs [25,26], this work further highlights the value of combining insights from domain experts, especially in the extraction of domain-specific features, and computational methods. Our approach is also the first to examine “neighboring posts” to leverage semantic connections between UGCs. Experiment results reveal that contents from such neighboring posts do contribute to the performance of smoking status identifications.

The proposed approach also has the potential to be applied to online communities about other addictions. In a different online community for another type of addiction, users' vocabulary may change, which means some domain-specific features could vary and may need to be updated with the help of frequent community users or via reading UGCs. Nevertheless, the other three types of features—text features of focal posts, author-based features, and thread-based features are all available in most online communities. For example, most online communities allow user to interact with each other via posting. No matter whether such interactions occur in the form of posts in the same threaded discussion or comments to the same blog post, they semantically connect UGCs, making it possible to leverage content of “neighboring” posts in the mining of a focal post.

This work also provides an exciting foundation for the development and evaluation of real-time interventions in online smoking cessation communities via personalized post recommendations. Such real-time tailoring can be directly integrated into the online platform through which the intervention is delivered. For example, a user whose post indicates that she is still preparing for a quit attempt could benefit from

receiving recommendations on skills training exercises or community content about what to expect during the first few days of a quit attempt. For another user who publishes a post to claim her abstinence, the community can recommend others' posts on how to cope with withdrawal and adjust to living smoke free to avoid relapse. As a proof-of-concept, this study focused on identifying smoking status from user-generated content. However, our approach could also be applied to other important drivers of cessation that may be amenable to real-time intervention. For example, knowing a smoker's attitudes toward or intent to use a medication [30] may present opportunities to dispel common myths and misperceptions [69–71] and encourage adherence [72]. Recent innovations in computer-tailored health communication systems have combined implicit data derived from user actions (e.g., website page views) with explicit data collected via user self-report to empirically tailor content [28,73]. We are not aware of any work to mine user-generated content from an online health community to drive a tailored intervention.

Another noteworthy finding is that among users who posted, 60% indicated at least one period of abstinence. Although some posts explicitly mentioned the duration of abstinence, our classification scheme of smoking status does not readily map to traditional outcome assessments (e.g., 7-day point prevalence [74]). From this perspective, the classification of “quit vs. not” may best represent an indicator of a thus-far-successful quit attempt. Future research will need to develop machine classification methods to mine user-generated content to identify the duration of abstinence, and to infer such duration for posts that do not include such time spans.

Several limitations of this work are worth noting. First, this approach is only applicable to community members who contribute content. Historically, online communities have followed the 1% rule, with 90% of users lurking (i.e., not posting content), 9% commenting on other posts, and 1% creating new content [75,76]. These trends have shifted in recent years as social media use has become more ubiquitous, with some now suggesting a 70-20-10 split [77] or even 55-25-20 [78]. As of 2018, approximately 10% of BecomeAnEX users contribute content, representing a clinically meaningful number of users who could benefit from future interventions developed from this approach. Second, we only evaluated the performance of six classification algorithms. Using other classification algorithms with more annotated posts may yield better performance.

Competing interests

The authors have read the journal's policy and have the following competing interests: ALG, SC, and MSA are employed by Truth Initiative which runs the BecomeAnEX website.

Funding

This research was funded by grant #R01 CA192345 (Graham/Zhao, MPI) from the National Cancer Institute (https://www.cancer.gov/) of the National Institutes of Health as part of a trans-NIH initiative known as Collaborative Research on Addiction (CRAN). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Amy Cohn is now at Battelle; Jennifer Pearson is now at the University of Nevada, Reno.

References

[1] American Cancer Society, Health Risks of Smoking Tobacco, (n.d.). https://www.cancer.org/cancer/cancer-causes/tobacco-and-cancer/health-risks-of-smoking-tobacco.html (accessed May 15, 2018).

[2] Healthways, QuitNet Tobacco Cessation: Dynamic, Multimodal Tobacco Cessation Programs, Healthways Center for Health Research, Franklin, TN, USA, n.d. http://www.healthways.com/hs-fs/hub/162029/file-691487149-pdf/Fact_Sheet/QuitNet_Fact_Sheet.pdf.

[3] A.L. Graham, M.S. Amato, Twelve million smokers look online for smoking cessation help annually: health information national trends survey data, 2005–2017, Nicotine & Tobacco Research. (n.d.). doi: https://doi.org/10.1093/ntr/nty043.

[4] K.J. Head, S.M. Noar, N.T. Iannarino, N. Grant Harrington, Efficacy of text messaging-based interventions for health promotion: a meta-analysis, Social Science & Medicine 97 (2013) 41–48, https://doi.org/10.1016/j.socscimed.2013.08.003.

[5] M.W. Kreuter, D.L. Oswald, F.C. Bull, E.M. Clark, Are tailored health education materials always more effective than non-tailored materials? Health Education Research 15 (2000) 305–315.

[6] C.S. Skinner, M.K. Campbell, B.K. Rimer, S. Curry, J.O. Prochaska, How effective is tailored print communication? Annals of Behavioral Medicine 21 (1999) 290–298.

[7] R.E. Petty, J.T. Cacioppo, D. Schumann, Central and peripheral routes to advertising effectiveness: the moderating role of involvement, Journal of Conflict Resolution 10 (1983) 135–146.

[8] S.M. Noar, C.N. Benac, M.S. Harris, Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions, Psychological Bulletin 133 (2007) 673–693, https://doi.org/10.1037/0033-2909.133.4.673.

[9] L. Morrison, R. Moss-Morris, S. Michie, L. Yardley, Optimizing engagement with internet-based health behaviour change interventions: comparison of self-assessment with and without tailored feedback using a mixed methods approach, British Journal of Health Psychology 19 (2014) 839–855, https://doi.org/10.1111/bjhp.12083.

[10] L. Donkin, H. Christensen, S.L. Naismith, B. Neal, I.B. Hickie, N. Glozier, A systematic review of the impact of adherence on the effectiveness of e-therapies, Journal of Medical Internet Research 13 (2011), https://doi.org/10.2196/jmir.1772.

[11] S.M. Kelders, R.N. Kok, H.C. Ossebaard, J.E. Van Gemert-Pijnen, Persuasive system design does matter: a systematic review of adherence to web-based interventions, Journal of Medical Internet Research 14 (2012), https://doi.org/10.2196/jmir.2104.

[12] J.R. Schubart, H.L. Stuckey, A. Ganeshamoorthy, C.N. Sciamanna, Chronic health conditions and internet behavioral interventions: a review of factors to enhance user engagement, Computers, Informatics, Nursing 29 (2011) TC9–20, https://doi.org/10.1097/NCN.0b013e3182155274.

[13] T. Lancaster, F.S. Stead, Self-help interventions for smoking cessation, Cochrane Database of Systematic Reviews 2005, (n.d.).

[14] G.M.J. Taylor, M.N. Dalili, M. Semwal, M. Civljak, A. Sheikh, J. Car, Internet-based interventions for smoking cessation, Cochrane Database of Systematic Reviews, John Wiley & Sons, Ltd, 2017, , https://doi.org/10.1002/14651858.CD007078.pub5.

[15] V.J. Strecher, J. McClure, G. Alexander, B. Chakraborty, V. Nair, J. Konkel, S. Greene, M. Couper, C. Carlier, C. Wiese, R. Little, C. Pomerleau, O. Pomerleau, The role of engagement in a tailored web-based smoking cessation program: randomized controlled trial, Journal of Medical Internet Research 10 (2008) e36, , https://doi.org/10.2196/jmir.1002.

[16] J. Balmford, R. Borland, P. Benda, Patterns of use of an automated interactive personalized coaching program for smoking cessation, Journal of Medical Internet Research 10 (2008) e54, , https://doi.org/10.2196/jmir.1016.

[17] W.F. Velicer, C.A. Redding, X. Sun, J.O. Prochaska, Demographic variables, smoking variables, and outcome across five studies, Health Psychology 26 (2007) 278–287, https://doi.org/10.1037/0278-6133.26.3.278.

[18] P. Krebs, J.O. Prochaska, J.S. Rossi, A meta-analysis of computer-tailored interventions for health behavior change, Preventive Medicine 51 (2010) 214–221, https://doi.org/10.1016/j.ypmed.2010.06.004.

[19] H. Liang, Y. Xue, B.A. Berger, Web-based intervention support system for health promotion, Decision Support Systems 42 (2006) 435–449, https://doi.org/10.1016/j.dss.2005.02.001.

[20] B.G. Danaher, J.R. Seeley, Methodological issues in research on web-based behavioral interventions, Annals of Behavioral Medicine 38 (2009) 28, https://doi.org/10.1007/s12160-009-9129-0.

[21] S. Ba, L. Wang, Digital health communities: the effect of their motivation mechanisms, Decision Support Systems (n.d.). doi: https://doi.org/10.1016/j.dss.2013.01.003.

[22] X. Wang, K. Zhao, W.N. Street, Social Support and User Engagement in Online Health Communities, Proceedings of the International Conference for Smart Health, Springer, Beijing, China, 2014, pp. 97–110, , https://doi.org/10.1007/978-3-319-08416-9_10.

[23] T. Nguyen, M.E. Larsen, B. O'Dea, D.T. Nguyen, J. Yearwood, D. Phung, S. Venkatesh, H. Christensen, Kernel-based features for predicting population health indices from geocoded social media data, Decision Support Systems 102 (2017) 22–31, https://doi.org/10.1016/j.dss.2017.06.010.

[24] A. Lamb, M.J. Paul, M. Dredze, Separating Fact from Fear: Tracking Flu Infections on Twitter, Proceedings of HLT-NAACL'13, Atlanta, GA, 2013, pp. 789–795.

[25] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting Depression via Social Media, Seventh International AAAI Conference on Weblogs and Social Media, Boston, 2013, pp. 128–137 http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6124 (accessed July 22, 2013).

[26] J.H.D. Cho, T. Gao, R. Girju, Identifying medications that patients stopped taking in online health forums, 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 2017, pp. 141–148, , https://doi.org/10.1109/ICSC.2017.24.

[27] W.T. Riley, D.E. Rivera, A.A. Atienza, W. Nilsen, S.M. Allison, R. Mermelstein, Health behavior models in the age of mobile interventions: are our theories up to

the task? Translational Behavioral Medicine 1 (2011) 53–71, https://doi.org/10.1007/s13142-011-0021-7.
[28] R.S. Sadasivam, S.L. Cutrona, R.L. Kinney, B.M. Marlin, K.M. Mazor, S.C. Lemon, T.K. Houston, Collective-intelligence recommender systems: advancing computer tailoring for health behavior change into the 21st century, Journal of Medical Internet Research 18 (2016), https://doi.org/10.2196/jmir.4448.
[29] K. Ingersoll, R. Dillingham, G. Reynolds, J. Hettema, J. Freeman, S. Hosseinbor, C. Winstead-Derlega, Development of a personalized bidirectional text messaging tool for HIV adherence assessment and intervention among substance abusers, Journal of Substance Abuse Treatment 46 (2014) 66–73, https://doi.org/10.1016/j.jsat.2013.08.002.
[30] N.K. Cobb, D. Mays, A.L. Graham, Sentiment analysis to determine the impact of online messages on smokers' choices to use varenicline, Journal of the National Cancer Institute. Monographs 2013 (2013) 224–230, https://doi.org/10.1093/jncimonographs/lgt020.
[31] S. Myneni, N.K. Cobb, T. Cohen, Finding meaning in social media: content-based social network analysis of QuitNet to identify new opportunities for health promotion, Studies in Health Technology and Informatics 192 (2013) 807–811.
[32] P. Selby, T. van Mierlo, S.C. Voci, D. Parent, J.A. Cunningham, Online social and professional support for smokers trying to quit: an exploration of first time posts from 2562 members, Journal of Medical Internet Research 12 (2010) e34, https://doi.org/10.2196/jmir.1340.
[33] T. van Mierlo, S. Voci, S. Lee, R. Fournier, P. Selby, Superusers in social networks for smoking cessation: analysis of demographic characteristics and posting behavior from the Canadian Cancer Society's smokers' helpline online and StopSmokingCenter.net, Journal of Medical Internet Research 14 (2012) e66, https://doi.org/10.2196/jmir.1854.
[34] M. Zhang, C.C. Yang, X. Gong, Social support and exchange patterns in an online smoking cessation intervention program, IEEE International Conference on Healthcare Informatics (ICHI), 2013 2013, pp. 219–228, , https://doi.org/10.1109/ICHI.2013.37.
[35] C.L. Brandt, P. Dalum, L. Skov-Ettrup, J.S. Tolstrup, “After all–it doesn't kill you to quit smoking”: an explorative analysis of the blog in a smoking cessation intervention, Scandinavian Journal of Public Health 41 (2013) 655–661, https://doi.org/10.1177/1403494813489602.
[36] S.J. Bondy, K.L. Bercovitz, “Hike up yer skirt, and quit.” what motivates and supports smoking cessation in builders and renovators, International Journal of Environmental Research and Public Health 10 (2013) 623–637, https://doi.org/10.3390/ijerph10020623.
[37] H. Cole-Lewis, A. Perotte, K. Galica, L. Dreyer, C. Griffith, M. Schwarz, C. Yun, H. Patrick, K. Coa, E. Augustson, Social network behavior and engagement within a smoking cessation Facebook page, Journal of Medical Internet Research 18 (2016), https://doi.org/10.2196/jmir.5574.
[38] M. Burri, V. Baujard, J.-F. Etter, A qualitative analysis of an internet discussion forum for recent ex-smokers, Nicotine & Tobacco Research 8 (Suppl. 1) (2006) S13–S19.
[39] H. Yang, C.C. Yang, Using health-consumer-contributed data to detect adverse drug reactions by association mining with temporal analysis, ACM Transactions on Intelligent Systems and Technology 6 (55) (2015) 1–55:27, https://doi.org/10.1145/2700482.
[40] C.C. Yang, H. Yang, L. Jiang, M. Zhang, Social Media Mining for Drug Safety Signal Detection, Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, ACM, New York, NY, USA, 2012, pp. 33–40, , https://doi.org/10.1145/2389707.2389714.
[41] B. Qiu, K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G.E. Greer, K. Portier, Get online support, feel better–Sentiment analysis and dynamics in an online cancer survivor community, Proceedings of the Third IEEE International Conference on Social Computing (SocialCom'11), Boston, MA, 2011, pp. 274–281.
[42] K. Zhao, J. Yen, G. Greer, B. Qiu, P. Mitra, K. Portier, Finding influential users of online health communities: a new metric based on sentiment influence, Journal of the American Medical Informatics Association 21 (2014) e212–e218, https://doi.org/10.1136/amiajnl-2013-002282.
[43] X. Wang, K. Zhao, N. Street, Analyzing and predicting user participations in online health communities: a social support perspective, Journal of Medical Internet Research 19 (2017) e130, , https://doi.org/10.2196/jmir.6834.
[44] M. Wen, Z. Zheng, H. Jang, G. Xiang, C.P. Rosé, Extracting Events with Informal Temporal References in Personal Histories in Online Communities, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Sofia, Bulgaria, 2013, pp. 836–842.
[45] K. Zhao, G. Greer, J. Yen, P. Mitra, K. Portier, Leader identification in an online health community for cancer survivors: a social network-based classification approach, Information Systems and e-Business Management 13 (2015) 629–645, https://doi.org/10.1007/s10257-014-0260-5.
[46] Q.T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S.N. Murphy, R. Lazarus, Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system, BMC Medical Informatics and Decision Making 6 (2006) 30, , https://doi.org/10.1186/1472-6947-6-30.
[47] C.-Y. Wu, C.-K. Chang, D. Robson, R. Jackson, S.-J. Chen, R.D. Hayes, R. Stewart, Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register, PLoS One 8 (2013) e74262, , https://doi.org/10.1371/journal.pone.0074262.
[48] Ö. Uzuner, I. Goldstein, Y. Luo, I. Kohane, Identifying patient smoking status from medical discharge records, Journal of the American Medical Informatics Association 15 (2008) 14–24, https://doi.org/10.1197/jamia.M2408.
[49] R. Wicentowski, M.R. Sydes, Using implicit information to identify smoking status in smoke-blind medical discharge summaries, Journal of the American Medical Informatics Association 15 (2008) 29–31, https://doi.org/10.1197/jamia.M2440.
[50] A.M. Cohen, Five-way smoking status classification using text hot-spot identification and error-correcting output codes, Journal of the American Medical Informatics Association 15 (2008) 32–35, https://doi.org/10.1197/jamia.M2434.
[51] R. Millington, Buzzing Communities: How to Build Bigger, Better, and more Active Online Communities, Feverbee, London, 2012.
[52] C. Smith, P. Stavri, Consumer Health Vocabulary, in: D. Lewis, G. Eysenbach, R. Kukafka, Z. Stavri, H. Jimison (Eds.), Consumer Health Informatics: Informing Consumers and Improving Health Care, Springer, 2005, pp. 122–128, , https://doi.org/10.1007/0-387-27652-1_10.
[53] M. Zhang, Social Media Analytics of Smoking Cessation Intervention: User Behavior Analysis, Classification, and Prediction, Dissertations & Theses - Gradworks, http://search.proquest.com/docview/1667470689, (2015) , Accessed date: 1 November 2017.
[54] T. Nguyen, R. Borland, J. Yearwood, H.-H. Yong, S. Venkatesh, D. Phung, Discriminative Cues for Different Stages of Smoking Cessation in Online Community, Web Information Systems Engineering – WISE 2016, Springer, Cham, 2016, pp. 146–153, , https://doi.org/10.1007/978-3-319-48743-4_12.
[55] A. Tamersoy, M. De Choudhury, D.H. Chau, Characterizing smoking and drinking abstinence from social media, HT ACM Conf Hypertext Soc Media, 2015, pp. 139–148, , https://doi.org/10.1145/2700171.2791247.
[56] M. Deady, Social influence, addictions and the internet: the potential of web 2.0 technologies in enhancing treatment for alcohol/other drug use problems, Journal of Addiction Research & Therapy (2012), https://doi.org/10.4172/2155-6105.S8-002.
[57] K.L. Mccausland, L.E. Curry, A. Mushro, S. Carothers, H. Xiao, D.M. Vallone, Promoting a web-based smoking cessation intervention: implications for practice, Cases in Public Health Communication & Marketing, 2011, pp. 3–26.
[58] K. Zhao, X. Wang, S. Cha, A.M. Cohn, G.D. Papandonatos, M.S. Amato, J.L. Pearson, A.L. Graham, A multirelational social network analysis of an online health community for smoking cessation, Journal of Medical Internet Research 18 (2016) e233, , https://doi.org/10.2196/jmir.5985.
[59] M.S. Amato, G.D. Papandonatos, S. Cha, X. Wang, K. Zhao, A.M. Cohn, J.L. Pearson, A.L. Graham, Inferring smoking status from user generated content in an online cessation community, Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco (2018), https://doi.org/10.1093/ntr/nty014.
[60] C.X. Zhai, S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, Association for Computing Machinery and Morgan & Claypool, http://dl.acm.org/citation.cfm?id=2915031, (2016) , Accessed date: 17 October 2017.
[61] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, ArXiv:1301.3781 [Cs], http://arxiv.org/abs/1301.3781, (2013) , Accessed date: 28 March 2018.
[62] Q.V. Le, T. Mikolov, Distributed Representations of Sentences and Documents, ArXiv:1405.4053 [Cs], http://arxiv.org/abs/1405.4053, (2014) , Accessed date: 23 April 2018.
[63] Y. Goldberg, O. Levy, word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method, ArXiv:1402.3722 [Cs, Stat], http://arxiv.org/abs/1402.3722, (2014) , Accessed date: 28 March 2018.
[64] G.D. Papandonatos, B. Erar, C.A. Stanton, A.L. Graham, Online community use predicts abstinence in combined internet/phone intervention for smoking cessation, Journal of Consulting and Clinical Psychology 84 (2016) 633–644, https://doi.org/10.1037/ccp0000099.
[65] A.L. Graham, G.D. Papandonatos, B. Erar, C.A. Stanton, Use of an online smoking cessation community promotes abstinence: results of propensity score weighting, Health Psychology 34S (2015) 1286–1295, https://doi.org/10.1037/hea0000278.
[66] P. Biyani, S. Bhatia, C. Caragea, P. Mitra, Thread Specific Features Are Helpful for Identifying Subjectivity Orientation of Online Forum Threads, COLING, 2012, pp. 295–310.
[67] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106, https://doi.org/10.1007/BF00116251.
[68] M. Pennacchiotti, A.M. Popescu, A Machine Learning Approach to Twitter User Classification, International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 2011.
[69] S.S. Fu, D. Burgess, M. van Ryn, D.K. Hatsukami, J. Solomon, A.M. Joseph, Views on smoking cessation methods in ethnic minority communities: a qualitative investigation, Preventive Medicine 44 (2007) 235–240, https://doi.org/10.1016/j.ypmed.2006.11.002.
[70] K. Bowker, K.A. Campbell, T. Coleman, S. Lewis, F. Naughton, S. Cooper, Understanding pregnant smokers' adherence to nicotine replacement therapy during a quit attempt: a qualitative study, Nicotine & Tobacco Research 18 (2016) 906–912, https://doi.org/10.1093/ntr/ntv205.
[71] A.L. Smith, S.M. Carter, S. Chapman, S.M. Dunlop, B. Freeman, Why do smokers try to quit without medication or counselling? A qualitative study with ex-smokers, BMJ Open 5 (2015) e007301, https://doi.org/10.1136/bmjopen-2014-007301.
[72] G.J. Hollands, M.S. McDermott, N. Lindson-Hawley, F. Vogt, A. Farley, P. Aveyard, Interventions to increase adherence to medications for tobacco dependence, Cochrane Database of Systematic Reviews (2015) CD009164, https://doi.org/10.1002/14651858.CD009164.pub2.
[73] R.S. Sadasivam, E.M. Borglund, R. Adams, B.M. Marlin, T.K. Houston, Impact of a collective intelligence tailored messaging system on smoking cessation: the perspect randomized experiment, Journal of Medical Internet Research 18 (2016) e285.
[74] J.R. Hughes, J.P. Keely, R.S. Niaura, D.J. Ossipklein, R.L. Richmond, G.E. Swan, Measures of abstinence in clinical trials: issues and recommendations, Nicotine & Tobacco Research Official Journal of the Society for Research on Nicotine & Tobacco. 5 (2003) 13.

[75] C. Arthur, What Is the 1% Rule?, the Guardian, http://www.theguardian.com/technology/2006/jul/20/guardianweeklytechnologysection2, (2006) , Accessed date: 16 October 2017.
[76] J. Nielsen, Participation Inequality: The 90–9-1 Rule for Social Features, https://www.nngroup.com/articles/participation-inequality/, (2006) , Accessed date: 16 October 2017.
[77] P. Schneider, Is the 90-9-1 Rule for Online Community Engagement Dead?, (n.d.). http://blog.higherlogic.com/2011/08/11/Is-the-90-9-1-Rule-for-Online-Community-Engagement-Dead-Data.
[78] T. McEnroe, Three Community Myths Busted – A SOCM 2016 Preview, (n.d.). https://communityroundtable.com/state-of-community-management/three-community-myths-busted/.

Xi Wang is an Assistant Professor in the School of Information at Central University of Finance and Economics, China. Her research focuses on user behavior analyses in social media, especially related to online health communities. She is interested in the metrics of social network analyses and data mining. Dr. Wang received her Ph.D. from The University of Iowa in 2017.

Kang Zhao is an Associate Professor of Management Sciences, with a joint appointment in Informatics, at the University of Iowa. His current research focuses on data science and social computing, especially the mining, modeling, and simulation of social, business, and scholarly networks. His research has been featured in public media from more than 25 countries, such as Washington Post, USA Today, Forbes, New York Public Radio, BBC, and Agence France-Presse. He also served as the Chair for INFORMS Artificial Intelligence Section 2014–2016. He earned his Ph.D. from Penn State University.

Sarah Cha, MSPH, is Senior Project Manager at Truth Initiative. She received a master's of science in public health in health education and health communications from Johns Hopkins Bloomberg School of Public Health, a bachelor's degree in health promotion and disease prevention studies and a bachelor's degree in psychology from the University of Southern California. Her expertise is in the management of large-scale research studies involving digital smoking cessation interventions.

Michael S. Amato is Methodologist and Research Investigator at Truth Initiative. His research focuses on the use of digital cessation interventions to reduce smoking in the population. He is interested in metrics of meaningful intervention engagement, text analytics for user generated content, and the unique methodological challenges and opportunities presented by observational studies of online behavior. Dr. Amato received his PhD in Psychology from the University of Wisconsin, focusing on quantitative methods and behavior change communication

Amy M. Cohn is a Senior Research Scientist at Battelle Memorial Institute and Associate Professor (Adjunct) in the Department of Oncology at Georgetown University Medical Center. She received her doctorate in Clinical Psychology from the University of Georgia and completed post-doctoral training in smoking cessation research at Brown University and in the assessment and treatment of alcohol use disorders at the Center of Alcohol Studies at Rutgers University. Her research focuses on mental health and substance co-morbidities with tobacco use and factors associated with substance use behavior change in young adult and adult populations.

Jennifer L. Pearson is an Assistant Professor in the Division of Health Sciences at the University of Nevada, Reno (UNR), an Adjunct Assistant Professor at the Johns Hopkins Bloomberg School of Public Health, and an Honorary Lecturer at the University of Sterling in Scotland. Her research focuses on how U.S. federal tobacco regulation influences consumer behavior, specifically concerning e-cigarettes and tobacco products labeled as “natural,” “organic,” or “additive-free.” Dr. Pearson has authored over 50 scientific articles on tobacco control topics, and published in high-impact journals such as the New England Journal of Medicine, Tobacco Control, and Nicotine & Tobacco Research. Dr. Pearson began her career in tobacco control as a Tobacco Education Coordinator for the American Lung Association of Nevada in 2004.

George D. Papandonatos received his Ph.D. in Statistics from the University of Minnesota (Twin Cities). As an Associate Professor of Biostatistics at Brown University, he has participated in numerous randomized controlled clinical trials assessing the effects of lifestyle interventions on improving health behaviors. As project biostatistician on over 40 funded NIH grants, he has collaborated with behavioral medicine researchers and epidemiologists in promoting smoking cessation and the prevention of relapse, drinking moderation, weight loss and the prevention of weight gain, and the adoption and maintenance of physical activity. He is particularly interested in causal inference and mediation modeling in the context of randomized trials affected by selection bias and informative dropout.

Amanda L. Graham is Senior Vice President of Innovations and a Research Investigator in the Schroeder Institute at Truth Initiative. She is also Professor of Oncology (Adjunct) at Georgetown University Medical Center/Lombardi Comprehensive Cancer Center. Dr. Graham's program of research focuses on the optimization and evaluation of broad-reach technology-based interventions for smoking cessation, with a particular focus on improving adherence and effectiveness. For the past 20 years, Dr. Graham has led numerous large-scale NIH-funded studies addressing digital smoking cessation interventions and the role of online social networks in health behavior change. She earned her PhD from the Chicago Medical School.