Exploratory Analysis of Marketing and Non-Marketing E-Cigarette Themes on Twitter

Sifei Han[2] and Ramakanth Kavuluru[1,2⋆]

[1] Division of Biomedical Informatics, Department of Internal Medicine
[2] Department of Computer Science
University of Kentucky, Lexington, KY
{sehan2,ramakanth.kavuluru}@uky.edu

Abstract. Electronic cigarettes (e-cigs) have been gaining popularity and have emerged as a controversial tobacco product since their introduction in 2007 in the U.S. The smoke-free aspect of e-cigs renders them less harmful than conventional cigarettes and is one of the main reasons for their use by people who plan to quit smoking. The US food and drug administration (FDA) has introduced new regulations early May 2016 that went into effect on August 8, 2016. Given this important context, in this paper, we report results of a project to identify current *themes* in e-cig tweets in terms of semantic interpretations of topics generated with topic modeling. Given marketing/advertising tweets constitute almost half of all e-cig tweets, we first build a classifier that identifies marketing and non-marketing tweets based on a hand-built dataset of 1000 tweets. After applying the classifier to a dataset of over a million tweets (collected during 4/2015 – 6/2016), we conduct a preliminary content analysis and run topic models on the two sets of tweets separately after identifying the appropriate numbers of topics using *topic coherence*. We interpret the results of the topic modeling process by relating topics generated to specific e-cig themes. We also report on themes identified from e-cig tweets generated at particular places (such as schools and churches) for geo-tagged tweets found in our dataset using the GeoNames API. To our knowledge, this is the first effort that employs topic modeling to identify e-cig themes in general and in the context of geo-tagged tweets tied to specific places of interest.

1 Introduction

Electronic cigarettes (e-cigs) are an emerging smoke-free tobacco product introduced in the US in 2007. 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 [12], an activity often referred to as *vaping*. The broad topic of e-cig use has become a major fault line among clinical, behavioral, and policy researchers who work on tobacco products. There are arguments on either side given their reduced harm aspect ([28] claims they are 95% less harmful than combustible

⋆ corresponding author

cigarettes) may help addicted smokers quit smoking [24] while the long term effects of e-cigs are not yet thoroughly understood. However, there is recent evidence that vaping is linked to suppression of genes associated with regulating immune responses [27]. Furthermore, based on recent news releases from the Centers for Disease Control (CDC) [37], there is an alarming 900% increase in e-cig use from 2011 to 2015 by middle and high school students who might be acquiring nicotine dependence albeit through the new e-cig product. There is also recent evidence that never smoking high school students are at increased risk of moving from vaping to smoking [2]. In light of these findings, the FDA has recently introduced a final deeming rule [13] that went into effect on 8/8/2016 when regulations were extended to many electronic nicotine delivery systems including e-cigs. In this context, surveillance of online messages on e-cigs is important both to monitor the spread of false/incomplete information [22] about them and to gauge prevalence of any adverse events related to their use [6, 36] as disclosed online.

For an emerging product like e-cigs, the follower-friend connections and "hashtag" functionality offer a convenient way for Twitter users (or "tweeters") to propagate information and facilitate discussion. An official quote we obtained earlier this year from Twitter Inc. indicates there are over 30 million public tweets on e-cigs since 2010. In our prior effort [19], we found that there is a 25 fold increase in e-cig tweets from 2011 to 2015 indicating the popularity of e-cig messages on Twitter. A major amount of chatter on e-cigs on Twitter surrounds their marketing by vendors making it generally difficult to analyze regular e-cig tweets that are not dominated by marketing noise. As such, building and using a classifier that separates marketing tweets is an important pre-processing step in several efforts. We are aware of at least four such efforts [10, 14, 18, 20] on building automatic classifiers for e-cig marketing tweets for various end-goals. Other researchers who studied e-cig tweets focused on sentiment analysis [14, 30] and diffusion of messages from e-cig brands on Twitter [8]. In our current effort

1. We manually estimated the proportion of marketing and non-marketing tweets to be 48.6% (45.5–51.7%) : 51.4% (48.3–54.5%) from a sample of 1000 randomly selected tweets selected from over one million e-cig tweets collected through Twitter streaming API between 4/2015 and 6/2016 (Section 2). The ranges in parentheses show 95% confidence intervals of the proportions calculated using Wilson score [38].
2. We built a classifier that achieves an accuracy of 88% (Section 3) in identifying marketing and non-marketing tweets using a variety of approaches ranging from traditional linear text classifiers to recent advances in classification with convolutional neural networks based on word embeddings [21]. Prior efforts [18, 20] that seem to report similar or slightly superior ($< 2.5\%$) results estimate the proportion of marketing tweets in the dataset to be 80%–90%[1], which we find unrealistic in the current situation (based on our own

[1] Although achieving high F-scores for the minority class is generally difficult in heavily skewed datasets, they typically lend themselves to building classifiers with high overall accuracy across all classes or high F-score for the majority class.

assessment mentioned earlier) as public awareness and their participation in the conversation have increased.

3. After applying the binary classifier to over a million e-cig tweets, we conducted a rudimentary analysis of differences in content and user traits in both subsets (Section 4). We then ran topic modeling algorithms tailored for short texts [7] on the two separate subsets by determining the ideal numbers of topics using average topic coherence scores [32]. We manually examined the topics generated to identify themes in general and also based on subsets of geotagged tweets at popular places of interest as identified through the GeoNames geographical database (`http://www.geonames.org`). Although prior efforts identified broad themes through manual analyses [9], we believe our current effort is the first to employ topic modeling to discover more specific e-cig themes (Section 5). Thus, rather than having investigators predetermine which themes to look for in the dataset, our approach lets the dataset determine the prominent themes.

2 Dataset and Annotation

We used the Twitter streaming API to collect e-cig related tweets based on following key terms: `electronic-cigarette, e-cig, e-cigarette, e-juice, e-liquid, vape` and `vaping`. Variants of these terms with spaces instead of hyphens or those without the hyphens (for matching hashtags) were also used. A total of 1,166,494 tweets were obtained through the API calls from 4/2015 to 6/2016. From this dataset we randomly chose 1000 tweets to manually annotate them as marketing or non-marketing. For our purposes, marketing tweets are those that

- promote e-cig sales (coupons, free trials, offers),
- advertise new e-cig products (liquid nicotine or vaping devices), or
- review different flavors or vaping devices aiming to sell.

We (both authors) independently annotated the 1000 tweets. The labels matched for 87.3% of the tweets with an inter-annotator agreement of $\kappa = 0.726$, indicating substantial agreement [23]. Conflicts for the 127 tweets where we chose different labels were resolved based on a subsequent face to face discussion resulting in a consolidated labeled dataset of 1000 tweets. Disagreements occurred when the marketing/advertising intent is not explicit or clear. For example, a simple message that encourages the followers to also follow the tweeter's Instagram account is not explicitly promoting e-cigs in and of itself but is nevertheless aimed toward marketing. Conflicts also occurred with reviews/recommendations when it was not clear whether a user is genuinely recommending a particular flavor that he/she has tried or whether it is the message from a manufacturer simply drawing followers' attention to their product line. While the former is not a marketing tweet, the latter would definitely fit our notion of such a message. Our final consolidated dataset has 486 marketing and 514 non-marketing tweets.

3 Marketing Tweet Classifier

The measure of performance used in this effort is accuracy, which is essentially the proportion of correctly classified tweets. We did not use the popular F-measure given we wanted to give equal importance to both classes given our aim is to study themes in both subsets of tweets. We first used linear classifiers such as support vector machines (SVM) and logistic regression (LR) classifiers as made available in the scikit-learn [33] machine learning framework. Tweet text was first preprocessed to replace all hyperlinks with the token URL and user mentions with the token TARGET. This is to minimize sparsity of very specific tokens having to do with links and user mentions and is in line with other efforts [1]. Besides uni/bi-grams we also used as features, counts of emoticons, hashtags, URLs, user mentions, sentiment words (positive/negative), and different parts of speech in the tweet. These additional features were useful in our prior efforts in tweet sentiment analysis [15] and spotting e-cig proponents [19] on Twitter. However, in this effort, considering average accuracy over hundred distinct 80%-20% train-test splits of the dataset, we did not observe any improvements with these additional features. So our final mean and 95% confidence intervals for accuracies are 88.10 ± 0.40 with LR and 87.14 ± 0.45 with SVM.

Recent advances in deep learning approaches specifically convolutional neural networks (CNNs) have shown promise for text classification [21]. Given our own positive experiences in replicating those approaches for biomedical text classification [35], we also applied CNNs with word embeddings to generate feature maps for marketing tweet classification. The main notion in CNNs is of so called *convolution filters* (CFs) that are traditionally used in signal processing. The general idea is to learn several CFs which are able to extract useful features from a document for the specific classification task based on the training dataset. In the training phase, the inputs to the CNN are projections of constituent word vectors (which are typically randomly initialized) from a fixed size sliding window over the document. Model parameters to learn include the word vectors, the convolution filters (which are typically modeled as matrices), and the connection weights from the convolved intermediate output to the two nodes (for binary classification) in the output layer. Due to the nature of this particular paper, we refer the readers to our recent paper [35, Section 3] for a detailed description of CNN models including specifics of parameter initialization and drop-out regularization (to prevent overfitting). Averaging the $[0, 1]$ probability estimates of the corresponding classes from several (typically ten) CNNs seems to help in getting a more robust model. We ran ten such models (each with ten CNNs, so a total of 100 CNNs) on ten different 80%-20% train-test splits of the dataset. The corresponding accuracies were: 89, 88.5, 85.5, 86, 87, 90.5, 87.2, 88.5, 90.5, and 89 with an average of 88.17%, which is only slightly better than the mean accuracy obtained using logistic regression.

4 Characteristics of Marketing/Non-Marketing Tweets

As discussed earlier, although the ability to separate marketing tweets from those that do not have that agenda is of interest in and of itself, in this effort, we wanted to study themes evolving from both subsets of the dataset. We applied all three classifiers (SVM, LR, and CNN) built in Section 3 using all hand-labeled tweets to all 1,166,494 tweets in our full dataset. We considered those tweets for which all three classifiers predicted the same label, which turned out to be for 1,021,561 (87.56% of the full dataset) of which 456,290 (44.66%) were predicted to be marketing and 565,271 (55.34%) belonged to the other class. To get a basic idea of the tweet content, we simply counted and sorted the words in each subset in descending count values. The top 20 words in both subsets are

- *Marketing*: win, vaporizer, free, mod, get, enter, giveaway, new, premium, code, shipping, bottles, USA, use, box, promo, kit, available, follow, DNA
- *Non-Marketing*: smoking, new, use, rips, like, cigarettes, via, man, get, tobacco, health, video, study, FDA, ban, one, smoke, people, news, explodes

Even with this simple exercise, we notice that the marketing tweets are dominated by e-cig promotions and sales terms or devices for vaping (mod, vaporizer, kit). On the other hand, terms in the non-marketing tweets are about tobacco smoking, health studies, and FDA regulations.

Table 1: Content and user characteristics of the datasets

	Marketing	Non-marketing
E-cig flavors	25472	4612
Harm reduction	19	2256
Smokefree aspect	553	3201
Smoking cessation	6363	22421
Contain "FDA"	204	18297
Number of unique users	66,957	231,982
User handles containing e-cig terms	4777 (7.1%)	3859 (1.7%)
Avg. # tweets per user	6.81 ($\sigma = 197$)	2.44 ($\sigma = 84$)

Next, we look at specific content and user characteristics of both subsets. In our prior work [19], we analyzed the tweets generated by e-cig proponents tweeters along four well known broad themes. We developed regular expressions (please see [19, Section 5.3]) in consultation with a tobacco researcher to capture tweets belonging to these themes. As part of the preliminary analysis, in this effort, we applied those regexes to the two subsets of tweets and obtained the corresponding numbers of thematic tweets shown in the first four rows of Table 1. Except for e-cig flavors, which are a well known major selling point,

the non-marketing datasets contain more tweets in the three other themes (even after accounting for the slight variation in dataset sizes). It is still disconcerting to see the 6363 (1.4%) marketing tweets discussing smoking cessation when long term consequences of e-cig use are still being investigated. We also looked at how many tweets mention FDA and as expected the majority belong to the non-marketing class.

The last three rows of Table 1 deal with user characteristics of both datasets. We notice that there are 3.5 times as many unique tweeters in the non-marketing set as in the marketing class (row 6). We clarify that some users can belong to both the marketing and non-marketing class if they generate tweets in both datasets. In fact, the top non-marketing tweeter `@ecigitesztek` has 37,949 such tweets but is also ranked 2nd among tweeters in the marketing group with 27,019 tweets. A cursory examination of this public profile indicates that it belongs to a Hungarian vaping aficionado who almost exclusively tweets about e-cigs and at the time of this writing (re)tweeted over 153,000 times. However, with 11,186 tweeters common to both datasets corresponding to counts from row 6, the Jaccard similarity coefficient is only 0.03. Given marketers tend to use appealing user handles that indicate their purpose, we counted the number of user handles that contain e-cig popular terms such as ecig, vapor, vapour, vape, vaping, eliquid, ejuice, and smoke as substrings of the user handle. 15 out of the top 20 tweeters in both datasets contain one of these terms as a substring. From row 7, we see that more than 7% of the marketing profiles satisfy this compared with only 1.7% from the other class.

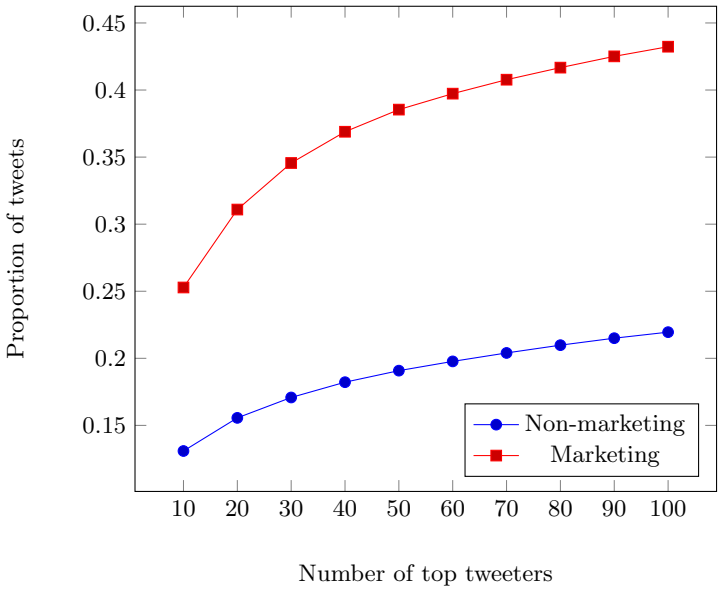

Fig. 1: Proportion of tweets via top $10, \ldots, 100$ tweeters

The final row indicates the average number of tweets per user with standard deviations in parentheses; the difference in the averages is not surprising but the standard deviation magnitude in the marketing set being more than twice that in the other class is revealing in that few users are responsible for many marketing tweets. To further examine this phenomenon, we plotted the cumulative proportion of tweets in the corresponding datasets contributed by the top $10, \ldots, 100$ tweeters in Figure 1. It is straightforward to see that the top tweeters in the marketing dataset generate twice the proportion of tweets as generated by the corresponding top users in the non-marketing dataset. Although the Jaccard coefficient between tweeter sets from both datasets in only 0.03, when considering only top 100 tweeters from both datasets, 84 of the top 100 marketing tweeters have generated non-marketing tweets; 88 of the top 100 non-marketing tweeters also authored marketing tweets.

5 Themes in Marketing/Non-Marketing Tweets

To dig more into these two subsets of tweets, we applied the Biterm Topic Modeling (BTM) [7] approach, which is specifically designed for short text messages like tweets, to these marketing and non-marketing tweets subsets separately. Given recent results that demonstrate that aggregating short text messages such as tweets can lead to better modeling [17], we partitioned the datasets into groups of ten tweets each where each such group is treated as a short document before applying BTM. Besides using the same tweet pre-processing techniques used for classification, we additionally removed commonly occurring terms from the tweets such as stop words and frequent terms such as the key words used to search for e-cig tweets (e.g., e-cig, vape, vaping, vapor, eliquid) given we already know the tweets are on the general topic of e-cigs.

5.1 Topic Modeling Configuration

Most topic modeling approaches have the inherent requirement that the user suggest the number of topics k to fit to the corpus. It is often tricky to pick a specific k, which is generally chosen by trial and error based on human examination of topics generated with different settings of k. We circumvented this potentially tedious and subjective exercise by using a recently introduced measure of topic coherence by O'Callaghan et al. [32] based on neural word embeddings. Topic coherence is a direct measure of intrinsic quality of a topic. For each topic T generated, let $w_1^T, \ldots, w_N^T$ be the set of top N words according to the $P(w|T)$ distribution resulting from the topic modeling process. Then the coherence of T parameterized by N is

$$\mathcal{C}_N^T = \frac{1}{\binom{N}{2}} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \cos(\boldsymbol{w_i^T}, \boldsymbol{w_j^T}),$$

where $\boldsymbol{w_i^T} \in \mathbb{R}^d$ is the dense vectorial representation for the corresponding words learned through the continuous bag-of-words (CBOW) word embedding

approach, which is part of the popular word2vec package [29]. We picked dimensionality $d = 300$ and word window size of five for the CBOW configuration in word2vec and ran it on the full corpus of e-cig tweets. Given this definition of average coherence, the idea is to pick $k \in \{10, 20, 30, \ldots, L\}$ that maximizes the weighted average coherence (WAC) across k topics

$$\sum_{i=1}^{k} P(T_i) \cdot \mathcal{C}_N^{T_i}, \tag{1}$$

where $P(T_i)$ is the probability estimate of the prominence of topic T_i in the corpus (from BTM output), N is the number of top few terms chosen per topic (typically 10 or 20, the latter is used in this paper), and L is chosen to be 50. Note that cosine similarity measure we use here scores term pairs that are semantically similar higher than pairs of words related in a different fashion. This does not, however, affect the validity of our topic coherence approach given topics that contain highly similar words are generally more coherent and simpler to interpret than those that contain words that are related in a more associative manner.

5.2 Prominent E-cig Themes

We recall that topic models output several parameters [3] including a distribution of topics per document (topic proportions: $P(T|d)$) in the corpus and also the distribution of words per topic (per-topic term probabilities: $P(w|T)$), where T is a topic, d is a document, and w is a word. In general, a topic is visualized by displaying the top N (the variable in equation 1) words w according to $P(w|T)$. However, a human agent still needs to look at the top N terms of the topic and identify/interpret a semantic *theme*. This is the distinction we use in this effort too – a topic is a group of N words/terms sorted in descending order according to $P(w|T)$ and a theme is a semantic interpretation of what the topic represents based on our manual review. Even though topic modeling research has come a long way, interpretation of resulting topics for exploratory purposes involves significant manual effort, albeit guided by output distributions mentioned earlier. The rest of this paper involves such exploration to grasp the underlying themes.

Based on our experiments, we found that $k = 10$ maximizes the WAC in equation 1 for the marketing tweets in the corpus. The corresponding value for non-marketing tweets is $k = 50$. This is not surprising given marketing tweets are expected to contain fairly predictable themes that are favorable to e-cigs in general encouraging tweeters to buy/try them or sign-up for more offers. However, the non-marketing subset is more diverse given it is essentially a catch-all for all other topics about e-cigs. Next, we discuss some topics from both subsets.

Marketing Themes: Upon manual examination of the ten topics from the marketing tweets (MT) corpus, we notice a few that are clear and reflect expected themes from this subcorpus. Here we show three of those topics enumerating some of the top 20 words in the topic. The words are rearranged slightly to

better reflect the theme on hand. (However, all words are still from the list of top 20 terms for the topic; otherwise, our analysis would be self-deceiving.)

MT1: free, shipping, code, promo, win, purchase, prizes, enter, giveaway
MT2: vaporizer, pen, mod, kit, battery, portable, starter, electronic, atomizer
MT3: premium, line, lab, certified, AEMSA, cleanliness, consistency, wholesale

The first topic represents the theme of promotional activities involved in marketing e-cigs. The second theme involves vape pens or devices that actually vaporize the liquid nicotine to be inhaled by vapers. The third topic surfaces an unexpected theme of marketing activities that also highlight the quality of the e-liquid products through independent lab certifications offered by the registered nonprofit organization American E-liquid Manufacturing Standards Association (AEMSA), which was established in 2012 for the purpose of promoting safety and standardization in manufacturing liquid nicotine products.

Non-Marketing Themes: The following is the list of major topics in the non-marketing tweets (NT) corpus.

NT1: lungs, cells, flavors, toxic, effects, exposure, study, damage, aerosols
NT2: FDA, poisonings, calls, surge, skyrocket, nicotine, poison, children
NT3: explodes, coma, teen, mouth, burns, injured, suffers, neck, hole, hospital
NT4: FDA, tobacco, industry, market, regulation, product, ban, deeming, rule
NT5: tobacco, laws, CASAA, smoke, healthier, alternative, FDA, grandfather
NT6: quit, smoking, help, current, smokers, cigarette, users, NHS, review
NT7: teen, smoking, CDC, study, middle, school, students, tripled, fell
NT8: ban, Wales, government, public, enclosed, spaces, pushes, ahead
NT9: gateway, drug, doing, cocaine, bathroom, lines, puffin, Wendy, heroin

Note that we only report nine topics here because we found these to be most interesting and also given several others seemed very similar to these nine. There are also a few that do not seem to indicate a specific non-trivial theme and hence were excluded. The first theme NT1 is about toxic effects of e-cigs. An examination of biomedical articles with the search terms `e-cigarettes` AND `toxic` AND `lungs` returned several articles discussing experiments that demonstrated how flavoring agents of e-cigs, and not the liquid nicotine itself, are responsible for toxic effects of inhaling e-cig vapors. NT2's theme relates to a news piece that diffused through Twitter about FDA receiving many calls involving poisoning complaints by e-cig users. NT3 and several related topics (not displayed here) discussed explosions of the vaping devices while in use resulting in burns and hospitalizations [36]. NT4 represents a general theme involving FDA regulatory activities and the new deeming rule [13], which was thought to be impending throughout the past few years.

In NT5, we see a very specific theme that involves the non-profit organization Consumer Advocates for Smokefree Alternatives Association (CASAA) and

the general harm-reduction perspective of e-cigs as an healthier alternative to cigarettes for people who want to quit smoking. The last term 'grandfather' in NT5 refers to new regulations extending to any product introduced/modified on or after the so called grandfather date set to 2/15/2007 by the FDA [13]. This date is critical to many e-cig businesses as all those products (already in market) will now be subject to the new FDA regulations and hence need to be approved by it. NT6 represents the theme of using e-cigs as an aid to smoking cessation. The term NHS refers to UK's National Health Service, which has taken a favorable stance to e-cig use for treating addicted smokers [28]. NT7 is about research reports by the CDC indicating tripling of current e-cig use by middle and high school students from 2013 to 2014 [4]. NT8 highlights another news piece on Wales (of UK) government passing a law to ban vaping in enclosed spaces.

Fig. 2: Tweet leading to topic NT9 on e-cigs as a gateway drug

The final topic NT9 is unusual and seems to indicate e-cigs as a gateway drug to use other more harmful products such as cigarettes, cocaine, and heroin. Although there is some evidence [2] to support this idea, this particular topic appeared atypical with words like bathroom, lines, and Wendy. A deeper examination revealed that most of the words in this topic are mostly coming from one tweet shown in Figure 2. As can be seen, this tweet was retweeted more than 1000 times. Given retweets are essentially a reasonable and natural mechanism to add more weight to a particular topic, we decided to not to delete them in our analysis. However, this particular topic led us to dig deeper into manifestations of topics of this nature. There were two other non-marketing topics like this based on frequent retweets or many tweets involving some minor modifications of a very specific tweet: one involved a picture of film actor Ben Affleck vaping after getting a traffic rule violation ticket (the topic had words Ben, Affleck, and ticket) and another involved the URL of an online petition offering support to the then UK prime minister David Cameron and other politicians trying to block certain e-cig regulations in the UK.

Effect of Excluding Retweets: Given this observation involving NT9, we wanted to study the effect of retweets on topic modeling. We found that 36% of marketing tweets and 43% of non-marketing tweets were due to retweets.

Thus we see that retweets constitute a significant proportion of the full datasets. We generated new topic models with these subsets excluding all retweets to see if there is a noticeable difference in the themes. Although the themes did not change significantly, the words used to represent the topics have changed slightly in most cases. For example, the theme in this new set of topics corresponding to NT9 had the following top words: gateway, drug, smoking, heroin, cocaine. None of the specific words (bathroom, doing, puffin, lines, Wendy) from the highly retweeted message in Figure 2 showed up in the new topic. There were no other topics indicating a gateway theme. There was no topic involving Ben Affleck's traffic ticket but the petition related topic involving former UK prime minister David Cameron was apparent with slightly different words. All other themes NT1–NT8 were evident in the new set of topics. There was only one new theme that wasn't already in the topic set from the full dataset. This was mostly about vaporizer/e-liquid brand names with top terms including: sigelei, hexohm, flawless, ipvmini, districtf, tugboatrda, appletop, longislandbrewed. There was no major change in the themes for the marketing tweet subset.

Finally, we wanted to see who is tweeting on various themes identified through our approach. To this end, we picked two different non-marketing themes, NT6 (e-cigs for smoking cessation) and NT7 (CDC reporting on increasing teen vaping). For each of the corresponding topics T, we ranked all tweets s according to $P(T|s)$. Based on the authorship of the top 10,000 tweets according this ranking, we sorted tweeters in descending order based on the counts of top 10,000 tweets they authored. We manually examined the top few ranked tweeters in this list. For theme NT7, 11 out of 20 top tweeters are regular people tweeting about e-cigs but only 2 out of 20 top tweeters for NT6 are regular tweeters; the other tweeters being institutions or companies that have a clear positive stance for and commercial interest in e-cigs. This indicates that regular tweeters (even if they are in favor of e-cigs) are more inclined to tweet about news involving e-cigs, even when it is not favorable. Commercial tweeters tend to exclusively focus on propagating favorable news pieces besides promoting their products.

Overall, our effort offers a complementary approach by surfacing specific themes in comparison to manual coding [9, 19] where only broad topics such as smoking cessation, flavors, and safety are typically used. This is our main contribution – demonstrating the feasibility of topic modeling based thematic analysis of e-cig chatter on Twitter. Some of our extracted themes may already be common knowledge for tobacco researchers who regularly follow e-cig related news. But we believe the topic modeling approach can help surface a more comprehensive set of themes with less manual exploration burden. It also gives a better sense of the strength of a theme (as observed by the the corresponding topic's ranking) and main tweeters authoring the corresponding thematic tweets.

5.3 Themes in Geotagged Tweets

Geotagged tweets with the associated latitude and longitude information offer a different lens to understand e-cig messages. There have been very few

studies examining the locations where e-cigs are used. There is only particular study [20] that we are aware of where prepositional phrase patterns were used over tweet text to identify e-cig use in a class, school, room/bed/house, or bathroom. In our effort, we are not necessarily concerned about e-cig use, but are generally interested in knowing themes from tweets generated near different types of places of interest. Our dataset has a total of 3208 geotagged tweets which is less than 1% of the full dataset. Using the GeoNames API (`http://www.geonames.org/export/web-services.html`), we identified the nearest *toponym* for each of the corresponding geocodes using the `findNearby` method. In our dataset, the average distance between the geo-code and nearest toponym was 300 meters. Toponyms can be names of larger geographical areas such as cities or rivers, but can also refer to small locations such as a school, hospital, or a park. Each toponym (e.g., University of Kentucky) is associated with a corresponding feature code (e.g., UNIV).

We aggregated tweets based on feature codes (`http://www.geonames.org/export/codes.html`) of the toponyms returned and obtained the following distribution (top ten codes) where counts are shown in parentheses:

hotel (596), populated-place (411), church (314), school (311), building (286),

mall (158), park (109), lake (91), library (80), and post office (74).

In addition to these we also considered, travel end-points (81) as a single class (airports, bus stations, and railway stations), restaurants (39), hospitals (45), museums (13), and universities (11). A simple string search revealed that in very few cases the geotagged tweet content actually made explicit connection to the corresponding feature code. We were able to find 2–3 tweets at hotels, schools, and airports indicating the location type as part of the tweet (e.g., "vaping in class" and "flight is full"). Except for schools, parks, restaurants, hospitals, and airports, all locations had more marketing tweets than regular tweets. Overall, 52% of geotagged tweets belonged to the marketing class, a 7.5% increase compared with the corresponding proportion in the full dataset as discussed in the beginning of Section 4.

For each of these different location types, we identified top topics by fitting topic models to the corresponding sets of tweets. Given marketing tweets have a clear agenda, we only look at non-marketing top topics. For clarity, we simply outline the theme without listing all the keywords

- Church: Ban on e-cigs for minors in Texas
- Hotel: E-cig use rising among young people
- Park: Pros and cons of E-cig regulations
- School: Smoking rates fall as e-cig use increases among teens

Other locations either did not have a significant number of tweets or had tweets without any dominant theme. We realize that our analysis in this section may not be precise in the sense that tweets originating from different types of places may not be from people who are visiting those places for relevant purposes;

tweeters might simply be around those places when they tweet. However, we believe with a large exhaustive dataset spanning multiple years, given we only look at top themes, we can arrive at themes that are representative of people visiting those places.

6 Conclusion

E-cigs continue to survive as a controversial tobacco product and are currently subject to new FDA regulations since 8/8/2016 with a grandfathering date set to 2/15/2007. The FDA, biomedical researchers, physicians, tobacco industry, and most important the nation's public are all key players whose activities will be affected with these products for the foreseeable future. Public health and tobacco researchers are split in their opinions regarding e-cig use by smokers who would otherwise continue with regular cigarettes. Computational social science and informatics approaches can offer a more objective lens through which the social media landscape of e-cigs can be gleaned for online surveillance of both product marketing practices and adverse events.

Although prior efforts exist in content analysis based on pre-determined broad themes, we do not see results on automatic extraction of themes from social media posts on e-cigs. We believe computational approaches provide an important avenue that can complement traditional survey based research efforts considering the cost and time factors involved in the latter case. Twitter in particular has been well studied in the context of public health informatics efforts and provides a major platform for e-cig chatter on the Web.

In this paper, we conduct thematic analysis experiments involving over a million e-cig tweets collected during a 15 month period (4/2015 – 6/2016). To deal with the major presence of marketing chatter, we first built a classifier that achieved an accuracy of over 88% in identifying marketing and non-marketing tweets based on a manually labeled dataset. We conducted preliminary content and user analysis of marketing and non-marketing tweets as classified by our model. Subsequently, we fit topic models to the two subsets of tweets and interpreted them to identify specific themes that were not apparent in manual efforts. This is not surprising given the fast changing discourse on e-cigs creates a corresponding rapidly evolving social media landscape. This, however, points to an important weakness of our approach – it is not *online*, where new e-cig tweets continuously collected through the Twitter streaming API are used to generate new topics as enough evidence accumulates. As part of future work, we plan to employ online topic models [16] and facilitate their exploration using well known topic browsing approaches [5, 26]. Nevertheless, here we provide what we believe is a first strong proof of concept for employing topic models to comprehend evolving e-cig themes on Twitter. Given gender, age group, race and ethnicity can be predicted with reasonable accuracy [11, 25, 31], an important future research direction is to use these methods to classify e-cig tweeters into these demographic categories and identify e-cig themes in tweets authored by specific subpopulations. For example, given african american teenagers are

an active group on Twitter [34], identifying popular e-cig themes authored by them (including retweets and favorites) may yield insights specific to that demographic segment. Similar analysis can also be conducted with tweets originating from rural areas given the typical firehose is dominated by urban tweeters.

Acknowledgements

We thank anonymous reviewers for constructive criticism that helped improve the presentation of this paper. 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. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011.
2. J. L. Barrington-Trimis, R. Urman, K. Berhane, J. B. Unger, T. B. Cruz, M. A. Pentz, J. M. Samet, A. M. Leventhal, and R. McConnell. E-cigarettes and future cigarette use. *Pediatrics*, page e20160379, 2016.
3. D. M. Blei and J. D. Lafferty. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Classification, Clustering, and Applications*, chapter 4, pages 71–93. Chapman and Hall, CRC Press, 2009.
4. 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`.
5. A. J.-B. Chaney and D. M. Blei. Visualizing topic models. In *International Conference of Weblogs and Social Media*, ICWSM '12, 2012.
6. I.-L. Chen et al. FDA summary of adverse events on electronic cigarettes. *Nicotine & Tobacco Research*, 15(2):615–616, 2013.
7. X. Cheng, X. Yan, Y. Lan, and J. Guo. BTM: Topic modeling over short texts. *Knowledge and Data Engineering, IEEE Transactions on*, 26(12):2928–2941, 2014.
8. K.-H. Chu, J. B. Unger, J.-P. Allem, M. Pattarroyo, D. Soto, T. B. Cruz, H. Yang, L. Jiang, and C. C. Yang. Diffusion of messages from an electronic cigarette brand to potential users through twitter. *PloS one*, 10(12):e0145387, 2015.
9. H. Cole-Lewis, J. Pugatch, A. Sanders, A. Varghese, S. Posada, C. Yun, M. Schwarz, and E. Augustson. Social listening: A content analysis of e-cigarette discussions on twitter. *Journal of medical Internet research*, 17(10), 2015.
10. H. Cole-Lewis, A. Varghese, A. Sanders, M. Schwarz, J. Pugatch, and E. Augustson. Assessing electronic cigarette-related tweets for sentiment and content using supervised machine learning. *J. of medical Internet research*, 17(8):e208, 2015.
11. A. Culotta, N. R. Kumar, and J. Cutler. Predicting the demographics of twitter users from website traffic data. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 72–78, 2015.

12. J.-F. Etter, C. Bullen, A. D. Flouris, M. Laugesen, and T. Eissenberg. Electronic nicotine delivery systems: a research agenda. *Tobacco Control*, 20(3):243–248, 2011.
13. Food and Drug Administration, HHS et al. 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. final rule. *Federal register*, 81(90):28973, 2016.
14. A. K. Godea, C. Caragea, F. A. Bulgarov, and S. Ramisetty-Mikler. An analysis of twitter data on e-cigarette sentiments and promotion. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 205–215. Springer, 2015.
15. S. Han and R. Kavuluru. On assessing the sentiment of general tweets. In *Canadian Conference on Artificial Intelligence*, pages 181–195. Springer, 2015.
16. M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent Dirichlet allocation. In *Advances in neural information proc. systems*, pages 856–864, 2010.
17. L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proc. of the 1st workshop on social media analytics*, pages 80–88. ACM, 2010.
18. J. Huang, R. Kornfield, G. Szczypka, and S. L. Emery. A cross-sectional examination of marketing of electronic cigarettes on twitter. *Tobacco control*, 23(suppl 3):iii26–iii30, 2014.
19. R. Kavuluru and A. Sabbir. Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on Twitter. *J. of biomedical informatics*, 61:19–26, 2016.
20. A. E. Kim, T. Hopper, S. Simpson, J. Nonnemaker, A. J. Lieberman, H. Hansen, J. Guillory, and L. Porter. Using twitter data to gain insights into e-cigarette marketing and locations of use: An infoveillance study. *Journal of Medical Internet Research*, 17(11):e251, 2015.
21. Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, October 2014.
22. E. G. Klein, M. Berman, N. Hemmerich, C. Carlson, S. Htut, and M. Slater. Online e-cigarette marketing claims: A systematic content and legal analysis. *Tobacco Regulatory Science*, 2(3):252–262, 2016.
23. J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
24. D. T. Levy, K. M. Cummings, A. C. Villanti, R. Niaura, D. B. Abrams, G. T. Fong, and R. Borland. A framework for evaluating the public health impact of e-cigarettes and other vaporized nicotine products. *Addiction*, 2016.
25. W. Liu and 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*, pages 10–16, 2013.
26. S. Malik, A. Smith, T. Hawes, P. Papadatos, J. Li, C. Dunne, and B. Shneiderman. Topicflow: visualizing topic alignment of twitter data over time. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 720–726. ACM, 2013.
27. E. Martin, P. W. Clapp, M. E. Rebuli, E. A. Pawlak, E. E. Glista-Baker, N. L. Benowitz, R. C. Fry, and I. Jaspers. E-cigarette use results in suppression of immune and inflammatory-response genes in nasal epithelial cells similar to cigarette smoke. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, pages ajplung–00170, 2016.
28. A. McNeill, L. Brose, R. Calder, S. Hitchman, P. Hajek, and H. McRobbie. E-cigarettes: an evidence update. *Report from Public Health England*, 2015.

29. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.
30. M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8), 2013.
31. D. Nguyen, R. Gravel, D. Trieschnigg, and 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)*, pages 439–448, 2013.
32. D. OCallaghan, D. Greene, J. Carthy, and P. Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, 2015.
33. 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, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
34. 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/`.
35. A. Rios and R. Kavuluru. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 258–267. ACM, 2015.
36. S. Rudy and E. Durmowicz. Electronic nicotine delivery systems: overheating, fires and explosions. *Tobacco control*, 2016.
37. T. Singh, R. Arrazola, C. Corey, C. Husten, L. Neff, D. Homa, and B. King. Tobacco use among middle and high school students – United States, 2011 – 2015. *MMWR Morbidity and mortality weekly report*, 65(14):361–367, 2016.
38. E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.