Toward Automated E-cigarette Surveillance: Spotting E-cigarette Proponents on Twitter

Ramakanth Kavuluru[a,b], AKM Sabbir[b]

[a]*Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, 230E MDS Building, 725 Rose Street, Lexington KY 40536, USA*
[b]*Department of Computer Science, University of Kentucky, Davis Marksbury Building, 329 Rose Street Lexington, KY 40506, USA*

Abstract

Background: Electronic cigarettes (e-cigarettes or e-cigs) are a popular emerging tobacco product. Because e-cigs do not generate toxic tobacco combustion products that result from smoking regular cigarettes, they are sometimes perceived and promoted as a less harmful alternative to smoking and also as means to quit smoking. However, the safety of e-cigs and their efficacy in supporting smoking cessation is yet to be determined. Importantly, the federal drug administration (FDA) currently does not regulate e-cigs and as such their manufacturing, marketing, and sale is not subject to the rules that apply to traditional cigarettes. A number of manufacturers, advocates, and e-cig users are actively promoting e-cigs on Twitter.

Objective: We develop a high accuracy supervised predictive model to automatically identify e-cig "proponents" on Twitter and analyze the quantitative variation of their tweeting behavior along popular themes when compared with other Twitter users (or tweeters).

Methods: Using a dataset of 1000 independently annotated Twitter profiles by two different annotators, we employed a variety of textual features from latest tweet content and tweeter profile biography to build predictive models to automatically identify proponent tweeters. We used a set of manually curated key phrases to analyze e-cig proponent tweets from a corpus of over one million e-cig tweets along well known e-cig themes and compared the results with those generated by regular tweeters.

Results: Our model identifies e-cig proponents with 97% precision, 86% recall, 91% F-score, and 96% overall accuracy, with tight 95% confidence intervals. We find that as opposed to regular tweeters that form over 90% of the dataset, e-cig proponents are a much smaller subset but tweet two to five times more than regular tweeters. Proponents also disproportionately (one to two orders of magnitude more) highlight e-cig flavors, their smoke-free and potential harm reduction aspects, and their claimed use in smoking cessation.

Conclusions: Given FDA is currently in the process of proposing meaningful regulation, we believe our work demonstrates the strong potential of informatics approaches, specifically machine learning, for automated e-cig surveillance on Twitter.

Keywords: electronic cigarettes, text mining, text classification

1. Introduction

Electronic cigarettes (e-cigarettes or simply e-cigs) were introduced in the United States (US) in 2007 [1] and are currently a popular emerging tobacco product across the world. An e-cig essentially consists of a battery that heats up liquid nicotine available in a cartridge into a vapor that is inhaled by the user [2]. E-cig users are termed vapers and the process of using an e-cig is called vaping. E-cigs are similar to conventional tobacco cigarettes with regards to visual, sensory, and behavioral aspects and hence were observed to reduce craving [3]. Owing to their recent introduction, there are very few studies on e-cig safety, risk of abuse, and their efficacy as a smoking cessation aid especially about long term use effects. In fact, currently the search phrase `electronic nicotine delivery systems OR e-cigarette OR electronic cigarette` with its plural/hyphenated variants yields 1794 articles in the PubMed search system out of which 1459 ($\approx 81\%$) had dates of publication in 2014 or 2015. Because e-cigs do not generate toxic combustion products that are produced with tobacco cigarettes, they are perceived and also sometimes marketed as suitable alternatives for smoking cessation [4]. However, scientific research to verify these claims is limited and is often inconclusive. On one hand there are studies that indicate comparable or superior effectiveness of e-cigs in smoking cessation [5, 6]. However, there are also results [7, 8] that show no such associations exist between e-cig use and quitting or reduced conventional cigarette consumption. Another recent effort [9] also indicates that passive exposure to e-cigs increases the desire to smoke both regular cigarettes and e-cigs. Nevertheless, current research seems to indicate that they are less harmful than traditional cigarettes [10].

The ongoing healthy scientific debate around e-cigs is welcomed by the society, especially by regular smokers who are interested in quitting or adopting less harmful alternatives. However, lack of FDA regulation (except for therapeutic use) has heavily increased marketing of e-cigs on the Web [11] and through television ads [12] even if individual states have recently started to enact their own regulations to limit sales, marketing, and use [13]. According to a 2013 Centers for Disease Control and Prevention (CDC) report [14], e-cig consumption doubled in middle and high school students from 2011 to 2012. Furthermore, 9.3% of middle and high school ever e-cig users in 2012 have never smoked conventional cigarettes. Alarmingly, this percentage goes up to 20.3% when considering only middle school students. A more recent CDC report [15] shows that e-cig use tripled from 2013 to 2014 among middle and high school students. Since long term safety of e-cigs has not been thoroughly studied yet, the prospect of adolescents developing nicotine dependence could be detrimental to public health in future generations. When considering adult smokers, however, the significant increase in e-cig awareness has reduced their perception of e-cigs as being less harmful compared with regular cigarettes [16]. Since Web based advertising and discussion still plays a major role in e-cig marketing and use and given one in four online US teenagers uses Twitter [17], we believe it is critical to study the landscape of e-cig messages and their authors on Twitter.

Although e-cig message themes and author classification might be highly granular, in this pilot project we take a simpler approach to tweet author classification – each tweeter is either a "proponent" or not for our purposes. Proponents are tweeters who represent e-cig

Email addresses: `ramakanth.kavuluru@uky.edu` (Ramakanth Kavuluru), `akm.sabbir@uky.edu` (AKM Sabbir)

sales or marketing agencies, individuals who advocate e-cigs, or tweeters who specifically identify themselves as vapers in their profile bio. Essentially these tweeters are generally more inclined to support e-cigs regardless of their specific motivation (e.g., business, lobbying, smoking cessation). In this paper, based on a hand-labeled dataset of 1000 tweeter profiles, we build machine learned models to automatically identify proponents. We subsequently use this model to analyze the content of tweets generated by proponents in comparison with other tweeters along several well known e-cig themes (e-cig flavors, harm reduction, smoke-fee aspect, and smoking cessation) using straightforward text processing. We demonstrate that proponents are many times as likely to highlight the attractive (and sometimes scientifically not yet verified) aspects of e-cigs compared with regular tweeters. To our knowledge this is the first attempt in identifying proponents and a first step in building a framework for automatic surveillance of e-cig related chatter.

2. Background and Related Work

Since its introduction in 2006, Twitter has grown into one of the top 15 visited websites [18] in the world with 100 million daily active users who generate over 500 million tweets per day [19]. The asymmetric network structure of Twitter inherently supports information diffusion and given that a recent study [20] reveals that over 95% of Twitter profiles are public, mining tweets is a practical tool to measure user engagement with various events and products. Since users are not required to publicly declare personal information, several recent studies have focused on identifying user demographic attributes such as age groups and life stages [21], gender [22], and race and ethnicity [23].

In the context of public health, Twitter based automatic syndromic surveillance has been shown to have high correlation with traditional surveillance methods [24, 25] with the added advantage of near real time access to trends, especially in the early epidemic stages [26, 27]. Recent efforts also noted Twitter's suitability for promoting health literacy [28], encouraging fitness activity [29], and monitoring drug safety [30]. In the context of tobacco control advocacy, researchers found significant reach through Twitter in obtaining signatures for an online petition to drop tobacco sponsorship for an international music concert in Indonesia [31]. Another recent Twitter based study focused on emerging tobacco products [32] found high prevalence of positive sentiment for hookah and e-cig. It also successfully demonstrated the application of machine learning methods in automatically identifying tobacco related tweets, constituent themes, and sentiments.

The most relevant effort in the context of our paper is from Huang et al. [33] who automatically identify "commercial" tweets from a corpus of nearly 73,000 tweets collected in the months of May and June in 2012. For their purposes tweets that contain links to sales websites and promotional messages are all commercial regardless of who posts them (regular tweeters vs e-cig marketers/advocates). They use DiscoverText, a cloud based commercial text analytics software program, to semi-automatically (Naive Bayes with additional heuristics) classify tweets as commercial or not. They report that 90% of their tweets are commercial and nearly 10% of such tweets mention smoking cessation. In our effort, instead of identifying commercial tweets, we identify e-cig proponent Twitter profiles using supervised machine learning. We believe this is a more direct approach that aids in electronic surveillance efforts needed in the immediate future to monitor e-cig marketing/publicity

practices and as such complements Huang et al.'s effort. Furthermore, compared with 2012, there is an order of magnitude increase in the number of e-cig tweets (based on an official quote from Twitter Inc.) and our experiments are conducted on a corpus of over one million e-cig related tweets. After identifying proponent profiles, we conducted analyses along well known e-cig themes to see differences in tweeting behaviors of proponents and regular tweeters.

3. Datasets and Annotation

We used two different datasets of e-cig related tweets: the first set of 224,000 tweets was obtained using rate limited Twitter streaming API during the months of September to December 2013 and a second dataset of nearly one million tweets (purchased from the exhaustive Twitter firehose[1]) for the month of March 2015 that match the query terms: `electronic-cigarette, e-cig, e-cigarette, e-juice, e-liquid, vape-juice,` and `vape-liquid`. Variants of these terms with spaces instead of hyphens or just without the hyphens (for matching hashtags) were also used in the query. These terms were chosen in consultation with a faculty member in the College of Nursing at the University of Kentucky (UKY) who works on tobacco policy research. They are specific enough and empirically shown to result in a 99% match to actual e-cig related tweets [33]. The juice/liquid terms represent the liquid nicotine cartridges that need to be refilled for the vaping devices. The smaller dataset represents a free sample curated during a four month period and the second larger dataset constitutes a full dataset from one month. Thus, we get a recent estimate of around 31,000 e-cig tweets per day (as of March 2015) compared with about 1200 such tweets per day in mid 2012 [33], indicating a 25 fold increase.

Our supervised prediction of proponents on Twitter relies on each tweeter's Twitter username, profile bio/description, and their latest tweets. The username or handle is unique to each tweeter and can be up to 15 characters in length. Optionally, each tweeter can also choose to write a profile bio of at most 160 characters to characterize his/her persona on Twitter. If a tweeter never tweets about e-cigs (matches none of our query terms) during a period of surveillance, we automatically assume that they are not a proponent. Thus only those tweeters who have authored at least one e-cig tweet are considered as candidates for classification. Due to reasons that we elaborate later, we classify profiles with empty bios differently compared with those that have non-empty profile descriptions. Based on a sample of 34,000 unique users with at least one e-cig tweet during our four month sample collection, we determine that approximately 20% of e-cig tweeters have empty profile descriptions where the tweeters choose not to provide a bio. Our supervised approach is designed for the 80% of e-cig tweeters with non-empty bios. We use a simpler lexical pattern matching approach for tweeters with empty bios. Next, we outline the training dataset creation for classifying tweeters with non-empty bios.

We randomly chose 1000 tweeter profiles with non-empty bios from the 2013 e-cig tweet dataset. Two annotators independently annotated each of those profiles with a positive

[1]Twitter's terms of use for purchased datasets allow for reporting of aggregate analyses but not presentation of full tweets. Hence all examples shown in this paper are from datasets collected through their rate limited API calls.

(proponent) label if the bio indicates that they are in e-cig sales/marketing or if they are advocates or vapers. This seemed reasonable given our manual overview of the dataset indicated that bio text is often a strong indicator of tweeter perception of e-cigs. For example, the following are a few real bios from our dataset.

- "a vaper trying to help other find that sweet state of vape."
- "I love e-cigs. If anyone wants to buy, I can hook you up."
- "a Finnish vaping advocate who semi occasionally enjoys rambling on youtube"
- "Proud vaper and constitutionalist. I support vaping and the constitution as is was written."
- "Dedicated to bringing the best deals on vaping needs."
- "Manufacturer and distributor of American made premium vape Juice. We also sell e-cigarettes and supplies from around the world."

If the bio does not give enough evidence to make a conclusive decision, the annotators looked at the recent 200 tweets generated from that profile to see if they are predominantly favoring e-cigs. We note that tweeters who retweet e-cig favoring tweets are also considered proponents even if they are not original authors of those tweets. We reiterate that for our purposes proponents are tweeters who are generally more inclined to support e-cigs regardless of their specific motivation (e.g., business, lobbying, smoking cessation). We obtained "almost perfect" inter annotator agreement with Cohen's $\kappa = 0.88$ based on agreement ranges from Landis and Koch [34]. After ignoring the 43 profiles where the annotators disagreed, we have 957 profiles in the final dataset with 216 proponents and 741 in the negative class.

4. Proponent Classification Model

Given the short size of the profile bio we just used the logistic regression (LR) classifier in the Python Scikit-Learn [35] machine learning library. As indicated in Section 3, we use features extracted from the username, bio, and latest tweets to build the classification model. We split our dataset into 70% training and 30% test splits with stratified sampling of both classes. Using average four fold cross-validation F-scores computed over 200 distinct shuffles on the training dataset, we identified the best feature combination from all the feature types we experimented with as shown in Table 1. Next, we present details of our feature space.

4.1. Tweeter Bio and Tweet Text Features

Unigrams and bigrams are traditional n-gram features typically used in text classification. Besides this, we also use parts of speech of different tokens in both bio and latest tweets' text. In addition to this we also used sentiment score features from bio and latest tweets. For a given bio (or a set of latest tweets taken as a single text blob), our features are the average positive and negative scores of the text computed over all sentiment words in it using the scores available in SentiWordNet 3.0 [36]. Each of the sentiment word scores is a real number in $[0, 1]$ where the higher the value the higher the degree of sentiment (positive or

Feature description	Presence in best combination
Unigrams and bigrams from bio and latest tweets	✓
Part of speech tags of bio and latest tweets	✓
Average positive and negative polarity scores of the bio and latest tweets	✓
Topic distribution of bio: Based on 10 topics for bios and 20 topics for latest tweets generated using LDA based on training dataset profiles	✓
Presence of user mentions, URLs, and punctuation marks in the bio	✗
Length of the bio and average length of latest tweets; also binarized versions of these features: whether the lengths are above the averages determined on training data	✗
Presence of the following terms as substrings in user name: vape, vapor, vapour, vaping, ecig, eliquid, ejuice (e.g., `@ecighunter`, `@askavaper`, `@vapeclub`)	✓

Table 1: Feature groups explored for predicting e-cig proponents

negative) expressed. Thus the average positive and negative scores are also in that range and are indicative of the degree of both types of sentiments present in the bio or latest tweets.

We also included topic modeling based features using bios and latest tweets. The central idea is to apply latent Dirichlet allocation [37] modeling using the MALLET [38] toolkit to the bio text from all profiles in the training dataset and at the test time "fold in" a new bio into the model to infer probabilities $P(t_i|b)$ of the topic t_i given the new bio b. Based on experiments we determined ten is the ideal number of topics for the bios and twenty topics for modeling the latest tweets. This means that for each bio, we will have as features $P(t_1|b), \ldots, P(t_{10}|b)$ for the new bio b. Similarly, for each set of latest tweets we have twenty features based on topic modeling. The ideal number of topics was selected based on the value of $k \in \{10, \ldots, 100\}$, the number of topics, that maximizes average 5-fold cross validation F-score over 200 distinct shuffles of the training dataset.

We also used binary features to incorporate presence of user mentions, URLs, and certain punctuation marks (e.g., !, ?) in the bio. However, these features did not end up in best feature combination. Similarly length of the bio and average length of the latest tweets also did not make it into the final model based on cross validation experiments. Another important binary feature is the presence of e-cig related keywords (specifically vape, vapor, vapour, vaping, ecig, eliquid, ejuice) appearing as substrings of the tweeter user name as shown in the final row of Table 1. For example, consider a candidate tweeter with username `@vapeclub`. Given this username contains "vape" as a substring, this Boolean feature will

fire for this tweeter at training and test times.

4.2. Weighted Model Averaging Ensemble

A key finding in our initial experiments was that simply using latest tweets based and profile bio based features together in the same model is not very helpful since crucial predictive signal in the bio text (being short, about 160 characters) was getting drowned out by features from the tweeter's latest tweets even when we considered only 10 or 20 latest tweets in our experiments. To address this, we applied χ^2 feature selection [39, Chapter 13.5] method for n-grams resulting from the joint model. Without any feature selection we obtained an F-score of 0.78 and applying feature selection by selecting the top $m\%$ of the features with $m = 90, \ldots, 10$ reduced the F-score to 62%. Thus feature selection did not help in this scenario and the original F-score of 0.78 turns out to be much lower than what were able to obtain using a weighted ensemble model described next.

To counter the performance issues in our earlier experiments, we also trained two separate LR models: one based on just the bio and username based features (say, M^b) and the second one based only on latest tweets' features (say, M^r). In our experiments to obtain the best M^r, we conducted cross-validation experiments on training data with various numbers of recent tweets (specifically, with 10, 20, ..., 100 tweets) included in the model and found that latest 20 tweets gave the best performance. The final prediction model $\mathcal{M}$ is the weighted average of positive class probability estimates output by both models. That is

$$P_{\mathcal{M}} = \alpha P_{M^b} + (1 - \alpha) P_{M^r},$$

where $\alpha \in [0, 1]$. We determined that $\alpha = 0.85$ and the recent 20 tweets to be the best configuration for this weighted averaging approach by maximizing cross validation F-scores on the training data. Specifically, to find optimum α we did a simple grid search with 0.1 increments starting with $\alpha = 0.1$ and noticed a maximum F-score at a value of 0.8 for α with dips on either side at 0.7 and 0.9. Then we conducted another grid search experiment with 0.01 increments in the range $0.7 < \alpha < 0.9$ and found a new maximum F-score value at $\alpha = 0.85$. Thus we use $\alpha = 0.85$ in all our experiments in the rest of the paper. Intuitively, this means that profile bio is nearly six times more important than latest tweets in spotting proponents, which is not surprising given our single model with both feature types resulted in very poor performance. The latest tweets for test profiles are collected immediately before making a prediction using Twitter's rate limited API calls. We believe this is a natural set up for this scenario given predictions need to be made with the snapshot of what can be readily obtained, at the time, that would more appropriately reflect the proponent status of a tweeter. The full pipeline for the proponent classification approach described thus far is shown in Figure 1. At training time the model at the right most end is built while at the test time it is used to make predictions.

4.3. Overall Results and Feature Ablation Experiments

Using the best feature combination as represented in Table 1 and the weighted model averaging approach with $\alpha = 0.85$, we trained a model on the training data and tested on 30% test set to obtain a F-score of 0.896 (with precision 0.96 and recall 0.84). Since this is the score on a single run, we considered 500 distinct shuffles of our full dataset. For each

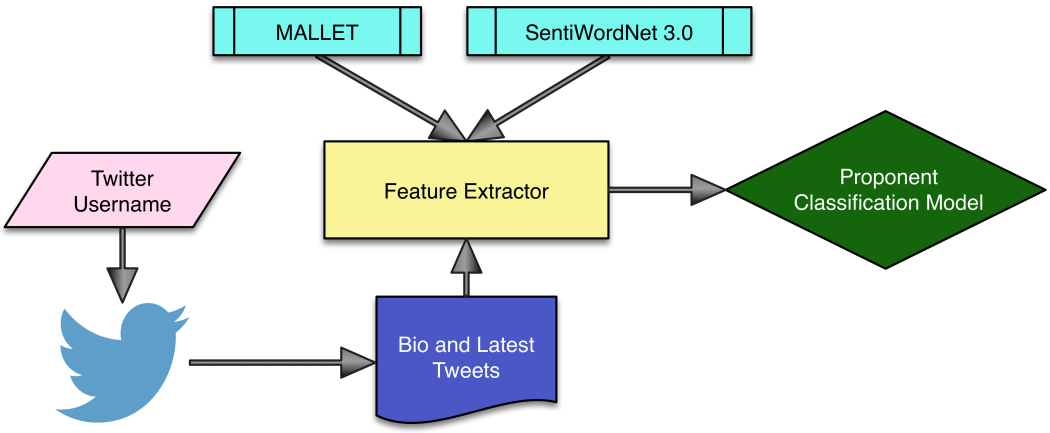

Figure 1: E-cigarette proponent classification pipeline

shuffle, we used stratified sampling (maintaining class proportions) to split it into 80%-20% train-test sets and using the best combination, trained on the 80% set and tested on the 20% set. Using this approach and descriptive statistics, we obtain a mean F-score 0.9152 (0.9147 – 0.9157), mean precision 0.9721 (0.9716 – 0.9726), mean recall 0.8663 (0.8656 – 0.8671), and mean overall accuracy 0.9598 (0.9588 – 0.9608) with 95% confidence intervals shown in parentheses. The simple baseline for accuracy of 77% obtained by always picking the non-proponent class is 18% lower than the lower bound of the 95% confidence interval for that measure, which further warrants the application of our methods.

Using 500 distinct shuffles of the dataset, we ran feature ablation experiments where we removed one feature at a time to measure the performance drop incurred, which indicates the contribution of that feature to the overall model. From ablation results shown in Table 2 we see that dropping latest tweet features causes biggest drop in recall and dropping the username substring feature (last row of Table 1) causes the biggest drop in precision, F-score, and accuracy. POS tag ablation causes a small drop in performance compared with polarity score and topic distribution score removal. This is not surprising because the tweets and bios of proponents seemed to largely revolve around e-cig themes while regular tweeters discuss varied topics and are not focused on e-cigs. For all features, the drop in recall is always larger than the drop in precision and accuracy does not drop as much as F-score.

Overall our efforts have resulted in a very high precision model with reasonable recall for profiles with non-empty bios. Specifically, the feature that incorporates e-cig related terms as substrings of tweeter username turns out to be a powerful predictor for spotting proponents. We use this specific lexical pattern matching feature to identify proponents for tweeters with empty bios (about a fifth of all e-cig tweeters). That is, for profiles with empty bios we simply see if certain e-cig related terms (last row of Table 1) are in the username and if they are present, we classify corresponding tweeters as proponents. A manual examination of 500 profiles with user names that have these terms as substrings reveals a 99% precision of this approach in identifying proponents for tweeters with empty bios. However, at this point, we don't have a methodical way to do a supervised approach for identifying proponents with empty bios that do not fit this simple criterion. Still, most users whose bios are empty and

	Precision	Recall	F-score	Accuracy
Full model	**0.9721**	**0.8663**	**0.9152**	**0.9598**
– POS tags	0.9596	0.8458	0.8979	0.9474
– Polarity scores	0.9361	0.7838	0.8514	0.9319
– Topic scores	0.9342	0.7869	0.8524	0.9393
– Username	0.8900	0.7802	0.8295	0.9232
– Latest tweets	0.9229	0.7733	0.8394	0.9342

Table 2: Feature ablation results with averages computed over 500 distinct shuffles of the dataset

do not match this user name based rule appear to be regular non-proponent tweeters based on a manual examination of a sample of such profiles.

5. Analysis of Proponent Tweets

In this section, we analyze the tweets generated by proponents and other tweeters along familiar e-cig themes. Before we proceed, we introduce a slight modification to the way we apply the model built in Section 4. We analyzed the confusion matrix of the test set predictions of our model and noticed that there were several false negatives which could have been captured by applying the simple user name substring match approach used for empty profiles (last row of Table 1). Given this particular substring based classification yields near perfect precision (see Section 4), for the rest of the paper we apply our model to non-empty profiles that do not match the user name substring match. However, we note that this user name based identification approach is not comprehensive because proponents do not always use such user names. In our experiments with two datasets in the rest of the section, our full model yields 25-35% (depending on the dataset used: sample vs exhaustive) more proponents compared with the simpler user name based identification.

5.1. Proportions of Proponents and their Tweets

Dataset	Tweeters		Tweets	
	Total	Proponents	Total	Proponents
2013 Sep-Dec sample	34,000	2540 (7.5%)	224,000	32,682 (14.6%)
2015 March subset	100,000	4359 (4.3%)	349,401	72,384 (20.7%)

Table 3: Proponent and corresponding tweet counts in both datasets

To analyze the tweets from proponents compared with the rest of the users, we apply our model to the tweeters in two different datasets introduced in Section 3. The first dataset is a free rate limited sample from a four month period (September to December of 2013) and the

second dataset is an exhaustive dataset of nearly one million tweets for the month of March 2015. This newer one month dataset has nearly 360,000 tweeters. Given our model depends also on latest tweets and Twitter imposes prohibitive rate limits for collecting latest tweets in a timely fashion, we chose to look at 100,000 randomly chosen tweeters and their tweets (nearly 350,000) in the new dataset. The old free API based dataset has 34,000 tweeters and 224,000 tweets in the sample generated by them. Although we consider a subset of the tweeters in the new dataset, we believe this captures a different perspective given we are more likely to hit frequent tweeters with the rate limited free sampling approach but are likely to incorporate more infrequent tweeters with the selected subset from the one month dataset. After applying our classification model to these tweeters in both datasets, we obtain results as shown in Table 3. We notice that the percentage of proponents is 3% more in the older sample compared to the subset of the exhaustive one month dataset. This is not surprising given the subset of the exhaustive sample is more likely to capture tweets from infrequent tweeters who are occasionally tweeting about e-cigs.

The final two columns of Table 3 show the total number of tweets and corresponding sizes of the subsets generated by the proponents. From the first row, for the rate limited sample from 2013, we notice that on average a proponent generates twice as many tweets as other tweeters. Based on the second row we notice that in the 2015 sample, on average proponents tweet more than fives time as other tweeters. We compute this by simply comparing the average number of tweets by proponents ($72384/4359 = 16.6$) with those by regular tweeters: $(349401 - 72384)/(100000 - 4359) = 2.9$. We indicated in Section 3 that the number of tweets per day on e-cigs has increased 25 times compared with the rate in 2012. Here we also observed that the tweets by proponents are also increasing considerably compared with those by regular tweeters. Although it is not surprising that proponents tweet more often, it is nevertheless interesting to quantify this and and to see the dramatic increase from 2012 to 2015.

5.2. Sentiment Analysis of Proponent Tweets

We looked at the mean average positive and negative sentiment scores of tweets (in the 2013 dataset) by the proponents and others groups from our hand-labeled dataset. This was done by obtaining average positive and negative scores per tweet (based on the constituent word scores form SentiWordNet 3.0) and then finally averaging these scores across sets of tweets by proponents and others. The mean positive scores are 0.92 and 0.79 and the mean negative scores are 0.01 and 0.03, for tweet sets from proponents and others, respectively. It has been shown previously [40] that English language tends to have a positive bias on social media (including Twitter) and that also seems to hold for our datasets. However, the proponent tweets conveyed higher positive scores (0.92) compared with others' tweets (0.79), which is not unexpected because their tweets often contain positive words used in the context of e-cigs.

5.3. Thematic Analysis of E-cig Tweets

In this section, we focus on tweet content analysis based on four popular e-cig themes shown in the first column of Table 4. These themes were identified based on consultation with a researcher who works on tobacco policy at UKY. Before we get into details of various themes, we briefly describe different elements of Table 4. The 'total' column in the table

indicates the total number of tweets in the dataset for a specific theme and the 'by props' column indicates the number of those tweets arising from proponents identified through our methods in Section 4. As we can see, although proponents form a very small percentage (third column of Table 3) of the full set of tweeters, they generate a significant proportion of tweets for many of these popular e-cig themes. To compare the tweeting behavior of proponents compared with others, we use the measure

$$\text{rate ratio} = \frac{\dfrac{\text{proportion of thematic tweets by proponents}}{\text{proportion of proponents}}}{\dfrac{\text{proportion of thematic tweets by others}}{\text{proportion of others}}},$$

where the proportions of thematic tweets by proponents are obtained from Table 4 (columns 3 and 6) and proportions of proponents over the full datasets are obtained from Table 3 (column 3). So this is the ratio of the average number of thematic tweets by a proponent to the average number of such tweets by a non-proponent. For example, consider the smoking cessation theme for the 2015 dataset. The proportion of thematic tweets for proponents is 65% (from column 6 of Table 4) and for others is $100 - 65 = 35\%$. Similarly, proportion of proponents is 4.3% in 2015 (column 3 in Table 3) and others is $100 - 4.3 = 95.7\%$. So the rate ratio is $\frac{65/4.3}{35/95.7} \approx 41$ as shown in the last row and last column of Table 4.

E-cig theme	2013 Sep-Dec sample tweets			2015 March subset tweets		
	Total	By props	Rate ratio	Total	By props	Rate ratio
Flavors	4018	2258 (56%)	15	10,855	5207 (48%)	20
Harm reduction	374	193 (51%)	13	1527	1118 (73%)	60
Smoke-free aspect	1902	1033 (54%)	14	11,220	7380 (66%)	42
Smoking cessation	5228	1923 (37%)	7	5863	3820 (65%)	41

Table 4: Thematic distribution of e-cig tweets for proponents vs other tweeters

5.3.1. E-cig Flavors Theme

One of the distinctive features of e-cigs compared with traditional cigarettes is the multitude of flavors available ranging from fruits to desserts. We went through websites of three popular e-cig company websites (Blu, Njoy, and VaporFi as identified in [33]) and curated a set of 22 popular flavors including menthol, strawberry, blueberry, cola, cherry, and mint in the order of their frequency. We simply searched for these flavor names in our e-cig tweets to obtain our counts shown in Table 4. As we notice from the rate ratio, proponents are 15 times more likely to tweet about e-cig flavors based on the 2013 sample and are 20 times more likely to do that based on recent data. FDA might regulate the use of flavors as it had done for regular cigarettes and at least as of now it appears that proponents are heavily tweeting the flavor rich aspect of e-cigs. A caveat here is that some of the tweets mentioning e-cigs might also be discussing consuming the various berries as fruits instead of using

flavored e-cig. However, our manual examination of several hundred flavor containing e-cig tweets revealed that it is very rare and only happens with the chocolate flavor given users often mentioned drinking hot chocolate while vaping. Given this disambiguation issue, the chocolate flavor was ignored in our analysis.

Before we move ahead, we note that our thematic tweet identification was based on filtering with Python regular expressions that model lexical constraints. For the sake of clarity, we present more intuitive lexical expressions in the rest of this section. Furthermore, our focus was on precision of identifying tweets belonging to a particular theme and hence lexical expressions seemed more appropriate given the tweets are already related to e-cigs. These expressions were determined in consultation with a faculty member who works in tobacco policy research at UKY.

5.3.2. Harm Reduction Theme

Another popular theme in e-cig discussions is perceived harm reduction compared with traditional cigarettes. Although there might be merit in claims that e-cigs are not as harmful, publicity of this nature might point youth and other non-smokers to another gateway to form nicotine addiction. Hence this may not be an appropriate way to promote e-cigs in general. We identified e-cig tweets on this topic by searching with the expressions: harm reduction, reduced harm, less harmful, safe[r] than, safe[r] alternative, and healthy/healthier alternative. Note that these searches are only limited to our tweets that are already filtered using e-cig related keywords as discussed in Section 3. From Table 4, we see that in the recent data sample, proponents are 60 times more likely to discuss this aspect compared with other users. The ratio increased by $60/13 = 4.6$ times from late 2013 to early 2015 for this theme.

5.3.3. E-cigs' Smoke-free Aspect

A major hindrance to using traditional cigarettes is the smoke they generate and the smoking bans in place due to that reason. Second hand smoke related consequences might also discourage smokers to reduce their consumption especially in public places and in the presence of family members. Hence e-cig proponents are more likely to highlight the smoke free aspect of vaping. We applied further filtering on our e-cig tweets using the following expressions: 1. smoke[-]free, 2. smoke[-]less, and 3. tokens 'tobacco' or 'smoking' and the word 'alternative' in a tweet. We used optional hyphen or white space for the first two expressions and the third expression requires the occurrence of either tobacco or smoking in the tweet along with word alternative. From the third row of Table 4 we can see that the rate ratio has increased three times from 2013 to 2015. It is also extremely high, at 42, as of early 2015.

5.3.4. Smoking Cessation Theme

Our final theme is smoking cessation with the aid of e-cigs. Given evidence is still being gathered and clinical studies are being conducted to actually test these claims, it may not be appropriate to publicize e-cigs as means to quit smoking. To estimate the popularity of this theme, we filter our e-cig tweets based on the following tweet text constraints

- stop/stopped/stopping smoking

- smoking cessation
- give up, giving up, or given up smoking/tobacco
- quit/quitting and tobacco/smoking in the tweet
- kick/kicked/kicking his/her/their/my/your smoking/tobacco

For this theme, from the last row of Table 4, there is a staggering six fold (the highest among four themes) increase in rate ratio compared with the older dataset. However, we also notice that the absolute volume of cessation related tweets has decreased given the newer dataset has 125,000 more tweets than the 2013 sample.

Although we have been careful in the paper to convey that the thematic tweets simply match a few specific lexical patterns, we would be remiss if we did not also discuss an important shortcoming of this approach. In our filtering we did not account for negated mentions or more generally speaking the polarity of statements that discuss a given theme. For example, our dataset has tweets matching our cessation patterns that mention a research study where e-cigs were not shown to aid in cessation. Similarly tweets that say e-cigs are not less harmful are also included in the harm reduction theme. However, in this work we essentially identified tweets that discuss a theme but not necessarily their polarity, which is a crucial next step in our efforts on this topic. This does not necessarily take away from our analysis because proponents rarely discuss negative aspects of e-cigs especially with regards to harm reduction and smoking cessation. However, it would be interesting to understand the polarity of tweets by the others group along these two themes.

6. Concluding Remarks

E-cigs are a popular emerging tobacco product currently not regulated by the FDA. As such, their sales and marketing are not subject to the stricter rules typically applied to regular cigarettes although individual states have recently enacted laws to regulate them to some extent. In this paper, to aid automated surveillance of e-cigs on social media, we conducted what we believe is the first study to automatically identify e-cig proponents on Twitter. Using a hand-labeled dataset, we built a classification model with features based on tweeter bio, latest tweet text, and user name. Our model achieves a precision of 97% with recall of 86% and can be used to classify new unseen profiles. We applied our model to two different datasets with complementary characteristics collected in late 2013 and March 2015. Our experiments showed that e-cig proponents on Twitter constitute a very small percentage of the tweeters who write about e-cigs. However, they tweet more often (two to five times) compared with other users and are tens of times more likely than others to highlight favorable, but not often scientifically corroborated, aspects of e-cig use. Based on this feasibility study we believe automated surveillance of e-cigs on Twitter is an important research direction that has tremendous application potential especially in the immediate future in the context of impending FDA initiated regulations.

We identify several new research directions that can advance automated surveillance of e-cigs. Most of these tasks involve human annotation of user profiles and tweets to generate training data.

1. In this effort we focused on identifying proponents using a broad definition. However, an important future research direction is to identify fine grained classes, such as sales/marketing profiles, individual e-cig advocates who are not affiliated with any companies, regular e-cig users (even if they don't explicitly advocate e-cigs), and pro-regulation representatives.

2. Given gender, age group, race and ethnicity can be predicted with reasonable accuracy [21, 22, 23], an important immediate future research direction is to use these methods to classify e-cig tweeters into these demographic categories and study e-cig themes in tweets by specific subpopulations. For example, given teenagers, and especially african american teens, are an active group on Twitter [17], studying this specific subpopulation with regards to popular e-cig topics may yield crucial insights into their usage patterns and perceptions.

3. Polytobacco is the practice of simultaneously using multiple forms of tobacco including regular cigarettes, e-cigs, hookah, and snus, which can lead to dangerous nicotine dependence. Another important question is to understand prevalence of polytobacco by spotting tweets that discuss such usage and identifying other forms being used along with e-cigs. Additionally, usage of addiction forming substances such as alcohol, illicit drugs, and prescription drugs along with e-cigs can also be studied by a more refined analysis of tweet content.

4. Another important direction is to identify "popular" tweets and factors contributing to the popularity of different types of e-cig related tweets where popularity is assessed in terms of retweets, replies, and favorites. For example, what tweet characteristics (such as presence of images, URLs, hashtags, numbers of followers) drive the popularity of e-cig sales/marketing tweets vs pro-regulation tweets. For tweets that gathered significant retweet/favorite/reply activity from teenagers, identify factors for such popularity including the proportion of their friends who contributed to such activity before them, their gender/and race. We believe this will not only aid in surveillance, but also in developing strategies to maximize the diffusion of results of scientific research and recommendations from FDA to a broader audience on Twitter, which will be critical to raise awareness.

Acknowledgements

We are grateful to anonymous reviewers whose comments have helped improve the quality of this paper. Many thanks to Ellen Hahn of the College of Nursing at UKY for general discussions on e-cig themes and search terms. This research was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, US National Institutes of Health (NIH), through Grant UL1TR000117 and the Kentucky Lung Cancer Research Program through Grant PO2-415-1400004000-1. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] A. K. Regan, G. Promoff, S. R. Dube, R. Arrazola, Electronic nicotine delivery systems: adult use and awareness of the e-cigarette in the USA, Tobacco Control 22 (1) (2013) 19–23.

[2] J.-F. Etter, C. Bullen, A. D. Flouris, M. Laugesen, T. Eissenberg, Electronic nicotine delivery systems: a research agenda, Tobacco Control 20 (3) (2011) 243–248.

[3] C. Bullen, H. McRobbie, S. Thornley, M. Glover, R. Lin, M. Laugesen, Effect of an electronic nicotine delivery device (e cigarette) on desire to smoke and withdrawal, user preferences and nicotine delivery: randomised cross-over trial, Tobacco Control 19 (2) (2010) 98–103.

[4] R. A. Grana, P. M. Ling, Smoking revolution: A content analysis of electronic cigarette retail websites, American journal of preventive medicine 46 (4) (2014) 395–403.

[5] J. Brown, E. Beard, D. Kotz, S. Michie, R. West, Real-world effectiveness of e-cigarettes when used to aid smoking cessation: a cross-sectional population study, Addiction 109 (9) (2014) 1531–1540.

[6] C. Bullen, C. Howe, M. Laugesen, H. McRobbie, V. Parag, J. Williman, N. Walker, Electronic cigarettes for smoking cessation: a randomised controlled trial, The Lancet 382 (9905) (2013) 1629–1637.

[7] R. Grana, L. Popova, P. Ling, A longitudinal analysis of electronic cigarette use and smoking cessation, JAMA Internal Medicine 174 (5) (2014) 812–813.

[8] K. A. Vickerman, K. M. Carpenter, T. Altman, C. M. Nash, S. M. Zbikowski, Use of electronic cigarettes among state tobacco cessation quitline callers, Nicotine and Tobacco Research 15 (10) (2013) 1787–1791.

[9] A. C. King, L. J. Smith, P. J. McNamara, A. K. Matthews, D. J. Fridberg, Passive exposure to electronic cigarette (e-cigarette) use increases desire for combustible and e-cigarettes in young adult smokers, Tobacco control.

[10] K. E. Farsalinos, R. Polosa, Safety evaluation and risk assessment of electronic cigarettes as tobacco cigarette substitutes: a systematic review, Therapeutic advances in drug safety 5 (2) (2014) 67–86.

[11] A. Slomski, Report shows e-cigarette marketing aimed at youth, JAMA 311 (22) (2014) 2264.

[12] M. McCarthy, Youth exposure to e-cigarette advertising on US television soars, BMJ: British Medical Journal 348.

[13] M.-C. Tremblay, P. Pluye, G. Gore, V. Granikov, K. B. Filion, M. J. Eisenberg, Regulation profiles of e-cigarettes in the united states: a critical review with qualitative synthesis, BMC medicine 13 (1) (2015) 130.

[14] Centers for Disease Control and Prevention, Notes from the field: Electronic cigarette use among middle and high school students – united states, 2011-2012, Morbidity and Mortality Weekly Report Sept.

[15] Centers for Disease Control, E-cigarette use triples among middle and high school students in just one year, http://www.cdc.gov/media/releases/2015/p0416-e-cigarette-use.html.

[16] A. S. Tan, C. A. Bigman, E-cigarette awareness and perceived harmfulness: Prevalence and associations with smoking-cessation outcomes, American journal of preventive medicine.

[17] Pew Research Internet Project, Part 1: Teens and social media use, http://www.pewinternet.org/2013/05/21/part-1-teens-and-social-media-use/.

[18] Alexa, Inc, Alexa top 500 global sites, http://www.alexa.com/topsites (2014).

[19] Twitter, Inc, Registration with United States securities and exchanges commission, http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm (2013).

[20] Y. Liu, C. Kliman-Silver, A. Mislove, The tweets they are a-changin': Evolution of twitter users and behavior, in: Proceedings of the Eighth AAAI International Conference on Weblogs and Social Media (ICWSM), 2014.

[21] D. Nguyen, R. Gravel, D. Trieschnigg, T. Meder, "how old do you think i am?" a study of language and age in twitter., in: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM), 2013, pp. 439–448.

[22] W. Liu, D. Ruths, What's in a name? using first names as features for gender inference in twitter., in: Proceedings of the AAAI Spring Symposium: Analyzing Microtext, 2013, pp. 10–16.

[23] A. Culotta, N. R. Kumar, J. Cutler, Predicting the demographics of twitter users from website traffic data, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 72–78.

[24] P. Velardi, G. Stilo, A. E. Tozzi, F. Gesualdo, Twitter mining for fine-grained syndromic surveillance, Artificial Intelligence in Medicine 61 (3) (2014) 153–163.

[25] M. J. Paul, M. Dredze, You are what you tweet: Analyzing twitter for public health., in: Proceedings of the Fifth AAAI International Conference on Weblogs and Social Media (ICWSM), 2011, pp. 265–272.

[26] M. Dredze, How social media will change public health, Intelligent Systems, IEEE 27 (4) (2012) 81–84.

[27] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: Detecting influenza epidemics using twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, Association for Computational Linguistics, 2011, pp. 1568–1576.

[28] H. Park, S. Rodgers, J. Stemmle, Analyzing health organizations' use of twitter for promoting health literacy, Journal of health communication 18 (4) (2013) 410–425.

[29] R. Teodoro, M. Naaman, Fitter with twitter: Understanding personal health and fitness activity in social media, in: Proceedings of the Seventh AAAI International Conference on Weblogs and Social Media (ICWSM), 2013, pp. 611–620.

[30] C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, N. Dasgupta, Digital drug safety surveillance: Monitoring pharmaceutical products in twitter, Drug Safety 37 (5) (2014) 343–350.

[31] M. Hefler, B. Freeman, S. Chapman, Tobacco control advocacy in the age of social media: using facebook, twitter and change, Tobacco control 22 (3) (2013) 210–214.

[32] M. Myslín, S.-H. Zhu, W. Chapman, M. Conway, Using twitter to examine smoking behavior and perceptions of emerging tobacco products, Journal of medical Internet research 15 (8).

[33] J. Huang, R. Kornfield, G. Szczypka, S. L. Emery, A cross-sectional examination of marketing of electronic cigarettes on twitter, Tobacco control 23 (suppl 3) (2014) iii26–iii30.

[34] J. Landis, G. Koch, The measurement of observer agreement for categorical data., Biometrics 33 (1) (1977) 159–174.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[36] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining., in: LREC, Vol. 10, 2010, pp. 2200–2204.

[37] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.

[38] A. K. McCallum, MALLET: A machine learning for language toolkit, http://mallet.cs.umass.edu (2002).

[39] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[40] I. M. Kloumann, C. M. Danforth, K. D. Harris, C. A. Bliss, P. S. Dodds, Positivity of the english language, PloS one 7 (1) (2012) e29484.