7-1-2016

Vaporous marketing: Uncovering pervasive electronic cigarette advertisements on twitter

Eric M. Clark
University of Vermont

Chris A. Jones
University of Vermont

Jake Ryland Williams
Computational Story Lab

Allison N. Kurti
University of Vermont

Mitchell Craig Norotsky
University of Vermont

See next page for additional authors

Recommended Citation

Clark EM, Jones CA, Williams JR, Kurti AN, Norotsky MC, Danforth CM, Dodds PS. Vaporous marketing: uncovering pervasive electronic cigarette advertisements on Twitter. PLoS One. 2016 Jul 13;11(7):e0157304.

PLOS ONE

RESEARCH ARTICLE

Vaporous Marketing: Uncovering Pervasive Electronic Cigarette Advertisements on Twitter

Eric M. Clark[1,2,3,4]*, Chris A. Jones[4,5,6], Jake Ryland Williams[2,3,7], Allison N. Kurti[5], Mitchell Craig Norotsky[4], Christopher M. Danforth[1,2,3], Peter Sheridan Dodds[1,2,3]

1 Department of Mathematics & Statistics, University of Vermont, Burlington, VT, United States of America, **2** Computational Story Lab, Burlington, VT, United States of America, **3** Complex Systems Center, University of Vermont, Burlington, VT, United States of America, **4** Department of Surgery, University of Vermont, Burlington, VT, United States of America, **5** Vermont Center for Behavior and Health, University of Vermont, Burlington, VT, United States of America, **6** Global Health Economics Unit of the Vermont Center for Clinical and Translational Science, University of Vermont, Burlington, VT, United States of America, **7** University of California Berkeley, School of Information, Berkeley CA, United States of America

* eric.clark@uvm.edu

Citation: Clark EM, Jones CA, Williams JR, Kurti AN, Norotsky MC, Danforth CM, et al. (2016) Vaporous Marketing: Uncovering Pervasive Electronic Cigarette Advertisements on Twitter. PLoS ONE 11(7): e0157304. doi:10.1371/journal.pone.0157304

Editor: Raymond Niaura, Legacy, Schroeder Institute for Tobacco Research and Policy Studies, UNITED STATES

Received: September 16, 2015

Accepted: May 30, 2016

Published: July 13, 2016

Data Availability Statement: All relevant data are included in the manuscript and Supporting Information files. Data are also available on a public server provided by the University of Vermont (http://www.uvm.edu/storylab/share/papers/clark2016a/TwitterEcigMarketing/).

Funding: The authors wish to acknowledge the Vermont Advanced Computing Core which provided High Performance Computing resources contributing to the research results. EMC was supported by the UVM Complex Systems Center; PSD was supported by NSF Career Award #0846668. PSD and CMD

Abstract

Background

Twitter has become the "wild-west" of marketing and promotional strategies for advertisement agencies. Electronic cigarettes have been heavily marketed across Twitter feeds, offering discounts, "kid-friendly" flavors, algorithmically generated false testimonials, and free samples.

Methods

All electronic cigarette keyword related tweets from a 10% sample of Twitter spanning January 2012 through December 2014 (approximately 850,000 total tweets) were identified and categorized as Automated or Organic by combining a keyword classification and a machine trained Human Detection algorithm. A sentiment analysis using Hedonometrics was performed on Organic tweets to quantify the change in consumer sentiments over time. Commercialized tweets were topically categorized with key phrasal pattern matching.

Results

The overwhelming majority (80%) of tweets were classified as automated or promotional in nature. The majority of these tweets were coded as commercialized (83.65% in 2013), up to 33% of which offered discounts or free samples and appeared on over a billion twitter feeds as impressions. The positivity of Organic (human) classified tweets has decreased over time (5.84 in 2013 to 5.77 in 2014) due to a relative increase in the negative words 'ban', 'tobacco', 'doesn't', 'drug', 'against', 'poison', 'tax' and a relative decrease in the positive words like 'haha', 'good', 'cool'. Automated tweets are more positive than organic (6.17 versus 5.84) due to a relative increase in the marketing words like 'best', 'win', 'buy', 'sale',

were supported by NSF Big Data Grant #1447634. CJ and AK are supported in part by the National Institute of Health (NIH) Research awards R01DA014028 & R01HD075669, and by the Center of Biomedical Research Excellence Award P20GM103644 from the National Institute of General Medical Sciences.

Competing Interests: The authors have declared that no competing interests exist.

'health', 'discount' and a relative decrease in negative words like 'bad', 'hate', 'stupid', 'don't'.

Conclusions

Due to the youth presence on Twitter and the clinical uncertainty of the long term health complications of electronic cigarette consumption, the protection of public health warrants scrutiny and potential regulation of social media marketing.

Introduction

Electronic Nicotine Delivery Systems, or e-cigs, have become a popular alternative to traditional tobacco products. The vaporization technology present in e-cigarettes allows consumers to simulate tobacco smoking without igniting the carcinogens found in tobacco [1]. Survey methods have revealed widespread awareness of e-cigarette products [2, 3]. The health risks [4–7], marketing regulations [8], and the potential of these devices as a form of nicotine replacement therapy [9–11] are hotly debated politically [12] and investigated clinically [13, 14]. The CDC reports that more people in the US are addicted to nicotine than any other drug and that nicotine may be as addictive as heroin, cocaine, and alcohol [15–18]. Nicotine addiction is extremely difficult to quit, often requiring more than one attempt [18, 19], however nearly 70% of smokers in the US want to quit [20]. Data mining can provide valuable insight into marketing strategies, varieties of e-cigarette brands, and their use by consumers [21–25].

Twitter, a mainstream social media outlet comprising over 230 million active accounts, provides a means to survey the popularity and sentiment of consumer opinions regarding e-cigarettes over time. Individuals post tweets which are short text based messages restricted to 140 characters. Using data mining techniques, roughly 850,000 tweets containing mentions of e-cigarettes were collected from a 10% sample of Twitter's garden hose feed spanning from January 2012 though December 2014. This analysis extends a preliminary study [26] which analyzed all e-cigarette related tweets spanning May through June 2012.

As Twitter has become a mainstream social media outlet, it has become increasingly enticing for third parties to gamify the system by creating self-tweeting automated software to send messages to organic (human) accounts as a means for personal gain and for influence manipulation [27]. We recently introduced a classification algorithm that is based upon three linguistic attributes of an individual's tweets [28]. The algorithm analyzes the average hyperlink (URL) count per tweet, the average pairwise dissimilarity between an individual's tweets, and the unique word introduction decay rate of an individual's tweets.

All tweets mentioning e-cigarettes were categorized using a two-tier classification process. Tweets containing an abundance of marketing slang ('free trial', 'starter kit', 'coupon') are immediately categorized as automated. All of the tweets from individuals that have mentioned an e-cigarette keyword are collected in order to classify the remaining tweets per individual as either organic or automated. The machine learning classifier was trained on the natural linguistic cues from human accounts to identify promotional and SPAM entities by exclusion.

The manipulative effects, agendas, and ecosystem of generalized social media marketing campaigns have been identified and extensively studied [29–31]. Other work, [32], has distinguished between purely automated accounts, or "robots", and human assisted automated accounts referred to as "cyborgs". On Twitter, these campaigns have also been characterized using Markov Random Fields to classify accounts as either promotional or organic [33]. This

study was able to achieve very high classification accuracy, but was working under a much shorter time frame (1 month) and was trained on all relevant tweets authored within this time window. Our study compiled a 10% sample of tweets over a three-year period, so we relied on a classifier that was trained on smaller samples of tweets per individual.

The emotionally charged words that contribute to the positivity of various subsets of tweets from each category were quantitatively measured using hedonometrics [34, 35]. Outliers in both the positivity and frequency time-series distributions correspond to political debates regarding the regulation of e-cigarettes. Recent studies [36–40] report an alarmingly rapid increase in the youth awareness and consumption of electronic cigarettes; a Michigan study found that the use of e-cigarettes surpass tobacco cigarettes among teens [41]. The CDC reports that "the number of never-smoking youth increased three-fold from approximately 79,000 in 2011 to 263,000 in 2013" [42]. During this time-period there has also been a substantial (256%) increase in youth exposure to electronic cigarette television marketing campaigns [43]. Due to the high youth presence on Twitter [44] as well as the clinical uncertainty regarding the risks associated with e-cigarettes, understanding the effect of promotionally marketing vaporization products across social media should be immediately relevant to public health and policy makers.

Materials and Methods

Data Collection

An exhaustive search from the 10% "garden hose" random sample from Twitter's streaming API spanning 2012 through 2014 yielded approximately 850,000 tweets mentioning a keyword related to electronic cigarettes including: e(-)cig, e(-)cigarette, electronic cigarette, etc. All tweets were tokenized by removing punctuation and performing a case insensitive pattern match for keywords. Using time zone meta-data the tweets were converted into their local post time, in order for a more accurate ordinal sentiment analysis. The language, reported by Twitter, and user features were also collected and analyzed. The data from our study was collected via a program developed by Dodds et al, that pings Twitter's streaming API and complies with Twitter's Terms of Service. Our study collected each account's unique twitter user identification number in order to classify them as either Automated or Organic, however our published data has been anonymized by replacing Twitter's UserIDs with placeholder values.

Automation Classification

As reported in [26] there is a high prevalence of automation among e-cigarette related tweets. Many of these messages were promotional in nature, offering discounted or free samples or advertising specific electronic cigarette paraphernalia. A human detection algorithm defined and tested in [28] was implemented to classify accounts as either automated or organic (human in nature). The original classifier was trained on 1000 accounts—752 were identified as humans and 248 as automated accounts. The classifier operates by isolating organic linguistic characteristics and identifies automated accounts by exclusion. All tweets from each individual appearing in our dataset were collected for the classifier. For each individual, the average URL count, average tweet dissimilarity, and word introduction decay rate were calculated for the individuals with at least 25 sampled tweets.

The majority (94%) of commercial e-cigarette tweets collected by [26] contain a hyperlink (URL). The average URL count per tweet has been demonstrated to be a strong feature for detecting robotic accounts [45–47]. Many algorithmically generated tweets contain similar structures with minor character replacements and long chains of common substrings, as opposed to Organic content. The Pairwise Tweet Dissimilarity of tweets t_i, t_j from a particular

individual was estimated by subtracting the length (number of characters) of the longest common subsequence, $|LCS(t_i, t_j)|$ from the length of both tweets, $|t_i| + |t_j|$ and normalizing by the total length of both tweets:

$$\mathrm{D}(t_i, t_j) = \frac{|t_i| + |t_j| - 2 \cdot |LCS(t_i, t_j)|}{|t_i| + |t_j|}.$$

For example, given the two tweets:

(t_1, t_2) = (I love tweeting, I love spamming). Then $|t_1| = 16$, $|t_2| = 15$, $LCS(t_1, t_2) = |\text{I love }| = 7$ (including whitespace) and we calculate the pairwise tweet dissimilarity as:

$$D(t_1, t_2) = \frac{16 + 15 - 2 \cdot 7}{16 + 15} = \frac{17}{31}.$$

The average tweet dissimilarity of the individual was then estimated by finding the arithmetic mean of each individual's calculated pairwise tweet dissimilarity. Since automated and promotional accounts have a structured and limited vocabulary, the unique word introduction decay rate introduced in [48] serves as another useful attribute to detect automated accounts. Using these attributes, the calibrated human detection algorithm, tested in [28], detected over 90% of automated accounts from a mixed 1000 user sample with less than a 5% false positive rate.

The Human Detection Algorithm was calibrated for a range of tweet sample sizes from hand classified Organic accounts. Ordinal samples of collected tweets from each account were binned into partitions of 25 ranging from 25 to a maximum of 500 tweets. Table 1 below lists the number of automated and organic classified accounts per year. Individuals with less than 25 sampled tweets were not classified with the detection algorithm.

To benchmark the accuracy of the detection algorithm on this sample of tweets, a random sample of 500 accounts algorithmically classified as automatons and 500 classified as Organic were hand classified. All collected tweets were hand coded by two evaluators. Tweets were reviewed until the evaluator noticed the presence of automation. If no subset of tweets appeared to be algorithmically generated, the individual was coded as human. Both evaluators had prior experience distinguishing algorithmic versus organic tweets. Refer to the supplementary materials in [28] for a detailed explanation of this annotation process.

In Fig 1, features of each of these 1000 sampled individuals are plotted in three dimensions. Organic features (green) are densely distributed, while the automated features (red points) are more dispersed. The black lines illustrates the organic feature cutoff for the classifier; individuals with features falling outside of the box are classified as automatons. On this sampled set of accounts, the classification algorithm exhibited a 94.6% True Positive rate with a 12.9% False Positive Rate.

Table 1. Electronic Cigarette Tweet Category Counts and Twitter Account Classification.

Year	Tweet Categorization				Account Classification		
	Total	Automated	Organic	Discarded	Automated	Organic	N/A*
2012	107,918	85,546	13,492	8,880	12,715	12,052	19,512
2013	426,306	339,111	76,037	11,158	64,874	59,376	120,142
2014	316,424	234,972	68,698	12,754	54,033	63,289	48,528

*Accounts with less than 25 tweets were not classified.

doi:10.1371/journal.pone.0157304.t001

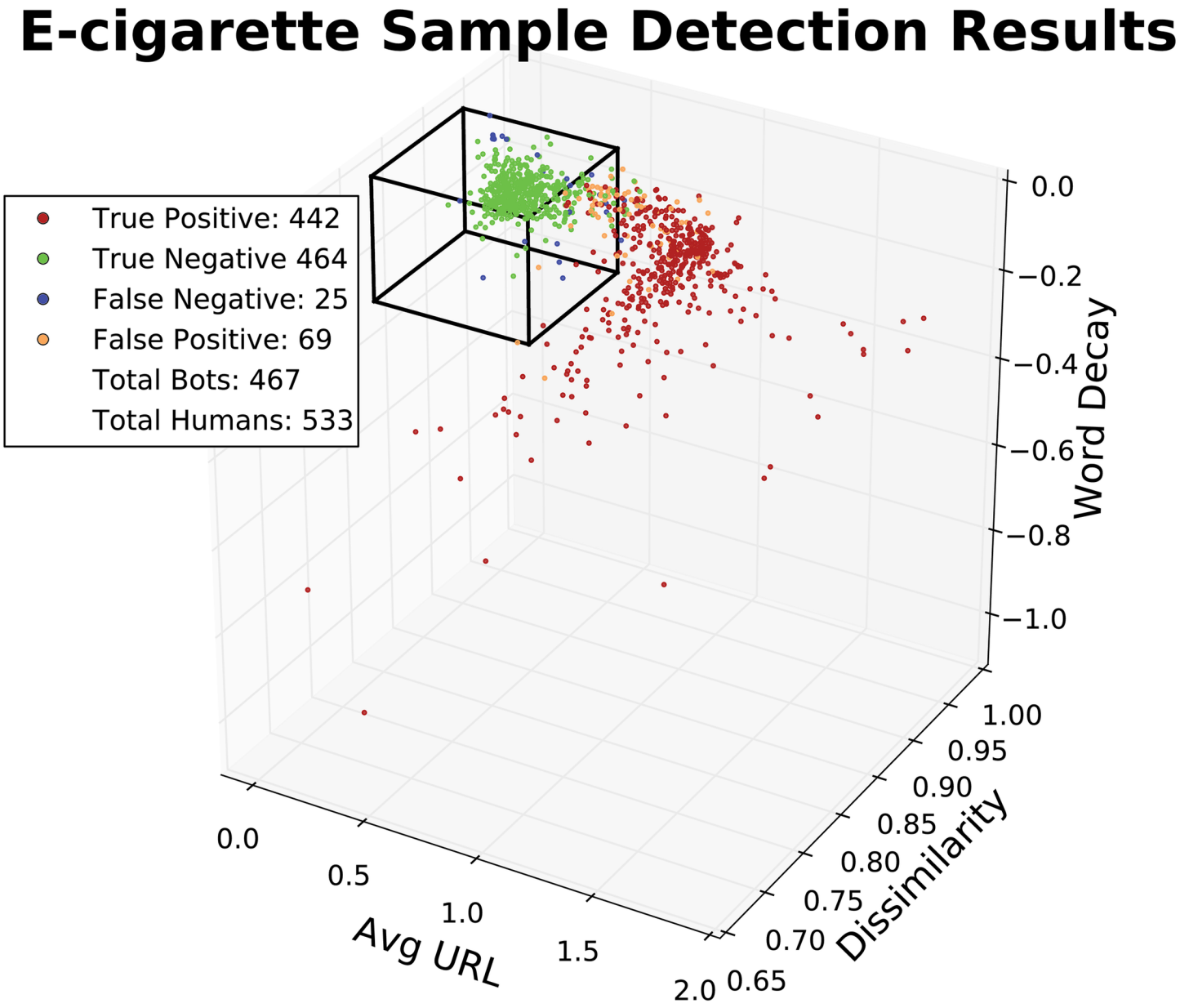

Fig 1. Tweets from a random sample of 500 organic classified and 500 automated classified accounts were hand coded to gauge the accuracy of the detection algorithm. The feature set of each sampled individual is plotted in three dimensions. The traced box indicate the organic feature cutoff. True Positives (red) are correctly identified automatons, True Negatives (green) are correctly identified Humans, False Negatives (blue) are automatons classified as humans and False Positives (orange) are humans classified as automatons.

doi:10.1371/journal.pone.0157304.g001

Categorization by Topics

Tweets with at least 3 advertising jargon references (e.g. coupon, starter kit, free trial) were immediately classified as automated. All posts from users with at least 10 marketing classified tweets were also flagged as automated. As noted in [26], some Organic users could retweet promotional content for rewards (e.g. winning free samples or discounts). All of these tweets were still classified as automated, but the user was not flagged as such. The remaining tweets were classified as either automated or organic by the human detection algorithm. Posts from users

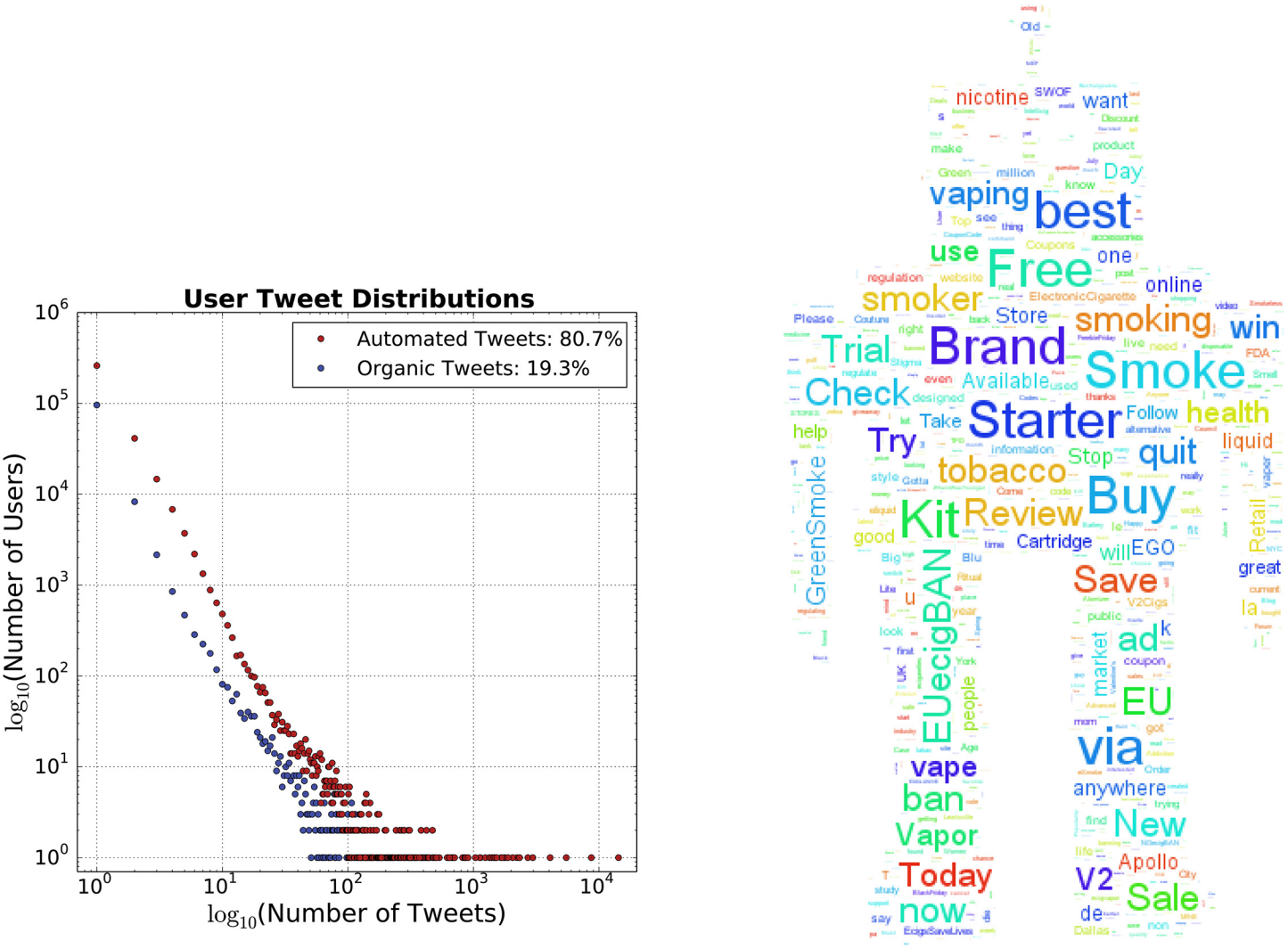

Fig 2. Left: Binned User E-cigarette Keyword Tweet Distribution (2012-2014). Right: 2013 Automated Tweet Rank-Frequency Word Cloud. High frequency stop words ('of', 'the', etc.) are removed from the rank-frequency word distribution.

doi:10.1371/journal.pone.0157304.g002

who had an insufficient number of sampled tweets (<25) to algorithmically classify and who hadn't posted commercial content were classified as Organic. Due to the high prevalence of hyperlinks included in tweets from promotional accounts, Tweets with URLs whose user had insufficient tweets to classify algorithmically were discarded (3.85% total tweets). A final list with each tweet classification coding is created by merging the commercial keyword classification with the results from the Human Detection Algorithm.

Results and Discussion

The number of automated, and in particular promotional, tweets vastly overwhelm (80.7%) the organic (see Fig 2). The identified automated accounts tweet e-cigarette content with much higher frequency than the Organic users. The average number of automated tweets per user was 1.96 with a standard deviation of 35.06 and a max of 14,310. Average organic posts per

Table 2. Automated Tweet Subcategory Counts.

Subcategory	Count	Percentage	Impressions	Relevance*	Year
Commercial	53,471	62.51%	59.74M	88.4%	'12
	283,677	83.65%	195.25M		'13
	149,333	63.55%	951.03M		'14
Cessation	6,392	7.47%	8.59M	90.8%	'12
	6,599	1.95%	25.64M		'13
	8,386	3.57%	42.72M		'14
Discount	26,596	31.09%	27.02M	89.8%	'12
	112,720	33.24%	38.21M		'13
	37,735	16.06%	160.49M		'14
Flavor	1,685	1.97%	2.24M	81%	'12
	2,715	0.80%	4.79M		'13
	6,133	2.61%	17.51M		'14

*Relevant percentage of 500 randomly sampled tweets

doi:10.1371/journal.pone.0157304.t002

user were 1.44 with a standard deviation of 4.01 and max of 356 tweets. A total of 607,446 Automated Tweets provided a URL (92.09%).

Frequency WordClouds (see Fig 2) illustrate the most frequently used words by the Automated category. The size of the text reflects the ranked word frequencies. Marketing key words (Free Trial, Brand, Starter Kit, win, Sale) and brand names (V2, Apollo) are prevalent, illustrating commercial intent. Many automated tweets also refer to the health benefits of switching to electronic cigarettes (#EcigsSaveLives), even though they have not been officially approved as such by the Food and Drug Administration, [49, 50]. See Table 2 for sub categorical counts of the automated tweets.

Tweet Sentiment Analysis

Hedonometrics are performed on the organic subset of electronic cigarette tweets to quantify the change in user sentiments over time. Using the happiness scores of English words from LabMT [34], along with its multi-language companion [35] the average emotional rating of a corpus is calculated by tallying the appearance of words found in the intersection of the word-happiness distribution and a given corpus, in this case subsets of tweets. A weighted arithmetic mean of each word's frequency, f_{word}, and corresponding happiness score, h_{word} for each of the N words in a text yields the average happiness score for the corpus, $\bar{h}_{\text{text}}$:

$$\bar{h}_{\text{text}} = \frac{\sum_{w=1}^{N} f_{\text{w}} \cdot h_{\text{w}}}{\sum_{w=1}^{N} f_{\text{w}}}$$

The average happiness of each word, h_{avg} lies on a 9 point scale: 1 is extremely negative and 9 is extremely positive. Neutral words ($4 \leq h_{avg} \leq 6$), aka 'stop words', were removed from the analysis to bolster the emotional signal of each set of tweets.

Fig 3 shows that automated electronic cigarette tweets are using very positive language to promote their products. The average happiness of the Organic tweets are much more stable,

and are becoming slightly more negative over time. Both distributions have a sudden drop in positivity during December 2013, around a debate regarding new e-cigarette legislation by the European Union. These tweets, labeled #EuEcigBan, are investigated separately in the next section. The words that have the largest contributions to changes in sentiments are investigated with Word-shift graphs.

Word-shift graphs, introduced in [34], illustrate the words causing an emotional shift between two word frequency distributions. A reference period (T_{ref}), creates a basis of the emotional words being used to compare with another period, (T_{comp}). The top 50 words responsible for a happiness shift between the two periods are displayed, along with their contribution to shifting the average happiness of the tweet-set. The arrows (↑, ↓) next to a word indicate an increase or decrease, respectively, of the word's frequency during the comparison period with respect to the reference period. The addition and subtraction signs indicate if the word contributes positively or negatively, respectively, to the average happiness score.

Marketing accounts that delivered personalized advertising by attempting to impersonate organic users were prevalent among these commercial entities. These accounts, along with the traditional marketing robots, were diluting the data with extremely positive sentiments regarding their products. Using hedonometrics, we distinguish the emotionally charged words that influence a shift in computed average word happiness between these types of accounts. The sentiment analysis helps to characterize the thematic differences between Organic and Automated entities.

In Fig 3, below, Word-shift graphs compare the change in Organic sentiments over time, as well as the difference in sentiments between automated and organic tweets. On the left, the

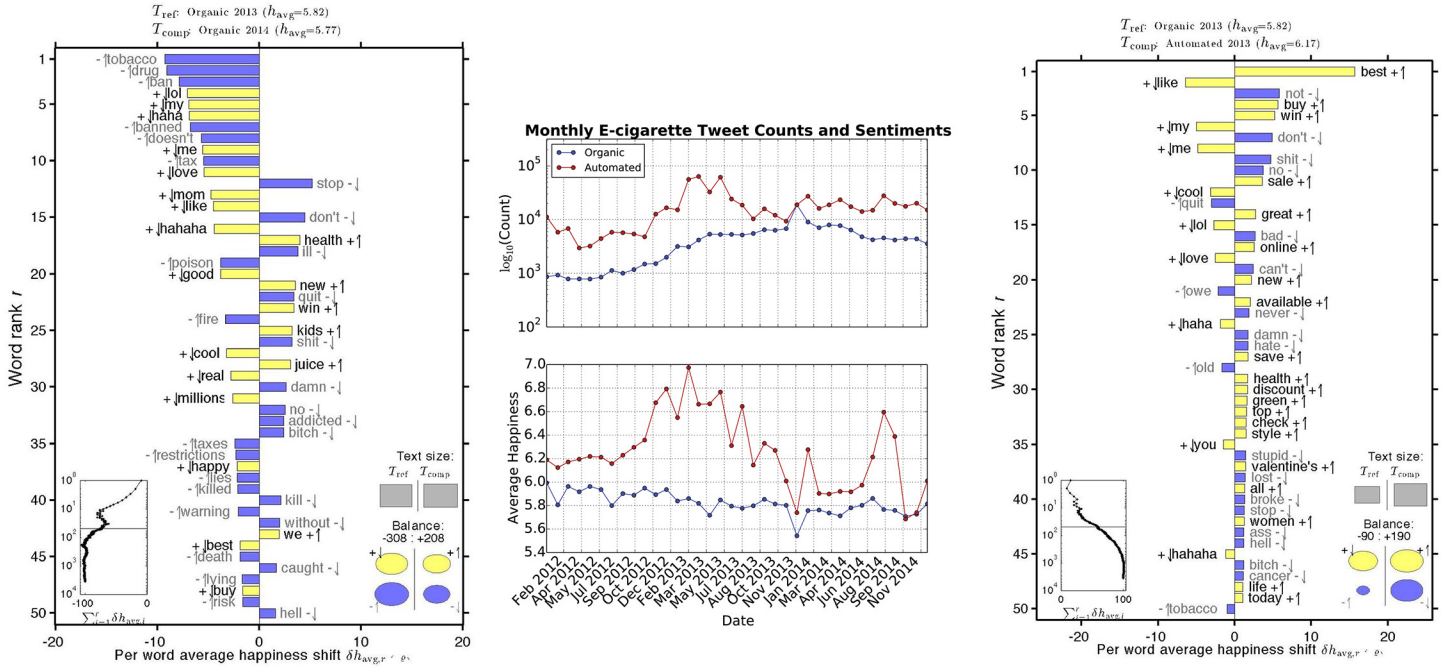

Fig 3. Categorical Tweet Word-shift Graphs: On the left, Organic Tweets from 2013 are the reference distribution to compare sentiments of Organic Tweets made in 2014 where we see a negative shift in the calculated average word happiness. Due to tweets tagged #EUEcig Ban, January 2014 and December 2013 are omitted. The computed average happiness (h_{avg}) decreases from 5.82 to 5.77 due to both an increase in the negative words 'tobacco', 'drug', 'ban', 'poison', and a decrease in the positive words 'love', 'like', 'haha', 'cool' among others. On the right, Organic Tweets from 2013 are the reference distribution to compare Automated Tweets from 2013. The words 'free' and 'trial' are excluded from the graph, since their high frequency and happiness scores distorts the image. With these key words included the the automated tweet h_{avg} increases from 6.17 to 6.59.

doi:10.1371/journal.pone.0157304.g003

2013 Organic Tweet distribution is used as a reference to compare sentiments from 2014 Organic Tweets. December 2013 and January 2014 are removed to dampen the effect of tweets mentioning the #EUecigBan (see S1 Fig). The average happiness score decreases from 5.84 in 2013 to 5.77 in 2014. This decrease in the average happiness score is due to a relative increase in the negative words 'ban', 'tobacco', 'doesn't', 'drug', 'against', 'poison', 'tax'; a relative decrease in the positive words 'haha', 'good', 'cool'. Notably, there is also relatively less usage of the words 'quit', 'addicted', and an increase in 'health', 'kids', 'juice'. On the right, Organic tweets from 2013 is the reference distribution to compare Automated tweets from the same year. Automated tweets are more positive (6.17-6.59 versus 5.84) due to a relative increase in the marketing words 'best', 'win', 'buy', 'sale', 'health', 'discount', etc and a relative decrease in the negative words 'bad', 'hate', 'stupid', 'don't', among others.

Sub-Categorical Tweet Topics

Pertinent topics related to e-cigarette marketing regulation include kid-friendly flavors, smoking cessation claims, and price reduction (including free trials, and starter kits). The commercialized, smoking cessation claims, and discounts were primary topics in the foundational study [51] that identified these campaigns over a 2 month time window. We included the kid-friendly flavors topic in this list due to recent studies reporting their prevalence [10, 24] as well as its current spotlight in political controversy.

Keywords from each of these topics are used to sub-classify the automated tweet set per year, see Table 2 below. Purely commercial tweets were those with any marketing keywords including: 'buy', 'save', 'coupon(s)', 'discount', 'price', 'cost', 'deal', 'promo', 'money', 'sale', 'purchase', 'offer', 'review', 'code', 'win(ner)', 'free', 'starter kit(s)', 'premium'. The URL from each tweet was also analyzed for promotional keywords. Any URL with at least three mentions of the above keywords was enough to classify the tweet as commercial.

When an individual on Twitter 'follows' another account, posts from these users appear on the 'timeline' of the individual. We quantify the social reach of each of these sub-categorical tweets by counting the total number of accounts' 'timelines' who could have been exposed to the advertisement. To approximate this, we sum the number of followers from each individual's tweets. The total number of impressions from the commercial category increases from 195.25 million to 951.03 million between 2013 to 2014, even though the total count has dropped from 283k to 149k. This implies that promotional accounts that are successful in deceiving Twitter's SPAM detector may be gaining many more social links to broadcast their commercial context.

In order to gauge the accuracy of these sub-categorical tweet topics, 500 tweets were randomly sampled from each category and were evaluated separately by two people to determine the relevance of the tweet to its categorization. The evaluators had a high level of concordance (84.8%) and the discrepancies were resolved and merged into a final list. Sampled tweets were highly relevant per category, the percentage for each is given in Table 2 below.

Many automated tweets mentioned using electronic cigarettes as a cessation device, or as a safe alternative. Over 20,000 tweets were classified as cessation related, which potentially appeared on over 76.8 million individual's Twitter feed as impressions. Although electronic cigarettes have not been conclusively authorized as an effective cessation device, [11] has demonstrated the infectiveness of electronic cigarettes to suppress nicotine cravings. It is also notable that these affiliate marketing accounts are advertising electronic cigarettes as a completely safe alternative to analog tobacco use, contrary to recent studies [52–55]. Cessation tweets were tallied using the keywords 'quit', 'quitting', 'stop smoking', 'smoke free', 'safe', 'safer', 'safest'. Many of the purely commercialized tweets mentioned discounts or even free samples. These

Discount tweets were categorized with the keywords 'free trial', 'coupon(s)', 'discount(s)', 'save', 'sale', 'free (e)lectronic (cig)arette'. Tweets advertising flavors were tallied using the keywords 'flavor(s)' and 'flavour(s)' along with an extensive list of popular electronic cigarette flavors compiled from a distributor's website (https://crazyvapors.com/e-liquid-flavor-list/).

A noteworthy class of E-cigarette commercial-bots, are those that are masquerading as Organic users to spam pseudo-positive messages towards potential consumers. These "cyborgs", as defined in [28, 33, 45], spam a positive message regarding a personal experience. One class of these automatons are sending contrived testimonies that e-cigarettes have successfully allowed them to quit smoking cigarettes. These messages are very intentionally structured and tend to swap a few words to appear organic. These messages also target specific individuals as a more personal form of marketing. The general tweet structure from a sample cyborg marketing strategy is given below:

@USER {I,We} {tried,pursued} to {give up, quit} smoking. Discovered BRAND electronic cigarettes and quit in {#} weeks. {Marvelous,Amazing,Terrific}! URL

@USER It's now really easy to {quit,give up} smoking (cigarettes).—these BRAND electronic cigarettes are lots of {fun,pleasure}! URL

@USER electronic cigarettes can assist cigarette smokers to quit, it's well worth the cost URL

@USER It's {incredible,amazing}—the (really) {easy,painless} {answer,method} to quit cigarette smoking through BRAND electronic cigarettes URL

I managed to quit smoking with these e-cigarettes, I highly recommend them: URL @USER

@USER Its {amazing, extraordinary}—I (really) quit smoking after {#} yrs thanks to BRAND electronic cigarettes! URL

Using cyborgs to mimic Organic Users for marketing purposes should be analyzed heavily, to gauge their impact and effectiveness on consumers.

Conclusion

Our study has identified an abundance of automated, and in particular, promotional tweets, and consequent organic sentiments. The collected categorized tweet data from this analysis is available for follow-up analyses into e-cigarette social media marketing campaigns. Future work can perform a deeper analysis on the URL content, similar to [23], posted by promotional accounts to get a better sense of the smoking cessation, flavor mentions, and discount prevalence. We take care not to downplay the well recognized health benefits from smoking cessation including: decreased risk of coronary artery disease, cerebrovascular disease, peripheral vascular disease, decreased incidence of respiratory symptoms such as cough, wheezing, shortness of breath, decreased incidence of chronic obstructive pulmonary disease, and decreased risk of infertility in women of childbearing age [15, 18, 56]. The greatest concern of promotional e-cigarette marketing on Twitter is the risk of enticing younger generations who otherwise may never have commenced consuming nicotine. Due to the unknown but unignorable long-term adverse health effects of electronic cigarettes and the alarmingly increased youth consumption, monitoring and potentially regulating social media commercialization of these products should be immediately relevant to public health and policy agendas.

Supporting Information

S1 Fig. European Union E-cigarette Ban Political Debate (#EUecigBan). (Left) Word shift graph comparing tweets tagged #EUecigBan against 2013 English Organic User Tweets (untagged). (top-right) The automated and Organic tagged tweet distributions are plotted. A histogram displays the counts per language and user class. (bottom-right) Word clouds compare ranked-word frequencies across language and user type. Each categorical time-series exhibits a severe negative trend occurring between December 2013 and January 2014. There is an inverse relationship with the average happiness scores during this time period. This was during the time that the EU was debating strict regulation and a possible ban on specific e-cigarette products [12]. Hashtags (#) allow users to categorize the content of their tweets. During this period, 13,227 sampled tweets were tagged with #EUecigBan. In S1 Fig, a word shift graph (left) visualizes the sentiments from English Organic users using #EUecigBan versus the remaining Organic tweets from 2013. English Tweets tagged #EuEcigBan are the comparison distribution in reference to all other tweets from 2013. Tweets containing #EuEcigBan are on average much more negative (h_{avg} 5.81 versus 5.37) due to an increase in the negative words 'ban', 'stop', 'no', 'not', 'fight', 'against', 'disaster', 'death', 'corruption', 'tobacco', 'kills', etc. The positive words also disfavor the legislation, with the words 'save', 'millions', 'lives', 'support', 'healthy' occurring more frequently. English, French, and German tagged tweets were the most prevalent, and word clouds help visualize themes between language and user class. This shows that Twitter sentiments can be useful in gauging public opinion toward regulation of electronic cigarettes. There is also a heavy automated tweet presence in each language with a similar attitude regarding the legislation, as depicted in the word clouds. Future work should also investigate if and how automated users can impact organic opinion on legislation.
(PDF)

S1 Table. Electronic Cigarette Table of Key Words. List of all key words used in the analysis. Flavors compiled from https://crazyvapors.com/e-liquid-flavor-list/ Keywords other than 'General Twitter Scrape' were applied to categorize automated account tweets.
(PDF)

S2 Table. Twitter IDs. List of all Twitter IDs appearing in the analysis.
(TXT)

Acknowledgments

The authors wish to acknowledge the Vermont Advanced Computing Core which provided High Performance Computing resources contributing to the research results. EMC was supported by the UVM Complex Systems Center, PSD was supported by NSF Career Award # 0846668. PSD and CMD were supported by NSF Big Data Grant #1447634. CJ, AK is supported in part by the National Institute of Health (NIH) Research awards R01DA014028 & R01HD075669, and by the Center of Biomedical Research Excellence Award P20GM103644 from the National Institute of General Medical Sciences.

Author Contributions

Conceived and designed the experiments: EC CJ PSD CD MN. Performed the experiments: EC JRW. Analyzed the data: EC JRW AK. Wrote the paper: EC JRW AK MN PSD CD.

References

1. Cobb NK, Byron MJ, Abrams DB, Shields PG. Novel nicotine delivery systems and public health: the rise of the ?e-cigarette? American journal of public health. 2010; 100(12):2340–2342. doi: 10.2105/AJPH.2010.199281 PMID: 21068414
2. Zhu SH, Gamst A, Lee M, Cummins S, Yin L, Zoref L. The use and perception of electronic cigarettes and snus among the US population. PloS one. 2013; 8(10):e79332. doi: 10.1371/journal.pone.0079332 PMID: 24250756
3. Pearson JL, Richardson A, Niaura RS, Vallone DM, Abrams DB. e-Cigarette awareness, use, and harm perceptions in US adults. American journal of public health. 2012; 102(9):1758–1766. doi: 10.2105/AJPH.2011.300526 PMID: 22813087
4. Vansickel AR, Cobb CO, Weaver MF, Eissenberg TE. A clinical laboratory model for evaluating the acute effects of electronic cigarettes: nicotine delivery profile and cardiovascular and subjective effects. Cancer Epidemiology Biomarkers & Prevention. 2010; 19(8):1945–1953. doi: 10.1158/1055-9965.EPI-10-0288
5. Goniewicz ML, Knysak J, Gawron M, Kosmider L, Sobczak A, Kurek J, et al. Levels of selected carcinogens and toxicants in vapour from electronic cigarettes. Tobacco control. 2014; 23(2):133–139. doi: 10.1136/tobaccocontrol-2012-050859 PMID: 23467656
6. Callahan-Lyon P. Electronic cigarettes: human health effects. Tobacco control. 2014; 23(suppl 2):ii36–ii40. doi: 10.1136/tobaccocontrol-2013-051470 PMID: 24732161
7. Kosmider L, Sobczak A, Fik M, Knysak J, Zaciera M, Kurek J, et al. Carbonyl compounds in electronic cigarette vapors: effects of nicotine solvent and battery output voltage. Nicotine & Tobacco Research. 2014; 16(10):1319–1326. doi: 10.1093/ntr/ntu078
8. Trtchounian A, Talbot P. Electronic nicotine delivery systems: is there a need for regulation? Tobacco control. 2011; 20(1):47–52. doi: 10.1136/tc.2010.037259 PMID: 21139013
9. Kandra KL, Ranney LM, Lee JG, Goldstein AO. Physicians? Attitudes and Use of E-Cigarettes as Cessation Devices, North Carolina, 2013. PloS one. 2014; 9(7):e103462. doi: 10.1371/journal.pone.0103462 PMID: 25072466
10. Grana R, Benowitz N, Glantz SA. E-cigarettes a scientific review. Circulation. 2014; 129(19):1972–1986. doi: 10.1161/CIRCULATIONAHA.114.007667 PMID: 24821826
11. Eissenberg T. Electronic nicotine delivery devices: ineffective nicotine delivery and craving suppression after acute administration. Tobacco control. 2010; 19(1):87–88. doi: 10.1136/tc.2009.033498 PMID: 20154061
12. Keating D. Battle over e-cigarettes dominates negotiations on tobacco legislation; 2013. Available from: http://www.europeanvoice.com/article/battle-over-e-cigarettes-dominates-negotiations-on-tobacco-legislation
13. Palazzolo DL. Electronic cigarettes and vaping: a new challenge in clinical medicine and public health. A literature review. Frontiers in public health. 2013; 1. doi: 10.3389/fpubh.2013.00056 PMID: 24350225
14. Avdalovic MV, Murin S. Electronic cigarettes: no such thing as a free lunch? Or puff. CHEST Journal. 2012; 141(6):1371–1372. doi: 10.1378/chest.12-0205
15. CONTROL CFD, PREVENTION, et al. The Health Consequences of Smoking–50 Years of Progress: A Report of the Surgeon General. Rockville, MD: US DEPARTMENT OF HEALTH AND HUMAN SERVICES. 2014; p. 171.
16. National Institute on Drug Abuse. Research Report Series: Is Nicotine Addictive? Bethesda (MD): National Institutes of Health, National Institute on Drug Abuse. 2012;.
17. American Society of Addiction Medicine. Public Policy Statement on Nicotine Addiction and Tobacco. Chevy Chase (MD): American Society of Addiction Medicine. 2008;.
18. US Department of Health and Human Services and others. How tobacco smoke causes disease: the biology and behavioral basis for smoking-attributable disease: a report of the Surgeon General. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. 2010;2.
19. US Department of Health and Human Services and others. Reducing tobacco use: a report of the Surgeon General. US Department of Health and Human Services; 2000.
20. Centers for Disease Control and Prevention (CDC and others. Quitting smoking among adults–United States, 2001-2010. MMWR Morbidity and mortality weekly report. 2011; 60(44):1513. PMID: 22071589
21. Kim AE, Hopper T, Simpson S, Nonnemaker J, Lieberman AJ, Hansen H, et al. Using Twitter Data to Gain Insights into E-cigarette Marketing and Locations of Use: An Infoveillance Study. Journal of medical Internet research. 2015; 17(11). doi: 10.2196/jmir.4466

22. Yip H, Talbot P. Mining data on usage of electronic nicotine delivery systems (ENDS) from YouTube videos. Tobacco Control. 2013; 22(2):103–106. doi: 10.1136/tobaccocontrol-2011-050226 PMID: 22116832

23. Grana RA, Ling PM. Smoking revolution: a content analysis of electronic cigarette retail websites. American journal of preventive medicine. 2014; 46(4):395–403. doi: 10.1016/j.amepre.2013.12.010 PMID: 24650842

24. Zhu SH, Sun JY, Bonnevie E, Cummins SE, Gamst A, Yin L, et al. Four hundred and sixty brands of e-cigarettes and counting: implications for product regulation. Tobacco control. 2014; 23(suppl 3):iii3–iii9. doi: 10.1136/tobaccocontrol-2014-051670 PMID: 24935895

25. Aphinyanaphongs Y, Lulejian A, Brown DP, Bonneau R, Krebs P. Text Classification for Automatic Detection of E-Cigarette Use and Use for Smoking Cessation from Twitter: A Feasibility Pilot. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. vol. 21. NIH Public Access; 2016. p. 480.

26. Huang J, Kornfield R, Szczypka G, Emery SL. A cross-sectional examination of marketing of electronic cigarettes on Twitter. Tobacco control. 2014; 23(suppl 3):iii26–iii30. doi: 10.1136/tobaccocontrol-2014-051551 PMID: 24935894

27. Harris D. Can evil data scientists fool us all with the world's best spam?; 2013. https://gigaom.com/2013/02/28/can-evil-data-scientists-fool-us-all-with-the-worlds-best-spam/

28. Clark EM, Williams JR, Galbraith RA, Danforth CM, Dodds PS, Jones CA. Sifting Robotic from Organic Text: A Natural Language Approach for Detecting Automation on Twitter. arXiv preprint arXiv:150504342. 2015;.

29. Lee K, Tamilarasan P, Caverlee J. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media. In: ICWSM; 2013.

30. Ranganath S, Hu X, Tang J, Liu H. Understanding and Identifying Advocates for Political Campaigns on Social Media;.

31. Wang G, Wilson C, Zhao X, Zhu Y, Mohanlal M, Zheng H, et al. Serf and turf: crowdturfing for fun and profit. In: Proceedings of the 21st international conference on World Wide Web. ACM; 2012. p. 679–688.

32. Chu Z, Gianvecchio S, Wang H, Jajodia S. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? Dependable and Secure Computing, IEEE Transactions on. 2012; 9(6):811–824. doi: 10.1109/TDSC.2012.75

33. Li H, Mukherjee A, Liu B, Kornfield R, Emery S. Detecting campaign promoters on twitter using markov random fields. In: Data Mining (ICDM), 2014 IEEE International Conference on. IEEE; 2014. p. 290–299.

34. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PloS one. 2011; 6(12):e26752. doi: 10.1371/journal.pone.0026752 PMID: 22163266

35. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, Williams JR, et al. Human language reveals a universal positivity bias. Proceedings of the National Academy of Sciences. 2015; 112(8):2389–2394. doi: 10.1073/pnas.1411678112

36. Dutra LM, Glantz SA. Electronic cigarettes and conventional cigarette use among US adolescents: a cross-sectional study. JAMA pediatrics. 2014; 168(7):610–617. doi: 10.1001/jamapediatrics.2013.5488 PMID: 24604023

37. Cho JH, Shin E, Moon SS. Electronic-cigarette smoking experience among adolescents. Journal of Adolescent Health. 2011; 49(5):542–546. doi: 10.1016/j.jadohealth.2011.08.001 PMID: 22018571

38. Pepper JK, Reiter PL, McRee AL, Cameron LD, Gilkey MB, Brewer NT. Adolescent males' awareness of and willingness to try electronic cigarettes. Journal of Adolescent Health. 2013; 52(2):144–150. doi: 10.1016/j.jadohealth.2012.09.014 PMID: 23332477

39. Goniewicz ML, Zielinska-Danch W. Electronic cigarette use among teenagers and young adults in Poland. Pediatrics. 2012; 130(4):e879–e885. doi: 10.1542/peds.2011-3448 PMID: 22987874

40. Wills TA, Knight R, Williams RJ, Pagano I, Sargent JD. Risk factors for exclusive e-cigarette use and dual e-cigarette use and tobacco use in adolescents. Pediatrics. 2015; 135(1):e43–e51. doi: 10.1542/peds.2014-0760 PMID: 25511118

41. Johnston LD, Bachman JG, et al. Monitoring the Future: National results on adolescent drug use: Overview of key findings. 2014;.

42. Bunnell RE, Agaku IT, Arrazola R, Apelberg BJ, Caraballo RS, Corey CG, et al. Intentions to smoke cigarettes among never-smoking US middle and high school electronic cigarette users, National Youth Tobacco Survey, 2011-2013. Nicotine & Tobacco Research. 2014; p. ntu166.

43. Duke JC, Lee YO, Kim AE, Watson KA, Arnold KY, Nonnemaker JM, et al. Exposure to electronic cigarette television advertisements among youth and young adults. Pediatrics. 2014; 134(1):e29–e36. doi: 10.1542/peds.2014-0269 PMID: 24918224

44. Brenner J, Smith A. 72% of online adults are social networking site users. Washington, DC: Pew Internet & American Life Project. 2013;.

45. Chu Z, Gianvecchio S, Wang H, Jajodia S. Who is Tweeting on Twitter: Human, Bot, or Cyborg? In: Proceedings of the 26th Annual Computer Security Applications Conference. ACSAC'10. New York NY, USA: ACM; 2010. p. 21–30. Available from: http://doi.acm.org/10.1145/1920261.1920265

46. Lee K, Caverlee J, Webb S. The Social Honeypot Project: Protecting Online Communities from Spammers. In: Proceedings of the 19th International Conference on World Wide Web. WWW'10. New York NY, USA: ACM; 2010. p. 1139–1140. Available from: http://doi.acm.org/10.1145/1772690.1772843

47. Lee K, Caverlee J, Webb S. Uncovering Social Spammers: Social Honeypots + Machine Learning. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'10. New York, NY, USA: ACM; 2010. p. 435–442. Available from: http://doi.acm.org/10.1145/1835449.1835522

48. Williams JR, Bagrow JP, Danforth CM, Dodds PS. Text mixing shapes the anatomy of rank-frequency distributions: A modern Zipfian mechanics for natural language. CoRR. 2014;

49. Zezima K. Cigarettes Without Smoke, or Regulation; 2009. Available from: http://www.nytimes.com/2009/06/02/us/02cigarette.html?_r=2&

50. Ashley D, Burns D, Djordjevic M, Dybing E, Gray N, Hammond S, et al. The scientific basis of tobacco product regulation. World Health Organization technical report series. 2007;(951:):1–277.

51. Huang J, Kornfield R, Szczypka G, Emery SL. A cross-sectional examination of marketing of electronic cigarettes on Twitter. Tobacco control. 2014; 23(suppl 3):iii26–iii30. doi: 10.1136/tobaccocontrol-2014-051551 PMID: 24935894

52. Sussan TE, Gajghate S, Thimmulappa RK, Ma J, Kim JH, Sudini K, et al. Exposure to Electronic Cigarettes Impairs Pulmonary Anti-Bacterial and Anti-Viral Defenses in a Mouse Model. PloS one. 2015; 10 (2):e0116861. doi: 10.1371/journal.pone.0116861 PMID: 25651083

53. Lerner CA, Sundar IK, Yao H, Gerloff J, Ossip DJ, McIntosh S, et al. Vapors Produced by Electronic Cigarettes and E-Juices with Flavorings Induce Toxicity, Oxidative Stress, and Inflammatory Response in Lung Epithelial Cells and in Mouse Lung. PloS one. 2015; 10(2):e0116732. doi: 10.1371/journal.pone.0116732 PMID: 25658421

54. Cameron JM, Howell DN, White JR, Andrenyak DM, Layton ME, Roll JM. Variable and potentially fatal amounts of nicotine in e-cigarette nicotine solutions. Tobacco control. 2014; 23(1):77–78. doi: 10.1136/tobaccocontrol-2012-050604 PMID: 23407110

55. Williams M, Villarreal A, Bozhilov K, Lin S, Talbot P. Metal and silicate particles including nanoparticles are present in electronic cigarette cartomizer fluid and aerosol. PloS one. 2013; 8(3):e57987. doi: 10.1371/journal.pone.0057987 PMID: 23526962

56. US Department of Health and Human Services and others. The health consequences of smoking: a report of the Surgeon General. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. 2004;62.