An Analysis of Twitter Data on E-cigarette Sentiments and Promotion

Andreea Kamiana Godea[1], Cornelia Caragea[1], Florin Adrian Bulgarov[1], and Suhasini Ramisetty-Mikler[2]

[1] Computer Science and Engineering, University of North Texas Denton, TX, US
[2] Biostatistics and Epidemiology, University of North Texas Health Science Center, Fort Worth, TX, US
andreeagodea@my.unt.edu, ccaragea@unt.edu, florinbulgarov@my.unt.edu, Suhasini.Ramisetty-Mikler@Unthsc.edu

Abstract. We investigate general sentiments and information dissemination concerning *electronic cigarettes* or *e-cigs* using Twitter. E-cigs are relatively new products, and hence, not much research has been conducted in this area using large-scale social media data. However, the fact that e-cigs contain potentially dangerous substances makes them an interesting subject to study. In this paper, we propose novel features for e-cigs sentiment classification and create sentiment dictionaries relevant to e-cigs. We combine the proposed features with traditional features (i.e., bag-of-words and SentiStrength features) and use them in conjunction with supervised machine learning classifiers. The feature combination proves to be more effective than the traditional features for e-cigs sentiment classification. We also found that Twitter users are mainly concerned with sharing information (33%) and promoting e-cigs (22%). Although a low percentage of users share opinions, the majority of these users have positive opinions about e-cigs (11% positive, 3% negative).

Keywords: E-cigs, Sentiment analysis, Social networking sites, Twitter

1 Introduction

Much has been written concerning the effects of tobacco smoking on people's health. Research has shown that smoking is harmful to almost every organ in the human body and can cause people's deaths [1, 2]. In particular, tobacco smoking results in more than 480,000 deaths every year in the United States [1–3]. Smoking is associated with many diseases such as cardiovascular and respiratory diseases, and cancer. Quitting smoking can help reduce the risk of such diseases and could boost people's lifespan. Electronic nicotine delivery systems (ENDS) such as *electronic cigarettes* or *e-cigs* have been recently introduced as an alternative way to using tobacco products. E-cigs provide a nicotine-containing aerosol to users by heating glycerol, nicotine, and flavoring agents [4].

It is not fully known yet if e-cigs are safer than tobacco cigarettes or if they are simply a way to develop addiction to nicotine, and hence, a gateway leading non-smokers into smoking tobacco habits. However, recent progress has been

made to reveal the negative effects of e-cigs on human health. For example, it has been shown that glycerol can produce mouth and throat irritation and dry cough [5]. Furthermore, despite that nicotine can help people feel calmer and more relaxed by reducing stress and anxiety, it has paradoxical effects, acting as a depressant [6]. Nicotine also has a negative impact on insulin resistance, which raises the risk of developing diabetes and heart diseases [7]. Despite these potentially negative effects of *e-cigs* on health, e-cigs have become a popular product in recent years. Their increasing popularity is in part due to the social context in which they occur (e.g., among friends), their availability in attractive flavors, and the perception of youth that e-cigs are safer than other nicotine products. The Centers for Disease Control and Prevention perform surveillance to monitor trends in the health of populations. National surveys such as National Youth Tobacco Survey and Youth Risk Behavior Surveys have recently begun to assess new and evolving health risk behaviors. However, trending behaviors are not comprised in such surveys, consequently resulting in the delay in recognizing the problem and its impact on the population health. Social networking sites appear to have embraced the attention in a unique way, being used as a tool for personal expression and freedom. Technological advancements allow online users to instantaneously share their experiences, sentiments and beliefs via blogs, micro-blogs (e.g., Twitter), discussion boards, etc., with peers and the public. Such social networking sites provide researchers a great opportunity to utilize alternate ways to analyze trending behaviors and sentiments shared by users.

In this study, we employ natural language processing and text-mining techniques to analyze the sentiment polarities expressed by Twitter users towards e-cigs and investigate how information is disseminated about these relatively new products. Twitter is now in the top 10 most visited Internet sites, which makes it an attractive platform to analyze sentiments and information spread.

2 Related Work

Sentiment analysis has been an actively researched area due to its importance in mining, analyzing and summarizing user opinions from online sites such as product review sites, forums, Facebook, and Twitter [8, 9]. Sentiment analysis focuses on identifying the polarity (positive/negative) of a piece of text (often tied to a particular target). Here, we survey several sentiment analysis works.

Pang et al. [10] used supervised learning techniques on lexical features (e.g., unigrams, bigrams, part-of-speech tags) for sentiment analysis of movie reviews. Previous approaches based on lexicon rules exist that aim at aggregating sentiments for an entity [9, 11]. Sentiment analysis has also been recently used in online health communities (OHCs). For example, Biyani et al. [12] performed sentiment classification of posts in a Cancer Survivors' Network to discover sentiment change patterns in its members and to detect factors affecting the change.

E-cigs have recently become the subject of several studies. For example, Bullen et al. [13] explored the effectiveness of e-cigs compared to nicotine patches in helping smokers to quit. The authors found no significant difference between e-cigs over nicotine patches. Popova and Ling [14] analyzed the correlation between e-cig usages among tobacco smokers and quit attempts. They found no

substantiated evidence that e-cigs lead to smoking cessation. Myslín et al. [15] analyzed sentiments towards tobacco products from Twitter. They found that they are generally positive and are correlated with social image, personal experience, and recently popular products like hookah and e-cigs.

In contrast, our work is significantly different from the previous works. In this paper, we aim to study "*the voice of population*" concerning e-cigs by using Twitter data. Specifically, we seek to identify the general sentiments and information dissemination related to e-cigs, by implementing a supervised approach. For our task, we created domain-dependent sentiment dictionaries and designed new features based on polarity measures, user information and tweet structure.

3 Materials and Methods

Data Source and Analysis. For our study, we used Twitter data. We collected 105,605 tweets between March and April 2014 (using Twitter API), based on the following words: *e-cigs, electronic cigarette, vapor, vaping, e-juice, e-liquid and personal vaporizer*. We refer to these words as keywords throughout the paper. From the dataset, we manually annotated 1200 random tweets with the categories: *advertising, informational, opinion (positive and negative)* and *other*. The *advertising* category contains tweets shared with a commercial purpose, whereas the *informational* type refers to the ones providing general information. The *opinion* class was subdivided into *positive* and *negative* based on the overall sentiment about e-cigs. The *other* category consists of neutral or irrelevant tweets that could not be associated with any category above. Table 1 shows categories' names, the number of tweets in each category, the username and examples of tweets from each class. The keywords are encoded with bold.

Table 1. Examples of tweets from each category and categories distribution

Category	#Tweets	Username	Tweet content
Advertising	356	*TheVapeBook*	Great deals on Starter Kits. Save 10% - **Clearomizers** on Sale too! http://t.co/AMANIrXzte #ecig
Informational	339	*Forest_Smoking*	**E-cigs** now banned on all Dublin buses: "The news will come as a breath of fresh air for commuters"
Positive opinion	81	*SonnHardesty*	Oh how I enjoy the pleasure of **vaping**.
Negative opinion	54	*ksquarl*	**E-CIGS** ARE THE STUPIDEST THING I HAVE EXPERIENCED IN MY 15 YEARS OF EXISTENCE
Other	370	*Ernesto_Calva*	**Vaping** at the movies with the crew #BTC #betyoucandoitlikeme

Domain-dependent sentiment dictionaries. Many previously proposed sentiment analysis approaches made use of dictionaries that contain words considered positive or negative in a general context (i.e. domain-independent dictionaries). However, our analysis indicated that many words that express a sentiment in a general context become neutral when used in the e-cigs context. Hence, using domain-independent dictionaries in our experiments may introduce noise. For example, the word *victory* is generally considered positive, but used in the e-cigs context, it does not express any sentiment. Considering this word to be positive will bring the tweet "Victory has the best flavors." closer to the positive class, although *victory* has no sentiment attached to it.

We built e-cigs dictionaries based on the tweets' contexts and the syntactical relationships obtained using Stanford Lexicalized Parser (SLP) [16]. We considered two steps: 1) extract opinion words associated with keywords by leveraging direct and indirect dependencies; 2) extract opinion words from hashtags (i.e., words/phrases prefixed with the character "#").

For the first step, we started with an empty set D. Words that are linked to keywords in a tweet via direct dependencies are added to D. Each tweet is represented as a dependency tree using SLP, based on the grammatical relationships between the words. From the resulting tree, we used the following direct relations: *nsubj (nominal subject), acomp (adjectival complement), dobj (direct object), xcomp (open causal complement), ccomp (clausal complement), prep_about (words linked through the preposition about), prep_in (preposition in), prep_of (preposition of)*. For example, the tweet "I enjoy the pleasure of vaping." contains two direct dependencies: [enjoy - pleasure]^{dobj} and [pleasure - vaping]$^{prep\_of}$. We add "pleasure" to D due to its linkage with the keyword "vaping." Next, we used direct dependencies to find indirect dependencies between keywords and other words. Specifically, for each tweet, we identified two or (possibly) more direct dependencies that are linked by a word, such that at least one dependency contains a keyword. From these direct dependencies, we extracted all words that were not already in D. For the above example, from the direct dependencies [enjoy - pleasure]^{dobj} and [pleasure - vaping]$^{prep\_of}$, linked by "pleasure," we inferred the indirect dependency [enjoy - vaping] and added "enjoy" to D. After extracting all the words using the above procedure (i.e., the compilation of D), we determined the polarity of each word in D using SentiStrength [17]. For each word, SentiStrength returns two scores: positive and negative. A word was added to a dictionary (positive or negative) based on the maximum absolute value between the scores; in case of equality, the word was considered neutral and was not added to the dictionaries. For example, "victory" has the SentiStrength scores 1 and -1 and is considered neutral, although it appears in domain-independent dictionaries. We ended this first step by using a domain-independent dictionary, developed by Hu and Liu [8]. Specifically, we added all word forms. For example, for the word *love*, different forms were added from the domain-independent dictionaries: *loved, loveliness, lovely, lover, loves.*

For the second step, we collected all the hashtags from the dataset and employed SentiStrength and the domain-independent dictionary to create positive and negative dictionaries. Finally, we concatenated the sentiment dictionaries obtained from both steps and acquired a positive dictionary of 260 words and a negative dictionary with 353 words. The dictionaries are available upon request.

Feature Engineering. We designed novel features and used them in conjunction with traditional sentiment features to improve the performance of sentiment classifiers targeted to e-cigarettes. These features, described in Table 2, are compiled based on polarity measures, user information and tweet structure.

Next, we provide some intuition that led to the design of several features as well as some implementation details. The features *noOfPositiveWords* and *noOfNegativeWords* were designed considering the negations, i.e., negations pre-

ceding the sentiment in a window of 2 words reverse the sentiment's polarity. For the feature *checkIfHasECigSenti*, we checked if there was a grammatical dependency between a keyword and sentiment words in a tweet. The intuition behind this feature is to identify if the sentiment is related to e-cigs. Further, to identify user's opinions towards e-cigarettes, we created *personalECigSenti* feature. Specifically, we used dependency trees and checked if the subject of a sentence with a sentiment inside is a keyword or a first-person singular/plural pronoun; e.g., "I love vaping" and "E-cigs smell good." The sentiments shared in these tweets are those of the writer, and can express a personal opinion, based on his/her experience. On the contrary, a tweet such as "You love vaping" does not show the writer's opinion i.e., no personal sentiment is attached, but it reflects his/her opinion that someone else could possibly love e-cigs.

Table 2. Features' description.

FEATURES	DESCRIPTION
SENTIMENT FEATURES	
TRADITIONAL FEATURES	
noOfPositiveWords	Number of positive words.
noOfNegativeWords	Number of negative words.
noOfPositiveEmoticons	Number of positive emoticons.
noOfNegativeEmoticons	Number of negative emoticons.
SentiStrength – positive	Positive SentiStrength score.
SentiStrength – negative	Negative SentiStrength score.
+NEWLY DESIGNED	
checkIfHasECigSenti	Checks for opinions about e-cigs.
personalECigSenti	Checks if the opinion is personal.
+USER INFORMATION	
userHasKeyword	Checks if the username contains e-cigs' specific words.
noOfRetweetsOverAVG	Checks if the user has more retweets than the average.
+TWEET STRUCTURE	
hasQuestion	Checks the presence of questions.
noOfWords	Number of words in a tweet.
noOfKeywords	Number of e-cigs' related words.
hasLink	Checks if a tweet has a link.
hasHashtag	Checks if a tweet has hashtags.
hasNumbersAndQuantities	Checks the presence of product details (e.g., price, discounts).
hasRepeatingCharacters	Checks for repeating characters.
hasSlangWords	Checks slang words' presence.
oneSentenceAndLink	Checks if a tweet contains a sentence and link.

Features based on user information proved to be effective for our task. According to statistics from the labeled data, 85 out of 356 advertising tweets were posted by a user whose username contains a keyword. Therefore, the username can be a good indicator of advertising tweets and we seize this aspect through the *userHasKeyword* feature. Also, the statistics based on all 105,605 tweets show that many tweets represent retweets. Hence, we extracted information about highly re-tweeted users, encoded in the feature *noOfRetweetsOverAVG*.

Intuitively, the features based on the tweet structure can also be effective, because each category of tweets can have a specific structure. The *informational* tweets usually contain general-interest information and hyperlinks to further information; *advertising* tweets usually contain products' details and, often, external links. The *opinion* tweets are generally more informal than informational or advertising tweets, e.g., they contain slang words or repeating characters.

4 Experimental Setting

Our experiments are designed around the following questions:

1. *How do models trained only on the proposed features perform compared with other models for sentiment analysis such as those trained on bag-of-words and SentiStrength? Does the combination of bag-of-words, SentiStrength features*

and our features result in better performance compared with that obtained using each feature type individually?
2. *What are the most informative categories of features for our task?*
3. *How can we characterize the entire dataset of 105,605 tweets in terms of informational, advertising, and opinion tweets, when automatically labeling them using our best classifiers?*

To answer the first question, we compared the performance of classifiers trained using our features with that of classifiers trained using bag-of-words (in a 10-fold cross-validation setting) and with SentiStrength [17] and report Precision, Recall and F1-score. SentiStrength [17] is an algorithm specifically designed to calculate sentiment strength of short informal texts in online social media. We experimented with four classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Decision Trees (DT). Although we show results only for SVM, NB and RF due to space constraints, DT followed the same trends in performance as the other studied classifiers. For bag-of-words, we removed stopwords, except pronouns, which were useful for our task. We used binary features, i.e., the presence/absence of a word from the vocabulary. Using SentiStrength, we identify tweet's polarity using positive and negative scores. We tested two multiplier values: 1.5 (SentiStrength's default) and 1 (equally weighting positive and negative scores) and obtained higher performance using the latter value. Therefore, a tweet was assigned to the class with the maximum absolute score and in case of equality, the tweet was neutral. Last, we combined all features and compared their performance with that of individual features, specifically, the proposed features, SentiStrength features and bag-of-words features.

To answer the second question, we incrementally added feature categories, starting with the traditional sentiment features. From the spectrum of these experiments, we show the classifiers' performance after the addition of a feature category, that yields an improvement in performance (from the smallest to the largest increase over the preceding setting).

To answer the third question, we employed our best performing classifiers trained on the combination of features to label the entire dataset and computed statistics with respect to each class.

5 Results

Classifiers' performance using various types of features. Table 3 shows the results obtained using our features in comparison with those obtained by bag-of-words and with their combination.

As can be seen from the table, our features outperform bag-of-words for the *positive, negative* and *other* classes, for all studied classifiers. Note that the number of our features is much smaller compared with the bag-of-words size (e.g., 19 vs. 3437, respectively). Although bag-of-words exceeds our approach for the *advertising* and *informational* classes, the combination of our features with bag-of-words performs best for all categories, regardless of the classifier used. For example, the combination of features used as input to the SVM classifier has the highest performance for all classes. Precisely, it improves substantially, reaching

Table 3. Results obtained with our approach, bag-of-words and the combined method.

		SVM			Naïve Bayes			Random Forest		
Categories	**Metrics**	Our approach	Bow	Bow+our approach	Our approach	Bow	Bow+our approach	Our approach	Bow	Bow+our approach
Informational	Precision	0.627	0.671	0.778	0.628	0.644	0.700	0.612	0.654	0.715
	Recall	0.587	0.723	0.809	0.606	0.682	0.761	0.652	0.717	0.817
	F1-score	0.603	0.693	**0.792**	0.614	0.660	**0.728**	0.628	0.680	**0.760**
Advertising	Precision	0.606	0.840	0.845	0.663	0.883	0.898	0.687	0.841	0.833
	Recall	0.691	0.772	0.822	0.659	0.661	0.715	0.703	0.717	0.749
	F1-score	0.642	0.803	**0.832**	0.659	0.753	**0.795**	0.692	0.772	**0.787**
Positive opinion	Precision	0.663	0.371	0.634	0.460	0.340	0.473	0.501	0.383	0.578
	Recall	0.315	0.267	0.545	0.551	0.359	0.622	0.360	0.191	0.419
	F1-score	0.410	0.303	**0.552**	0.493	0.336	**0.528**	0.403	0.250	**0.456**
Negative opinion	Precision	0.740	0.536	0.688	0.441	0.340	0.409	0.439	0.500	0.425
	Recall	0.281	0.333	0.514	0.508	0.500	0.578	0.388	0.173	0.210
	F1-score	0.368	0.366	**0.558**	0.446	0.378	**0.460**	**0.382**	0.242	0.257
Other	Precision	0.671	0.615	0.767	0.676	0.562	0.694	0.682	0.584	0.689
	Recall	0.752	0.707	0.819	0.670	0.640	0.693	0.693	0.740	0.765
	F1-score	0.705	0.652	**0.787**	0.669	0.590	**0.689**	0.681	0.648	**0.718**

almost a doubled performance compared with bag-of-words on the *positive* and *negative* classes. We conclude that each feature type from the combination captures some aspect of a tweet and hence, is important for the overall classification.

Comparison of SentiStrength with our approach. We computed the results obtained by SentiStrength and found that our approach performs better in terms of F1-score (i.e., SentiStrength achieves 0.251 for the *positive* class and 0.232 for the *negative* class). SentiStrength has better recall (i.e., 0.580 for *positive*; 0.666 for *negative*), but it achieves a low precision (lower than 20%, i.e., 0.160 for *positive*; 0.140 for *negative*), while our approach obtains a precision higher than 60% for both *positive* and *negative* classes. These results can be justified by the fact that SentiStrength does not follow sentiments towards specific entities, focusing mainly on the overall sentiment of a tweet. We conclude that our proposed features are fairly good indicators of sentiments toward e-cigs.

Classifiers' performance after adding different categories of features. Table 4 shows the performance achieved after we incrementally add each feature category from our approach, starting with traditional sentiment features. The reported results are computed employing 10 folds cross-validation and SVM.

Table 4. F1-scores obtained by sequentially adding the categories of features.

FEATURES	Informational	Advertising	Positive	Negative	Other
SENTIMENT FEATURES					
TRADITIONAL FEATURES	0.000	0.484	0.208	0.080	0.258
+NEWLY DESIGNED	0.041	0.447	0.367	0.324	0.231
+USER INFORMATION	0.055	0.377	0.376	0.324	0.517
+TWEET STRUCTURE	0.589	0.634	0.380	0.372	0.710

Our features are grouped in the following categories: *sentiment features* comprised of traditional features (*noOfPositiveWords, noOfNegativeWords, noOfPositiveEmoticons, noOfNegativeEmoticons*) and new features tightened to e-cigs (*personalECigSenti, checkIfHasECigSenti*); features extracted from *user information* (*userHasKeyword, noOfRetweetsOverAVG*) and *tweet structure* (*hasQues-*

tion, noOfWords, noOfKeywords, hasRepeatingCharacters, hasNumbersAndQuantities, hasLink, oneSentenceAndLink, hasSlangWords, hasHashtag).

As can be seen in the table, the traditional sentiment features provide relevant knowledge for *advertising, positive* and *other* classes. Combined with traditional features, newly designed features perform very well, significantly raising the opinion classes' performance. Specifically, after adding newly designed features (*checkIfhasECigSenti, personalECigSenti*), the performance increased for the positive class with more than 0.1 and for the negative class with more than 0.3. Adding user information shows better results for almost all classes, over the setting that does not use this information. Features from the tweet structure result in the highest increase in performance over the previous setting, mainly for the *informational, advertising* and *other* classes. More precisely, after combining this feature category with those previously used, the performance increased with almost 0.5 for the *informational* class, 0.3 for *advertising* and 0.2 for *other*. This feature type brings also a slight boost in performance for the *opinion* classes.

Twitter Data Characterization. We automatically labeled the entire set of tweets employing the combination of our features, SentiStrength and bag-of-words and the SVM classifier trained on the labeled dataset. Our dataset contains tweets that are not entirely written in English, e.g., "*vaping* aja atuh beroh..." and "『【最短でお届け!!】*FUMI E-JUICE* ブルをる[天]." A closer analysis of the predicted labels for non-English tweets showed that they were assigned to the *other* category, although they may express sentiments. In future, it would be interesting to process them using machine translation. We computed statistics from the predicted labels and found that the majority of tweets are spread with *informational* purpose (almost 33% of tweets). Since *e-cigs* are relatively new products, online users post many informational tweets or links to pages that contain information about e-cigs, making it a trending subject. The second larger category (almost 28%) comprises the tweets that do not express sentiments, information or advertising (i.e., the *other* category), followed by the *advertising* category that represents almost 26% of the collected data. The tweets that express an opinion represent a small fraction of data. We found that users are more likely to share positive opinions/experiences (11%) than negative (3%).

Further, we detected the influential spreaders related to e-cigs from our data. To this end, we built a network which leverages tweets' relationships. Specifically, a node represents a tweet and an edge links two tweets if one is a retweet for the other. Each node is represented giving its importance: the bigger the node (i.e., its degree), the more important the corresponding tweet is. We show the network in Figure 1.

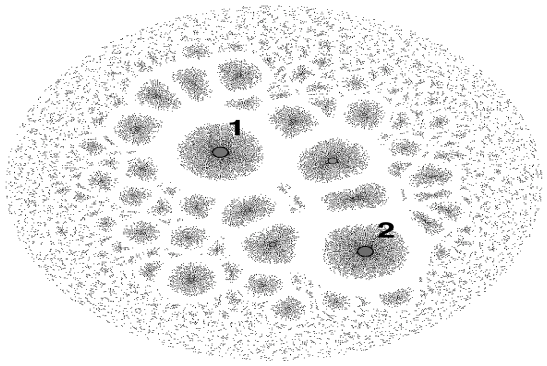

Fig. 1. Information spread network

As can be seen from the figure, there are two important nodes in the network (marked with bigger circles), which implies that there are two highly re-tweeted tweets and many other that are not so important, although they still have a fair amount of re-tweets. We identified the two most important tweets and the users who posted them originally. The first user was a regular user who posted substantial information about e-cigs. Specifically, the highest re-tweeted tweet (marked with "1" in figure) in the network provided a link to reviews. Because e-cigs are relatively new products, people are mainly interested in others' experiences, opinions or concerns. Not only they read the information, but they also share it further. The second highly re-tweeted tweet (marked with "2" in figure) was posted by a company for promotion. The tweet contains a link to information about company's products. An analysis of all tweets that receive a high number of re-tweets in our data show that highly shared tweets inside the network belong to our predominant categories (informational and advertising).

6 Discussion and Conclusion

In this paper, we proposed a supervised approach to identifying sentiments and information dissemination concerning e-cigs from Twitter data. Based on the results obtained using only a small portion of data (1200 labeled tweets), we see that our best setting (the combination of our features with bag-of-words, used as input to an SVM classifier) offers a clearer perspective on the e-cigs domain than other traditional sentiment analysis methods (SentiStrength and bag-of-words). That is, our approach identifies ≈80% of *advertising, informational* and *other* tweets, whereas the *opinion* tweets can be identified in a proportion of 55%.

We expanded our experiments to a more general case (i.e., all the collected dataset) and found that the majority of Twitter users share information concerning e-cigs or spread advertising tweets. Because e-cigarettes are relatively new products, users are mainly interested in sharing/finding meaningful information, while the producer companies are interested in promotion. Although opinions are shared in a small proportion, the general sentiments related to e-cigs tend to be positive. We also found that user information and the way a tweet is structured can be effectively used to automatically discover a tweet's purpose.

This study proposed an effective way of leveraging information posted in online social media to study public opinions and information dissemination related to e-cigarettes. Our method can facilitate the identification of trending behaviors concerning e-cigs, being able to use both current and past information from social networking sites. This approach can be used by various agencies to improve features of e-cigs or marketing strategy, based on public opinion. It is also a general-enough method, which can be easily adapted to study other entities. As future work, it would be interesting to study the type of information posted with respect to *e-cigs*, that is, if the posted information is in support for using e-cigs, or, quite on the contrary, this refers to the information that suggest the negative effects on human health. We further plan to extend our models, to analyze other entities that are legal, but potentially dangerous, such as energy drinks.

Acknowledgments. The authors would like to thank Kishore Neppalli for helping with the collection of the dataset.

References

1. U.S. Department of Health and Human Services. The health consequences of smoking—50 years of progress: A report of the surgeon general. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health 17 (2014).
2. U.S. Department of Health and Human Services. How Tobacco Smoke Causes Disease: What It Means to You. Atlanta: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health (2010).
3. Centers for Disease Control and Prevention. QuickStats: Number of Deaths from 10 Leading Causes—National Vital Statistics System, United States (2010). Morbidity and Mortality Weekly Report 2013:62(08);155.
4. R. Grana, N. Benowitz and S.A. Glantz. E-cigarettes: A scientific review. Circulation. Vol. 129, (2014), p.1972-1986.
5. P. Callahan-Lyon. Electronic cigarettes: human health effects. Tobacco control 23.suppl 2 (2014): ii36-ii40.
6. World Health Organization, `http://www.who.int/tobacco/publications/gender/en_tfi_gender_women_addiction_nicotine.pdf`.
7. D. Kapoor and T.H. Jones. Smoking and hormones in health and endocrine disorders. Eur. J. Endocrinol., 152 (2005), pp.491–499.
8. M. Hu and B. Liu. Mining and summarizing customer reviews. In: The tenth ACM SIGKDD international conference on Knowledge discovery and data mining, (2004).
9. X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In: International Conference on Web Search and Data Mining. ACM, (2008).
10. B. Pang, L. Lee and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. ACL-02 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 79-86, (2002).
11. X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang. Entity-centric topic-oriented opinion summarization in twitter. In: The 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 379-387 (2002).
12. P. Biyani, C. Caragea, P. Mitra, C. Zhou, J. Yen, G. E. Greer, and K. Portier. Co-training over Domain-independent and Domain-dependent Features for Sentiment Analysis of an Online Cancer Support Community. In: ASONAM, 2013.
13. C. Bullen, C. Howe, M. Laugesen, H. McRobbie, V. Parag, J. Williman, and N. Walker. Electronic cigarettes for smoking cessation: a randomised controlled trial. The Lancet, 382.9905: 1629-1637 (2003).
14. L. Popova and P. M. Ling. Alternative Tobacco Product Use and Smoking Cessation: A National Study. Am J Public Health, Vol.103,No.5, pp.923-930 (2013).
15. 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).
16. M.-C. De Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In: LREC, vol 6, pp. 449–454 (2006).
17. M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. J. Am. Soc. Inf. Sci. Technol, 163-173 (2012).