An analysis of e-cigarette sentiment and demographics 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 platforms, such as Twitter plays a big role in the society nowadays and they have become an important source for public health surveillance and others applications. Electronic cigarette is an outstanding topic in the health area nowadays, and one important question that arises is to know how people are dealing with this new idea of "smoking free", in other words, how they are concerned about e-cigs? In order to answer questions like that we need to know what people are saying about e-cigs, if it is positive or negative and how this sentiment changes over the time. In this project, we have been working with Twitter data and we have tried to find out the sentiment of the users related to electronic cigarettes. Through our analysis we have noticed interesting facts that can contribute to the marketing and health campaigns, and also that most of sentiment is negative (marketing or criticism).

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

Electronic cigarettes (e-cigs) provide a nicotine-coitaining aerosol with different flavors, glycol and other ingredients, that users smoke by heating up a solution [3]. There are some researches being done to confirm how harmful e-cigs can be, such as [10, 9], but e-cigs have not been still well studied and the effects of long-term use is not known yet [1]. King et al.[5] state that electronic cigarettes use is increasing a lot among adults, and [2, 6] also describes that the use among middle school students increased, and for high school students tripled in 2014. In this project we have been working with twitter data which was collected during a year period, from October 2012 through September 2013, looking for mentions and for people talking about e-cigs. The main goal of this project is analyze what Twitter users are talking about this subject, and try to get the sentiment of the users towards e-cigs, in order to find out how people are receiving the use of e-cigs and also to check the type of customers who are buying them.

2. PRIOR WORK

[7] collected twitter data and analysed tobacco-related posts from November 2011 through July 2012, right before our data collection which was from October 2012 through September 2013. They used three machine learning algorithms and varied parameters, such as n-grams and the number of features used. Their final data set is composed of 7362 tweets and our data set is composed of more than 900k tweets. They were manually classified and each one of them was assigned to several categories (genre, theme and sentiment). The authors found that those tweets are usually more positive (41%/30% positive/negative ratio) towards tobacco, specially tweets related to hooka and e-cigarettes. Fist-hand personal experience is more correlated with positive sentiment and opinion correlates more strongly with negative sentiment. They also found that social factors influence in the sentiment toward tobacco. Young users and social relationships are keys that emerge with positive sentiment, what tells that positive sentiment has a interaction between hooka and e-cigarettes, younger users and positive social experiences.

This paper focused on analysing tobacco-related tweets, positive and negative sentiment, but it does not report possible age and gender of the users. So, our study fills this gap by inferring gender and age of Twitter users in order to address the problem to monitor closely certain ages and also to drive public health campaigns to the right public.

3. METHODS

3.1 Data Collection

Tweets were collected through keywords about electronic cigarettes (e-cigarette,ecigarette, e-cig, or ecig). Same data collection process was done in this study as in the work [4]. A Support Vector Machine (SVM) algorithm was used to classify tweets in two classes commercial, which linked to commercial websites, branded promotional messages or spammer accounts, and non-commercial, which showed individual opinions and experiences. As result, 992,633 tweets

Table 2: Cross-validation classification accuracy

	Prec	Rec	F1	N
negative	0.84	0.83	0.84	1296
positive	0.70	0.72	0.71	704
avg	0.79	0.79	0.79	2000

were classified as organic (non-commercial), and all the collection add up to 4,639,885.

Aiming to analyze the sentiment of users we need to work just with organic tweets, so we have selected just organic tweets which represent (21.39%) of the data. Figure **??** shows the number of tweets (y axis) by month (x axis). The green line refers to organic tweets and blue line to all tweets. We can see that there is a big spike in March and also an increase in the months of May and September.

A random sample of 2000 collected tweets, from the set of organic tweets, were manually labelled as positive, negative and neutral. Each tweet was put in one class according to the following:

- tweets from merchandising companies, news about electronic cigarettes, and tweets belonging to other languages were classified as neutral class, they did not express any sentiment;
- tweets from users who expressed buying or use desire, or from users who tweeted about their e-cigars were considered positive class;
- and tweets from users who were complaining (or citing) about other people vaping were labelled as negative.

Table 1 describes some examples of our training data with tweets, their classes and also the distribution of each class. We can see that 52.45% of the tweets were labeled as neutral, 13.75% as negative and 33.80% as positive.

For this analysis we have joint negative and neutral tweets because we are more interested in the positive, in other words, telling that you use e-cigs is more important then just criticism or marketing campaigns. So, at this point we have 676 (33.8%) tweets classified as positive and 1324 (66.2%) as negative. The coded set of tweets was used to train a machine learning classifier using Logistic Regression algorithm, and the built model was applied to the full set of organic tweets.

3.2 Classification

A Logistic Regression (LG) classifier was built, trained it with coded tweets, and tested it with the organic set. Section 3.2.1 and 3.2.2 are going to show how the methodology of training and testing we followed.

3.2.1 Training

A Logistic Regression classifier was used in Python language, which is responsible to predict the labels of all organic tweets; the set of labelled tweets were used as our training set. In order to have a better accuracy for the classifier, we used a grid search to check the best values for "Penalty" and "C", which are parameters of LG. Running the grid search we found out that Penalty = L2 and C = 2.6 is the best configuration for our LG model. Figure **??** shows the accuracy versus C parameter, and we can see that the accuracy is higher when C is equal to 2.6.

Our best classifier configuration has the accuracy of 81.8%. The confusion matrix of the model is shown in the Figure **??**, and Table 2 shows precision, recall and F-score values for both classes. As we can notice the percentage of retrieved tweets that are relevant is 90% what is seen as a good value for our assignment.

3.2.2 Testing

Our testing set is defined as all organic tweets (900k tweets). The LG classifier was applied to them, as result we had 27% classified as positive and 73% as negative. What is comparable to our manual labeling which was 33.8% positive and 66.2% negative.

In order to check the quality of our LG model, 200 tweets were randomly selected and labeled. LG model was again applied to this same new set of tweets. The labels were used as true labels and the predicted labels (output from LG model) were compared to the true labels to get the accuracy. As result, the model had 79% of accuracy, so we conclude that our model is good for predicting the sentiment of tweets for this problem.

The top and bottom words of the classes are shown in the Table 3. Top coefficients refer to negative class and bottom coefficients show words related to positive class. Through Table 3 we can see that LG model learned pretty well because words in the Top Coefficients, such as THIS_IS_A_URL, RETAIL and STORES are related to marketing tweets, also HE, YOU, PEOPLE and SHE refer to other people, what represents tweets criticizing people who use e-cigs, thus being negative class. The bottom coefficients MY, I, VAPING, VAPE and BUY refer to users who tweet about getting or using e-cigs.

Using three-class classifier: predicted label distribution on raw tweets: [(neutral, 537474), (positive, 329919), (negative, 125240)]

4. ANALYSIS OF THE SENTIMENT DURING A YEAR (OCT/2012-SEP/2013)

We analyzed the amount of tweets by month according to their sentiment (negative or positive) toward e-cigs. In the Figure **??** we have considered all organic tweets. In the Figures **??** and **??** it is considered just users who tweeted once. Figure **??** one tweet per user is considered, and in the Figure **??** does not consider re-tweets.

In those Figures red represents positive tweets, green represents negative and blue all organic tweets. Figure **??** shows spikes in April, June and September for all sentiment lines, but Figures **??** and **??** have moved the first left spike from April to March, what tell us that the number of users was higher in March than April. Figure **??** and Figure **??** have similar shape (same for Figures **??** and **??**), they differ just in the number of tweets in the y axis.

Aiming to try to find out why we have those spikes we analyzed n-grams, which are a sequence of words in the tweet, and also the 10 top words for each month during the year, shown in the Table **??**. We can see that the words in the table are very correlated with the tweets posted in the respective month. For example, November has the word THANKSGIVING, CHRISTMAS appears in December, VALENTINE in February, and so on. These words can help us to understand the spikes, and what are the main topics for positive and negative tweets. Some of the spikes are due to some events that occurred during those months. Besides

Table 1: Training Data Summary

Class	#Tweets	Tweet Sample
Positive	676	I bought,a Ecig today
		@MikeeeDeeeLeon,Nice I tried to smoke electronic cigarette last Saturday, it was my first,time.
Negative	275	This kid,was hitting an e-cig during class. I look over and smokes coming out his,nose.
		e-cigs,are bad, mmkay? [URL]
Neutral	1049	#vapor,#smoking,#vaping - Vapor Joes - Daily,Vaping Deals: Boge ViVi Nova - $6.99 Free Shipping [URL],- #ecig #VapeDeal
		@SenKClark,Stop #H3639 by @jeffrey_sanchez , Millions have quit smoking with the #Ecig.,#H3639 promotes #cigarettes and #cancer !!

Table 3: Top Coefficients of Classifier

negative	URL, my, for, to, i, good, #euecigban, #ecigs, new, we, of, #ecig, my ecig, ban, la, de, and, buy, #vaping, or
neutral	you, smoking, smoking an, he, fuck, people, smokes, an, faggot, are, class, smoke, stupid, in, look, shit, her, pussy, sorry, one
, this, and, good, ecig	

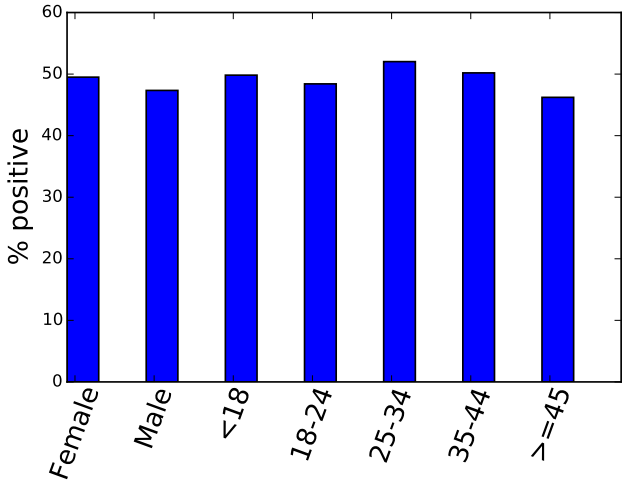


Figure 2: Percent of tweets from each group classified as positive sentiment.

the significance of each word, the Table **??** also shows the percentage of tweets that each word is present, what can describe the importance of them to the entire data. Following we have a brief analysis description of the spikes shown in the graphs.

March

Analysing Figure **??** which shows one tweet per user, March has 44788 negative and 17656 positive tweets. If we look in the Table **??**, the month of March shows the word ONEW, what refers to a Korean singer vaping. A lot of Twitter users tweeted about him, and most of tweets were negative. We also have the words VATICAN and POPE which represent the month of the choice of the pope, and these are also classified most negative.VAPERWARE is a store and most of the tweets are classified as negative. SWSW is a music, film and interactive festival happening in March and most tweets are negative classified. 300 tweets were saying: BLACKFRIDAY #BLACKFRIDAY IS TOO GOOD FOR ME BECAUSE I USED ELECTRONIC CIGARETTE WHICH I MISSED A LOT and they were classified as positive tweets.

This month had about 16.5% of users who posted positively according to Figure **??**.

April

Considering just positive tweets from Figure **??** and going through the n-grams of the month of April we found in total 17548 tweets. After analyzing the n-grams, we took a further look in those tweets and we found more than 600 re-tweets of: " @SIMONCOWELL: I HAVE CUT DOWN SMOKING. A LOT. USING ELECTRIC CIGARETTES TO HELP. ARE THEY BAD FOR YOU?" and they were labeled as positive.

According to Figures **??** and **??**, April was the month which people talked less in positive way toward e-cigs. In the Table **??**, we can see that is because some senators in US made regulations to restrict the sale of e-cigars [1], and also about Courtney TV commercial.[2]

Some other words from Table **??** are described below and most of the tweets containing them were classified as negative, what contributed for negativeness in this month.

#**welcometomyschoolwhere** PEOPLE SMOKE ELECTRONIC CIGS DURING CLASS

E-CIGARETTES **primarily** USED TO QUIT TOBACCO: STUDY HTTP://T.CO/L2G1DKSNJT #EUECIGBAN

Leachon: E-CIGARETTES MAY CONTAIN CARCINOGENS, INGREDIENTS NOT APPROVED BY FDA

September

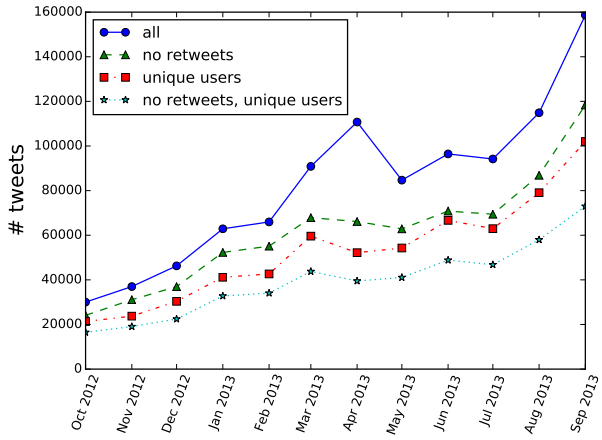
This month is the third most negative with 15% of positive tweets overall. 28938 tweets are positive and 78341 are

[1]http://www.enewspf.com/latest-news/health-and-fitness/42182-senators-call-on-fda-to-restrict-the-sale-distribution-and-marketing-of-e-cigarettes.html

[2]http://www.huffingtonpost.com/2013/04/03/courtney-love-njoy_n_3006660.html?ncid=edlinkusaolp00000003

Figure 1: Tweets by month and sentiment ≥.

(a) Tweets by month.



(b) Tweets by sentiment.

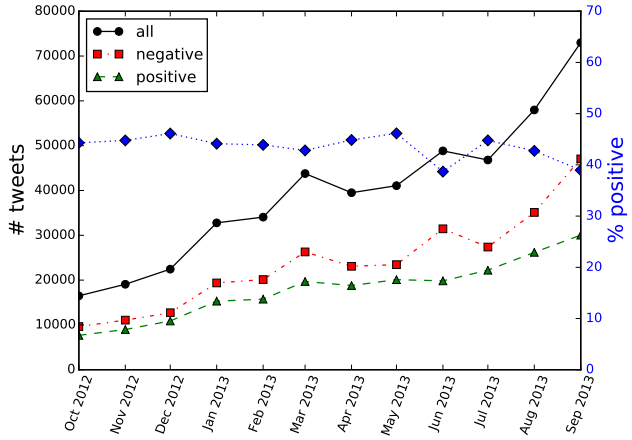
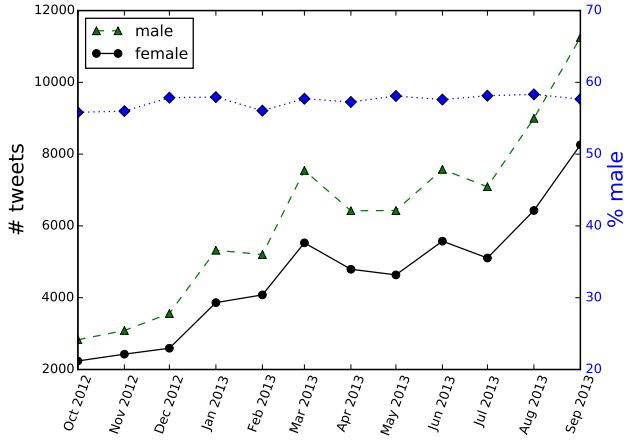
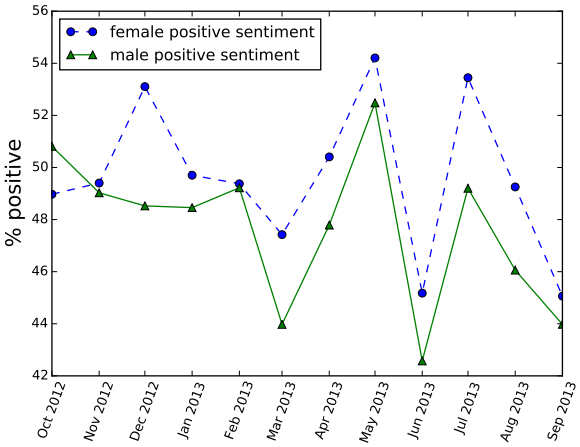


Figure 3: Tweets by gender and sentiment.

(a) Tweets by gender.



(b) Tweets by gender by sentiment.



negative according to Figure **??**.

The word **whisp** refers to "RT YABOYLILB: *GETS IN BOOSTER SEAT* *PUTS SEATBELT ON* *PUTS IN THE NEWEST KIDZBOP CD* *TAKES DRAG OF E-CIG* *LOWERS GLASSES* *WHISP" and it was classified as positive.

worldvapingday is a hashtag about http://www.world-vaping-day.com/, which is the day of vaping, and most of them were negative. Two main tweets appeared in the data:

1. THURSDAY IS #WORLDVAPINGDAY! TIME TO TELL THE WORLD HOW #ECIGS HAVE CHANGED YOUR LIFE!!!. It was classified as negative.

2. THURS. IS #WORLDVAPINGDAY! CONSERVATIVES NEED TO KNOW THE FDA WILL TRY TO BAN #ECIGS AGAIN IN OCT. STOP LIBERALS,SAVE LIVES. It was classified as positive.

PROPERLONGBOARD VAPING DALLAS PARTY GIFT E-JUICE & **Nude** SELFIES HTTP://T.CO/MSKYYKFZZW explains the appearance of the word nude, and they were negative tweets.

The word **attorneys** represents tweets saying about e-cigarette regulations. E.g.: VEGASNEWSNOW ATTORNEYS GENERAL CALL ON FDA TO REGULATE E-CIGARETTES [URL] #VEGAS. Most classified as negative.

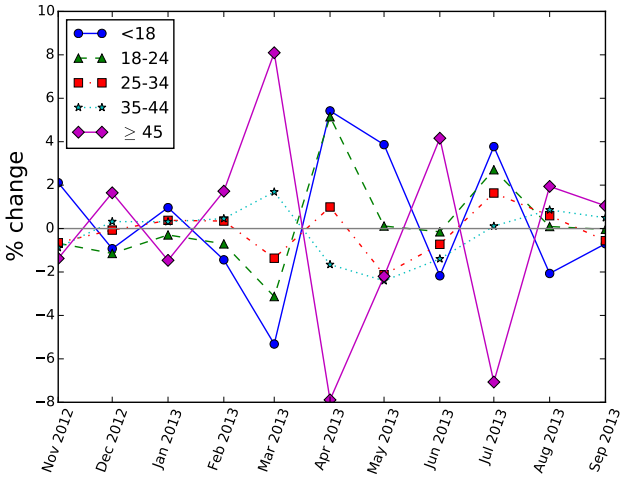
Vapefest occurs during September and most of tweets were negative. E.g.: CANT WAIT FOR #VAPEFEST NEXT WEEKEND. STOKED

Teens represents tweets talking about the grow of teens vaping, and they also represent negative class. E.g.: BAKO-COM STUDY SHOWS MORE TEENS TURNING TO ELECTRONIC CIGARETTES: #PALMSPRINGS #[URL]

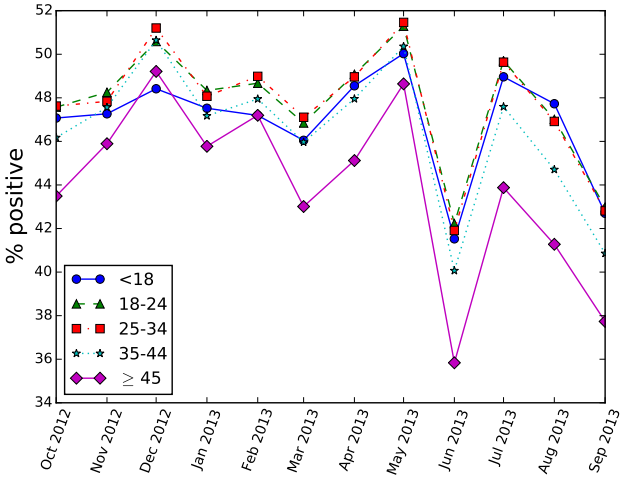
Figure **??** shows, in the right y axis, the percentage of positive tweets over all organic with LG model built with 2 classes (positive and negative). We can see that the percentage of positivism is higher during October, December and February, and then it goes down during March and April. May starts going up again, but for the rest of the months no one goes upper than May. The left y axis describes the number of users, and in the x axis, the bars show the number of users who are posting positively by month, from October

Figure 4: Tweets by age and sentiment.

(a) Tweets by age.



(b) Tweets by age by sentiment.



2012 through September 2013 in blue color, and we can notice that this number is much less than negative, what tell us that most of tweets are propaganda or criticism about someone using e-cigs.

In order to understand more Figure **??**, we have done an analysis considering the 3 classes explained in Section 3.1. Figure **??** plots the percentage of positive tweets over the year considering positive and negative tweets from LG model using 3 classes (positive, negative and neutral). As it is expected, the percentage of positivism increases. April still is the month with less positive tweets, October is the most positive month overall.

4.1 Gender Inference

U.S. Census provides list of names from population. We use the 1990's census[3] to compute the proportion of males and females for positive and negative classes. We keep 75% of each names list, eliminate ambiguous names, at the end the total names are 226 males and 518 females. Each tweet gets then its label, if the first name of the user is present in the male list, it is labeled as male or the same for female list. Figure **??** shows the distribution of genders in our data. 74% overall tweets have their gender unknown because the users did not have their names in the census data and majority of tweets belonged to users of companies (369127 unique users are labeled as "unknown"). We can see that 28-30% of users, independently of gender, express positive sentiment towards electronic cigarettes, but the majority express negativeness toward them.

4.2 Age Inference

In a recent study Silver and McCanc [8], showed how the age of a person can be indicated by her/his name. The Social Security Administration (SSA) provides a list of baby names[4] who are born in each year, from 1880 to 2014. Through this list we can know gender and number of babies born with a name by year. Another resource we use to infer age is a cohort life table also provided by SSA[5], that give us the estimate of how many people born in a given year are still alive.

We define the years in age brackets (under 18, 18-24, 25-34, 35-44 and 45+) and with the number of people in certain age and who are still alive we can compute the fraction of people who are still alive for each age bracket. For example, given a name we can see the fraction of people who are 18 years old, which is in the first bracket. Figure **??** shows an example for the name Elaine, the first column represents the brackets; "n" is the number of people who born during that period, "n_alive" is the number of people born in that period who are still alive, and "fraction" shows the percentage of people in that bracket who are still alive. In this figure we can see that the most common age estimate for Elaine is 45+, followed by 35-44, 25-34, 18-24 and 18-.

The age inference analysis was done by month, considering both classes of our data set. The next figures show the age brackets estimates by month during the year data, plotting separately by sentiment. Figures **??**, **??**, and **??** plot the sentiment by age for each age bracket. Figures **??**, **??**, and **??** show the differences in the estimates of age.

5. DISCUSSION

5.1 Relation to prior survey results

"Likewise, Pearson et al. (2012) found no difference in men and women's ever-use of e-cigarettes (men: 12.6

Younger and older individuals appeared equally likely to have ever used e-cigarettes in the U.S., but younger individuals were more likely to ever-use them in other regions.

6. CONCLUSION

As we could notice from our analysis the over all tweets increase over the time, but negative tweets (marketing, criticism) increases more than positive tweets. Most of tweets are negative and the positive tweets don't change much over the time, they remain around 16% and 23%. Most of the

[3]http://www2.census.gov/topics/genealogy/
[4]http://www.ssa.gov/oact/babynames/
[5]http://www.ssa.gov/oact/NOTES/as120/LifeTables_Tbl_7.html

Table 4: Significant terms by category.

Positive sentiment	i, my, me, ecig, got, vaping, i'm, e-cig, an, cig, and, it, want, get, too
Negative sentiment	URL, e-cigarettes, cigarettes, e-cigarette, via, de, la, electronic, electronique, markten, are, health, en, smokers, as
Female	cigarette, my, her, me, electric, mom, electronic, dad, smoking, amp, missed, blackfriday, i, #giveaway, really
Male	vaping, green, e-cigs, blu, bro, pope, e-cigarettes, @blucigs, the, @youtube, #vape, ban, stephen, game, new
<18	electric, cigarette, cig, class, an, lol, e, just, my, i, me, mom, cool, mask, really
18-24	cigarette, electric, those, lol, her, fait, class, me, electronique, cig, smh, i, blu, feel, hookah
25-34	cigarette, an, my, electric, i, me, cig, class, really, those, just, blackfriday, electronic, missed, bought
35-44	blackfriday, missed, cigarettes.its, dear, loved, hi, cigarette, electronic, hello, lot, #blackfriday, i, really, ones, #gotitfree
≥ 45	cigarettes.its, loved, hi, dear, blackfriday, hello, missed, ones, check, guys, friends, lot, which, #blackfriday, really
2012-10	electric, cigarette, electronic, cigarettes, green, those, free, quitsmoking, alternative, trial, class, obtain, lol, during, an
2012-11	election, cigarette, quitsmoking, electric, mask, elaborate, electronic, #quitsmoking, kristen, traditional, device, rob, halo, na, knee
2012-12	christmas, #blackfriday, cigarettes.its, electronic, dear, loved, blackfriday, electric, missed, cigarette, elaborate, mask, quitsmoking, ones, really
2013-01	prevent, electric, accessories, regulating, rolling, electronic, sale, cigarette, banning, na, fda, continue, #blackfriday, ng, undo
2013-02	amg, miracle, #blackfriday, menace, boat, racing, cigarette, electric, blackfriday, cigarettes.its, loved, missed, drive, dear, enjoys
2013-03	onew, green, utah, cigarette, electronic, #blackfriday, #gotitfree, gallagher, noel, muse, pope, moves, caught, drummer, criticises
2013-04	courtney, f-bomb, drops, ad, sexy, love, americans, bring, inside, @simoncowell, restrict, contact, electric, marketing, #gotitfree
2013-05	@lord_sugar, france, la, places, electronique, les, really, governo, logic, cigarettes.its, blackfriday, loved, my, missed, fortune
2013-06	e-smoking, package, restrictions, incredible, britain, rise, medicines, medicine, include, kits, website, parker, face, markten, experience
2013-07	markten, van, alle, thuis, vuse, vaporizer, #uk, my, e-cig, ook, op, i, hookah, me, ecig
2013-08	promise, markten, discusses, companies, benowitz, neal, considering, sales, echo, thuis, van, heyday, @bieberrkfc, cigs, alle
2013-09	doubles, among, patches, students, teens, cdc, e-cigarettes, use, effective, kids, flames, 9-foot, survey, shows, u.s

spikes that the Figures showed are due to a particular event during its respective month. So, we are sure that TV commercials and new campaigns are relevant and they help to make changes in the data.

7. 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] 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.

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

[4] 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.

[5] 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.

[6] 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`. Accessed 2015-06-01.

[7] 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.

[8] A. M. Nate Silver. 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/`. Accessed 2015-06-01.

[9] 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.

[10] 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 anti-viral defenses in a mouse model. *PloS one*, 10(2):e0116861, 2015.