Original Paper

# Classification of Twitter Users Who Tweet About E-Cigarettes

Annice Kim[1], MPH, PhD; Thomas Miano[2], MS; Robert Chew[2], MS; Matthew Eggers[3], MPH; James Nonnemaker[3], PhD

[1]Center for Health Policy Science and Tobacco Research, RTI International, Berkeley, CA, United States

[2]Center for Data Science, RTI International, Research Triangle Park, NC, United States

[3]Center for Health Policy Science and Tobacco Research, RTI International, Research Triangle Park, NC, United States

**Corresponding Author:**
Annice Kim, MPH, PhD
Center for Health Policy Science and Tobacco Research
RTI International
2150 Shattuck Ave, 8th Fl
Berkeley, CA, 94704
United States
Phone: 1 510 665 8237
Email: akim@rti.org

## Abstract

**Background:** Despite concerns about their health risks, e-cigarettes have gained popularity in recent years. Concurrent with the recent increase in e-cigarette use, social media sites such as Twitter have become a common platform for sharing information about e-cigarettes and to promote marketing of e-cigarettes. Monitoring the trends in e-cigarette–related social media activity requires timely assessment of the content of posts and the types of users generating the content. However, little is known about the diversity of the types of users responsible for generating e-cigarette–related content on Twitter.

**Objective:** The aim of this study was to demonstrate a novel methodology for automatically classifying Twitter users who tweet about e-cigarette–related topics into distinct categories.

**Methods:** We collected approximately 11.5 million e-cigarette–related tweets posted between November 2014 and October 2016 and obtained a random sample of Twitter users who tweeted about e-cigarettes. Trained human coders examined the handles' profiles and manually categorized each as one of the following user types: individual (n=2168), vaper enthusiast (n=334), informed agency (n=622), marketer (n=752), and spammer (n=1021). Next, the Twitter metadata as well as a sample of tweets for each labeled user were gathered, and features that reflect users' metadata and tweeting behavior were analyzed. Finally, multiple machine learning algorithms were tested to identify a model with the best performance in classifying user types.

**Results:** Using a classification model that included metadata and features associated with tweeting behavior, we were able to predict with relatively high accuracy five different types of Twitter users that tweet about e-cigarettes (average $F_1$ score=83.3%). Accuracy varied by user type, with $F_1$ scores of individuals, informed agencies, marketers, spammers, and vaper enthusiasts being 91.1%, 84.4%, 81.2%, 79.5%, and 47.1%, respectively. Vaper enthusiasts were the most challenging user type to predict accurately and were commonly misclassified as marketers. The inclusion of additional tweet-derived features that capture tweeting behavior was found to significantly improve the model performance—an overall $F_1$ score gain of 10.6%—beyond metadata features alone.

**Conclusions:** This study provides a method for classifying five different types of users who tweet about e-cigarettes. Our model achieved high levels of classification performance for most groups, and examining the tweeting behavior was critical in improving the model performance. Results can help identify groups engaged in conversations about e-cigarettes online to help inform public health surveillance, education, and regulatory efforts.

***(JMIR Public Health Surveill 2017;3(3):e63)*** doi:10.2196/publichealth.8060

**KEYWORDS**

electronic cigarettes; social media; machine learning

## Introduction

E-cigarettes have gained popularity among adults and youth in recent years. Following sustained increases in the use of e-cigarettes by adults from 2010 to 2013 [1], the prevalence of adult e-cigarette use plateaued at 3.7% in 2014 and was reported to be much higher among current cigarette smokers (15.9%) [2]. Despite the slight decline in the use of e-cigarettes by youth from 2014 to 2015, e-cigarettes remain the most commonly used tobacco product among the middle and high school students in the United States, with 16.0% reporting current use in 2015 [3,4]. Although the long-term health effects of e-cigarette use are largely unknown, e-cigarettes commonly contain nicotine, which has negative effects on the adolescent brain [5], along with a range of other chemicals that are harmful to human health [6-10]. In addition, youth who initiate nicotine use with e-cigarettes may transition to combustible tobacco use [11-14], which has been identified as the leading preventable cause of death in the United States [15].

Concurrent with the rapid rise in e-cigarette use, advertising and sharing of information about e-cigarettes have proliferated in recent years. Although advertisements for tobacco products have been banned on television since 1971 in the United States, e-cigarette advertising via television, magazines, outdoor, radio, and Web-based channels has increased dramatically between 2010 and 2013. Approximately 24 million adolescents were exposed to e-cigarette advertising in 2014 [16]. In addition to traditional advertising platforms, e-cigarette–related information and promotional material are widely available through e-cigarette user forums, Web-based marketing, branded websites, and user-generated content on social media sites such as Twitter and YouTube [17,18].

Social media has become a particularly important platform for sharing information about e-cigarettes. The majority of youth (81%) and adults (74%) in the United States use some form of social media [19-21], and the microblog, Twitter, has more than 316 million active users creating more than 500 million brief posts (called *tweets*) daily [22]. Twitter's pervasiveness makes it a convenient tool for e-cigarette manufacturers, enthusiasts, and advocates to promote e-cigarettes actively to a wide audience. Some studies of the content of e-cigarette–related tweets suggest that the overwhelming majority is commercial or promotional in nature [23-25], and many of these tweets offer discounts or free samples [24]. However, recent research suggests that many tweets reflect discussion of policies, personal experiences, and risks and benefits associated with e-cigarette use among individuals and e-cigarette proponents [26]. Another study found that although the majority of Twitter users engaged in social media conversations about e-cigarettes are not affiliated with the e-cigarette industry, e-cigarette proponents (ie, e-cigarette marketing or manufacturing representatives, advocates, and enthusiasts) tweet more frequently and are more likely to highlight favorable aspects of e-cigarette use [27].

Monitoring trends in e-cigarette–related social media activity requires timely assessment of the content of posts and the types of users generating the content to inform regulatory and surveillance efforts. In 2016, the Food and Drug Administration (FDA) finalized a rule extending the agency's authority to regulate e-cigarettes, which includes federal provisions requiring companies that sell e-cigarettes to include warning statements about nicotine on advertising and promotional materials, including content on digital/social media. To ensure that e-cigarette companies are complying with these advertising and labeling restrictions, FDA will need to identify and monitor websites and social media accounts maintained by these companies. Furthermore, as public health researchers continue to use social media data to track and understand emerging issues concerning e-cigarettes, they will need to be able to distinguish between the content from individuals who may be the target of Web-based e-cigarette advertising (eg, young adults) and the content from e-cigarette companies, marketers, or spammers who may be posting content for commercial purposes. Such information could also be useful in the development and targeting of social media campaigns to prevent e-cigarette use.

The proliferation and variety of Web-based information sharing about e-cigarettes presents challenges in differentiating content from different types of social media users. Previous studies have used a range of techniques to identify Twitter accounts that are purely automated (*robots*), human-assisted automated (*cyborgs*), or organic (ie, individuals) [28] and to distinguish between promotional and nonpromotional tweets [25,29]. Less is known about identifying the diversity of user types responsible for generating e-cigarette–related content on Twitter, including vape proponents, promotional marketers, automated spammers, public health agencies, news organizations, and individuals. In a recent study of tweets about e-cigarettes, Lazard and colleagues [26] analyzed clusters of e-cigarette topics (eg, marketing-focused personal experience) to categorize tweets as being generated by marketers, individual users, or e-cigarette proponents. However, this assessment was based on a review of the topics being discussed (eg, personal experience about e-cigarette use must be posted by individual users) and was not informed by analysis of user handles that were tweeting the content. Thus, Lazard and colleagues' attribution of message source may be limited. For example, Lazard and colleagues reported that tweets about e-cigarette policy bans (a common topic cluster identified in the study) were posted by e-cigarette proponents opposing the ban, but these tweets could have been posted by policy makers announcing or promoting the ban. Examining the topic of tweets may not be sufficient for attributing the source of the message. A more detailed assessment of Twitter users' profile and tweet metadata, in addition to the content of their tweets, could provide better insights into the types of users posting the content.

This study demonstrates a novel methodology for automatically classifying Twitter users who tweet about e-cigarette–related topics into five categories of users—individuals, vaper enthusiasts, informed agencies, marketers, and spammers. We used a supervised machine learning approach to predict different types of Twitter users based on their metadata and tweeting behavior. We tested different models, evaluated model performance, and discussed features that are most predictive of each user type. This study expands on previous research studying the content and the types of users who tweet about e-cigarettes [23-25,27] by providing a greater level of granularity in the

classification of users. Findings from this study provide insight into the composition and the characteristics of social media users posting about e-cigarettes, which can help inform future regulatory action.

## Methods

Using a supervised machine learning approach, we developed models to predict different types of Twitter users who tweet about e-cigarettes. First, a random sample of Twitter handles that have tweeted about e-cigarettes was obtained, and our trained human coders examined the handles' profiles and manually labeled a specific user type for each handle. Next, Twitter metadata and a sample of tweets for each labeled user were gathered, and features that reflect users' metadata and tweeting behavior were created. In the final steps, multiple machine learning algorithms to identify a model with the best classification performance were tested. Figure 1 illustrates our approach to developing the classification model, which we describe further in the sections below. This study was exempt from the institutional review board (IRB) review because it used publicly available Twitter data. Our approach to obtaining and analyzing the Twitter data was in compliance with Twitter's terms of service at the time of the study, such as removing tweets that were deleted or made private by the user.

### Phase 1: Twitter Data Source and Manual Annotation of User Types

Using Twitter's enterprise application programming interface (API) platform, Gnip, we collected e-cigarette–related tweets posted between November 2014 and October 2016. A comprehensive search syntax was developed with 158 keywords, including terms such as *ecig*, *vape*, and *ejuice*, as well as popular e-cigarette brands and hashtags, which resulted in approximately 11.5 million e-cigarette–related tweets from 2.6 million unique users. Next, a random sample of the users associated with these tweets was reviewed, and the content of their posts was examined to identify the range of entities tweeting about e-cigarettes. Using a grounded theory approach informed by literature review and guidance from subject matter experts, a protocol was developed for categorizing Twitter users who tweet about e-cigarettes according to the following types: (1) individual, (2) vaper enthusiast, (3) informed agency, (4) marketer, and (5) spammer (see Table 1).

Six coders were trained using the protocol and practice data to classify the user types manually. For each user, the coders reviewed the user's profile page on Twitter, which included a profile description and a sample of recent tweets on their timeline, which may have included e-cigarette and non-e-cigarette topics. Random samples of Twitter users were double coded until at least 300 labeled cases were obtained per user type. Coding discrepancies were resolved by an adjudicator. In total, 4897 users were manually classified according to the user type definitions (see Table 1 for coding results).

**Figure 1.** Approach to classifying Twitter users who tweet about e-cigarettes.

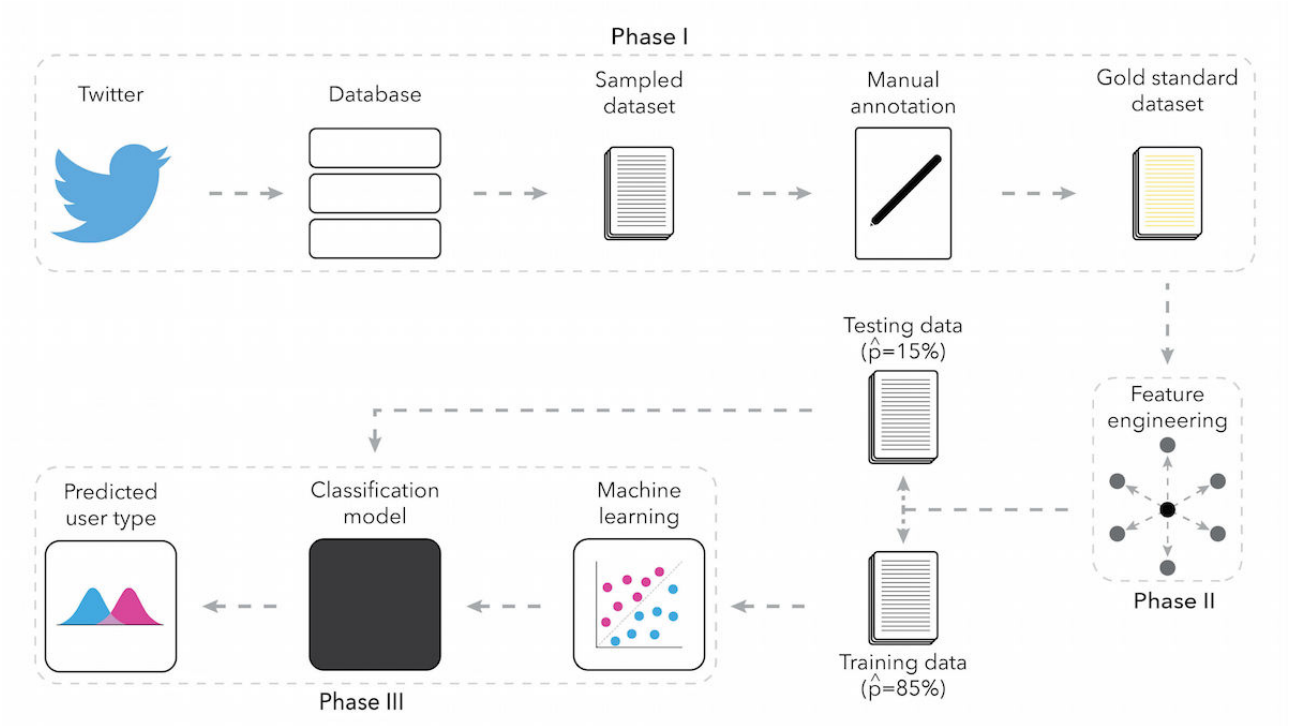


**Table 1.** Manual classification of Twitter users who tweet about e-cigarettes: user type definitions and proportion of each type in manually labeled sample.

| Type | Definition | Sample, N |
|---|---|---|
| Individual | The account of a real person whose Twitter profile information and tweets reflect their individual thoughts and interests. An individual is someone whose primary post content is not about vaping. | 2168 |
| Vaper enthusiast | The account of a person or organization whose primary content is related to promoting e-cigarettes but is not primarily trying to sell e-cigarettes or related products. | 334 |
| **Informed agency**[a] | | 622 |
| News media | The account of a newspaper, magazine, news channel, etc. News media does not include vaping-specific news sources. | |
| Health community | The account of a public health organization, coalition, agency, or credible individual affiliated with an organization. These may also be the accounts of organizations with authority on a topic that should be thought of as *trusted sources*. | |
| **Marketer**[a] | | 752 |
| Marketer | An account marketing e-cigarette or vaping products. These accounts can belong to a Web-based or brick-and-mortar retailer or an individual who is an affiliate marketer. | |
| Information aggregator | An account that primarily aggregates information about e-cigarettes/vaping and where most or all tweets are news articles related to e-cigarettes/vaping. This account could also aggregate vaping coupons or deals. | |
| **Spammer** | An account that does not fall into one of the other coding categories. These accounts often post on a broad range of topics unrelated to this project, and their content can be nonsensical. Anecdotally, it was observed that many of these accounts exhibited *bot* behaviors. | 1021 |

[a]During manual annotation of data, we initially categorized subtypes of informed agency (ie, news media and health community) and marketer (ie, marketer and information aggregator) user types, but we did not identify sufficient numbers of user handles for these subtypes to conduct meaningful analyses. Thus, during the feature selection and modeling phases, we collapsed across user subtypes to define five total user types.

## Phase 2: User Metadata Features and Derived Behavioral Features

Next, we built out the feature space for 4897 labeled users, extracting the metadata provided by the Twitter API and engineering our own features that were derived from the users' tweet text (see Multimedia Appendix 1).

### *User Metadata Features*

The Twitter API provides basic profile information about a user such as screen name, location, bio, number of friends, number of followers, and total number of tweets. The API also provides the actual tweet text and underlying metadata associated with tweet text that was used in this study to characterize the tweeting behavior (eg, retweet) of the users. These types of metadata features have a demonstrated utility in characterizing different types of users [30,31]. Using the Twitter API, 15 metadata features were obtained for each labeled user. Examples of metadata features include number of followers and the number of tweets favorited by the user (see Multimedia Appendix 1).

### *Derived Tweeting Behavior Features*

In addition to the metadata, the users' tweet text data were also examined to capture their tweeting behavior. It was hypothesized that tweeting behaviors would vary across different user types (eg, individuals are likely to tweet about more diverse topics than marketers). Studies have shown that linguistic content of social media posts is particularly useful because it illustrates the topics of interest to a user and provides information about their lexical usage that may be predictive of certain user types [32,33]. For each Twitter handle, the 200 most recent publicly available tweets were collected using the Twitter REST API. Previous studies have shown that 100 to 200 tweets are typically sufficient for predicting Twitter user characteristics [34,35]. These 200 tweets included tweets about e-cigarettes as well as non-e-cigarette–related topics. The non-e-cigarette–related tweets were included because most of the user types examined in this study (eg, news media agency, individuals, and public health agencies) do not tweet about e-cigarettes alone.

To capture the users' tweeting behavior, 58 features derived from the behavioral and linguistic content of the account profile and the tweet text were generated; summary statistics of sets of users' tweets were also created. To generate these features, a variety of text mining techniques were used to capture the distribution characteristics of the users' tweeting behavior and word usage. For example, the minimum, maximum, median, mean, and mode for how many times an e-cigarette keyword was used per tweet was calculated. A term frequency-inverse document frequency matrix of each user's corpus of tweets (up to 200 tweets) was also created, and the pair-wise cosine similarity between each tweet was calculated. For each user, the mean and standard deviation of the set of cosine similarity values, which provided a sense of the semantic diversity and consistency of a user's vocabulary, was calculated. After generating the behavioral features, we dropped nine features in our dataset that had more than 10% missing data. Then, a mean imputation was performed on the derived features that had 10% or less missing data.

### Phase 3: Predictive Models

To determine the best model for classifying the user types, several different algorithms were built and compared using the features described in phase 2. Before modeling, the data were split into a training set (85%) and a test set (15%), using stratified sampling to preserve the relative ratio of classes across sets. To construct our models, a stratified 10-fold cross-validation on the training set was first run and eight different classifiers as well as a *dummy classifier* were evaluated. The dummy classifier—which makes random guesses based on the known distributions of user types in the training data—served as a benchmark for evaluating the performance of our other models. The results from these analyses showed that $F_1$ scores were highest (82.5%) for the Gradient Boosting Regression Trees (GBRT) classifier and lowest for the dummy classifier (28.6%) (see Multimedia Appendix 2 for results of all classifiers).

On the basis of these results, the GBRT algorithm was used to classify the testing dataset. The GBRT approach builds an additive model in a forward stage-wise fashion [36]. The *boosting* technique combines an ensemble of many weak predictive models—in this case, shallow trees—into a single strong one [37]. Each weak model is weighted and trained to be an *expert* on the residuals of the preceding model [38,39] (see Multimedia Appendix 3 for additional information about GBRT and the other algorithms examined).

To determine the best tuning values for the hyperparameters in our model, a fourfold grid-search cross-validation on the training dataset was run (see Multimedia Appendix 3). Then, to evaluate the performance of our tuned GBRT model and the marginal impact of our derived features in improving class differentiation, two separate models were run—one composed of metadata features alone and the other composed of both metadata and derived features. These two separate models were used to evaluate the marginal impact of adding derived features as metadata features for user profile and tweets are easily obtainable, whereas derived features are more labor intensive to create. Finally, the extent of misclassification and the most important features for user types were examined.

## Results

### User Classification Model Results

Table 2 presents the GBRT model results for predicting different types of Twitter users who have tweeted about e-cigarettes. When the complete dataset (metadata + derived features) was tested, the model achieved an average $F_1$ score of 83.3% across all user types. The $F_1$ score was highest for predicting individuals (91.1%) and progressively lower for informed agencies (84.4%), marketers (81.2%), spammers (79.5%), and vaper enthusiasts (47.1%).

The metadata-only model (72.7%) achieved lower $F_1$ scores than the full model (83.3%) (Table 2). Including derived features in the full model improved classification results for each user type, with improvements in $F_1$ scores ranging from 7.5% for individuals to 30.9% for vaper enthusiasts.

**Table 2.** Classification of Twitter users who tweet about e-cigarettes: Gradient Boosting Regression Trees (GBRT) results comparing full model and metadata-only model.

| User type | Full model (metadata + derived data) | | | Metadata-only model | | |
|---|---|---|---|---|---|---|
| | $F_1$ score, % | Recall, % | Precision, % | $F_1$ score, % | Recall, % | Precision, % |
| Individual | 91.1 | 92.3 | 89.8 | 83.6 | 86.2 | 81.2 |
| Vaper enthusiast | 47.1 | 40.0 | 57.1 | 16.2 | 12.0 | 25.0 |
| Informed agency | 84.4 | 78.5 | 91.3 | 70.0 | 67.7 | 72.4 |
| Marketer | 81.2 | 85.9 | 77.0 | 65.6 | 72.6 | 59.9 |
| Spammer | 79.5 | 81.1 | 78.0 | 74.8 | 71.9 | 78.0 |
| Average | 83.3 | 83.7 | 83.3 | 72.7 | 73.7 | 72.3 |

### Misclassification

To further examine variations in the predictive performance across user types, a confusion matrix illustrating predicted and actual user types was generated. Figure 2 shows the distribution of predicted user types on the horizontal axis and actual user types from the manual coding on the vertical axis. To aid in interpretation, the predicted sample proportion for each user type is shaded from light (low proportion) to dark (high proportion). Darker shading in the cells along the diagonal indicates correct classification, whereas darker shading elsewhere indicates misclassification. For example, of the 325 users manually coded as individuals, 300 (92.3%) were correctly predicted to be individuals. In contrast, there was a high level of misclassification of vaper enthusiasts; only 20 of the 50 vaper enthusiasts (40.0%) were correctly predicted to be vaper enthusiasts, whereas 22 (44.0%) were misclassified as marketers.

A two-dimensional (2D) plot of the feature space was also constructed to better understand the extent to which the user types fall into naturally separated clusters (see Figure 3). To accomplish this, a dimensionality reduction method called t-distributed stochastic neighbor embedding (t-SNE) [40] was used to create a 2D representation of the 78-dimensional feature space (see Figure 3). The results of the t-SNE plot indicate that individuals, marketers, and informed agencies fall into fairly discrete clusters, with some users in each class falling closer to other clusters. The plot also shows that whereas spammers are also fairly distinct from other user types, this user type appears to comprise two to three clusters, perhaps suggestive of different subtypes of spammers. Vaper enthusiasts also comprise a distinct

cluster, but there appears to be a substantial overlap between vaper enthusiast and marketer clusters.

**Figure 2.** Distributions of manually labeled versus model-predicted classification of Twitter users who tweet about e-cigarettes.

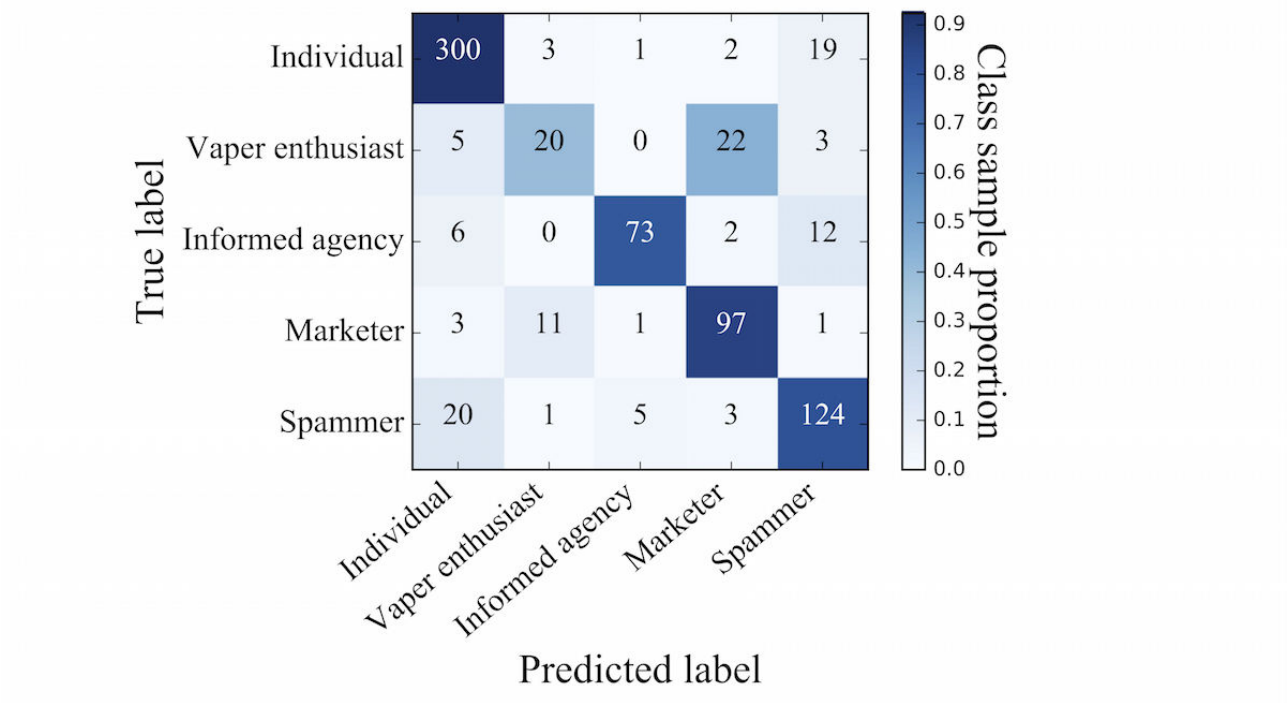


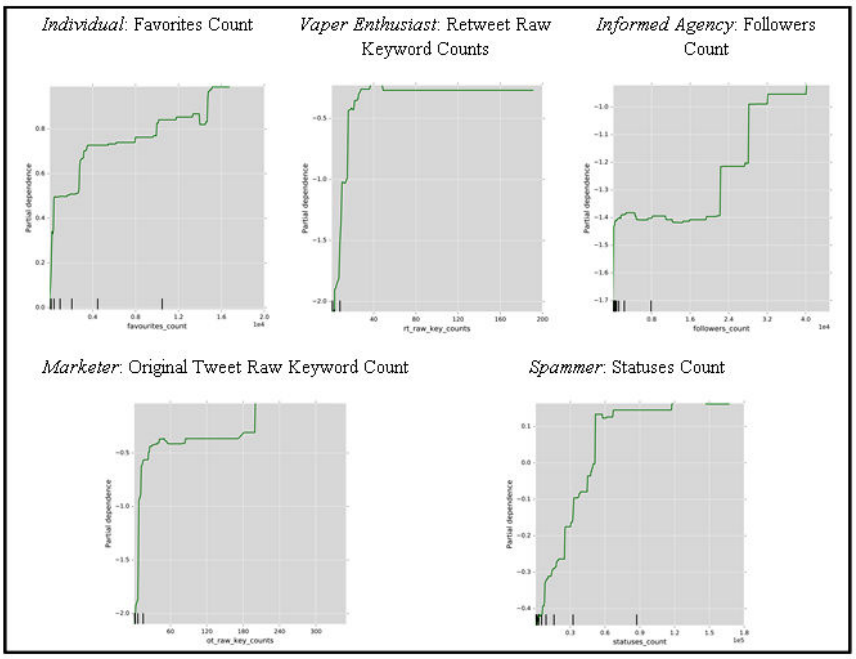

**Figure 3.** Two-dimensional t-SNE visualization of Twitter users who tweet about e-cigarettes.

**Table 3.** Ten most important features in predicting Twitter users who tweet about e-cigarettes across all user types.

| Features[a] | Proportion of feature importance among all variables, % |
|---|---|
| Statuses count | 5.1 |
| Followers count | 4.1 |
| Original tweet raw keyword count | 3.7 |
| Profile description keyword count | 3.3 |
| Original tweet cosine similarity mean | 3.2 |
| Retweet cosine similarity mean | 3.0 |
| Friends count | 3.0 |
| Retweet raw keyword count | 3.0 |
| Listed count | 2.9 |
| Original tweet URL count mean | 2.7 |
| Favorites count | 2.7 |

[a]Most important feature among each user type—Individual: favorites count (4.9%); Vaper enthusiast: retweet raw keyword count (8.3%); Informed agency: followers count (6.5%); Marketer: original tweet raw keyword counts (8.9%); Spammer: statuses count (8.1%).

**Figure 4.** Partial dependence plots of top features by user type for users who tweet about e-cigarettes.




## Feature Importance

To better understand the contribution of each variable in our modeling outcome, each variable was evaluated using Gini Importance, which is commonly used in ensembles of decision trees as a measure of a variable's impact in predicting a label that also takes into account estimated error in randomly labeling an observation according to the known label distributions [41]. Table 3 shows the top 10 most important features, ranked by the proportion of feature importance among all variables in the full model. Results show that two profile metadata features—statuses count and followers count—represent the most important features in the model, with values of 5.1% and 4.1%, respectively. Several derived data features were also important, including original tweet raw keyword counts (3.7%), profile description keyword count (3.3%), and original tweet cosine similarity mean (3.2%). The single most important feature varied among the user types. For individuals, the most important feature was favorites count (4.9%); for vaper enthusiasts, it was retweet raw keyword count (8.3%); for informed agencies, it was followers count (6.5%); for marketers, it was original tweet raw keyword count (8.9%); and for spammers, it was statuses count (8.1%). Feature importance scores for all features examined is available in Multimedia Appendix 4.

Partial dependence plots (PDPs) illustrate the dependence between a target function (ie, user type) and a set of target features. Figure 4 shows PDPs for each user type, illustrating the association between user type and the most important feature for that particular group. Figure 4 shows the most important

features for each user type, whereas Table 3 summarizes the most important features across all user types. For individuals, as the number of tweets the user has liked increases, a given user is more likely to be classified as an individual. For informed agencies, as the number of followers increases, a given user is more likely to be classified as an informed agency. For marketers, as the number of raw keyword counts increases in a given user's set of original tweets, that user is more likely to be classified as a marketer. This indicates that marketers tend to create original content using e-cigarette terms. For spammers, as the total number of statuses (original tweets and retweets) count increases, a given user is more likely to be classified as a spammer. For vaper enthusiasts, as the number of raw keyword counts increases in a given user's set of retweets, that user is more likely to be classified as a vaper enthusiast. This indicates that vaper enthusiasts tend to retweet content with e-cigarette terms.

## Discussion

### Principal Findings

In summary, we developed algorithms with relatively high performance in predicting different types of Twitter users that tweet about e-cigarettes. The rates of precision and recall for *most* user types ranged from 78% to 92%, which was well above the baseline dummy classification and serves as a new baseline for the future user type classification of users who tweet about e-cigarette content on Twitter. Although using metadata features alone in user classification demonstrates performance gains over dummy classification, the results of this study suggest that including additional tweet-derived features that capture tweeting behavior significantly improves the model performance—an overall $F_1$ score gain of 10.6%—beyond metadata features alone. Previous studies have shown that tweet linguistic patterns are strong predictors of social media user demographics [42]. This is the first study to show the predictive utility of tweeting behavior in classifying different types of users who tweet about e-cigarettes.

We achieved the best performance in predicting individuals, informed agencies (news media and health agencies), and marketers. In contrast, vaper enthusiasts were challenging to predict and were commonly misclassified as marketers. There are several reasons why this may be the case. First, it is possible that there were not enough labeled cases of vaper enthusiasts for the machine learning models; there were only 334 labeled cases of vaper enthusiasts (6.8% of all labeled users) compared with 622 to 2168 cases for the other classes. Second, vaper enthusiasts are an evolving group of individuals, and their tweeting behavior may therefore vary more than other established user types such as informed agencies (eg, news media and health agencies). Third, our definition of vaper enthusiasts may not have been distinct enough from marketers; a vaper enthusiast was defined as a user whose primary objective is to *promote but not sell* e-cigarette/vaping products, whereas a marketer was defined as a user whose primary objective is to *market and sell* e-cigarette/vaping products. The distinction of promoting but not selling may have been too subtle to pick up, as vaper enthusiasts promote e-cigarettes by using similar strategies that marketers employ to sell products, such as sharing information about new products, promoting giveaways, and posting product reviews. It is possible that having more labeled cases and extracting more than 200 tweets per handle could improve model performance and better discriminate vaper enthusiasts from marketers. Alternatively, not being able to distinguish vaper enthusiasts from marketers may signal that they share common interests and possible affiliations. With the rise of *social influencer marketing*, where brands incentivize influencers to promote products or subcultures on social media, it is possible that vaper enthusiast messaging may represent commercial marketing interests. The vagueness and ambiguity that was observed between the feature spaces of the vaper enthusiast and marketer classes warrants additional research that examines potential relationships between vaper enthusiasts and e-cigarette commercial entities.

Given the overlap between vaper enthusiasts and marketers, a possible strategy to improve predictive performance might be to combine the two groups. In fact, in their study, Kavuluru and Sabbir [27] classified e-cigarette proponents as "tweeters who represent e-cigarette sales or marketing agencies, individuals who advocate e-cigarettes, or tweeters who specifically identify themselves as vapers in their profile bio." They achieved a high level of accuracy in predicting these e-cigarette proponents (97% precision, 86% recall, and 91% F-score). Although combining these groups may help improve model performance, from a public health perspective, these are distinct groups whose Web-based behaviors have different implications for regulatory agencies. For example, FDA has the authority to regulate claims made by e-cigarette companies and will need to monitor e-cigarette brand social media handles to ensure that they are being compliant with regulatory policies (eg, not making cessation claims, posting warning statements about the harmful effects of nicotine) [43]. In contrast, FDA cannot regulate claims made by vaper enthusiasts because they are individuals and not companies selling e-cigarette products. Therefore, distinguishing vaper enthusiasts from marketers is critical to informing FDA compliance and enforcement efforts. Being able to distinguish vaper enthusiasts from marketers is also important with regard to public health education efforts because vaper enthusiasts have been known to undermine e-cigarette education campaigns. For example, when the California Department of Public Health launched its *Still Blowing Smoke* campaign to educate consumers about the potential harmful effects of e-cigarette use, vaper advocates launched a countercampaign (*Not Blowing Smoke*). By using both hashtags and creating new accounts, the countercampaign attacked the credibility of messages of the California Department of Public Health and effectively controlled the messaging on social media [44]. We would argue that classifying marketers and vaper enthusiasts separately is important for informing e-cigarette surveillance, regulatory, and education efforts; thus, future studies should build on our results and examine methods to improve classification of vaper enthusiasts.

In this study, the top features that were most predictive of each user type were also examined. Individuals like more tweets than nonindividuals; informed agencies have more followers than their counterparts; marketers use more e-cigarette words in their

original tweets than nonmarketers; vaper enthusiasts retweet e-cigarette content more than nonvaper enthusiasts; and more frequent tweeting behavior is indicative of spammers. Given the infancy of this research, the findings of this study should be viewed as an initial inquiry into classifying different types of users who tweet about e-cigarettes. Future studies should build on this work and examine other features that may be predictive of these classes of users. For example, other researchers have examined features such as sentiment of tweets [27] to classify certain subgroups of users who tweet about e-cigarettes.

### Limitations

Our study has several limitations. First, because of resource constraints, we only collected the 200 most recent tweets for the users in our dataset, and some users had less than 200 tweets in total. Previous studies examining Twitter metadata and linguistic features to predict sociodemographic characteristics of users (eg, gender and age) have extracted up to 3200 tweets per handle, but other researchers have also found that having more than 100 tweets per handle did not necessarily improve the model performance [34]. Additional studies are needed to determine whether increasing the number of tweets for each user would increase the importance of the behavioral features in our classification of user types. Second, the methodology involved manual feature engineering, which can be time intensive and is limited to researcher-defined categories. A neural network approach could enable more automated construction of other text-based features that may help in distinguishing user types. Whereas computational text mining methods make it easy to create a multitude of different features, having more features may not necessarily yield information that is useful for classification tasks [31]. Furthermore, issues about scalability and reproducibility should be considered. As social media data are increasingly being used in applied fields such as public health, we need to consider how to balance the resources to conduct this type of analysis with a high level of accuracy and methodological rigor against timeliness and usefulness of the data to inform surveillance and regulatory efforts. Third, the definitions used to classify Twitter users who tweet about e-cigarettes may not be generalizable. Some of the methodologies would be applicable in other contexts (eg, identifying marketers in other domains), but results may not generalize readily across domains.

### Comparison With Prior Work

This is the first study we are aware of that has examined methods to predict a broad set of different types of users tweeting about e-cigarettes. Previous studies have examined either the topic of e-cigarette tweets [23,24] or a single user type (eg, proponents of e-cigarettes vs nonproponents) [27]. In this study, five different categories of users who were involved in public discourse about e-cigarettes and groups that are of interest to inform public health surveillance, education, and regulatory efforts were examined. Second, multiple machine learning algorithms were tested and GBRT was used, which has not been used previously for this purpose. This is important, given the limited work in this area and the lack of existing methodology to build on. Third, in addition to analyzing Twitter metadata features, as prior studies have done, behavioral features that are shown to be important in performance gains were also examined. Finally, by using PDPs, evidence for how important features relate to a given user type was also provided.

### Conclusions

In conclusion, this study provides a method for classifying five different types of users who tweet about e-cigarettes. Our model achieved high levels of classification performance for most groups; examining tweeting behavior was critical in improving the model performance. The results of our approach can help identify groups engaged in conversations about e-cigarettes online to help inform public health surveillance, education, and regulatory efforts. Future studies should examine approaches to improve the classification of certain user groups that were more challenging to predict (eg, vaper enthusiasts).

## Acknowledgments

The authors thank Paul Ruddle II for his guidance on the study design and development of the data collection infrastructure. The authors also thank Geli Fei for his contributions in evaluating text-based features, Bing Liu for his feedback on the analytical approach, and Margaret Cress and Kelsey Campbell for leading the manual classification task. This work was supported by a grant from the National Cancer Institute (R01 CA192240). The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Authors' Contributions

AK conceptualized the study, secured funding, directed study implementation, and led the writing of the manuscript and the revisions. TM led the analysis, implemented the machine learning methods, interpreted the results, produced figures, wrote the Methods and Results sections, and revised the manuscript. RC contributed to the study design, data collection, analysis, and manuscript review. ME assisted with the writing of the manuscript. JN provided feedback on the analysis and the manuscript.

## Conflicts of Interest

None declared.

### Multimedia Appendix 1

List of Twitter metadata features and derived behavioral features used in models to classify Twitter users who tweet about e-cigarettes.

[PDF File (Adobe PDF File), 20KB - publichealth_v3i3e63_app1.pdf ]

### Multimedia Appendix 2

Modeling results of different machine learning algorithms to classify Twitter users who tweet about e-cigarettes.

[PDF File (Adobe PDF File), 16KB - publichealth_v3i3e63_app2.pdf ]

### Multimedia Appendix 3

Overview of machine learning algorithms examined.

[PDF File (Adobe PDF File), 214KB - publichealth_v3i3e63_app3.pdf ]

### Multimedia Appendix 4

Importance of features in models to predict Twitter users who tweet about e-cigarettes.

[PNG File, 196KB - publichealth_v3i3e63_app4.png ]

## References


1. King BA, Patel R, Nguyen KH, Dube SR. Trends in awareness and use of electronic cigarettes among US adults, 2010-2013. Nicotine Tob Res 2015 Feb;17(2):219-227. [doi: 10.1093/ntr/ntu191] [Medline: 25239961]
2. Schoenborn CA, Gindi RM. Electronic cigarette use among adults: United States, 2014. NCHS Data Brief 2015 Oct(217):1-8 [FREE Full text] [Medline: 26555932]
3. Singh T, Arrazola RA, Corey CG, Husten CG, Neff LJ, Homa DM, et al. Tobacco use among middle and high school students--United States, 2011-2015. MMWR Morb Mortal Wkly Rep 2016 Apr 15;65(14):361-367 [FREE Full text] [doi: 10.15585/mmwr.mm6514a1] [Medline: 27077789]
4. Arrazola RA, Singh T, Corey CG, Husten CG, Neff LJ, Apelberg BJ, Centers for Disease Control and Prevention. Tobacco use among middle and high school students - United States, 2011-2014. MMWR Morb Mortal Wkly Rep 2015 Apr 17;64(14):381-385 [FREE Full text] [Medline: 25879896]
5. Counotte DS, Smit AB, Pattij T, Spijker S. Development of the motivational system during adolescence, and its sensitivity to disruption by nicotine. Dev Cogn Neurosci 2011 Oct;1(4):430-443 [FREE Full text] [doi: 10.1016/j.dcn.2011.05.010] [Medline: 22436565]
6. Cheng T. Chemical evaluation of electronic cigarettes. Tob Control 2014 May;23(Suppl 2):ii11-ii17 [FREE Full text] [doi: 10.1136/tobaccocontrol-2013-051482] [Medline: 24732157]
7. Goniewicz ML, Knysak J, Gawron M, Kosmider L, Sobczak A, Kurek J, et al. Levels of selected carcinogens and toxicants in vapour from electronic cigarettes. Tob Control 2014 Mar;23(2):133-139 [FREE Full text] [doi: 10.1136/tobaccocontrol-2012-050859] [Medline: 23467656]
8. Orr MS. Electronic cigarettes in the USA: a summary of available toxicology data and suggestions for the future. Tob Control 2014 May;23(Suppl 2):ii18-ii22 [FREE Full text] [doi: 10.1136/tobaccocontrol-2013-051474] [Medline: 24732158]
9. Pellegrino RM, Tinghino B, Mangiaracina G, Marani A, Vitali M, Protano C, et al. Electronic cigarettes: an evaluation of exposure to chemicals and fine particulate matter (PM). Ann Ig 2012;24(4):279-288. [Medline: 22913171]
10. Williams M, Villarreal A, Bozhilov K, Lin S, Talbot P. Metal and silicate particles including nanoparticles are present in electronic cigarette cartomizer fluid and aerosol. PLoS One 2013;8(3):e57987 [FREE Full text] [doi: 10.1371/journal.pone.0057987] [Medline: 23526962]
11. Primack BA, Soneji S, Stoolmiller M, Fine MJ, Sargent JD. Progression to traditional cigarette smoking after electronic cigarette use among US adolescents and young adults. JAMA Pediatr 2015 Nov;169(11):1018-1023 [FREE Full text] [doi: 10.1001/jamapediatrics.2015.1742] [Medline: 26348249]
12. Leventhal AM, Strong DR, Kirkpatrick MG, Unger JB, Sussman S, Riggs NR, et al. Association of electronic cigarette use with initiation of combustible tobacco product smoking in early adolescence. J Am Med Assoc 2015 Aug 18;314(7):700-707 [FREE Full text] [doi: 10.1001/jama.2015.8950] [Medline: 26284721]
13. Wills TA, Knight R, Sargent JD, Gibbons FX, Pagano I, Williams RJ. Longitudinal study of e-cigarette use and onset of cigarette smoking among high school students in Hawaii. Tob Control 2017 Jan;26(1):34-39. [doi: 10.1136/tobaccocontrol-2015-052705] [Medline: 26811353]
14. Barrington-Trimis JL, Urman R, Berhane K, Unger JB, Cruz TB, Pentz MA, et al. E-cigarettes and future cigarette use. Pediatrics 2016 Jul;138(1). [doi: 10.1542/peds.2016-0379] [Medline: 27296866]

15. US Department of Health and Human Services. surgeongeneral.gov. 2012. Preventing tobacco use among youth and young adults: a report of the Surgeon General URL: https://www.surgeongeneral.gov/library/reports/preventing-youth-tobacco-use/ [accessed 2017-09-11] [WebCite Cache ID 6tOKpWAGy]
16. Duke JC, Lee YO, Kim AE, Watson KA, Arnold KY, Nonnemaker JM, et al. Exposure to electronic cigarette television advertisements among youth and young adults. Pediatrics 2014 Jul;134(1):e29-e36 [FREE Full text] [doi: 10.1542/peds.2014-0269] [Medline: 24918224]
17. Richardson A, Ganz O, Vallone D. Tobacco on the web: surveillance and characterisation of online tobacco and e-cigarette advertising. Tob Control 2015 Jul;24(4):341-347. [doi: 10.1136/tobaccocontrol-2013-051246] [Medline: 24532710]
18. Chu K, Sidhu AK, Valente TW. Electronic cigarette marketing online: a multi-site, multi-product comparison. JMIR Public Health Surveill 2015;1(2):e11 [FREE Full text] [doi: 10.2196/publichealth.4777] [Medline: 27227129]
19. Pew Research Center. Pewinternet. 2017 Jan 12. Internet/broadband fact sheet URL: http://www.pewinternet.org/fact-sheet/internet-broadband/ [accessed 2017-05-04] [WebCite Cache ID 6qDGzTmfa]
20. Pew Research Center. Pewinternet. 2017 Jan 12. Social media fact sheet URL: http://www.pewinternet.org/fact-sheet/social-media/ [accessed 2017-05-04] [WebCite Cache ID 6qDHBq5Nc]
21. Duggan M, Ellison NB, Lampe C, Lenhart A, Madden M. Pewinternet. 2015 Jan 09. Demographics of key social networking platforms URL: http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/ [accessed 2017-05-04] [WebCite Cache ID 6qDHHNrts]
22. Twitter. URL: https://about.twitter.com/company [accessed 2015-03-06] [WebCite Cache ID 6WplPQ8Cl]
23. Huang J, Kornfield R, Szczypka G, Emery SL. A cross-sectional examination of marketing of electronic cigarettes on Twitter. Tob Control 2014 Jul;23(Suppl 3):iii26-iii30 [FREE Full text] [doi: 10.1136/tobaccocontrol-2014-051551] [Medline: 24935894]
24. Clark EM, Jones CA, Williams JR, Kurti AN, Norotsky MC, Danforth CM, et al. Vaporous marketing: uncovering pervasive electronic cigarette advertisements on Twitter. PLoS One 2016 Jul;11(7):e0157304 [FREE Full text] [doi: 10.1371/journal.pone.0157304] [Medline: 27410031]
25. Kim AE, Hopper T, Simpson S, Nonnemaker J, Lieberman AJ, Hansen H, et al. Using Twitter data to gain insights into e-cigarette marketing and locations of use: an infoveillance study. J Med Internet Res 2015;17(11):e251 [FREE Full text] [doi: 10.2196/jmir.4466] [Medline: 26545927]
26. Lazard AJ, Saffer AJ, Wilcox GB, Chung AD, Mackert MS, Bernhardt JM. E-cigarette social media messages: a text mining analysis of marketing and consumer conversations on Twitter. JMIR Public Health Surveill 2016 Dec 12;2(2):e171 [FREE Full text] [doi: 10.2196/publichealth.6551] [Medline: 27956376]
27. Kavuluru R, Sabbir AK. Toward automated e-cigarette surveillance: spotting e-cigarette proponents on Twitter. J Biomed Inform 2016 Jun;61:19-26. [doi: 10.1016/j.jbi.2016.03.006] [Medline: 26975599]
28. Chu Z, Gianvecchio S, Wang H, Jajodia S. Detecting automation of Twitter accounts: are you a human, bot, or cyborg? IEEE Trans Dependable and Secure Comput 2012 Nov;9(6):811-824. [doi: 10.1109/Tdsc.2012.75]
29. Li H, Mukherjee A, Liu B, Kornfield R, Emery S. Detecting campaign promoters on Twitter using Markov random fields. 2014 Dec 14 Presented at: 14th IEEE International Conference on Data Mining (ICDM); 14-17 December 2014; Shenzhen, China.
30. Java A, Song X, Finin T, Tseng B. Why we Twitter: understanding microblogging usage and communities. 2007 Aug 12 Presented at: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis; August 12, 2007; San Jose, CA. [doi: 10.1145/1348549.1348556]
31. Rao D, Yarowsky D, Shreevats A, Gupta M. Classifying latent user attributes in Twitter. 2010 Oct 30 Presented at: SMUC '10 Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents; October 30, 2010; Ontario, Canada.
32. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One 2013;8(9):e73791 [FREE Full text] [doi: 10.1371/journal.pone.0073791] [Medline: 24086296]
33. Nguyen D, Gravel R, Trieschnigg D, Meder T. "How Old Do You Think I Am?": A Study of Language and Age in Twitter. 2013 Presented at: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media; July 8-10, 2013; Cambridge, Massachusetts.
34. Sap M, Park G, Eichstaedt J, Kern M, Stillwell D, Kosinski M, et al. Developing age and gender predictive lexica over social media. 2014 Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25-29, 2014; Doha, Qatar.
35. Volkova S, van Durme B, Yarowsky D, Bachrach Y. Social media predictive analytics. 2015 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies; May 31-June 5, 2015; Denver, CO. [doi: 10.3115/v1/N15-4005]
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(10):2825-2830.
37. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001 Oct 21;29(5):1189-1232. [doi: 10.1214/aos/1013203451]

38. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002 Feb;38(4):367-378. [doi: 10.1016/S0167-9473(01)00065-2]
39. MIT. Artificial intelligence URL: https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/ [accessed 2017-05-04] [WebCite Cache ID 6qDHQ8Kfe]
40. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11):2579-2605.
41. Louppe G, Wehenkel L, Sutera A, Geurts P. nips.cc. Understanding variable importances in forests of randomized trees URL: https://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf [accessed 2017-05-04] [WebCite Cache ID 6qDHWfLsk]
42. Preoţiuc-Pietro D, Volkova S, Lampos V, Bachrach Y, Aletras N. Studying user income through language, behaviour and affect in social media. PLoS One 2015;10(9):e0138717 [FREE Full text] [doi: 10.1371/journal.pone.0138717] [Medline: 26394145]
43. US Food and Drug Administration. Deeming tobacco products to be subject to the Federal Food, Drug, and Cosmetic Act, as amended by the Family Smoking Prevention and Tobacco Control Act; restrictions on the sale and distribution of tobacco products and required warning statements for tobacco products. Fed Regist 2016 May;81(90):28974-29106. [Medline: 27192730]
44. Allem J, Escobedo P, Chu KH, Soto DW, Cruz TB, Unger JB. Campaigns and counter campaigns: reactions on Twitter to e-cigarette education. Tob Control 2017 Mar;26(2):226-229. [doi: 10.1136/tobaccocontrol-2015-052757] [Medline: 26956467]

## Abbreviations

**API:** application programming interface
**FDA:** Food and Drug Administration
**GBRT:** Gradient Boosted Regression Trees
**IRB:** institutional review board
**PDPs:** partial dependence plots
**t-SNE:** t-Distributed Stochastic Neighbor Embedding
**2D:** two-dimensional

*Edited by T Sanchez; submitted 17.05.17; peer-reviewed by AE Aladağ, S Rose; comments to author 17.07.17; revised version received 31.07.17; accepted 14.08.17; published 26.09.17*

*Please cite as:*
*Kim A, Miano T, Chew R, Eggers M, Nonnemaker J*
*Classification of Twitter Users Who Tweet About E-Cigarettes*
*JMIR Public Health Surveill 2017;3(3):e63*
*URL: http://publichealth.jmir.org/2017/3/e63/*
*doi:10.2196/publichealth.8060*
*PMID:28951381*