Original Article

Healthc Inform Res. 2021 April;27(2):116-126.
https://doi.org/10.4258/hir.2021.27.2.116
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Social Network Analysis of an Online Smoking Cessation Community to Identify Users' Smoking Status

Adnan Muhammad Shah[1], Xiangbin Yan[2], Abdul Qayyum[3]

[1]Department of Management Science and Engineering, School of Management, Harbin Institute of Technology, Harbin, China
[2]School of Economics and Management, University of Science and Technology Beijing, Beijing, China
[3]Faculty of Management Sciences, Riphah International University, Islamabad, Pakistan

Objectives: Users share valuable information through online smoking cessation communities (OSCCs), which help people maintain and improve smoking cessation behavior. Although OSCC utilization is common among smokers, limitations exist in identifying the smoking status of OSCC users ("quit" vs. "not quit"). Thus, the current study implicitly analyzed user-generated content (UGC) to identify individual users' smoking status through advanced computational methods and real data from an OSCC. **Methods:** Secondary data analysis was conducted using data from 3,833 users of BcomeAnEX.org. Domain experts reviewed posts and comments to determine the authors' smoking status when they wrote them. Seven types of feature sets were extracted from UGC (textual, Doc2Vec, social influence, domain-specific, author-based, and thread-based features, as well as adjacent posts). **Results:** Introducing novel features boosted smoking status recognition (quit vs. not quit) by 9.3% relative to the use of text-only post features. Furthermore, advanced computational methods outperformed baseline algorithms across all models and increased the smoking status prediction performance by up to 12%. **Conclusions:** The results of this study suggest that the current research method provides a valuable platform for researchers involved in online cessation interventions and furnishes a framework for on-going machine learning applications. The results may help practitioners design a sustainable real-time intervention via personalized post recommendations in OSCCs. A major limitation is that only users' smoking status was detected. Future research might involve programming machine learning classification methods to identify abstinence duration using larger datasets.

Keywords: Smoking Cessation, Social Networking, Social Media, Machine Learning, Neural Networks

Submitted: July 14, 2020
Revised: January 19, 2021
Accepted: February 11, 2021

Corresponding Author
Adnan Muhammad Shah
Department of Management Science and Engineering, School of Management, Harbin Institute of Technology, Harbin, Heilongjinag 150001, China. Tel: +86-15004642903, E-mail: adnanshah486@gmail.com (https://orcid.org/0000-0002-9638-3514)

I. Introduction

Many current and former smokers use online smoking cessation communities (OSCCs) for smoking cessation every year. These users post about their smoking cessation journey, efforts to remain abstinent, and celebrations [1]. Participants' behaviors in these communities are meaningful in several ways. First, successful participation in an OSCC encourages active relationship-building with other members [2]. Second, active participation can better inform strategies to design successful OSCCs, which can encourage innovative treatments and a healthy lifestyle. Third, users' continuing involvement in OSCCs can help them receive social support, which reduces their stress and helps them cope with their disease [3,4]. Hence, a better understanding of OSCC users' engagement with these platforms can provide support for the design of effective OSCCs through high-quality community design, management, and user retention.

More than 12 million smokers search for online information about quitting smoking every year globally, of whom a majority participate in social networking sites for cessation [5]. To enhance the effectiveness of user-generated content (UGC) for smoking cessation, it is vital to use advanced computational techniques to reach a better understanding of how these efforts encourage users to quit smoking. Advanced computing techniques allow coders to analyze a large quantity of UGC, enabling research on common topics of discussion in online social networks for cessation [6-9]. Although the sentiments expressed in UGC have been extensively studied, very few studies have used advanced computational techniques to investigate the sentiments expressed by users who want to quit smoking.

Existing methods of data mining can be categorized as baseline and deep learning (DL) methods. In the former type of method, a classifier is used to assign a sentence to either a positive or negative class. Baseline classifiers such as support vector machines (SVMs) and logistic regression (LR) have successfully been applied in previous research [7,10-12]. However, those methods rely on natural language processing (NLP) tools, thereby augmenting the cost of research and increasing the inherent noise in the data, which may adversely affect models' efficiency.

In everyday conversation, opinions are expressed implicitly—that is, in a way that depends on domain and context. Identifying context-dependent features can also be useful in applications such as identifying users' smoking status and semantic searching. Although several challenges exist in monitoring and retrieving UGC, the clickstream data and metadata surrounding a user's participation in an online community are easy to retrieve. The extent to which UGC usually reflects user experiences makes it possible to gather additional details about the original post from how other users respond to it. In conjunction with the text of user blog posts, new features such as classifier inputs can be used to boost the output instead of relying solely on text. Moreover, methods using implicit and latent features have led to the emergence of DL models, which have demonstrated excellent performance in comparison to existing state-of-the-art methods [13,14]. In this study, we employed long short-term memory (LSTM)-DL methods that rely on latent features learned by neural network models. This study aimed to implicitly analyze the effectiveness of integrating various feature sets to identify users' smoking status (i.e., whether they have quit or not) using real textual data. We constructed forecasting models based on UGC that discriminated OSCC users who had quit smoking in comparison to those who had not quit yet. Previous studies have only investigated users who mentioned their quit date in their profiles [12,15]. In light of the fact that users' profile information may not be consistent in OSCCs (i.e., users may change their quit date, not mention their quit rate, or post false information), this study covered a larger population than those studied in other OSCCs to reduce sampling bias. Overall, mining UGC to detect users' smoking status could lead to the promotion of more convincing interventions designed in real time [16].

II. Methods

1. Data Collection and Data Labeling

This study collected data from BecomeAnEX.org, a publicly available web-based smoking cessation program composed of thousands of existing and former smokers who connect via several communication channels, such as private messages, public posts on member profile pages ("message boards"), group discussions, and blog posts [10,17]. This research focused on blogs and blog comments because these are the most common communication platforms and usually contain longer and more informative posts from users. A web crawler was developed in Python 3.6 to download 5,915 blog posts with 53,140 comments published by 3,833 users from January 2012 to May 2015. The overall analytical framework is shown in Figure 1.

To evaluate the performance of the machine learning classifiers, labeled data are needed for training and evaluation to learn the variation between instances from different classes. To label unstructured data, we availed ourselves of the services of three experts in the field of machine learning (ML) and data mining who had domain expertise and in-depth familiarity with user conversations in OSCCs. Sample posts along with their corresponding labels are shown in Table 1.

In addition, we performed several pre-processing steps on the raw data before performing ML-based basic classification and LSTM-based DL classification (Figure 1). This study used several feature sets to perform classification tasks: model 1 included the standard textual feature set; model 2 was based on feature set 2 (Doc2Vec); model 3 consisted of feature set 3; model 4 combined feature sets 3 and 4; model 5

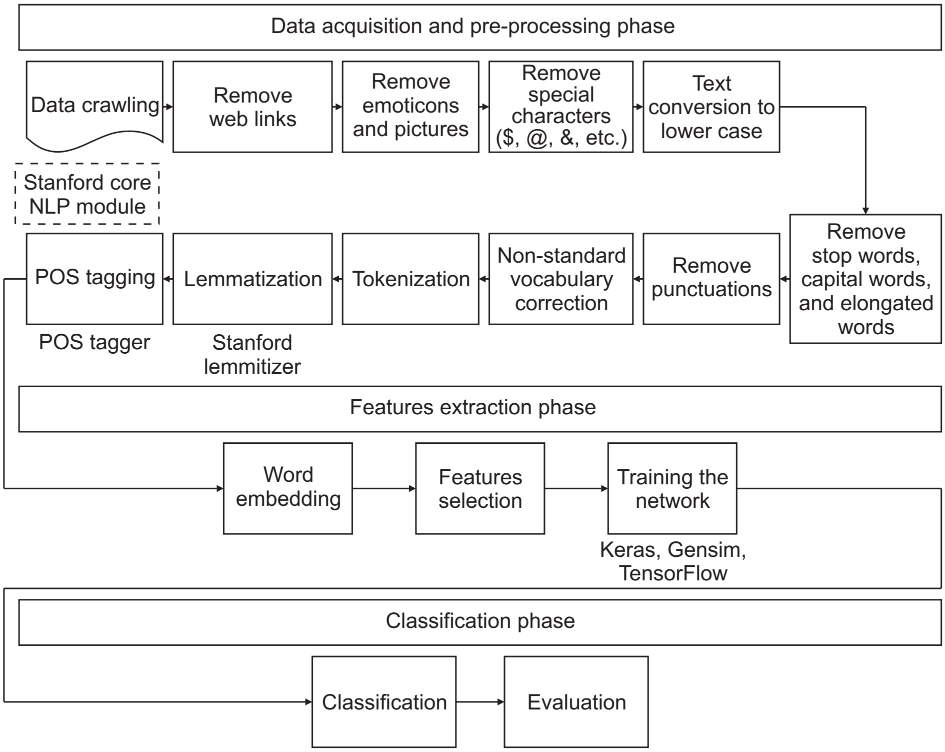
Data acquisition and pre-processing phase
Data crawling
Remove web links
Remove emoticons and pictures
Remove special characters ($, @, &, etc.)
Text conversion to lower case
Stanford core NLP module
Remove stop words, capital words, and elongated words
Remove punctuations
Non-standard vocabulary correction
Tokenization
Lemmatization
POS tagging
POS tagger
Stanford lemmitizer
Features extraction phase
Word embedding
Features selection
Training the network
Keras, Gensim, TensorFlow
Classification phase
Classification
Evaluation

Figure 1. The overall analytical framework of the study. NLP: natural language processing, POS: part-of-speech.

Table 1. Sample posts along with their corresponding labels

Post content	Label	Class label
Yay, I finally hit my week mark today. I am not going to lie. This weekend was tough but I made it through without smoking. Hope everyone else is doing well.	Obviously not smoking	Positive
Well, it's still only like 5.60 for a pack of reds here in Tulsa, Oklahoma. I have not quit yet but have my date set.	Obviously smoking	Negative
I work in a hospital with cancer patients and it still has not fazed me yet. It takes strength. You also have to want it. Good Luck.	Unidentified	Negative

added feature sets 3, 4, and 5; model 6 included feature sets 3, 4, 5, and 6; and lastly, model 7 comprised feature sets 3, 4, 5, 6, and 7 (Table 2).

2. Model Architecture

It is a major challenge to identify a learning algorithm for text analytics that takes individual word vectors as an input and transforms them into a feature vector. Several methods exist to generate word vectors, in which sentences are transformed into a matrix using word embedding [18,19].

In this study, DL methods were employed to process the sequential data. A recurrent neural network (RNN) for sequence encoding was used to input every word into the model and explain the overall meaning of each post [8]. LSTM networks contain a memory block, which includes input, forget, and output (gates) with self-recurrent connection neurons. LSTM networks reorganize computing nodes built on different RNNs [20]. To compute the similarity between words at different LSTM gates, an LSTM network employs two vectors (W_t and K_t) as long-term and short-term memory, respectively; these vectors can be depicted as semantic meanings and are upgraded as the RNN shifts between different words in sequence t [21]. The model used the embedding K_f at the last time step (i.e., for earlier word tokens in the post) as the feature representation for the textual content. The entire process explained the significance of specific content in user-generated posts.

3. Experimental Design

1) Training and test strategy

The algorithms in Stanford CoreNLP were trained using all blog posts and comments as a learning set. During the experiments, the dataset was split into a testing dataset (30%) and a training dataset (70%). To avoid sample biasing and

Table 2. Examples of feature sets alongside community posts

Feature sets	Post content
Feature set 1: Includes standard unigram text feature from a particular post.	-
Feature set 2: Contains bigrams of focal posts.	We used the BOWs approach by performing data pre-processing. Standard unigram and bigram text features are popular features for text classification tasks, and have been used previously to identify users' quit status. The second set contains the Doc2Vec feature set, a document embedding method in which each document is represented as a vector matrix.
Feature set 3: Includes social features influencing smoking behavior, such as family, social network, physicians, and social media sites.	Happy Milestones to Angelina. Love is good! Thank you doctor for your support in a less painless way. Thanks for my family and friends for helping me grow my sobriety.
Feature set 4: Contains domain-related features to highlight the smoking status of community members in OSCC. The entire first-person pronoun falls under this category, which indicates the author's own smoking status. These posts also contain the duration of authors' abstinence. We created the list of time span mentioned in these posts, including "hour," "day," "week," "month," and their possible abbreviations, such as "hrs" and "days." Moreover, the characteristics of the author who writes a post may also play a vital role in the classification task. Users who actively participate in an OSCC are often abstinent.	a) Almost Day 3 for me, and I' m worried about the weekend coming up, too. b) Today is my Day 7 still hasn't smoked or drinks.
Feature set 5: Includes focal post authors' activities in the community as author-based features.	For each post author, we separated the total duration of being a community member and the number of posts published by each author. Both these features were calculated since a user joined the community until s/he published the post.
Feature set 6: Comprises thread-based features, which investigate the entire "thread" that the post corresponds to.	For each particular post, we mined the length of each post (i.e., number of words), number of posted comments in the thread, number of individual users who posted to the thread, and duration of the thread activity.
Feature set 7: Includes all the replies and remarks to a post within the same thread, which were considered as adjacent posts. We include all posts and their neighboring comments, which clearly indicate abstinence.	Hey Joe, thanks for the encouragement. I have been an ex for 7 days. Today has not been too bad, and I keep exciting.

OSCC: online smoking cessation community, BOW: bag-of-words.

uneven distribution, we conducted various shuffling steps in the dataset. All ML algorithms, including LSTM models, used the same data split ratio between the training and test sets. Next, this study employed 10-fold cross-validation to reduce the bias associated with random sampling of the training data. The rationale behind this decision was that previous research successfully evaluated the performance of an algorithm using the above two criteria [20]. Following the initial fully connected layer, the dropout method for regularization was used to reduce the overfitting problem in the training dataset [22].

The sequential frameworks required for the optimization techniques and regularization parameters are listed in Table 3. The parameter selection depended on our engineering knowledge. Furthermore, dropout as a regularization method was used to diminish overfitting. In contrast, hyper-parameter optimization was performed using cross-validation. The Adam stochastic optimization algorithm [23] was used as the optimizer, and binary cross-entropy loss was applied to train the entire model [24]. Moreover, the sigmoid func-

Table 3. Hyperparameters for machine learning and LSTM algorithms

Algorithm	Hyperparameter	Value
AdaBoost	Number of estimators	250
	Base estimator	Decision stump
	Learning rate	0.1
GBDT	Number of estimators	250
	Learning rate	0.1
	Maximum depth	5
	Minimum samples at leaf node	2
XGBoost	Number of estimators	500
	Learning rate	0.001
	Maximum depth	3
	Regularization coefficient	0.0001
	Gamma	0.1
LSTM	Mini batch size	256
	Number of layers	2
	Optimization method	Adam
	Loss	Binary cross-entropy
	L2 regularization coefficient	1e-4
	Dropout	0.25
	Epochs	200
	Output activation	Sigmoid
	Learning rate	0.001

AdaBoost: adaptive boosting, XGBoost: eXtreme gradient boosting, GBDT: gradient boost decision tree, LSTM: long short-term memory.

tion was used as an output activation function to the final layer after merging the output from the hidden states and the inputs from each time point into a range of probabilities (0–1).

2) Classification tasks

We classified the users' posts with using ML algorithms—SVM, LR, adaptive boosting (AdaBoost), gradient boost decision tree (GBDT), and eXtreme gradient boosting (XGBoost)—and LSTM as a DL algorithm. These algorithms were implemented in Python using the sci-kit learn module. For the ML algorithms, we first created a document-word matrix. Each user post in the corpus was represented in the row matrix, whereas each column denoted a word occurring in the user post. Words other than nouns, verbs, adjectives, or adverbs were screened out. We chose highly unique words using the chi-square statistic [25]. From each category, only the top 10% of words were retained for further analysis. Finally, the algorithms were trained and tested.

For the LSTM model, we performed word embedding with the Skip-Gram model [26]. We used posts scraped from the smoking cessation community for word embedding. After removing characters that did not represent any word, the corpus included 797,150 words. The outcomes of word embedding contained a vocabulary of 85,753 words. Every was depicted by a vector of 300 elements, corresponding to the better-performance vector size. Each word of the user post was mapped to the parallel vector. Lastly, the LSTM algorithms were trained and tested. Let $U \in \Re^{N \times m}$, an output matrix obtained by LSTM. The attentive pooling layer output is expressed as follows:

$$H = \tan h(U) \tag{1}$$

$$\alpha = \text{softmax}(w^{\alpha T} H) \tag{2}$$

$$z = \alpha U^{T} \tag{3}$$

where $w^{a} \in \Re^{N}$ denotes the learning parameter, $\alpha \in \Re^{m}$ represents the attention weight vector, and $z \in \Re^{m}$ is the attentive pooling layer output. Opinions are expressed implicitly in everyday life; therefore, for each context word, the value of α would be different for every sentence in a post. The learned feature vectors were all combined for binary-class text classification ("quit or not quit") in the softmax layer.

III. Results

1. Concordance of Measures

Regarding the input feature sets of each of the seven models (models 1 to 7), feature sets 5, 6, and 7 were closely correlated with smoking status (Table 4, below diagonal entries). In contrast, measures with different types were poorly correlated. Smoking status was either not correlated or only somewhat correlated with feature sets 1 and 4 (Table 4, above diagonal entries). Feature sets 2 and 3 were moderately correlated with smoking status (Table 4, above diagonal entries).

The concordance of categorical smoking status with feature sets 1, 2, and 3 did not matched poorly (kappa = 0.04–0.23) (Table 5).

2. Identifying Smoking Status among OSCC Members

Of the 3,833 users in the study sample, 3,623 (94.52%) had written at least one post between the date of registration on the community website and the ending date of data collection, from which their smoking status was identified (average number of posts = 2). Posts suggesting that the user still smoked were usually written within a few days of users' registration (the average time until the first post on "smok-

Table 4. Correlations among input and output variables

	Smoking status	Feature set 1	Feature set 2	Feature set 3	Feature set 4	Feature set 5	Feature set 6	Feature set 7
Smoking status	1.00	−0.23	0.31	0.33	−0.15	-	-	-
Feature set 1	-	1.00	−0.23	−0.31	−0.06	-	-	-
Feature set 2	-	-	1.00	0.34	0.17	-	-	-
Feature set 3	-	-	-	1.00	0.30	-	-	-
Feature set 4	-	-	-	-	1.00	-	-	-
Feature set 5	0.72	-	-	-	-	1.00	-	-
Feature set 6	0.76	-	-	-	-	0.67	1.00	-
Feature set 7	0.82	-	-	-	-	0.71	0.75	1.00

Table 5. Concordance matrix for selected feature sets

			Quit	Not quit	Kappa
Feature set 1 (n = 41)	Smoking status	Quit	14	9	0.23
		Not quit	10	8	
Feature set 2 (n = 44)	Smoking status	Quit	15	13	0.04
		Not quit	9	7	
Feature set 3 (n = 47)	Smoking status	Quit	17	15	0.15
		Not quit	8	7	
Feature set 4 (n = 52)	Smoking status	Quit	19	17	0.10
		Not quit	9	7	

ing" was 2 days after registration). For 191 users, there were multiple posts indicating that they had quit, followed by a subsequent post indicating that they had resumed smoking.

Furthermore, 3,429 users wrote at least one post in which they reported a "quit" with information their quitting status (average number of posts = 2). On average, participants posted their first quit post 2 weeks after becoming members of the community (median = 14 days). The median interval of the quit posts was 5 days after the inferred date of quitting.

A total of 2,417 blog posts with 28,031 comments published by 1,733 users indicated that the author was not smoking at the time of the post. Thus, 57% of users (2,184/3,833) who authored a blog or blog comment wrote at least one post suggesting that they had quit smoking for at least a certain period.

3. Experimental Results

We conducted several experiments to show the classification performance for smoking status identification for each baseline and LSTM classifier using seven models. Sklearn. FeatureSelection, a well-known feature selection technique, was used for feature selection from the text. In addition, the performance of ML algorithms was calculated through various performance scores (i.e., accuracy, precision, recall, F-measure, and area under the receiver operating characteristic curve (AUC) in Table 6.

Overall, the study results are divided into two parts. For the ML algorithms, feature sets 3, 4, 5, 6, and 7 showed better execution of the classifier than when only the standard feature sets 1 or 2 were considered (Table 6). Model 7 achieved the best overall results, with accuracies ranging from 76.52% for the LR algorithm to 92.78% for the XGBoost algorithm, which was 9.3% higher than the same algorithm's performance in model 1 (0.834). The XGBoost algorithm achieved better predictive performance than the other methods, as shown by its values of accuracy (92.78%), precision (0.928), recall (0.927), F1-score (0.927), and AUC (0.931).

Furthermore, model 7 with the XGBoost algorithm had the best AUC (0.931), which was 8.5% higher than in model 1 (0.845). The next best algorithm across all models was GBDT, as shown by its accuracy (90.85%), precision (0.904), recall (0.905), F1-score (0.904), and AUC (0.916). The worst algorithm in terms of predictive value for model 7 was the

Table 6. Description of various measures used to evaluate algorithm performance

Model	Algorithm	Accuracy (%)	Precision	Recall	F1-score	AUC
Model 1	SVM	66.09	0.625	0.661	0.642	0.642
	LR	64.41	0.641	0.644	0.642	0.661
	AdaBoost	72.26	0.724	0.723	0.723	0.751
	GBDT	82.12	0.823	0.825	0.824	0.827
	XGBoost	83.45	0.833	0.835	0.834	0.845
	LSTM	85.51	0.859	0.855	0.857	0.823
Model 2	SVM	66.84	0.652	0.668	0.660	0.653
	LR	68.56	0.684	0.686	0.685	0.667
	AdaBoost	75.94	0.745	0.759	0.752	0.751
	GBDT	84.52	0.847	0.878	0.862	0.848
	XGBoost	84.65	0.842	0.845	0.843	0.847
	LSTM	87.68	0.876	0.874	0.875	0.871
Model 3	SVM	70.31	0.704	0.703	0.703	0.685
	LR	70.33	0.703	0.703	0.703	0.751
	AdaBoost	82.52	0.829	0.825	0.827	0.805
	GBDT	85.31	0.850	0.853	0.851	0.855
	XGBoost	85.78	0.856	0.858	0.857	0.857
	LSTM	89.68	0.897	0.896	0.896	0.892
Model 4	SVM	80.40	0.803	0.804	0.803	0.798
	LR	80.50	0.806	0.805	0.805	0.796
	AdaBoost	84.75	0.853	0.847	0.850	0.828
	GBDT	86.14	0.863	0.865	0.864	0.867
	XGBoost	86.25	0.864	0.866	0.865	0.888
	LSTM	90.42	0.901	0.903	0.902	0.907
Model 5	SVM	82.41	0.824	0.825	0.824	0.819
	LR	72.43	0.724	0.724	0.724	0.762
	AdaBoost	84.62	0.840	0.846	0.843	0.826
	GBDT	87.41	0.871	0.874	0.872	0.866
	XGBoost	87.88	0.877	0.868	0.872	0.857
	LSTM	92.13	0.922	0.921	0.921	0.923
Model 6	SVM	84.56	0.826	0.847	0.836	0.840
	LR	74.57	0.746	0.746	0.746	0.780
	AdaBoost	86.77	0.862	0.868	0.865	0.848
	GBDT	89.55	0.892	0.894	0.893	0.898
	XGBoost	89.95	0.898	0.899	0.898	0.875
	LSTM	94.15	0.943	0.942	0.942	0.944
Model 7	SVM	85.58	0.856	0.858	0.857	0.840
	LR	76.52	0.761	0.766	0.763	0.788
	AdaBoost	87.79	0.874	0.878	0.876	0.879
	GBDT	90.85	0.904	0.905	0.904	0.916
	XGBoost	92.78	0.928	0.927	0.927	0.931
	LSTM	97.56	0.974	0.971	0.972	0.977

AUC: area under the receiver operating characteristic curve, SVM: support vector machine, LR: logistic regression, AdaBoost: adaptive boosting, XGBoost: eXtreme gradient boosting, GBDT: gradient boost decision tree, LSTM: long short-term memory.

Table 7. Performance comparison between existing machine learning models and proposed models

Study, year	Accuracy (%)	Precision	Recall	F1-score
Cohn et al. [10], 2017	0.86	-	-	0.860
Pearson et al. [7], 2018	0.91	-	-	0.910
Nguyen et al. [12], 2016	75.40	-	-	-
Zhang and Yang [27], 2014	-	0.85	0.72	0.74
Myslin et al. [28], 2013	0.85	0.82	0.88	0.85
Rose et al. [11], 2017	73.60	-	-	-
Wang et al. [4], 2019	-	-	-	0.759
Proposed method	92.78	0.928	0.927	0.927

Table 8. Overall comparison between the proposed model and other deep learning methods

Method	Accuracy (%)	Precision	Recall	F1-score
Joint AB-LSTM [10]	-	74.47	64.96	69.39
Tree-LSTM[a]	-	79.30	67.20	72.70
Dep-LSTM[b]	-	72.53	71.49	72.00
Proposed method	97.56	0.974	0.971	0.972

AB-LSTM: attention-based bidirectional long short-term memory, Dep-LSTM: dependency-based long short-term memory.

[a] From Lim S, Lee K, Kang J. Drug drug interaction extraction from the literature using a recursive neural network. PLoS One 2018;13(1):e0190926.

[b] From Wang W, Yang X, Yang C, Guo X, Zhang X, Wu C. Dependency-based long short term memory network for drug-drug interaction extraction. BMC Bioinformatics 2017;18(Suppl 16):578.

LR algorithm, as shown by its accuracy (76.52%), precision (0.761), recall (0.766), F1-score (0.763), and AUC (0.788).

For the DL algorithm, the LSTM DL classifier outclassed all other baseline classifiers across all seven models (Table 6). For instance, a blog post, "Hey Joe, thanks for the encouragement. I have been an ex for 7 days. Today has not been too bad, and I keep exciting [*sic*]." In comparison with other algorithms, the performance of LSTM in model 7 in correctly predicting the smoking status of users was better than all other models, as shown by its accuracy (97.56%), precision (0.974), recall (0.971), F1-score (0.972), and AUC (0.977), which were higher than its accuracy (85.51%), precision (0.859), recall (0.855), F1-score (0.857), and AUC (0.823) in model 1, by approximately 12% overall.

The results also showed that adding the feature sets improved the prediction performance of the proposed algorithms to identify users' smoking status. The balance between positive and negative cases was 49.3:50.7 (with positive cases defined as posts for which the author was clearly not smoking). This implies that the proportion of threads containing positive cases was 49.3%. Moreover, the LSTM algorithm achieved the highest AUC of 0.977. Table 7 shows a comparison between previous research regarding social media analytics in OSCCs and our proposed method using baseline ML classifiers. Table 8 presents a comparison between our DL system and the state-of-the-art models that were applied to the same domain in previous research. The results of both tables reveal that our proposed model outperformed the other methods applied in earlier research.

IV. Discussion

The goal of our study was to predict the smoking status ("quit or not") of individual users who posted comments on an OSCC. Previous research has already described OSCC users' behaviors and engagement [17], peer sentiments [3], and social support for smokers trying to quit [6]. These studies either focused on traditional quantitative approaches to data collection (e.g., questionnaire-based surveys) or recent qualitative approaches (e.g., text messages, interviews, and social network analysis). Moving beyond those approaches, the current study addressed a significant gap in previous research to predict OSCC users' smoking status by using ML-based baseline algorithms and LSTM-based DL algorithms.

We added novel features along with user posts by considering social influence features, domain-specific aspects, author-based characteristics, thread-based features, and adjacent posts in our models. The addition of novel features to enhance the performance of our algorithms highlights the importance of these features in identifying users' smoking status in OSCCs. A high concordance between domain-dependent feature sets and smoking status identification supports the validity of those inferences [29].

This study provides several implications for practitioners. For designers of other online platforms such as Text2Quit, BecomeAnEX, Cancer.org, and Reddit, this work could suggest a recommendation system whereby a post could be recommended to a particular user as a real-time intervention in OSCCs. User needs and requirements can be identified automatically through the search combination of personalized post recommendations. Real-time interventions can be embedded into the online platform to deliver assistance to users. For example, if a user's post indicates that he or she has already quit, but still has some desire to smoke due to his or her social circle; the OSCC can recommend others' posts on dealing with quitting and adjusting to a smoking-free life to avoid relapse. Other users who want to lose weight after quitting could seek advice from an online recommendation system and community content on fitness, exercise, yoga, diet, and daily routines a few days after quitting [30].

Users' language may differ from one online community to another depending on the specific addiction being discussed, meaning that certain domain-specific characteristics can vary and may need to be modified with the aid of frequent group users or by reading UGC. However, in most online communities, social influence features are visible. For instance, users may communicate with other community members by posting comments. It does not matter whether such interactions take place in the form of posts suggested by others, allowing the content of "social influence" posts to be leveraged when mining a focal post. Therefore, practitioners should pay careful attention to social influence on their platforms.

This phenomenon might be associated with information overload, which exhausts users while making quick decisions [31]. As such, for smokers to make quick decisions while using online smoking cessation websites, practitioners must invest efforts into resolving the information overload problem. For instance, practitioners could allow users to rate the helpfulness of users' feedback or comments to a particular post in a thread and then present them in descending order based on their helpfulness score and search results. This process could filter and differentiate between helpful and non-helpful user comments. These actions and efforts are also important for tailoring successful online intervention programs in an OSCC.

There are a few limitations to our study. First, the UGC dataset contained a large volume of noisy text; thus, future research can perform different experiments and develop ML techniques that resolve the noisy text problem to improve the performance of classifiers. Second, in a few posts, users mentioned their duration of abstinence; however, our current classification algorithm only detected the current smoking status of users (i.e., quit or not). In further studies, researchers can program ML methods to identify the duration of abstinence using larger datasets.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

ORCID

Adnan Muhammad Shah (https://orcid.org/0000-0002-9638-3514)
Xiangbin Yan (https://orcid.org/0000-0002-8115-6191)
Abdul Qayyum (https://orcid.org/0000-0001-7707-6984)

References

1. Cruz TB, McConnell R, Low BW, Unger JB, Pentz MA, Urman R, et al. Tobacco marketing and subsequent use of cigarettes, e-cigarettes, and hookah in adolescents. Nicotine Tob Res 2019;21(7):926-32.
2. Graham AL, Zhao K, Papandonatos GD, Erar B, Wang X, Amato MS, et al. A prospective examination of online social network dynamics and smoking cessation. PLoS One 2017;12(8):e0183655.
3. Tucker JS, Stucky BD, Edelen MO, Shadel WG, Klein DJ. Healthcare provider counseling to quit smoking and patient desire to quit: the role of negative smoking outcome expectancies. Addict Behav 2018;85:8-13.
4. Wang X, Zhao K, Street N. Analyzing and predicting user participations in online health communities: a social support perspective. J Med Internet Res 2017;19(4):e130.
5. Graham AL, Amato MS. Twelve million smokers look online for smoking cessation help annually: health information national trends survey data, 2005-2017. Nicotine Tob Res 2019;21(2):249-52.
6. Selby P, van Mierlo T, Voci SC, Parent D, Cunningham JA. Online social and professional support for smokers

trying to quit: an exploration of first time posts from 2562 members. J Med Internet Res 2010;12(3):e34.

7. Pearson JL, Amato MS, Papandonatos GD, Zhao K, Erar B, Wang X, et al. Exposure to positive peer sentiment about nicotine replacement therapy in an online smoking cessation community is associated with NRT use. Addict Behav 2018;87:39-45.
8. Wang X, Zhao K, Cha S, Amato MS, Cohn AM, Pearson JL, Pet al. Mining user-generated content in an online smoking cessation community to identify smoking status: a machine learning approach. Decis Support Syst 2019;116:26-34.
9. Cobb NK, Mays D, Graham AL. Sentiment analysis to determine the impact of online messages on smokers' choices to use varenicline. J Natl Cancer Inst Monogr 2013;2013(47):224-30.
10. Cohn AM, Zhao K, Cha S, Wang X, Amato MS, Pearson JL, et al. A descriptive study of the prevalence and typology of alcohol-related posts in an online social network for smoking cessation. J Stud Alcohol Drugs 2017;78(5): 665-73.
11. Rose SW, Jo CL, Binns S, Buenger M, Emery S, Ribisl KM. Perceptions of menthol cigarettes among twitter users: content and sentiment analysis. J Med Internet Res 2017;19(2):e56.
12. Nguyen T, Borland R, Yearwood J, Yong HH, Venkatesh S, Phung D. Discriminative cues for different stages of smoking cessation in online community. In: Cellary W, Mokbel M, Wang J, Wang H, Zhou R, Zhang Y, editors. Web Information Systems Engineering – WISE 2016. Cham, Switzerland: Springer; 2016. p. 146-53.
13. Wang W, Yang X, Yang C, Guo X, Zhang X, Wu C. Dependency-based long short term memory network for drug-drug interaction extraction. BMC Bioinformatics 2017;18(Suppl 16):578.
14. Sahu SK, Anand A. Drug-drug interaction extraction from biomedical texts using long short-term memory network. J Biomed Inform 2018;86:15-24.
15. Tamersoy A, De Choudhury M, Chau DH. Characterizing smoking and drinking abstinence from social media. HT ACM Conf Hypertext Soc Media 2015;2015:139-48.
16. Wellman RJ, O'Loughlin EK, Dugas EN, Montreuil A, Dutczak H, O'Loughlin J. Reasons for quitting smoking in young adult cigarette smokers. Addict Behav 2018;77: 28-33.
17. Zhao K, Wang X, Cha S, Cohn AM, Papandonatos GD, Amato MS, et al. A multirelational social network analysis of an online health community for smoking cessation. J Med Internet Res 2016;18(8):e233.
18. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:135-46.
19. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 2013;26;3111-9.
20. Ma Y, Xiang Z, Du Q, Fan W. Effects of user-provided photos on hotel review helpfulness: an analytical approach with deep leaning. Int J Hosp Manag 2018;71: 120-31.
21. Graves A. Generating sequences with recurrent neural networks [Internet]. Ithaca (NY): arxiv.org; 2013 [cited at 2021 Apr 5]. Available from: https://arxiv.org/abs/1308.0850.
22. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1): 1929-58.
23. Kingma DP, Ba J. Adam: a method for stochastic optimization. Proceedings of the 3rd International Conference on Learing Representations (ICLR): 2015 May 7-9; San Diego, CA.
24. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge (MA): MIT Press; 2016.
25. Caropreso MF, Stan Matwin s, Sebastiani F. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In: Chin AG, editor. Text databases and document management: theory and practice. Hershey (PA): IGI Global; 2001. p. 78-102.
26. Mikolov T. Statistical language models based on neural networks [Internet]. Mountain View (CA): Google; 2012 [cited at 2021 Apr 5]. Available from: http://www.fit.vutbr.cz/~imikolov/rnnlm/google.pdf.
27. Zhang M, Yang CC. Classifying user intention and social support types in online healthcare discussions. Proceedings of 2014 IEEE International Conference on Healthcare Informatics; 2014 Sep 14-17; Verona, Italy. p. 51-60.
28. Myslin M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013; 15(8):e174.
29. Hughes JR, Oliveto AH, Riggs R, Kenny M, Liguori A, Pillitteri JL, et al. Concordance of different measures of nicotine dependence: two pilot studies. Addict Behav

2004;29(8):1527-39.

30. Cole-Lewis H, Perotte A, Galica K, Dreyer L, Griffith C, Schwarz M, et al. Social network behavior and engagement within a smoking cessation Facebook page. J Med Internet Res 2016;18(8):e205.

31. Zhang S, Zhao L, Lu Y, Yang J. Do you get tired of socializing? An empirical explanation of discontinuous usage behaviour in social network services. Inf Manag 2016;53(7):904-14.