Published in final edited form as:
Pac Symp Biocomput. 2016 ; 21: 480–491.

Text Classification for Automatic Detection of E-Cigarette Use and Use for Smoking Cessation from Twitter: A Feasibility Pilot

Yin Aphinyanaphongs,
NYU Langone Medical Center, New York, NY 10016, USA

Armine Lulejian,
NYU Langone Medical Center, New York, NY 10016, USA

Duncan Penfold Brown,
New York University Social Media and Political Participation lab, New York, NY 10012, USA

Richard Bonneau, and
Simons Center for Data Analysis, New York, NY, 10010, USA

Paul Krebs
NYU Langone Medical Center, New York, NY 10016, USA

Yin Aphinyanaphongs: yin.a@nyumc.org; Armine Lulejian: Armine.Lulejian@nyumc.org; Duncan Penfold Brown: dpenfoldbrown@gmail.com; Richard Bonneau: rb133@nyu.edu; Paul Krebs: paul.krebs@nyumc.org

Abstract

Rapid increases in e-cigarette use and potential exposure to harmful byproducts have shifted public health focus to e-cigarettes as a possible drug of abuse. Effective surveillance of use and prevalence would allow appropriate regulatory responses. An ideal surveillance system would collect usage data in real time, focus on populations of interest, include populations unable to take the survey, allow a breadth of questions to answer, and enable geo-location analysis. Social media streams may provide this ideal system. To realize this use case, a foundational question is whether we can detect ecigarette use at all. This work reports two pilot tasks using text classification to identify automatically Tweets that indicate e-cigarette use and/or e-cigarette use for smoking cessation. We build and define both datasets and compare performance of 4 state of the art classifiers and a keyword search for each task. Our results demonstrate excellent classifier performance of up to 0.90 and 0.94 area under the curve in each category. These promising initial results form the foundation for further studies to realize the ideal surveillance solution.

1. Introduction

1.1. E-cigarettes

The use of e-cigarettes has been rapidly increasing since their introduction onto the market a few years ago. Sales of e-cigs and refillable vaporizers more than doubled to $1.7 billion in 2013.[1] Indeed, the trend has become so popular that ‘vape’ was voted word of the year for

Sharing: The labeled classifications for both e-cigarette use and e-cigarette use for smoking cessation are available at https://github.com/yina/2015-amia-ecig-twitter-labeled as per [32].

2014 by the Oxford Dictionaries.[2] A limited, yet growing body of literature suggests that e-cigarettes and vaporizers can create potentially harmful byproducts including heavy metals[3] and formaldehyde,[4] and product failure can result in severe injury and burns. Very little is known, however, regarding the use, prevalence, and characteristics of e-cigarettes. Two surveys among youth have indicated rapid increases in use since 2011,[5] and recent results from the 2014 Monitoring the Future survey indicated that 17% of 12th graders have used an e-cigarette in the past 30 days, surpassing the number who used combustible cigarettes.[6] Even less information on adult use exists, with the only national data being one consumer-research web survey,[7] indicating that 8.5% of adults have tried e-cigarettes with a rate of 36% among combustible cigarette users. No large-scale surveys have yet assessed more in-depth opinions about e-cigarette use, such as reasons for use or beliefs about harm.

1.2. Surveillance

Survey results are necessary to understand usage trends, establish national and regional health goals and inform regulations and prevention campaigns. These surveys – while excellent in many ways – have several limitations. First, there is a time lag before new products of abuse are incorporated into the surveys.[8] For example, neither the BRFSS,[9] the National Health Interview Survey,[10] nor the National Survey on Drug Use and Health (NSDUH)[11] ask about e-cigarette use yet. Second, the time lag in collection and analysis may delay timely policy interventions. Third, the surveys are sized to capture general trends across demographics and may lack focus for specific populations. Fourth, surveys have limitations in detecting usage by minors as most are not allowed to take the surveys. Fifth, surveys may contain limited content for any specific question as every additional question competes against other questions for time and space in the survey. Sixth, surveys capture high level geo-located information of use. Continuing use of high-quality national surveys to inform prevention and treatment services is critical, yet new technologies may address some of these limitations.

An ideal surveillance solution could capture new drugs of abuse, collect data in real time, focus on populations of interest, include populations unable to take the survey, allow a breadth of questions to answer, and enable geo-location analysis. We believe that social media streams may provide one solution. Social media, in this case, specifically Twitter, may include up to date vernacular for drugs of abuse, is inherently real time in how Tweets are broadcast, includes many potential populations of interest and their demographic characteristics, has populations such as minors who may not qualify for surveys, contains Tweets that indicate other potentially risky behaviors, and includes geo-locations. To realize using social media for surveillance, a foundational question is whether we can detect drug use at all. This work addresses this foundational concern and reports two pilot tasks for e-cigarettes. In the first, we identify automatically e-cigarette Tweets that indicate *e-cigarette use*. In the second we identify automatically Tweets that indicate *e-cigarette use for smoking cessation*.

1.3. Our Contribution

This feasibility paper explores state of the art machine learning based text classification methodologies for identifying e-cigarette use tweets. This paper makes several key contributions:

1. Defines a novel classification task for identifying e-cigarette use.
2. Defines a novel classification task for identifying e-cigarette use for smoking cessation.
3. Defines a process for labeling tweets to identify e-cigarette use and use for smoking cessation.
4. Establishes baseline classification results for these tasks.
5. Distributes these labeled datasets for general use by the community.

2. Background

2.1. Twitter As a Data Source

Among social media platforms, Twitter offers unique potential to serve as a tool for tracking substance use. Twitter is a micro-blogging service (with posts limited to 140 characters) through which users can send messages to a set of followers. It has over 600 million users worldwide with 46% of users logging on daily. In a recent Pew Research survey conducted August-September 2013, 18% of US adults use Twitter. [12] A higher percentage of Blacks/ African-Americans (29%) use Twitter compared with Whites (16%) and Hispanics (16%). Of Twitter subscribers, 31% are 18-29 and 19% are 30-49 years old.[12] Interestingly, there are relatively no differences in use by education level, gender, or income suggesting that use cuts across socioeconomic differences.

2.2. Twitter and E-cigarettes

A few studies have specifically addressed e-cigarettes via Twitter. Clark et al.[13] used 700,000 tweets collected from January 2012 to July 2014 to survey the general popularity and sentiment of consumer opinions regarding e-cigarettes.[14] In a follow up publication, they focused on approximately 20,000 geo-located tweets to characterize density and sentiment surrounding tobacco and e-cigarette tweets and link prevalence of word choices to tobacco and e-cigarette use at various localities.[14] In another publication, Huang et al.[15] labeled 73,672 tweets related to e-cigarettes to characterize how e-cigarettes are marketed, and Harris et al.[16] conducted a manual content analysis of tweets related to Chicago's regulation of e-cigarettes. While these studies produced a one-time picture of e-cigarette sentiment, neither the methodology of identifying e-cigarettes with a simple term search nor using manual coding are useful for ongoing surveillance purposes. A system that harnesses social media posts could serve as a low-cost method of examining usage trends and attitudes toward particular products.

In these previous studies, a common theme for analysis is the manual labeling of tweets. Manual labeling requires (1) time, (2) expertise, and (3) consistency. In addition, the samples must be small enough to allow feasible manual labeling which inherently is limited to the snapshot in time when the tweets are collected. Our aim in this paper is to use text classification machine learning techniques to address these limitations and convert these manual classifications to automated classifications. With tweet volumes of nearly 500 million per day, automation is the only realistic and feasible solution.

3. Methods

3.1. Corpus Construction

A challenge for building a labeled training corpus from Twitter is the low prevalence of Tweets in a target category. To enrich the e-cigarette use target category, we filtered Tweets by e-cigarette brand followers and hashtags. Specifically, we downloaded a 28.6 million tweet collection in January 2015. The tweet collection represents the tweets of 29,410 followers of the largest e-cigarette brands @v2cigs, @VaporFl, @HaloCigs, @bluecigs, @NJOYVape, @KRAVEeCig, and @LogicECig. To further increase the probability of encountering tweets about e-cigarette use and e-cigarette use for smoking cessation, we filtered the 28.6 million tweets with the Boolean OR of #vape #mod_ #vapeing #vaping #flavhub #eliquid #ejuice #pureclass or #ecigs. This corpus had 5,435 Tweeters covering a time span from Jan 2010 to Jan 2015 representing 228,145 Tweets. From these Tweets, we build a final corpus consisting of 13,146 randomly selected Tweets to label for our classifiers as outlined in section 3.2. The remaining 214,999 Tweets were not used for this pilot work and limitations with labeler time constrained the number of labeled Tweets.

3.2. Corpus Labels

3.2.1. Task 1 – E-cigarette Use—We defined e-cigarette use according to a similar protocol we developed for alcohol use.[17] Specifically, we considered tweets as positive if they indicate intent to use, the act of using, or sequel from use. Table 1 (on the following page) outlines our labeling protocol with examples.

3.2.2. Task 2 – E-cigarette Use for Smoking Cessation—We defined e-cigarette use for smoking cessation as tweets that indicate use for smoking cessation. These tweets were by definition a subset of the e-cigarette use tweets. In other words, a tweet for e-cigarette use for smoking cessation is also a tweet for e-cigarette use. Some example tweets indicating smoking cessation are shown in Table 2.

3.3. Corpus Label Protocol

We implemented the same procedure for both categories, *(1) e-cigarette use, and (2) e-cigarette use for smoking cessation*. To label tweets initially, two authors (YA and PK) independently coded a random subsample of 1,000 tweets using a draft coding protocol. Both authors held a consensus meeting to discuss labels that did not agree, and to refine the coding protocol.

To confirm the quality of the coding protocol, both authors blindly labeled 1,000 additional tweets. Because of the widely varying prevalence in classes, we calculated Siegel & Castellan's bias adjusted kappa. The resulting kappa was calculated at 0.87. [18] This high kappa suggested that the protocol and task were sufficiently generalizable. Corpus statistics are listed in Table 3. Finally authors YA and PK split and independently labeled the remaining 11,146 tweets.

3.4. Tweet Preprocessing

We relied on several preprocessing steps used successfully in other Twitter classification studies. [19, 20] For each tweet, we removed screen names (e.g. @britney), and urls. We then produced 4 encodings of the tweets as shown in Table 4 using the libshorttext program. [21] Each tweet is represented as a feature vector of counts for each token.

3.5. Algorithms

We use one baseline text classification and three state of the art text classification algorithms. We chose the baseline algorithm to establish the general difficulty of the task and three of the most recent state of the art classification algorithms.

3.5.1. Naïve Bayes—This algorithm is the text classification algorithm that typically serves as a baseline measure of text classification performance. This algorithm directly applies Bayes theorem to the classification task and assumes that the probability distribution of a feature is independent of another feature, given the class labels. We used the Multinomial Naïve Bayes [22] implementation in the mallet package. [23]

3.5.2. Liblinear—We employed a linear Support Vector Machine (SVM) classification algorithm as implemented in the liblinear package. The linear SVM's calculate maximal margin hyperplane(s) separating the two classes of the data. For text data, the linear SVMs demonstrated superior text classification performance compared to other methods [24], and this motivated our use of them. The liblinear implementation is an optimized version of the support vector machine optimized for quickly finding a linear separating hyperplane. We used liblinear as implemented in libSVM v1.96. [25] We used the default solver of L2-regularized L2-loss and the default penalty parameter of 1.

3.5.3. Bayesian Logistic Regression—We employed Bayesian logistic regression. This algorithm demonstrated superior performance in text classification benchmarks and thus motivates our inclusion of them. This algorithm constrains the coefficients using a Laplace prior and thus allows an efficient solution to the convex optimization. We used the bbrtrain [26] implementation for this study. We used the autosearch option to optimize the regularization parameter. This option does a grid search using 10 fold cross validation across the lambda parameters of 0.01 to 316 in multiples of the square root of 10.

3.5.4. Random Forests—We employed the random forest implementation in the fest [27] program. Random forests [28] are an ensemble classification method. The method produces a classification tree at each iteration. This classification tree is built from a random subset of the data, and at each node in the tree, a random subset of predictor variables are selected. Multiple trees are constructed in this fashion until at test time, the classification of these individual trees are combined to form a final prediction. We use the default settings that produces 100 trees with a maximum depth of 1000.

3.5.5. Keyword Comparisons—We compared the machine learning models to a simple keyword based approach for identifying tweets. Based on our definition, we asked PK to look at our protocol and the portion of the dataset that he reviewed and generate a Boolean keyword set that would provide a relative non machine learning baseline for this classification task. We added this analysis to address whether this task is difficult and to counter the claim that a human could craft a wordset that performs as well as the machine learning models. To simplify the comparison, we compare the keywords to the "unigram" encoding in one 10% split of the data. We used the keyword "OR" searches shown in Table 5.

4. Results/ Discussion

4.1. Task 1 – Ecigarette Use

4.1.1. Learning Algorithms for Task 1—Table 6 shows the results for identifying e-cigarette use. Bayesian Logistic Regression, Liblinear, and Random Forests perform with high area under the receiver operating curve. The performances of each classifier are in line with expected text classification performances as in prior studies. [29]

These results highlight performance differences in encoding the tweets. Using unigram or bigram representations demonstrate little performance differences within each classifier. Including stopwords increases performances as shown by comparison between the top 4 and bottom 4 rows . Stemming seems to have a marginal effect in this classification task. In addition, the ranges across the 10 folds are relatively stable likely reflecting the homogeneity of content for this labeled task.

4.1.2. Keyword Comparisons for Task 1—Keyword based searches are inferior to the machine learning methods. The keyword search returns a sensitivity and specificity while the machine learning methods return a ranked result. To make the comparison, we take one split from the 10 fold cross validation and obtain the sensitivity and specificity for the keyword search. The keyword search has a sensitivity of 0.75 and a specificity of 0.36. At 0.75 sensitivity, the best learning algorithm performs at 0.87 specificity (compared to 0.36). At 0.36 specificity, the best algorithm performs at 0.99 sensitivity (compared to 0.75). Task 1 of defining e-cigarette use through a keyword search benefits from machine learned models.

4.2. Task 2 – E-cigarette Use for Smoking Cessation

4.2.1. Learning Algorithms for Task 2—Table 7 (on the following page) shows the results for identifying e-cigarette use for smoking cessation. Random Forests perform with high area under the receiver operating curve. The high performance of this classifier for this task is possibly attributable to ensembling [30] that captures non-linearities and interactions effectively.

These results highlight performance differences in encoding the tweets. Using unigram or bigram representations demonstrate little performance differences within each classifier. Including stopwords increases performances as shown by comparison between the top 4 and bottom 4 rows that reflect keeping and removing stopwords respectively. Stemming seems to have a marginal effect in this classification task.

In contrast to the previous task, the ranges across the 10 folds are wide. These results likely reflect the small positive sample size of 73 in this dataset (even less in each fold) and the suspected heterogeneity (e.g. the many ways of communicating e-cigarette use for smoking cessation) in this labeled task.

For both tasks, retaining stopwords improves performance. This observation runs contrary to most other text classification tasks where removing stopwords typically does not affect performance. Stopwords can make a difference and prior researchers have shown that these words can affect performance depending on the task. [31] Further study is needed to examine which stopwords are important for classification in these tasks.

4.2.2. Keyword Comparisons for Task 2—Keyword based searches are inferior to the machine learning methods. The keyword search returns a sensitivity and specificity while the machine learning methods return a ranked result. To make the comparison, we take one split from the 10 fold cross validation and obtain the sensitivity and specificity for the keyword search. The keyword search has a sensitivity of 0.29 and a specificity of 0.99. At 0.29 sensitivity, the best learning algorithm performs at 0.99 specificity (compared to 0.99). At 0.99 specificity, the best algorithm performs at 0.37 sensitivity (compared to 0.29). Task 2 of defining e-cigarette use for smoking cessation through a keyword search is not trivial and benefits from machine learned models.

5. Limitations

5.1. Generalizability

The models we built focused on Tweets from followers of e-cigarette brands that contain the specific hashtags. The excellent classification performance in both tasks lay the groundwork for a much larger study that will sample from the larger Tweet population and consider Tweets that do not contain the hashtags or whom are not followers of the top e-cigarette brands.

5.2. Validity

This study does not establish directly the validity of the e-cigarette use behavior. Because someone Tweets about use does not mean they actually used. An additional study would survey the Tweeters and about their use habits. We could then compare the tweets about use to actual reported behavior by these Tweeters.

5.3. Users versus tweets

In this study, we focused on identifying Tweets automatically. We did not uncover the users associated with the e-cigarette use Tweets. A logical next step will identify the users whom Tweet the use. It is theoretically possible that many Tweets about use originate from a small number of users. Further analysis is necessary.

5.4. Applications

In this study, we did not explore specific applications. The eventual driver of performance will dictate the necessary performance. This performance depends on the specific application. For example, if we used the tweet classifications to identify the tweet locations (from the subset of geo-located tweets), we could be more liberal in choosing a threshold that allows false positives as the classifier over time will identify tweets and thus locations of e-cigarette use.

5.5. Comprehensiveness

In this pilot, we focused on testing the feasibility of automatic Tweet classification for this task. In later work, we would aim to produce the best classifiers or comprehensively compare classifier performance.

6. Conclusion

This pilot shows that we can successfully build models to identify tweets indicating e-cigarette use and e-cigarette use for smoking cessation. These promising initial results form the foundation to build an ideal surveillance system that can collect data in real time, focus on populations of interest by place and characteristic, include populations unable to take the survey, allow a breadth of questions to answer, and enable geo-location analysis

Acknowledgments

Supported in part by the NYU CTSA grant UL1TR000038 from NCATS, NIH.

References

1. E-Cig Sales Slide as Regular Smokers Return to Real Thing. [http://www.bloomberg.com/news/articles/2014-07-16/e-cig-sales-slide-as-regular-smokers-return-to-real-thing]
2. Chappell, B. The Two Way. NPR; 2014. Take It In: ‘Vape’ Is The Oxford Dictionaries Word Of The Year.
3. Goniewicz ML, Knysak J, Gawron M, Kosmider L, Sobczak A, Kurek J, Prokopowicz A, Jablonska-Czapla M, Rosik-Dulewska C, Havel C, et al. Levels of selected carcinogens and toxicants in vapour from electronic cigarettes. Tobacco control. 2014; 23(2):133. [PubMed: 23467656]
4. Jensen RP, Luo W, Pankow JF, Strongin RM, Peyton DH. Hidden Formaldehyde in E-Cigarette Aerosols. New England Journal of Medicine. 2015; 372(4):392–394. [PubMed: 25607446]
5. CDC. Notes from the field: electronic cigarette use among middle and high school students - United States, 2011-2012. MMWR Morbidity and mortality weekly report. 2013; 62(35):729–730. [PubMed: 24005229]
6. E-cigarettes surpass tobacco cigarettes among teens. [http://www.monitoringthefuture.org/data/14data.html-2014data-cigs]
7. King BA, Patel R, Nguyen K, Dube SR. Trends in Awareness and Use of Electronic Cigarettes among U.S. Adults, 2010-2013. Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco. 2014
8. Richtel, M. New York Times. New York, NY: New York Times; 2014. E-Cigarettes, by Other Names, Lure Young and Worry Experts.
9. Behavioral Risk Factor Surveillance System Questionnaire. [http://www.cdc.gov/brfss/questionnaires.htm]
10. National Health Interview Survey. [http://www.cdc.gov/nchs/nhis/quest_doc.htm]

11. Results from the 2010 National Survey on Drug Use and Health: Summary of National Findings. [http://www.samhsa.gov/DATA/NSDUH/2K10NSDUH/2K10RESULTS.HTM-APPB]
12. Social Media Update 2013. [http://www.pewinternet.org/2013/12/30/social-media-update-2013/]
13. Clark EM, Jones C, Gaalema D, Redner R, White TJ, Kurti A, Schneider A, Couch M, Dodds P, Danforth C. Electronic Cigarettes and Twitter: Sentiments, Categorization, and Hedonometrics. 2014
14. Clark EM, Jones C, Gaalema D, White TJ, Redner R, R E, Dodds P, Couch M, Danforth C. SoCial Media MeetS PoPulation health: a SentiMent and deMoGRaPhiC analySiS oF tobaCCo and e-CiGaRette uSe aCRoSS the “tWitteRSPheRe”. Value in Health. 2014:A603.
15. Huang J, Kornfield R, Szczypka G, Emery SL. A cross-sectional examination of marketing of electronic cigarettes on Twitter. Tobacco control. 2014; 23
16. Harris J, Moreland-Russell S, Choucair B, Mansour R, Staub M, Simmons K. Tweeting for and Against Public Health Policy: Response to the Chicago Department of Public Health's Electronic Cigarette Twitter Campaign. J Med Internet Res. 2014; 16(10):e238. [PubMed: 25320863]
17. Aphinyanaphongs, Y.; Ray, B.; Statnikov, A.; Krebs, P. WICSOC. Redmond City, CA: 2014. Text Classification for Automatic Detection of Alcohol Use Related Tweets.
18. Siegel, S. CN: Nonparametric statistics for the behavioral sciences. NY: McGraw Hill; 1988.
19. Kouloumpis E, Wilson T, Moore J. Twitter Sentiment Analysis: The Good the Bad and the OMG! Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. 2011
20. Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation. 2010
21. Yu C-HH HF, Juan YC, Lin CJ. LibShortText: A Library for Short-text Classification and Analysis. 2013
22. Kibriya EF AM, Pfahringer B, Holmes G. Multinomial naive bayes for text categorization revisited. Lecture notes in computer science. 2004
23. MALLET: A Machine Learning for Langauge Toolkit. [http://mallet.cs.umass.edu]
24. Joachims, T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms (The Springer International Series in Engineering and Computer Science. 1. Springer; 2002. p. 228
25. Fan R, Chang K, Hsieh C. LIBLINEAR: A library for large linear classification. The Journal of Machine. 2008
26. Genkin A, Lewis DD, Madigan D. Large-Scale Bayesian Logistic Regression for Text Categorization. Technometrics. 2007; 49:291–304.
27. Fast Ensembles of Sparse Trees (FEST). [http://lowrank.net/nikos/fest/]
28. Breiman L. Random forests. Machine Learning. 2001; 45:5–32.
29. Aphinyanaphongs Y, Fu LD, Li Z, Peskin ER, Efstathiadis E, Aliferis CF, Statnikov A. A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. J Assn Inf Sci Tec. 2014
30. Seni, G.; Elder, JF. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predicitons. Morgan and Claypool Publishers; 2010.
31. Riloff E. Little words can make a big difference for text classification. Proceedings of the 18th annual international ACM. 1995
32. McCreadie, R.; Soboroff, I.; Lin, J.; Macdonald, C.; Ounis, I.; McCullough, D. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM; 2012. On building a reusable twitter corpus; p. 1113-1114.

Table 1

Defining Tweets that indicate e-cigarette use (@mentions and urls removed)

Definitions of E-cigarette Use	Tweet	Label
Current behavior of using	#vaping with my new #vamo…carto tank filled with Cups 'O Peanut Butter from So good.	Positive
Owning or discussing paraphernalia and products	Loving the Nautilus Mini. Definitely my work/driving setup. No more stopping to drip when I'm dry! #Vapelif…	Positive
Entering contests to get products	I just entered to win 120mL of #ejuice from #vapelife #vape #vapers #ecig #vaping #ecigs # Check it out!	Positive
Use to quit smoking	See Table 2	Positive
Liking a brand or product	So far I'm loving my new #iStick #Eleaf #vape #vapelife #malibu	Positive
Asking others to gather to use	Vaper meet this saturday in London, team smokium will see you all there! Message me for details if your interested in coming. #ukvape #vape	Negative
Advertising	Kanger pro tank 3 (duel coil) is now in stock at our Jarrow Shop. #kanger #vapourvapourjarrow #ecigs	Negative
Announcing news reports	Louisville Delays Vote to Ban E-cigarettes in Outdoor Public Places #Spinfuel #Vape #Vaping #ecig #ecigs	Negative
Promotion of other's use	Happy Friday JJuice Friends! Stay Rad and Enjoy Yourselves this weekend. #vape #ecig…	Negative

Table 2

Tweets that indicate e-cigarette use (@mentions and urls removed)

Tweet
170 days no real cigarettes #vaping
I haven't used an ashtray in about 1 1/2 years………………………………because #VAPING!
I'm an Ex-Smoker now thanks to #Ecigs. Public Health > #H3639 #EcigsSaveLives #mapoli

Table 3

Tweet Corpus Labeled

Descriptor	Value
Number of tweets labeled	13,146
Number of tweeters in labeled set	2,147
Number of tweets for task 1 (e-cigarette use)	728
Number of tweets for task 2 (e-cigarette use for smoking cessation)	73

Table 4

Encoding of Datasets

Encoding Name	Word Tokens Stemmed?	Stopwords Removed?	Unigram or Bigram	Number of Features
unigram	No	No	Unigram	17,371
bigram	No	No	Bigram	109,213
stem_unigram	Yes	No	Unigram	14,509
stem_bigram	Yes	No	Bigram	101,617
stop_unigram	No	Yes	Unigram	17,021
stop_bigram	No	Yes	Bigram	90,037
stop_stem_unigram	Yes	Yes	Unigram	14,186
stop_stem_bigram	Yes	Yes	Bigram	82,464

Table 5

Keyword Searches

Category	Keyword Searches
E-cigarette Use[*$]	vape OR ecig OR ecigarette OR vaping OR ejuice OR vapers OR (drip AND tip) OR dripping OR (eliquid AND flavor) OR (e AND juice) OR (e AND liquid)
E-cigarette Use for Smoking Cessation[*]	(smoke AND free) OR (off AND cigarettes) OR (ex AND smoker) OR (no AND analogs) OR (I AND quit)

[*] note that we do not consider phrases in these keyword searches. We assume that a bigram phrase is equivalently represented as a Boolean AND. This assumption seems reasonable considering how short tweets are.

[$] the query (e AND juice) is the preprocessed version of the word token “e-juice” with punctuation removed.

Table 6

E-cigarette Use - 10 Fold Cross Validation Area Under the Receiver Operating Curve Performance (Range of Performances Across 10 folds)

Encoding Name	Naïve Bayes	Liblinear	Bayesian Logistic Regression	Random Forests
unigram	0.80 (0.77-0.86)	0.86 (0.83-0.90)	0.90 (0.86-0.92)	0.89 (0.86-0.92)
bigram	0.79 (0.75-0.83)	0.86 (0.83-0.90)	0.90 (0.87-0.93)	0.88 (0.85-0.91)
stem_unigram	0.82 (0.78-0.85)	0.87 (0.85-0.89)	0.90 (0.85-0.92)	0.88 (0.85-0.91)
stem_bigram	0.78 (0.73-0.82)	0.87 (0.84-0.90)	0.90 (0.87-0.93)	0.88 (0.85-0.91)
stop_unigram	0.79 (0.74-0.84)	0.83 (0.81-0.85)	0.86 (0.84-0.89)	0.83 (0.80-0.88)
stop_bigram	0.77 (0.72-0.80)	0.83 (0.78-0.88)	0.86 (0.84-0.89)	0.83 (0.80-0.86)
stop_stem_unigram	0.79 (0.75-0.85)	0.83 (0.79-0.86)	0.85 (0.82-0.87)	0.84 (0.77-0.88)
stop_stem_bigram	0.77 (0.71-0.81)	0.83 (0.80-0.86)	0.86 (0.83-0.89)	0.83 (0.82-0.87)

Table 7

E-cigarette Use for Smoking Cessation - 10 Fold Cross Validation Area Under the Receiver Operating Curve (AUC) Performance (Range of Performances Across 10 folds)

Encoding Name	Naïve Bayes	Liblinear	Bayesian Logistic Regression	Random Forests
unigram	0.57 (0.45-0.71)	0.78 (0.68-0.92)	0.88 (0.74-1.0)	0.94 (0.81-0.97)
bigram	0.53 (0.38-0.68)	0.75 (0.65-0.87)	0.87 (0.73-0.95)	0.93 (0.86-0.98)
stem_unigram	0.59 (0.51-0.78)	0.80 (0.55-0.90)	0.89 (0.67-0.98)	0.93 (0.82-0.99)
stem_bigram	0.50 (0.38-0.71)	0.76 (0.65-0.88)	0.89 (0.83-0.95)	0.94 (0.86-0.97)
stop_unigram	0.59 (0.49-0.74)	0.71 (0.40-0.91)	0.87 (0.76-0.97)	0.90 (0.81-0.96)
stop_bigram	0.59 (0.42-0.70)	0.69 (0.44-0.82)	0.86 (0.71-0.98)	0.88 (0.81-0.98)
stop_stem_unigram	0.60 (0.51-0.72)	0.70 (0.41-0.86)	0.85 (0.72-0.95)	0.86 (0.79-0.96)
stop_stem_bigram	0.57 (0.46-0.66)	0.69 (0.41-0.90)	0.83 (0.37-0.94)	0.87 (0.80-0.97)