A demographic and sentiment analysis of e-cigarette messages on Twitter

Elaine Cristina Resende and Aron Culotta
Department of Computer Science
Illinois Institute of Technology
Chicago, IL
eresende@hawk.iit.edu, culotta@cs.iit.edu

ABSTRACT

Social media provide a potentially useful new data source to understand emerging public health behaviors. In this paper, we study messages about e-cigarettes posted to Twitter.com. We apply methods to classify messages by sentiment and to estimate the gender and age of users. We apply our approach to nearly one million messages about e-cigarettes posted from October 2012 to September 2013. We find that overall volume of e-cigarette tweets increased five-fold (from 30K per month to 150K per month); and that males and younger users were more likely to post positive messages about e-cigarettes. A qualitative analysis also reveals several trends, such as negative sentiment toward people who smoke in class; females giving e-cigarettes to relatives to help them quit smoking; and spikes in people using e-cigarettes to quit smoking in January.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications–Text processing; H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Web mining, social media, public health

1. INTRODUCTION

Understanding evolving behaviors and attitudes related to alcohol, tobacco, and other drugs is a critical public health goal. Typically, such topics are investigated by survey research; however, surveys can be costly and time-consuming, making them ill-suited to rapidly changing environments. Furthermore, survey response rates have fallen precipitously in recent years (Kohut et al. 2012), leading researchers to seek non-traditional data sources. A promising new methodology is social media analysis, which analyzes online content expressing health-related behaviors and opinions to provide insights into population-level trends. Online content offers several potential advantages over traditional data: one can in real-time measure how behaviors and attitudes change in response to rare events such as legal changes, new products, and marketing campaigns, and the open-ended nature of the content can provide more diverse data. Such approaches have been used to study depression [4], insomnia [10], and other health topics [5].

In this paper we present a descriptive study of attitudes towards electronic cigarettes (e-cigs) expressed on Twitter. E-cigs provide a nicotine-containing aerosol with different flavors, glycol and other ingredients that users smoke by heating up a solution [8]. E-cig adoption has increased rapidly in the United States recently [11, 6]; by one estimate, usage by high school students tripled in 2014 [12]. Despite this rapid growth, there is still considerable debate over the health impact of e-cigs [23, 15, 1], also reflected in consumer surveys [18], leading to uncertainty as to how they should be regulated.

We analyze nearly one million tweets posted from October 2012 to September 2013 that contain keywords related to e-cigarettes. We use a supervised classification algorithm to annotate each message by sentiment (positive, negative, or neutral). Additionally, using the first name of each user, we derive estimates of age and gender to further stratify results. Our main findings are as follows:

- Overall volume of e-cig tweets grew five-fold in our year sample, from 30K tweets in October 2012 to 150K in September 2013.
- Of tweets expressing sentiment, 65% are classified as positive sentiment (tweets either advocating for e-cigs or indicating that the user has tried e-cigs). This value ranges by month from 61% in March 2013 to 74% in July 2013.
- We find positive sentiment to be slightly higher for males than females (63% vs. 61%), and highest for users estimated to be 18-24 years old (67%).

We additionally perform a qualitative analysis to provide a more fine-grained insight into the different ways people discuss e-cigs and how that is related to sentiment and demographics. For example, in November-January we find a spike in messages from users wishing for e-cigs as a Christmas present to help them stop smoking. We also find that many tweets from young female users mention smoking e-cigs with their parents or buying e-cigs for their parents; whereas male users are more likely to ridicule those using e-cigs.

Table 1: Training Data Summary

Class	#Tweets	Tweet Sample
Positive	707	I bought a Ecig today Electric cigarettes are better than regular cigarettes!!
Negative	279	#IHatePeopleThat smoke ecigs e-cigs are bad, mmkay?
Neutral	1014	what is an electronic cigarette? A homeless guy just asked me if he could bum an e-cigarette

Table 2: Top Coefficients of Classifier

negative	you, smoking, smoking an, he, fuck, people, smokes, an, faggot, smoke, class, stupid, are, in, look, pussy, her, sorry, one, his
neutral	URL, e-cigarettes, de, la, 99, retail, URL retail, ni, e-cigarette, markten, store, by, cigarette, of, dallas, smokers, 9999, @vaper_trail, electronic, may
positive	my, i, vaping, #vaping, #ecig, my ecig, me, #vape, we, got, e-cig, my e-cig, #euecigban, i'm, my e, vape, good, and, this, i need

The remainder of the paper is organized as follows: Section 2 reviews related work; Section 3 describes the Twitter data and our method of analysis; Section 4 describes our main results; Section 5 discusses implications, limitations, and future work.

2. PRIOR WORK

Myslín et al. [14] manually classified 4K tobacco-related tweets along 30 dimensions, including sentiment, theme, and genre, finding that tweets about e-cigs and hookahs tended to have more positive sentiment. Here, we focus specifically on e-cigs, expanding upon this initial work with a much larger set of tweets (1M) over a longer time span, and further stratifying by gender and age.

Huang et al. [9] analyzed 74K tweets related to e-cigs and, using a supervised classifier, found that 90% were commercial tweets, and about 10% mentioned smoking cessation. We build on this prior work, using their classifier to first filter out commercial tweets.

Other work has found a prevalence of smoking cessation accounts on Twitter [19], and has attempted to track the progress of cessation attempts by following Twitter accounts of identified users [13].

Very recently, Godea et al. [7] analyzed 106K tweets containing e-cig related terms collected over a two month period. They built a sentiment classifier using a hand-engineered lexicon of over 600 terms, achieving a positive/negative F1-score of 55-56%.

Compared to this prior work, our main contributions consist of (1) an analysis of a much larger sample of e-cig tweets than has been done previously, consisting of nearly one million tweets written over one year; (2) a sentiment classifier tuned for precision that identifies positive tweets with 96% precision and negative tweets with 70% precision; (3) an analysis of temporal trends in sentiment and demographics.

3. METHODS

In this section we describe the data collected and the methods used for classifying messages by sentiment, age, and gender.

3.1 Data Collection

We use the data collection process described in Huang et al. [9]. Using the full Twitter Firehose, tweets were collected from October 2012 to September 2013 using a set of keywords identified by expert consensus (*e-cigarette*, *ecigarette*, *e-cig*, *ecig*). Additional tweets were identified that matched the query: (*cig* OR *cigarette*) AND (*electronic* OR *blu* OR *njoy*) (the latter two terms referring to the top-selling e-cig brands in the U.S.). This resulted in 4,639,885 tweets. Using the classifier of Huang et al. [9], we retained only those tweets identified as "organic," defined as "those reflecting individual opinions or experiences or linked to non-promotional content." This left 992,633 tweets.

On Twitter it is common for very similar messages to be posted many times, either as retweets or as part of a coordinated marketing campaign. Additionally, a number of accounts in our data are primarily focused on e-cigarettes, either as an official corporate account, or as "astroturf" accounts that are created to artificially inflate the perceived sentiment towards e-cigs [20]. As we are primarily interested in the sentiment of genuine, ordinary users, we further filtered the data as follows: (1) we removed all retweets; (2) we removed all tweets whose content was duplicated in other tweets; (3) we retained only the first tweet from each user. Figure 2a shows the number of tweets per month for each filtering step. After all filtering, 455,648 tweets remain.

3.2 Sentiment Classification

We next wished to classify each tweet by sentiment towards e-cigs. To do so, we fit a supervised classifier (logistic regression) to a collection of manually annotated tweets. From the 455K tweets above, we uniformly sampled 2,000 tweets and manually categorized them as positive, negative or neutral as follows:

- tweets from users who expressed buying or use desire, or from users who tweeted about their e-cigs, or who expressed support of e-cigs were labeled as **positive**;
- tweets from users who were expressing complaints or antipathy towards e-cigs were labeled as **negative**.
- tweets from merchandising companies, news about electronic cigarettes, and tweets belonging to other languages were labeled as **neutral**; they did not express any sentiment.

Figure 1: Precision-recall curves derived by classifying the labeled data using 10-fold cross-validation.

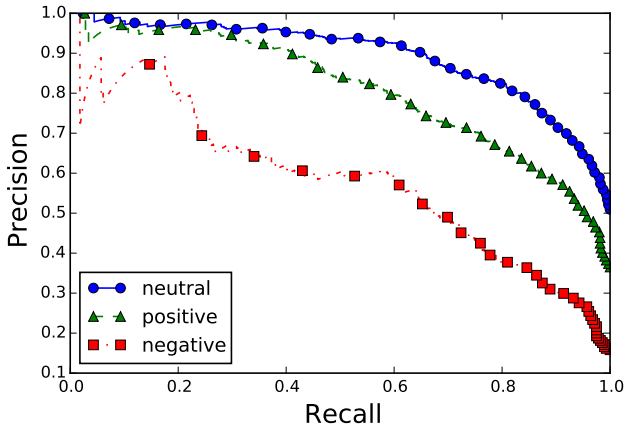


Table 3: Cross-validation classification accuracy

	Prec	Rec	F1	N
negative	0.60	0.55	0.57	279
neutral	0.78	0.84	0.81	1014
positive	0.75	0.68	0.72	707
avg	0.74	0.75	0.74	2000

Note that our definition of **positive** is a bit different from typical sentiment classification; by including tweets indicating possible usage of e-cigs, we hoped to capture general notions of popularity of e-cigs, as opposed to simply opinions about e-cigs.

Table 1 describes some examples of our training data with tweets, their classes and also the distribution of each class. We can see that roughly 51% of the tweets were labeled as neutral, 14% as negative and 35% as positive.

We fit a logistic regression classifier to the labeled tweets, using L2 regularization. After a series of pilot experiments, we adopted the following tokenization scheme: (1) convert to lower case; (2) maintain hashtags and mention terms; (3) remove punctuation (except for internal punctuation like hyphens and apostrophes); (4) remove characters repeated more than twice consecutively; (5) collapse all URLs to the same term; (6) collapse all digits to the same character. We retain both unigrams and bigrams, removing any term that does not occur in at least two tweets. We represent each tweet by a tf-idf vector (dividing term frequency by document frequency), normalized to unit length.[1]

Figure 1 displays the precision-recall curves for the logistic regression classifier using 10-fold cross-validation on the 2,000 hand-labeled tweets; Table 3 reports the precision, recall, and F1 metrics for each class; and Table 2 reports the top-weighted coefficients per class.

The F1 score averaged over each class is .74; the classifier is most accurate on the neutral class (.81 F1) and least accurate on the negative class (.57 F1).

Table 2 suggests that negative tweets are mostly ridiculing other people who use e-cigs, while positive tweets are typically first-person accounts of wanting or using e-cigs. Many of the neutral tweets are marketing related (that were missed by prior filters) or informative, as indicated by the presence of links, typically to news stories.

While the classifier accuracy is higher than that reported in prior work on a related task (Godea et al. [7] report F1 scores of 55-56% for positive/negative classes), we desire a high precision classifier to strengthen the validity of the conclusions drawn from its application to the remaining unlabeled tweets. Fortunately, we can use confidence thresholds to improve the precision. The precision-recall graph indicates that if we restrict our classifications to the 25% that the classifier is most confident in, the precision values for the positive and negative classes are .96 and .70, respectively. To classify each of the 455K unlabeled tweets, then, we apply the classifier trained on the 2,000 labeled tweets, setting confidence thresholds to achieve these levels of precision (the confidence threshold is .65 for the positive class and .5 for the negative class). Classifications below these thresholds are placed in the neutral class. Thus, we reduce the number of tweets labeled as sentiment-bearing, but increase the precision of the remaining classifications.

3.3 Gender and Age Inference

Following the work of Pavalanathan and Eisenstein [16] and Silver and McCann [22], we use name statistics to estimate the gender and age of each user in the data. We extract the first token from the *name* field in the user's profile, where available. We then compare this to government statistics regarding the gender and age distribution for each name.

For gender, we collect names from the census that comprise 75% of the population (to remove rare names that may produce false matches, such as *The*). We additionally remove names that appear both as male and female names. This results in 226 male and 518 female names. We use these two lists to assign each user a gender label, when possible.

For age, we use data from the Social Security Administration indicating baby names by year, along with life expectancy tables, to estimate the age distribution of a person with a given name. We define the years in age brackets (under 18, 18-24, 25-34, 35-44 and 45+). Thus, for each user with a matching name, we produce a probability distribution over the five age brackets. See Silver and McCann [22] for more details.

4. RESULTS

Here we report results of sentiment, age, and gender inference.

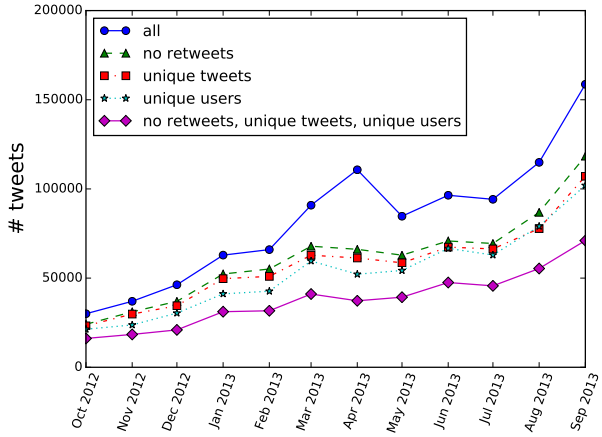
4.1 Sentiment

The sentiment classifier returns 103,103 positive and 56,652 negative tweets from the 455K unlabeled tweets. Figure 2b plots the sentiment per month. Additionally, in the second y-axis, we report the percentage of non-neutral tweets that are labeled as positive (i.e., $\frac{\#\text{positive}}{\#\text{positive} + \#\text{negative}}$). Overall, 65% of the non-neutral tweets are labeled as positive. We emphasize that this does not mean that 65% of Twitter users like e-cigs; rather, when a Twitter user posts a message about e-cigs, it is more likely to be in the positive class (e.g., about using or wanting to use e-cigs) than in the negative

[1]All code to reproduce our analysis is available at `https://github.com/tapilab/chs-2015-ecig`.

Figure 2: Tweets by month and sentiment.

(a) Tweets by month.



(b) Tweets by sentiment.

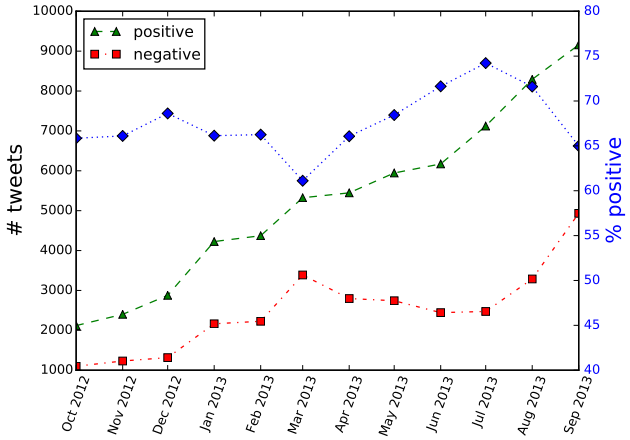
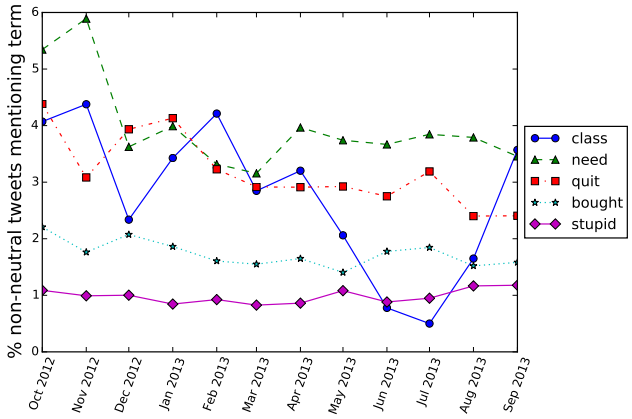


Figure 3: Percent of non-neutral tweets mentioning each term.



class (e.g., ridiculing the use of e-cigs).

There appears to be a mild increasing trend in percent of positive sentiment during this time period (a linear fit yields a slope of .41), but there is an obvious dip in March 2013 and spike in July 2013. In March, the South Korean singer Onew was photographed smoking an e-cig. Due to his reputation as a role model for his young fans, this news led to many critical tweets. Even after our filtering steps to remove retweets and duplicate tweets, this still resulted in 1,127 unique tweets mentioning Onew (out of 41K tweets in March). Of these 1,127 tweets, 470 were classified as positive and 2 were classified as negative. A representative negative tweet is: "*He smokes too WTF ONEW WHYYY! Not a cig but an electronic one at that!*"

The spike in sentiment in July does not appear to be related to one event. There is a gradual increase from March through July, then a swift decline in August and September. We believe this is due in large part to a common type of negative tweet in which a student criticizes another student for smoking e-cigs in a classroom, e.g., "*Who uses an e-cig during class? #Idiot.*"

To further investigate this, Figure 3 plots by month the percentage of all sentiment-bearing tweets containing a hand-selected set of keywords. For example, in November 2012, over 4% of all tweets classified as positive or negative contained the term "class." We can see that this plot closely matches the U.S. academic calendar, with drops in December, June, and July. There also appear to be spikes at the start of each school session (January/February and August/September), perhaps suggesting that either students stopped using e-cigs in class or it became a less notable incident.

Another interesting trend in Figure 3 concerns the terms "need" and "quit." In December and January, there are a number of users who report that they are interested in using e-cigs to help them quit smoking traditional cigarettes, e.g., "*Also, quitting smoking lasted a good 14 hours. I think I really need to get myself an e cig :(* ." This spike in the start of the year is likely due to the fact that smoking cessation is a common New Year's resolution.

Other terms that correlate with negative sentiment (e.g., "stupid") appear to comprise a consistent portion of sentiment tweets, with a small increase in the final three months.

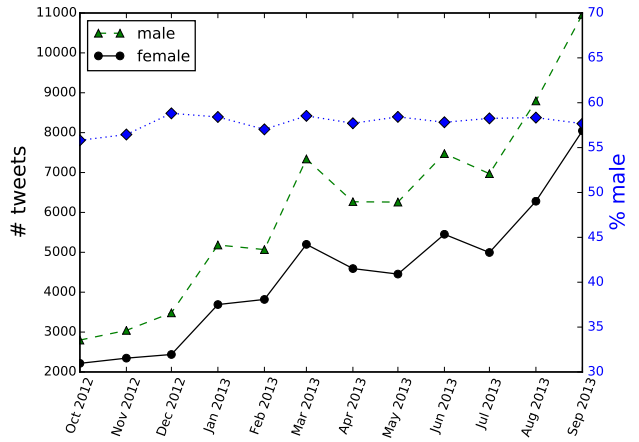
To further examine the salient terms used in positive and negative tweets, we grouped all tweets labeled by the classifier as positive or negative. We then computed Chi-Squared statistics for each feature, indicating how strongly each term correlates with the positive or negative class. This allows us to summarize the differences between these groups in the sample of 455K tweets. The top terms are shown in the first two rows of Table 4. Similar to Table 2, positive tweets tend to be first-person accounts of e-cig usage, while negative tweets tend to be second-person criticisms. We can see that "class" is the term most correlated with negative sentiment, in line with Figure 3.

Table 4 also displays a similar analysis to identify the most distinctive terms per month. A number of topics emerge, including regulation ("restrictions", "regulation"), celebrities (Onew, Courtney Love), and news reports ("cdc," "patches," referring to a report comparing the effectiveness of e-cigs and nicotine patches for cessation). We will highlight these topics in more detail below.

4.2 Gender

Figure 5: Tweets by gender and sentiment.

(a) Tweets by gender.



(b) Tweets by gender by sentiment.

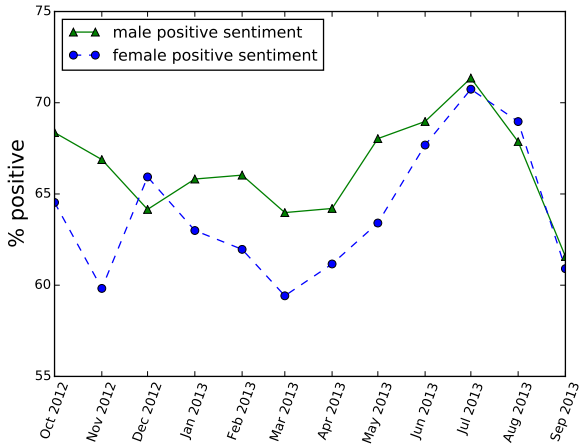
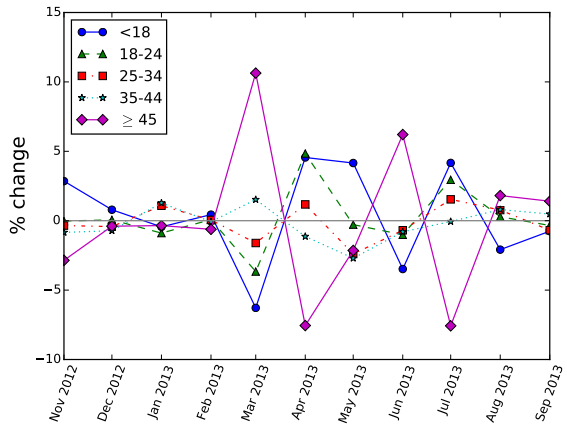
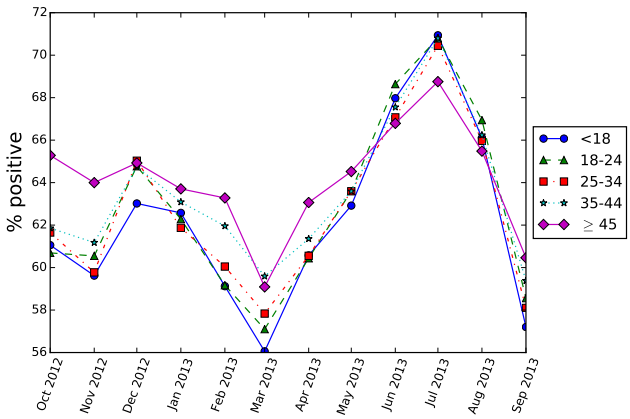


Figure 6: Tweets by age and sentiment.

(a) Tweets by age.



(b) Tweets by age by sentiment.



Overall, we identified 73,647 male and 53,528 female tweets from the filtered set of 455K tweets. We were unable to identify gender for the remaining tweets; many users leave the name field blank or enter a fake name. Figure 5a plots the gender distribution by month, and Figure 5b plots the percentage of non-neutral tweets labeled as positive by gender.

Males consistently tweets about cigarettes more than females for all months. Moreover, male sentiment is typically more positive than female sentiment, and the sentiment trends between the two groups are mostly comparable. An exception is December 2012, in which male sentiment decreases, but female sentiment increases, surpassing male sentiment. Examining positive, female tweets in December, we observe a number of tweets indicating attempts to get family members to quit smoking using e-cigs, e.g., "*Ima buy my dad an electronic cigarette for Christmas, no more tobacco pal*"; "*About to make my dad's day with an early Christmas present. I got him an E-Cig so he'll quit smoking.*"; "*My mom got an E-smoke for christmas. Anything to get her to quit.*" There are also a smaller number of tweets indicating parents smoking e-cigs with their children, e.g., "*Smoking an e-cig with my dad #uh.*"

As in the previous section, we computed Chi-Squared statistics for each feature to identify the top terms correlating with each gender. The second and third rows of Table 4 further confirm that females are more likely to discuss e-cigs in reference to other family members, whereas males are more likely to discuss their own e-cigs ("my") as well as to use profanities to insult e-cig users. Furthermore, the appearance of "class" on the female list suggests that many of the negative tweets about smoking in class (see the previous section) are authored by female users.

4.3 Age

Unlike for gender, for age we compute a distribution over brackets for each user. When we compute the average of these distributions for all users, we get the following estimated age distribution: under 18: 32%, 18-24: 13%, 25-34: 16%, 35-44: 14% and 45+: 25%. Due to well-known sample bias in Twitter users, we expect that this result overestimates the prevalence of the very young and the very old (as in Pavalanathan and Eisenstein [16]). Thus, we focus on

Table 4: Most significant terms by category using Chi-squared tests.

Positive sentiment	my, i, got, ecig, need, love, e-cig, me, i'm, cig, lost, mom, want, vaping, bought, dad, broke, so, just, get
Negative sentiment	class, smoking, an, you, in, kid, cool, electric, people, guy, look, smoke, he, you're, stupid, if, fuck, dude, cigarette, are
Female	cigarette, my, her, electric, mom, dad, smoking, me, amp, electronic, an, class, omg, #giveaway, so, x, his, boyfriend, i, onew
Male	vaping, green, e-cigs, bro, blu, pope, e-cigarettes, @blucigs, ban, #vape, @youtube, stephen, the, game, pussy, new, a, any, vapers, gay
<18	electric, cig, class, an, cigarette, e, lol, just, mom, cool, those, guy, haha, was, hookah, mask, elaborate, dad, shit, my
18-24	electric, cigarette, those, lol, cig, class, her, fait, electronique, blu, hookah, got, smh, cutting, laugh, quand, feel, me, mom, commercial
25-34	an, cigarette, electric, cig, class, my, just, those, i, bought, guy, so, her, girl, his, mom, kid, dad, me, cigs
35-44	cigarette, #gotitfree, electric, those, blu, i, cigs, just, cig, bought, an, these, was, lol, electronic, luck, her, it, what, smoking
$\geq$ 45	electric, thuis, markten, cigarette, alle, van, hq, those, blu, via, e-smoking, e-cigarette, elaborate, @lord_sugar, mask, just, cigs, mistic, reviews, was
2012-10	electric, cigarette, electronic, cigarettes, green, free, alternative, trial, class, those, obtain, an, during, enjoys, liked, looks, obama, photo, stephen, lol
2012-11	election, cigarette, electric, mask, elaborate, electronic, kristen, traditional, device, rob, halo, na, brand, knee, touching, kit, cigarettes, lol, was, class
2012-12	christmas, electric, elaborate, cigarette, mask, electronic, @overlymanlymann, na, ko, addiction, advertising, ako, ny, ng, cigarettes, hq, allowed, calif, report, those
2013-01	prevent, electric, accessories, regulating, rolling, sale, cigarette, banning, electronic, na, fda, continue, ng, ko, launch, tv, those, haha, cigarettes, ni
2013-02	amg, miracle, menace, boat, racing, electric, drive, cigarette, elaborate, mask, prevent, class, decision, 99/99, crave, continues, electronic, including, opinion, @e_swisher
2013-03	onew, green, utah, cigarette, gallagher, noel, muse, pope, drummer, caught, moves, criticises, electronic, electric, #gotitfree, risks, smoke, mistic, tax, an
2013-04	courtney, f-bomb, drops, ad, love, @simoncowell, electric, njoy, #gotitfree, inside, bring, commercial, cigarette, blu, class, an, lol, those, i, watch
2013-05	@lord_sugar, france, la, electronique, les, governo, logic, places, electric, ecig, une, my, an, public, c'est, electrique, lol, il, mais, cigarette
2013-06	e-smoking, package, restrictions, incredible, britain, rise, medicines, include, kits, medicine, website, parker, face, markten, experience, easy, perfect, alle, thuis, reynolds
2013-07	markten, van, alle, thuis, vuse, vaporizer, #uk, e-cig, ook, my, op, hookah, ecig, tweets, being, i, ik, regulation, blu, me
2013-08	promise, markten, companies, benowitz, neal, thuis, echo, van, cigs, heyday, alle, york, e, stadium, literally, merthyr, professor, never, bloomberg's, ook
2013-09	doubles, among, patches, students, teens, cdc, e-cigarettes, use, kids, effective, flames, 9-foot, survey, shows, markten, middle, u.s, teach, e-cigarette, study

the month-by-month differences in these distributions. Figure 6a shows the percent change for each bracket in each month.

The large spike in 45+ users in March 2013 appears to be due in part to a word-of-mouth marketing campaign promoting e-cigs; #gotitfree and #giveaway are used disproportionately more by 45+ users in this month. The spike in young users in April and May appears to be in part due to a viral video of the musician Courtney Love promoting NJOY brand e-cigs in a commercial containing profanity. This was disproportionately shared and discussed by young users.

To visualize the relation between age and sentiment, we assigned each user the most likely age bracket based on their estimated distribution. We then computed the percentage of positive sentiment tweets by age bracket. Figure 6b plots these values by month. The spike in positive sentiment for young users in June-August 2013 again appears to be due in part to the drop in tweets about smoking in class, as discussed in Section 4.1.

Table 4 also shows the terms with the strongest correlation to each age bracket. Intuitively, the term "class" tends to be indicative of younger users, as are references to "mom" and "dad." Older users tend to use fewer abbreviations (e.g., "electronic cigarettes" vs. "ecigs"); they also appear to be more likely to participate in word-of-mouth marketing campaigns ("#gotitfree"). It is possible that these older users may in fact be part of a coordinated "astroturf" campaign on behalf of e-cig companies, though further analysis is required to verify this.

Finally, Figure 4 shows the percentage of non-neutral tweets from each demographic category that are labeled as positive. We observe somewhat more positive sentiment among males, and the highest sentiment among 18-24 year olds. The higher sentiment of this age group is in line with some

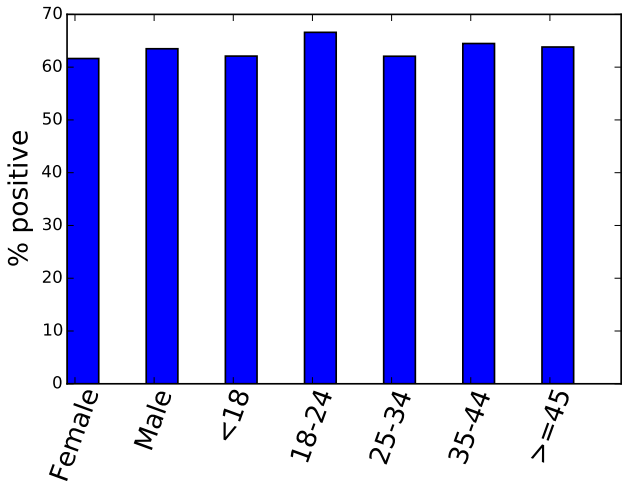


Figure 4: Percent of non-neutral tweets from each group classified as expressing positive sentiment towards e-cigarettes.

survey results of e-cig usage. A meta-analysis found ever-use highest among young adults (20-28 year olds) [2]; while Regan et al. [21] found 18-24 year olds to have the highest rate of ever-use. Pearson et al. [17] also find higher usage among males.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have quantified trends in e-cig messages posted to Twitter. Overall, the volume of tweets mentioning e-cig terms has grown five-fold from October 2012 to September 2013. A sentiment analysis classifier indicates that males and young users are more likely to post positive tweets, which either support e-cigs or mention their usage. The open-ended nature of the data provides opportunity for additional analysis that may be difficult with traditional surveys; for example, we find many users posting negative comments about fellow students smoking in class; many female users mentioning buying e-cigs to help family members stop smoking; and older users participating in promotional campaigns. Such observations may suggest avenues for future research and interventions.

There are a number of limitations and difficulties using such a noisy data source. While some trends match the conclusions of traditional surveys (e.g., younger, male users were most likely to post positive messages), the tweet volume is often influenced by rare events, such as a viral commercial (Courtney Love) or photograph (Onew smoking an e-cig). These events lead to a spike in messages that may not reflect true sentiment toward e-cigs. While we have taken steps to mitigate these (e.g., removing duplicate tweets), new methods are required to disentangle "genuine" tweets versus "pop culture" tweets. Additionally, it may be fruitful to investigate ways to measure the long-term impact of such one-time events; for example, did the Courtney Love video lead to an overall increase in e-cig awareness?

Additionally, while our analysis used a year of data, a multi-year study may be required to remove some of the observed cyclical effects (e.g., the academic year and end-of-year holidays appear to influence observed sentiment). Finally, we have used only first name statistics to infer age and gender. In future work, we will consider more sophisticated approaches to infer a wider range of attributes with higher precision and recall [3].

6. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under grant #IIS-1526674. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] A. Bhatnagar, L. P. Whitsel, K. M. Ribisl, C. Bullen, F. Chaloupka, M. R. Piano, R. M. Robertson, T. McAuley, D. Goff, and N. Benowitz. Electronic cigarettes a policy statement from the american heart association. *Circulation*, 130(16):1418–1436, 2014.

[2] S. L. C. Chapman and L.-T. Wu. E-cigarette prevalence and correlates of use among adolescents versus adults: a review and comparison. *Journal of psychiatric research*, 54:43–54, 2014.

[3] A. Culotta, N. R. Kumar, and J. Cutler. Predicting the demographics of twitter users from website traffic data. In *Twenty-ninth National Conference on Artificial Intelligence (AAAI)*, 2015.

[4] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *ICWSM*, 2013.

[5] M. Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84, 2012. ISSN 1541-1672. doi: 10.1109/MIS.2012.76.

[6] C. for Disease Control, P. (CDC, et al. Notes from the field: electronic cigarette use among middle and high school students-united states, 2011-2012. *MMWR. Morbidity and mortality weekly report*, 62(35): 729, 2013.

[7] A. K. Godea, C. Caragea, F. A. Bulgarov, and S. Ramisetty-Mikler. An analysis of twitter data on e-cigarette sentiments and promotion. In *Artificial Intelligence in Medicine - 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015. Proceedings*, pages 205–215, 2015.

[8] R. Grana, N. Benowitz, and S. A. Glantz. E-cigarettes a scientific review. *Circulation*, 129(19):1972–1986, 2014.

[9] 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):26–30, 2014.

[10] S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Lawson. "I can't get no sleep": Discussing #insomnia on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1501âĂŞ1510, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2208612. URL `http://doi.acm.org/10.1145/2207676.2208612`.

[11] B. A. King, S. Alam, G. Promoff, R. Arrazola, and S. R. Dube. Awareness and ever-use of electronic cigarettes among us adults, 2010–2011. *Nicotine & Tobacco Research*, 15(9):1623–1627, 2013.

[12] K. Leonard. E-cigarette use triples among teens: The cdc chief blames aggressive marketing for the devices' growing use among young people. http://www.usnews.com/news/articles/2015/04/16/e-cigarette-use-triples-among-teens, 2015. Accessed 2015-06-01.

[13] E. L. Murnane and S. Counts. Unraveling abstinence and relapse: Smoking cessation reflected in social media. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1345–1354, New York, NY, USA, 2014. ACM.

[14] 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.

[15] J. Pauly, Q. Li, and M. B. Barry. Tobacco-free electronic cigarettes and cigars deliver nicotine and generate concern. *Tobacco Control*, 16(5):357–357, 2007.

[16] U. Pavalanathan and J. Eisenstein. Confounds and consequences in geotagged Twitter data. In *EMNLP*, 2015.

[17] J. L. Pearson, A. Richardson, R. S. Niaura, D. M. Vallone, and D. B. Abrams. E-cigarette awareness, use, and harm perceptions in us adults. *American journal of public health*, 102(9):1758–1766, 2012.

[18] J. K. Pepper, S. L. Emery, K. M. Ribisl, C. M. Rini, and N. T. Brewer. How risky is it to use e-cigarettes? smokersâĂŹ beliefs about their health risks from using novel and traditional tobacco products. *Journal of behavioral medicine*, 38(2):318–326, 2015.

[19] J. J. Prochaska, C. Pechmann, R. Kim, and J. M. Leonhardt. Twitter= quitter? an analysis of twitter quit smoking social networks. *Tobacco control*, 21(4):447–449, 2012.

[20] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

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

[22] N. Silver and A. McCann. How to tell someone's age when all you know is her name. http://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name/, 2014. Accessed 2015-06-01.

[23] T. E. Sussan, S. Gajghate, R. K. Thimmulappa, J. Ma, J.-H. Kim, K. Sudini, N. Consolini, S. A. Cormier, S. Lomnicki, F. Hasan, et al. Exposure to electronic cigarettes impairs pulmonary anti-bacterial and antiviral defenses in a mouse model. *PloS one*, 10(2): e0116861, 2015.