

Social Media Meets Population Health: A Sentiment and Demographic Analysis of Tobacco and E-Cigarette Use Across the Twittersphere

Eric M. Clark^{1,2,3,4,5,10}, Chris Jones^{5,8,9,10}, Diann Gaalema^{6,7,8,10}, Ryan Redner^{6,8,10}, Thomas J. White^{6,8,10}, Allison Kurtz^{6,8,10}, Andrew Schneider^{8,10}, Peter Dodds^{1,2,3,4,10}, Mitchell Norotsky^{5,10}, and Chris Danforth^{1,2,3,4,10}

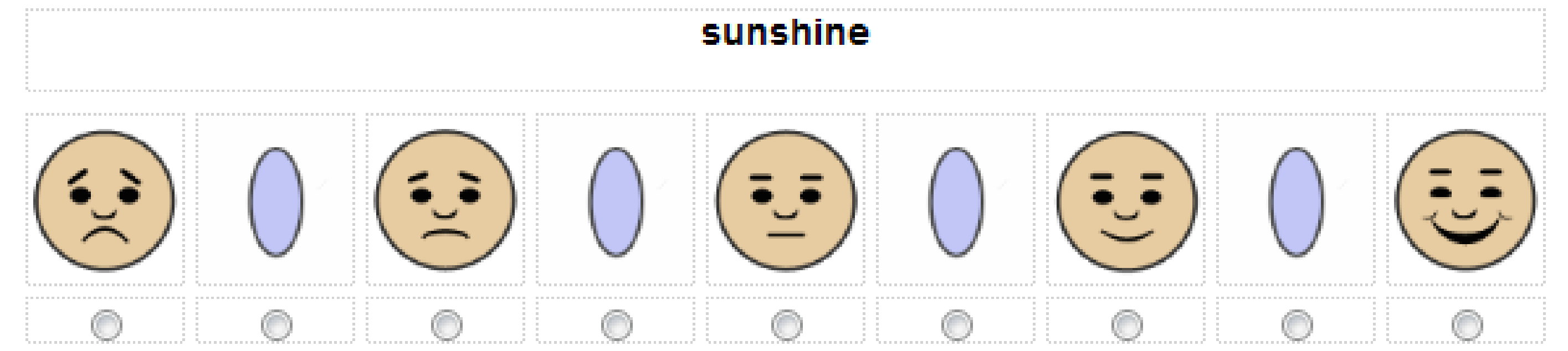
Computational Story Lab¹, Department of Mathematics & Statistics², Vermont Complex Systems Center³, Vermont Advanced Computing Core⁴, Department of Surgery⁵, Department of Psychiatry⁶, Department of Psychology⁷, Vermont Center on Behavior and Health⁸, Global Health Economics Unit of the Vermont Center for Clinical and Translational Science⁹, & The University of Vermont¹⁰



Abstract

BACKGROUND: Twitter, a popular social media outlet, has become a useful tool for the study of social behavior through user interactions called tweets. The geo-location and timestamp of tweets along with message content provide invaluable social and demographic information for an applied comparison of social behaviors across the world. **OBJECTIVES:** To determine the density and sentiments surrounding tobacco and e-cigarette tweets and link prevalence of word choices to tobacco and e-cigarette use at various localities. **METHODS:** All tweets with geo-spatial coordinates are salvaged from the twitter-feed, representing approximately 1% of the entire twitter-sphere, along with tweets mentioning Tobacco or Electronic Cigarettes from a 10% sample of twitter spanning 2012-2013. Pattern matching by tobacco and e-cigarette related keywords yield approximately 20,000 affiliated geo-tweets per month from North America. The emotionally charged words that contribute to the positivity of various subsets of regional tweets are quantitatively measured using hedonometrics. We examined the density of these behavioral tweet indicators by region and tested the relationship between tweeted smoking sentiments and time-space-type coordinates over a 6-month span using geo-data, as well as the change in sentiments over a two year span. **RESULTS:** For states with a high twitter prevalence, the ratio of tobacco tweets per state correlate to state smoking rate estimates. Over a 6-month span, the density of tobacco related tweets correlate to the CDC estimates of state smoking rates. Tobacco related tweets were collected over a two year span, and converted to each user's local time using time-zone meta data in order to map the daily cycle of tobacco use in terms of frequency and happiness. Tweets mentioning electronic cigarette were predominately commercialized ($\approx 80\%$). These tweets were categorized in order to investigate the relationship between commercialized tweets and their effect on organic users. Our results illustrate significant variation in smoking sentiments by state and at varying regional scopes as well as over time. **CONCLUSIONS:** It is anticipated that real-time analysis of health behavior using twitter feeds will allow for more targeted forms of health policy planning and intervention. Regional density of tobacco and e-cigarette related tweets yield insight to the prevalence of tobacco usage per capita. Sentiment analysis across the twitter-sphere can help illuminate hazardous health behavioral trends and allows the possibility to help mediate poor health habits and potentially a number of health interventions in order to improve health consciousness and target medical interventions towards maximizing population health.

Hedonometrics: Measuring the Happiness of a Text

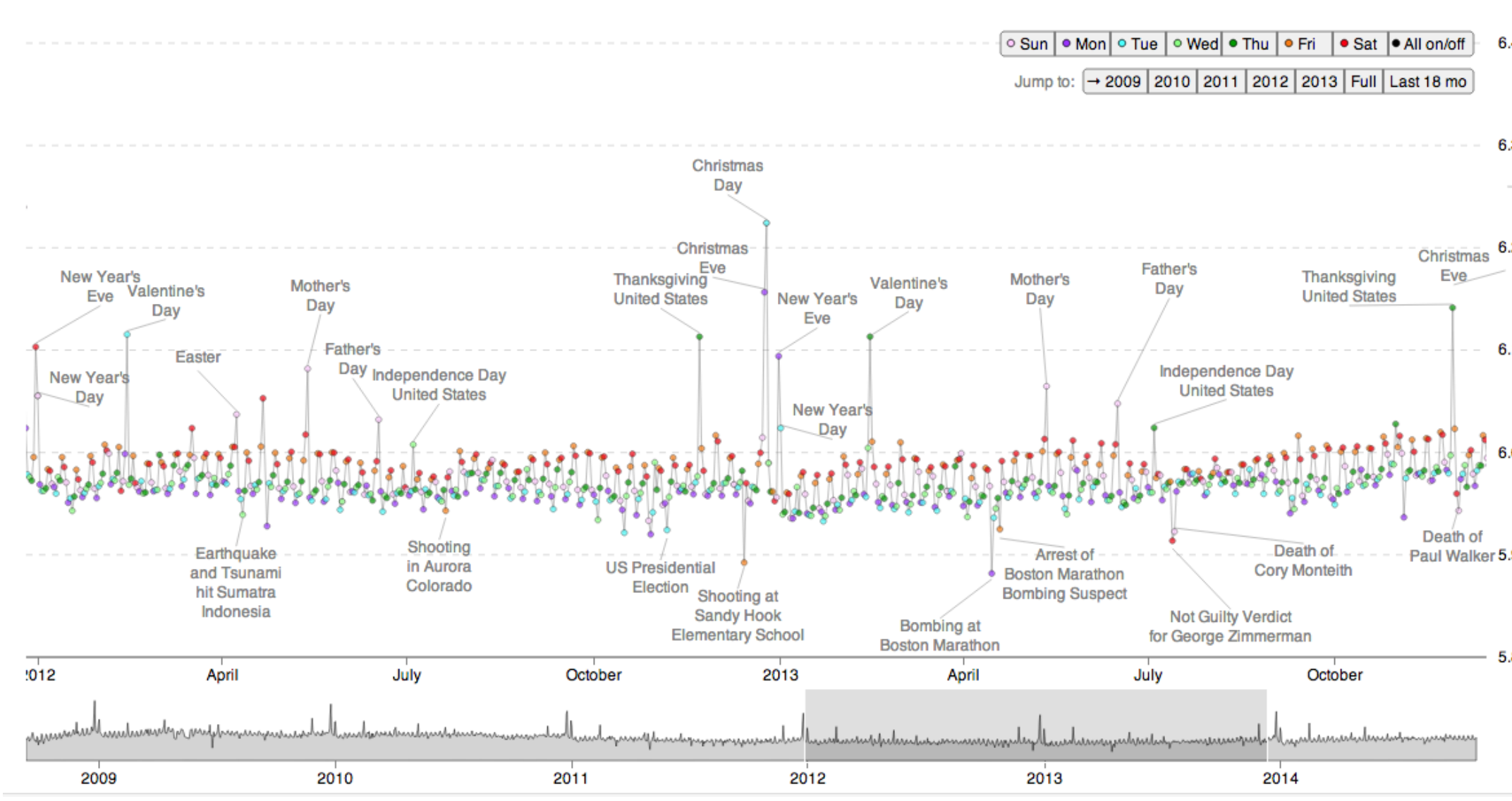


LabMT is a happiness distribution of the most frequently occurring 10,000 English words that were compiled through frequency distributions from literature, (Google Books), websites (Google Web Crawl), and Twitter. Surveys were created mimicking the self affective mannequin method, a sample of which is given above. Fifty participants were recruited using the online survey tool, Amazon Mechanical Turk, to identify the face that best matched the emotional response elicited by each word, which were then converted to a 9 point scale. On the numeric scale, 1 corresponded to the face with the largest frown and 9 to the face with the largest smile. The average happiness score, h_{avg} , for each word was then calculated via the arithmetic mean of 50 user reported ratings per word. Using the average happiness scores of each word, the average positivity of a subset of tweets can be quantified and used to compare different tweet distributions. To increase the emotional signal, neutral words ($4 \leq h_{avg} \leq 6$) are removed from the analysis. The standard approach to perform a hedonometric analysis on twitter is to create a happiness time-series. Outliers on the time series correspond to time-periods containing an overabundance of emotionally charged words. These outliers can then be investigated with word-shift graphs to help illuminate what is driving the emotional shift.

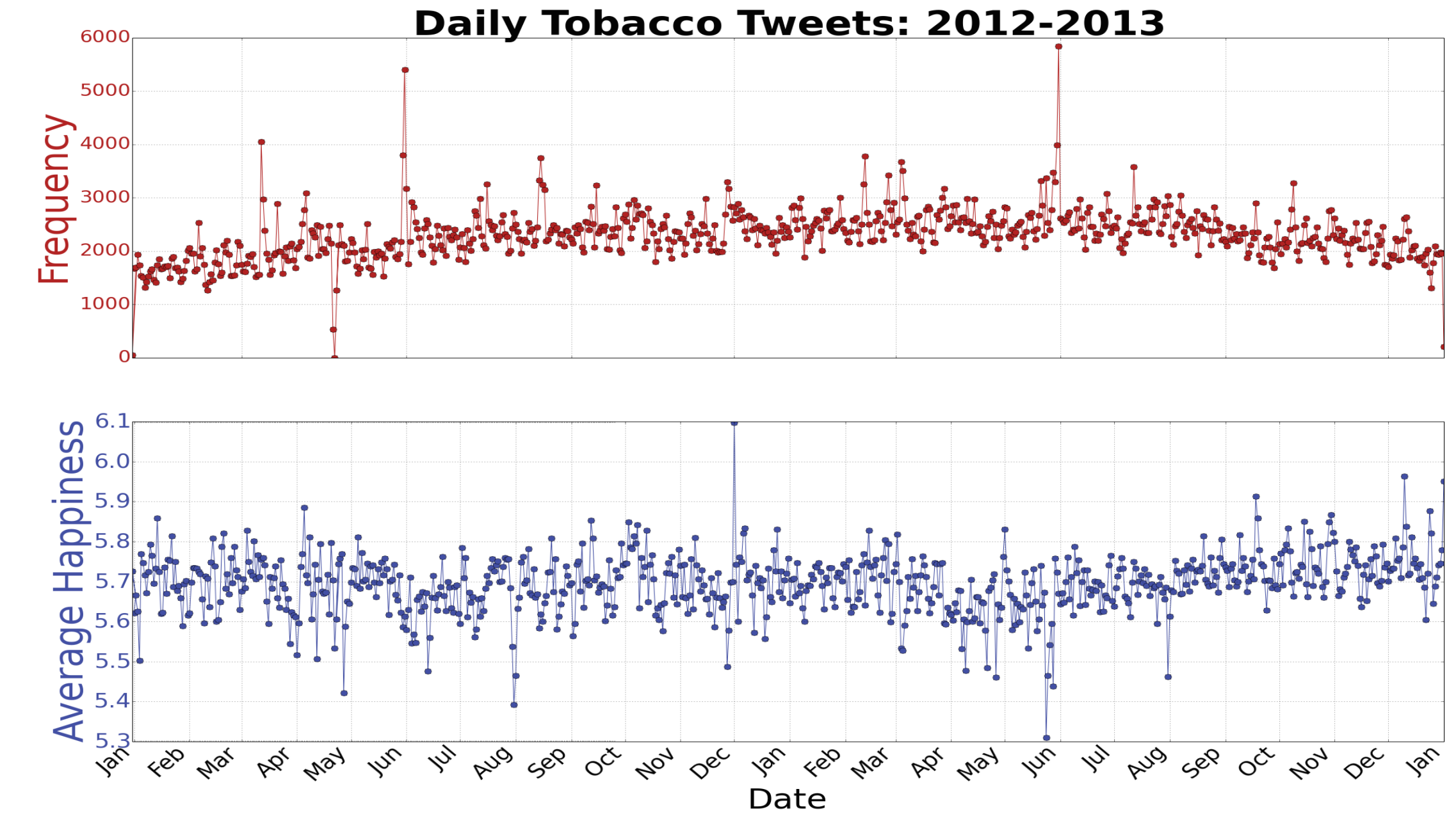
2012-2013 Twitter Happiness Time-series

Using the happiness scores from LabMT, the average emotional rating of a corpus is calculated by tallying the appearance of words found in the intersection of the wordlist 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} :

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

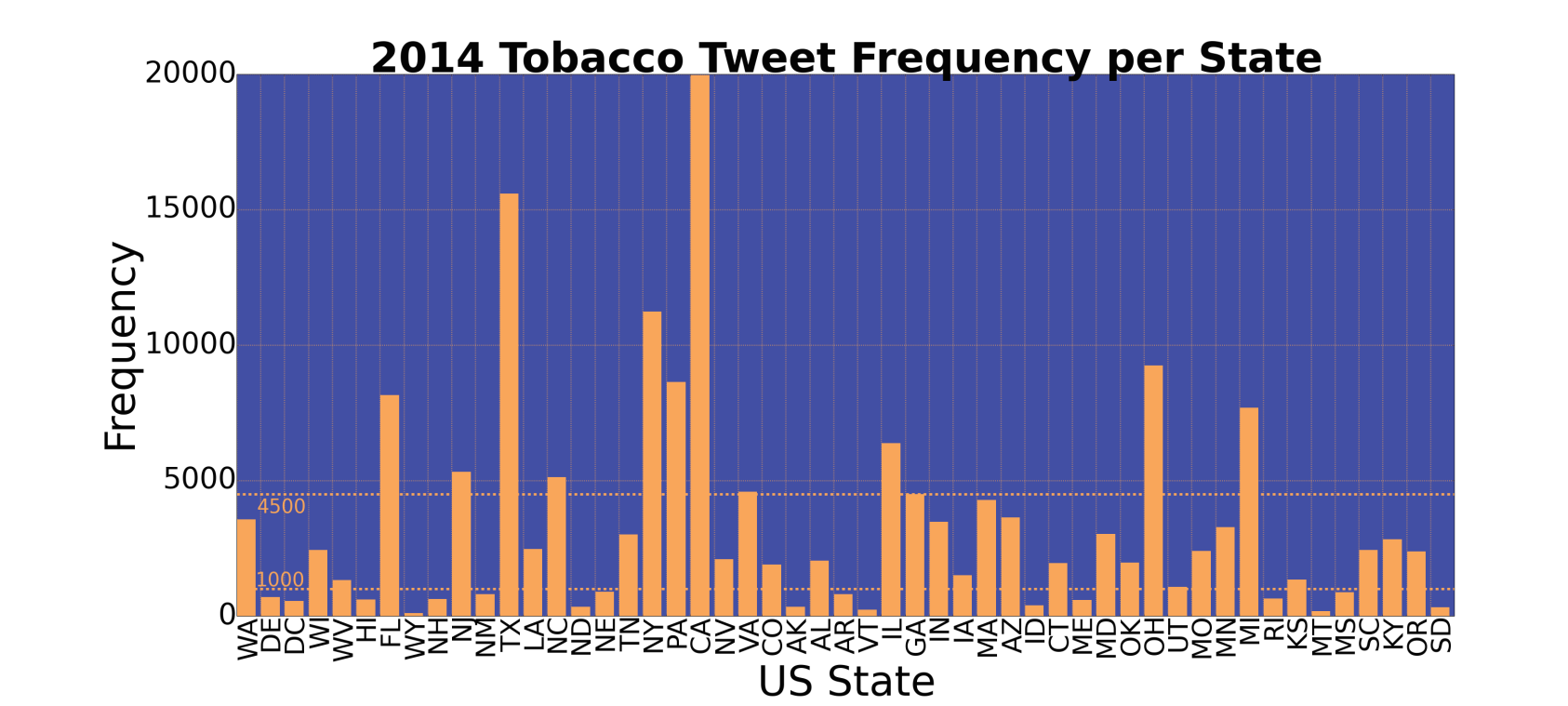
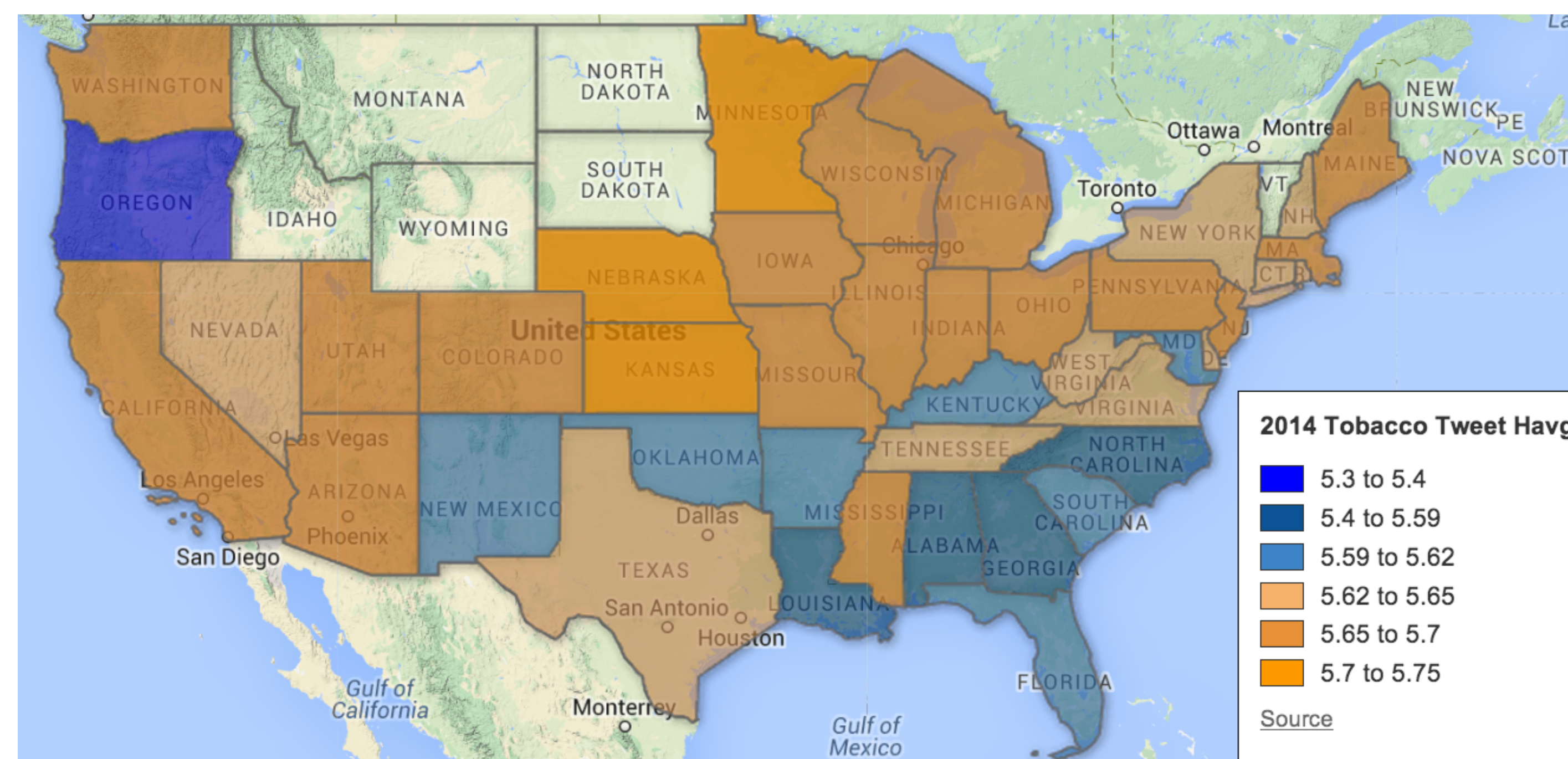
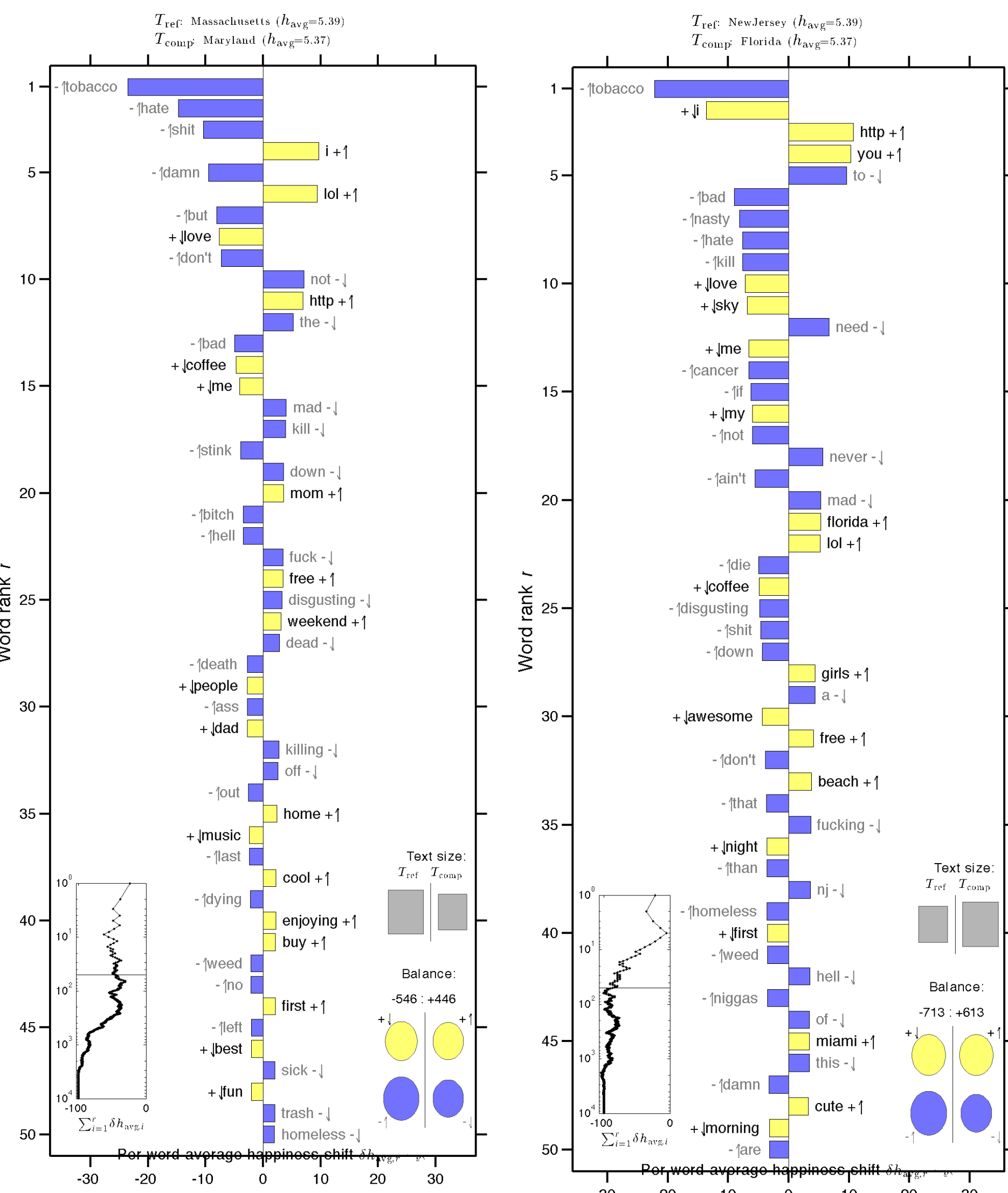


(left) The happiness average from all tweets spanning 2012 through 2013 are displayed, courtesy of hedonometer.org. Although there is a minor fluctuation of happiness per day, tweets average out around 6.0 on the happiness scale, which can be considered slightly positive. Outliers on this happiness curve can be attributed to worldly events (e.g. Christmas, Thanksgiving, etc.). (right) Tweets from this time-period with tobacco related keywords were collected and plotted by their daily frequencies and happiness averages. Tobacco related tweets are slightly more negative when compared to the entire twitter sphere.



US Geo-tagged Tobacco Tweets: March-September 2014

Approximately 1% of all tweets report the geo-location to within ten meters of accuracy of the user. This Geo-tagged data-set allows for regional comparisons of tobacco cigarette keyword mentions across the United States. A happiness heat map of the 50 States (center) helps visualize the happiness of tobacco related tweets at the regional level. States with less than 500 tweets are removed from the analysis.



Tobacco Tweet Density vs 2012 Smoking Rate

Type	Rho	P value	# States	Min Tweets
Adult	0.580	0.048	12	4500
Youth Insig.	NA	NA	NA	NA

2012 Smoking Rate vs. Tweet Happiness

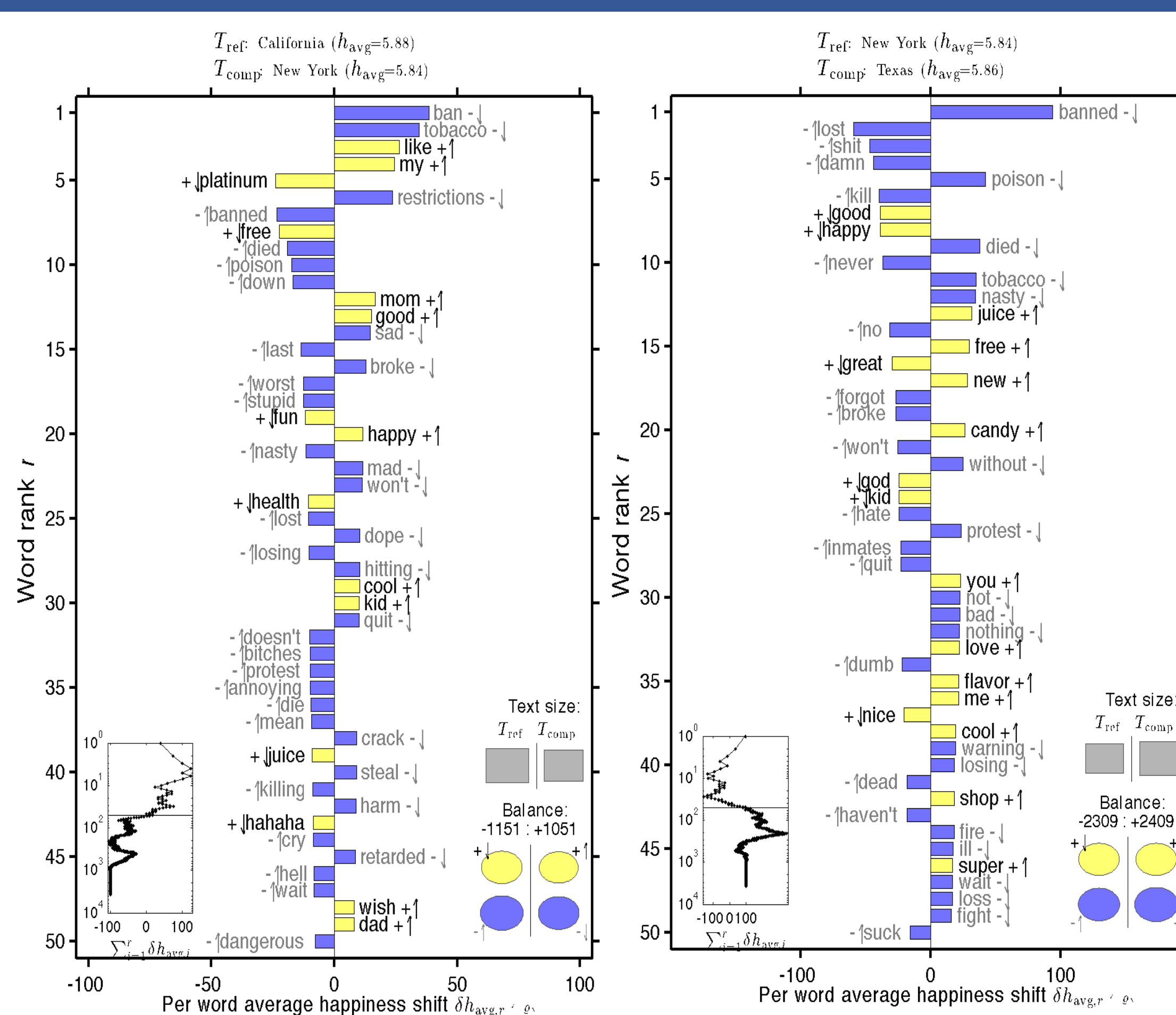
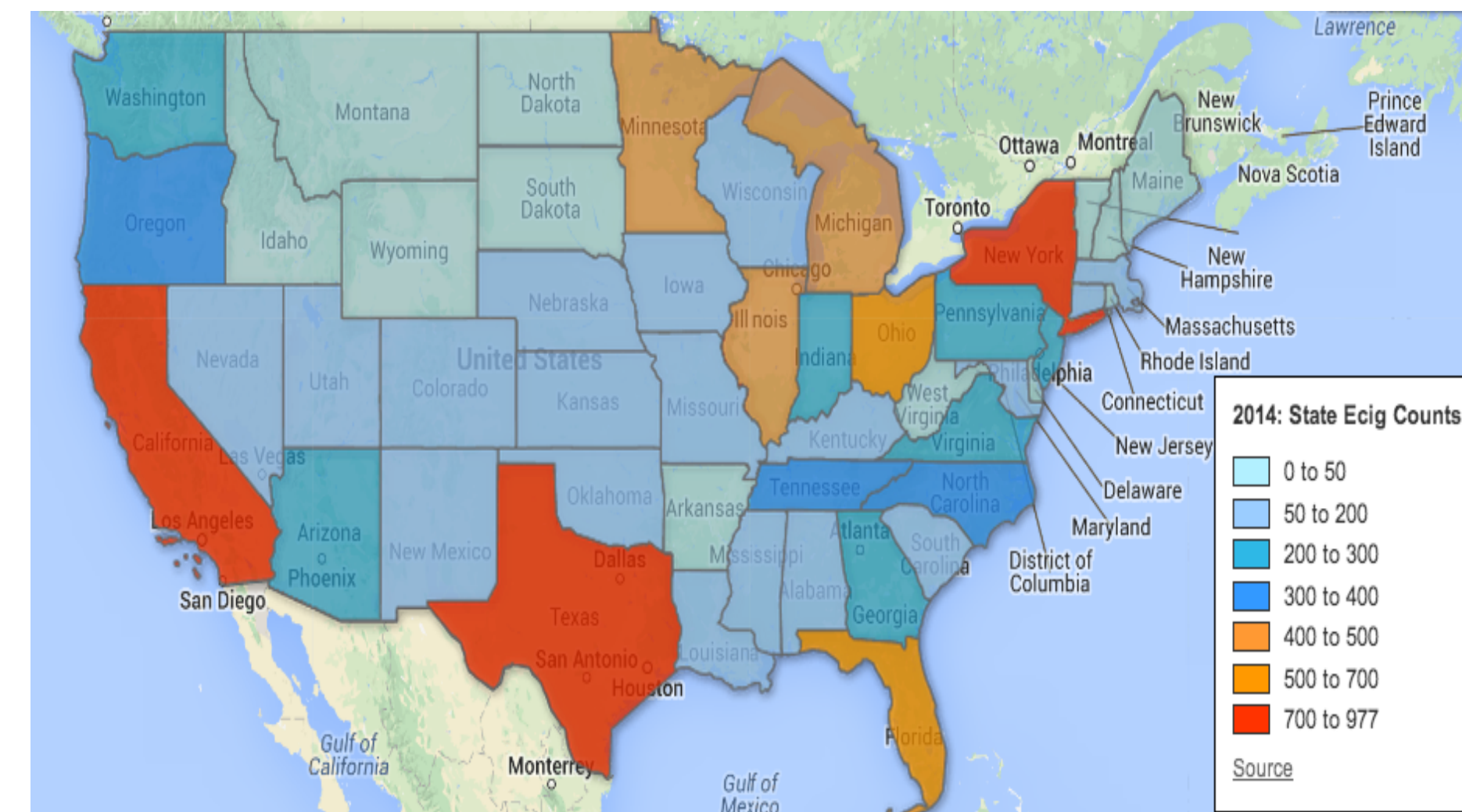
Type	Rho	P value	# States	Min Tweets
Adult	-0.348	0.044	34	1000
Youth	-0.465	0.015	27	1000

(left) Word-shift graphs illustrate the happiness differences of two separate 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 (\uparrow, \downarrow) 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. Here we can identify the words contributing to the change in happiness between each state. The leftmost shift compares the tweets from Maryland relative to Massachusetts where there is more 'hate', 'stink', 'disgusting', which caused a slightly negative shift in the tweet average happiness. On the right Florida is compared to New Jersey where there is more 'bad', 'nasty', 'hate', 'kill', which causes a negative shift in happiness.

(Above) Tables show the correlations between the happiness of tweets, relative tobacco tweet density, and smoking rate per state (provided by the CDC 2012 estimates). There is a high Spearman Correlation between the adult smoking rate and density of tobacco related tweets in states with at least 4500 tweets ($N = 12$). Using states with a large number of tweets this opens up the possibility to infer rates of tobacco use through the relative twitter activity. There is a high negative Spearman Correlation (second table) between both the adult and youth smoking rates versus the tobacco tweet average happiness for states with at least 1000 tweets ($N = 27$).

US Geo-tagged Electronic Cigarette Tweets: March-August 2014

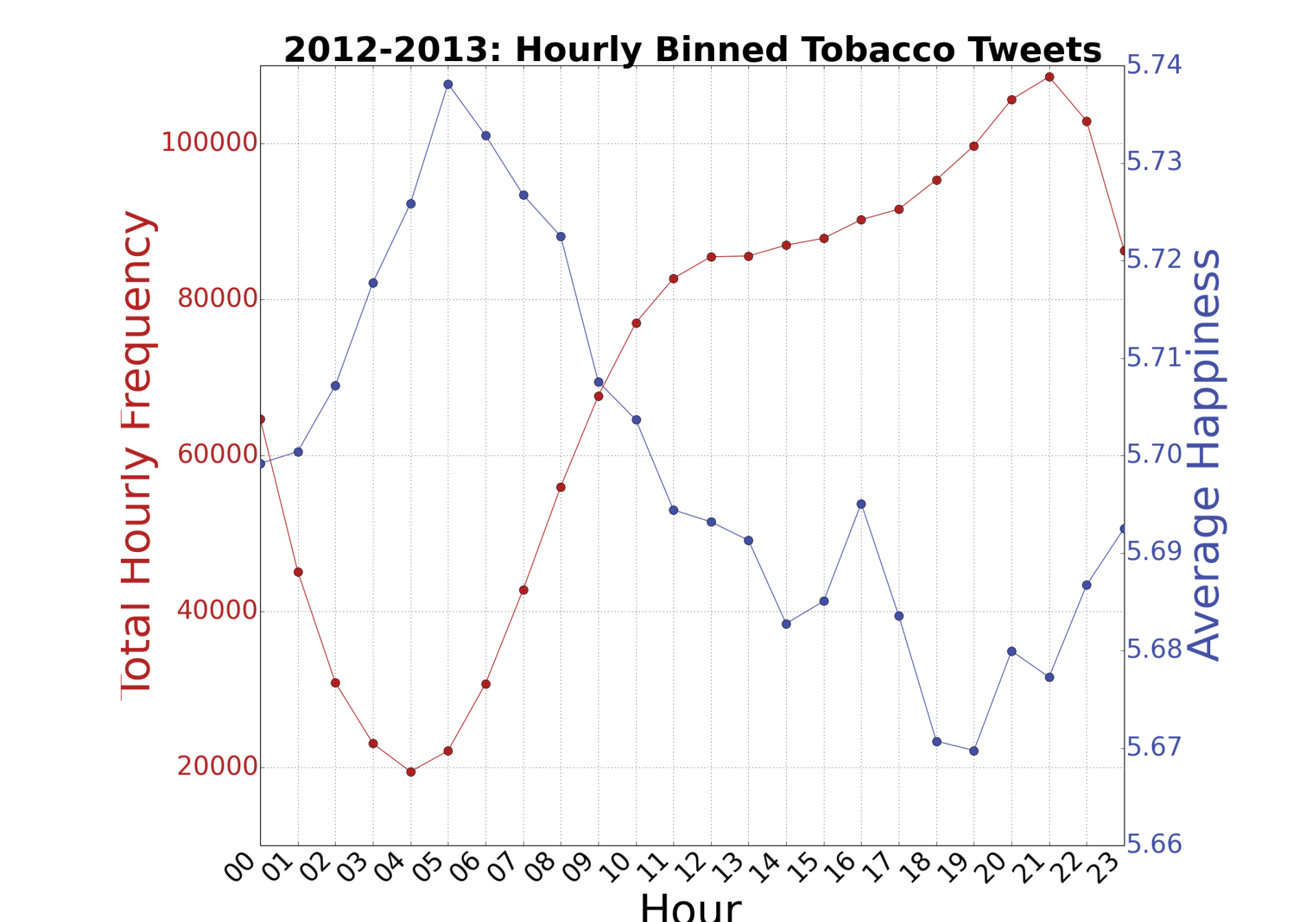
All tweets that mention electronic cigarette keywords between March and August 2014 were collected and binned by U.S. state. Below is a heat map of the counts of e-cig tweets per each state. A substantial number of tweets is required to perform a meaningful hedonometric analysis of a region. Since there are substantially less tweets per state mentioning e-cigs in comparison to tobacco, only a few states can be compared with hedonometrics.



The word shift graphs (right) compare the states with the most e-cig related tweets. The leftmost shift compares the tweets from New York relative to California. In New York there are less occurrences of the negative words 'ban', 'restrictions', a higher occurrence of the negative words 'poison', 'died', 'worst', 'stupid'. On the right, tweets from Texas are compared to tweets from New York. In Texas there are less occurrences of 'banned', 'poison', 'protest', and 'nasty', and more occurrences of 'juice', 'flavor', 'candy', and 'quit'.

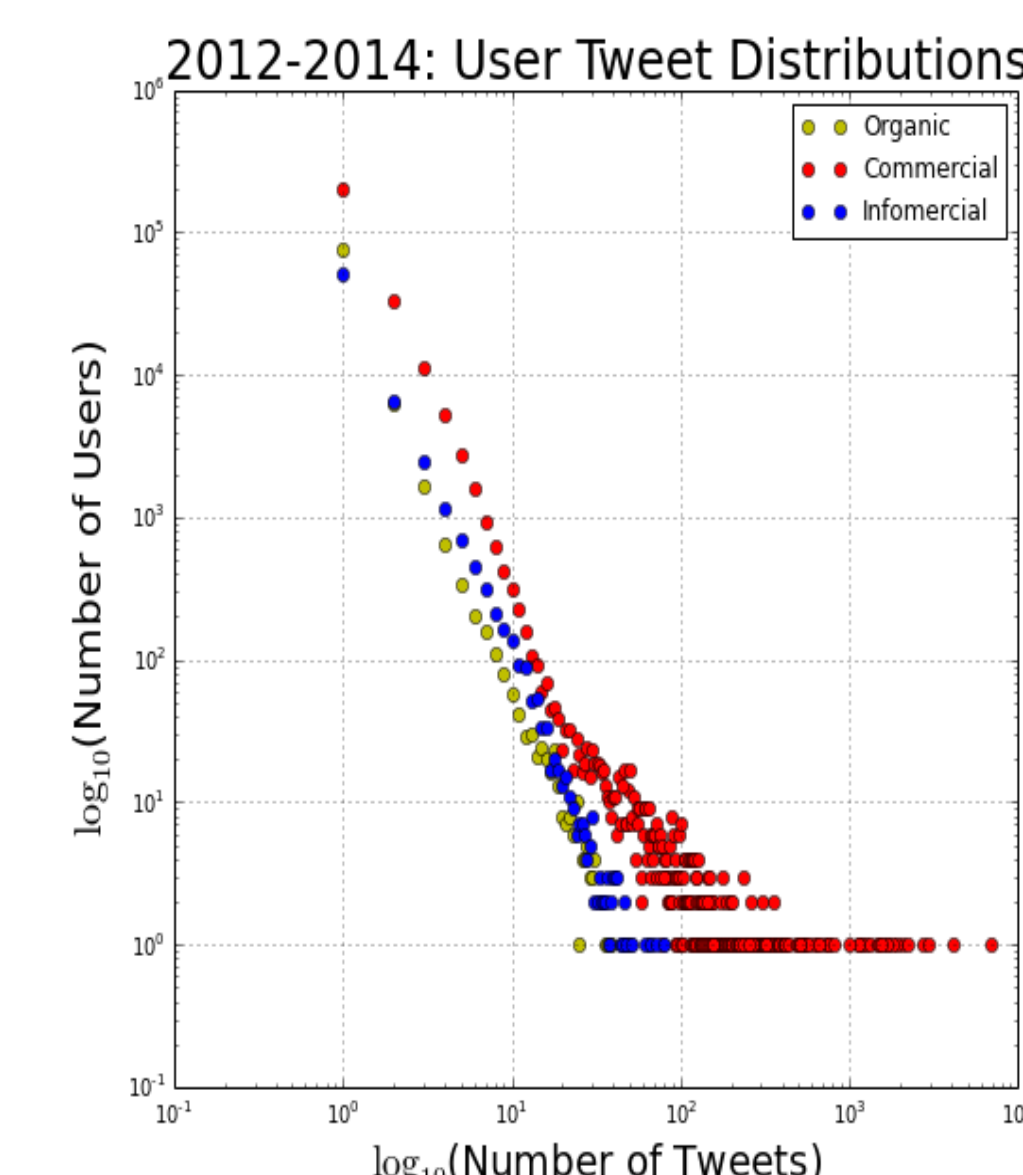
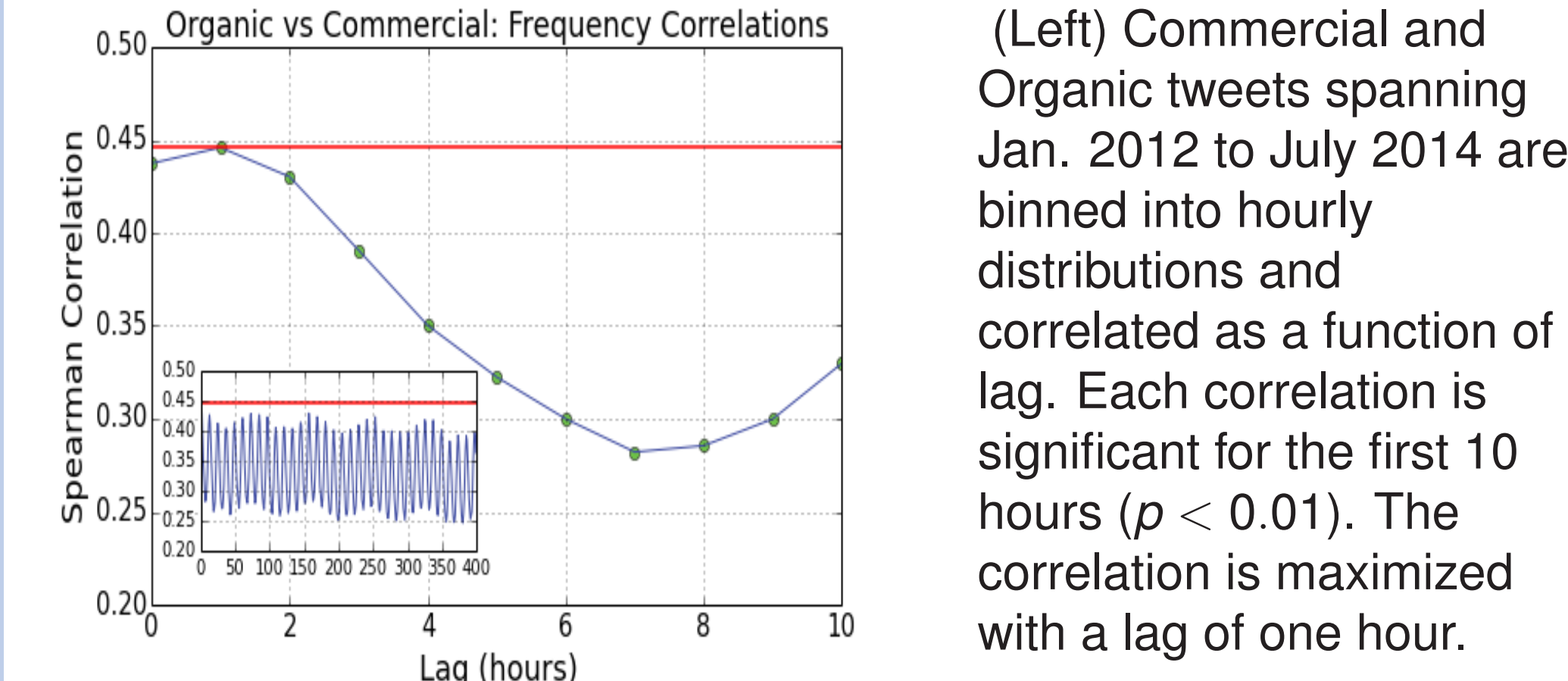
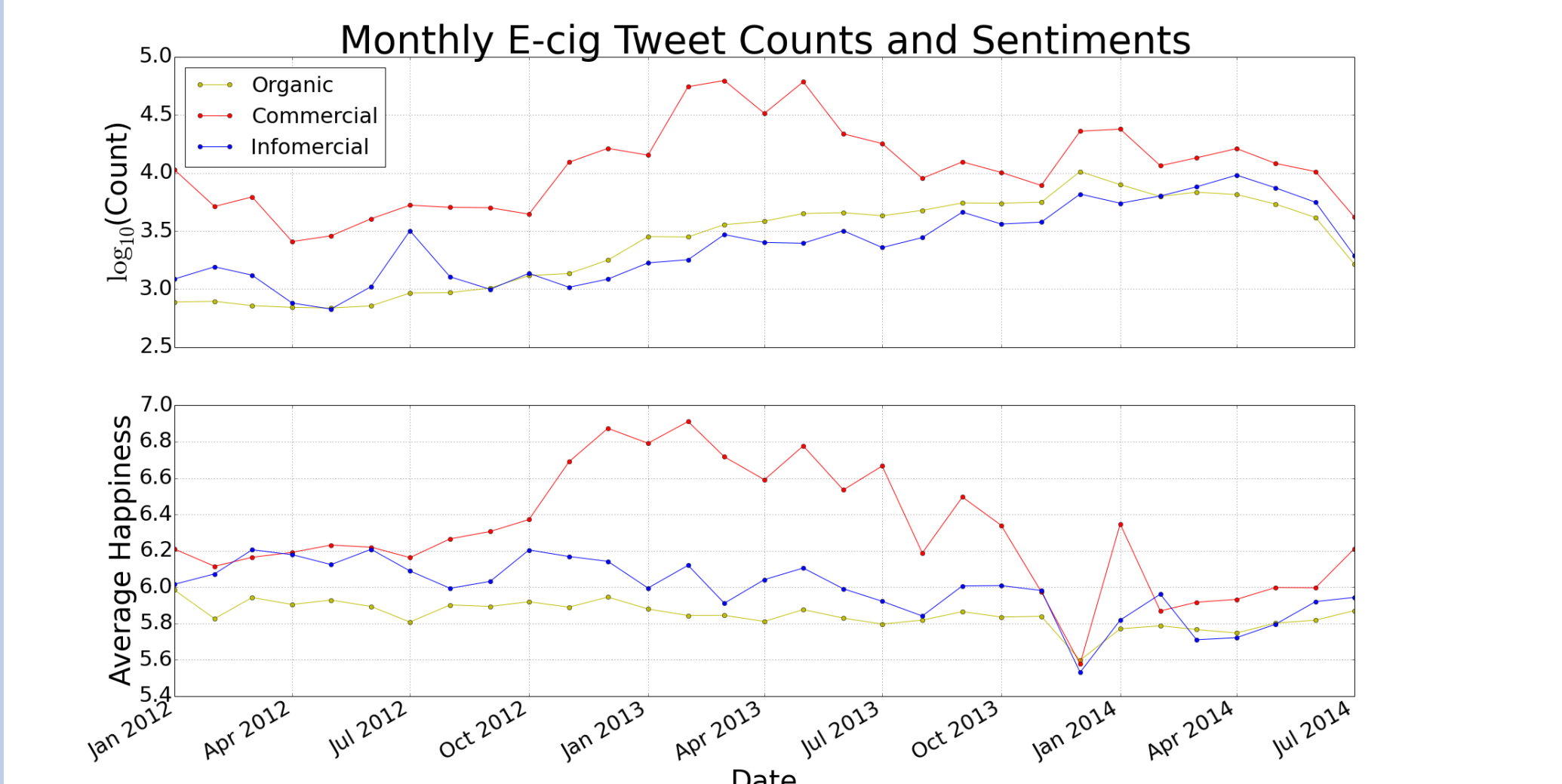
Tobacco Tweets: Daily Cycle

Tobacco related tweets from 2012-2013 were also collected from the Twitter firehose, a 10% random sample of the twittersphere. Using the time-zone metadata provided by twitter, each of these tobacco related tweets were converted to their local time and then binned by the hour of the day (0-23). The frequency and happiness distributions of these hourly binned tweets demonstrate a pronounced inverse relationship and are plotted on a twin axis below. This visualization demonstrates the 'daily cycle' of tobacco tweets over two years. In order to avoid a negative bias, the cigarette keywords implemented to scrape the tweets were removed from the hedonometric analysis.

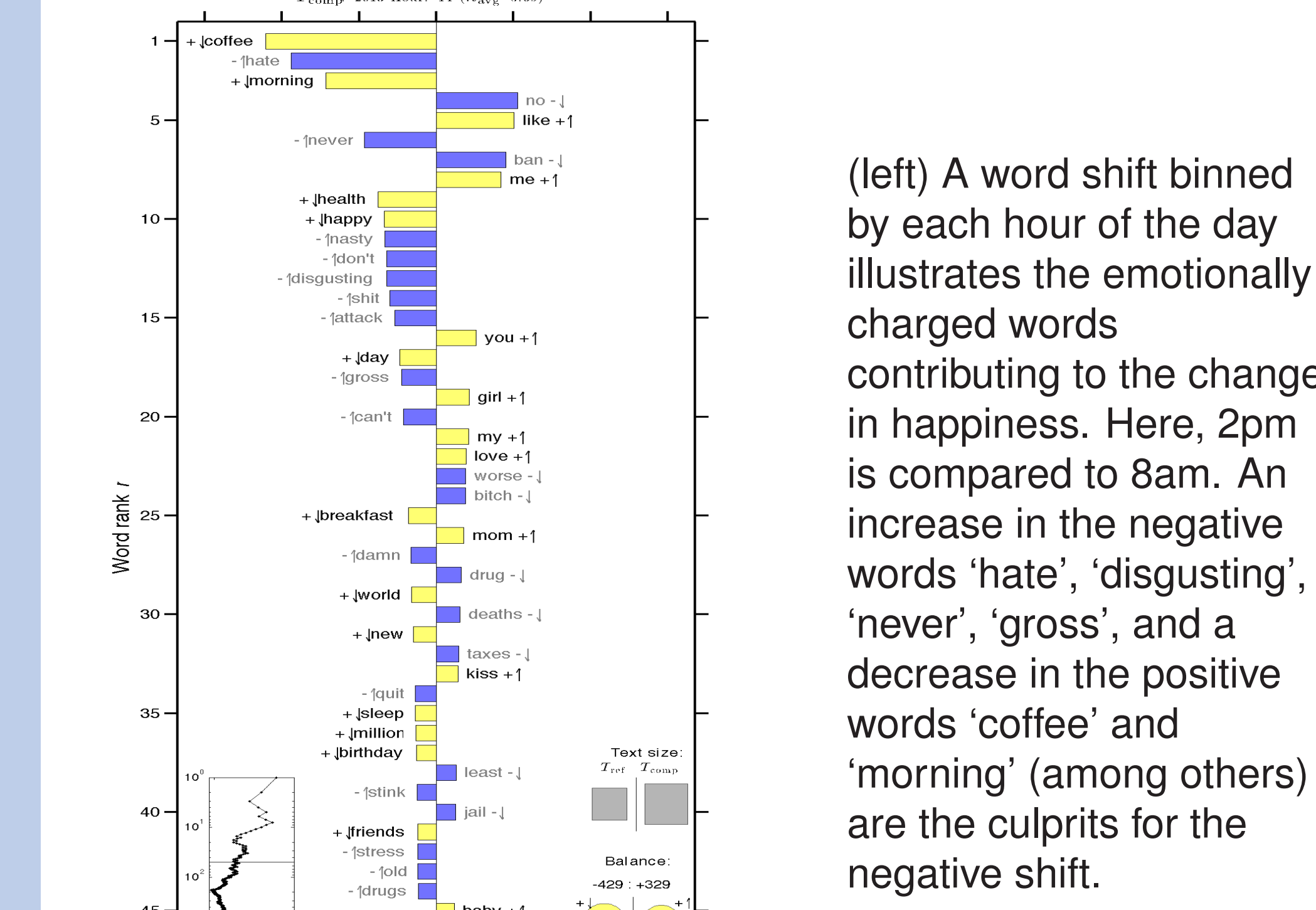
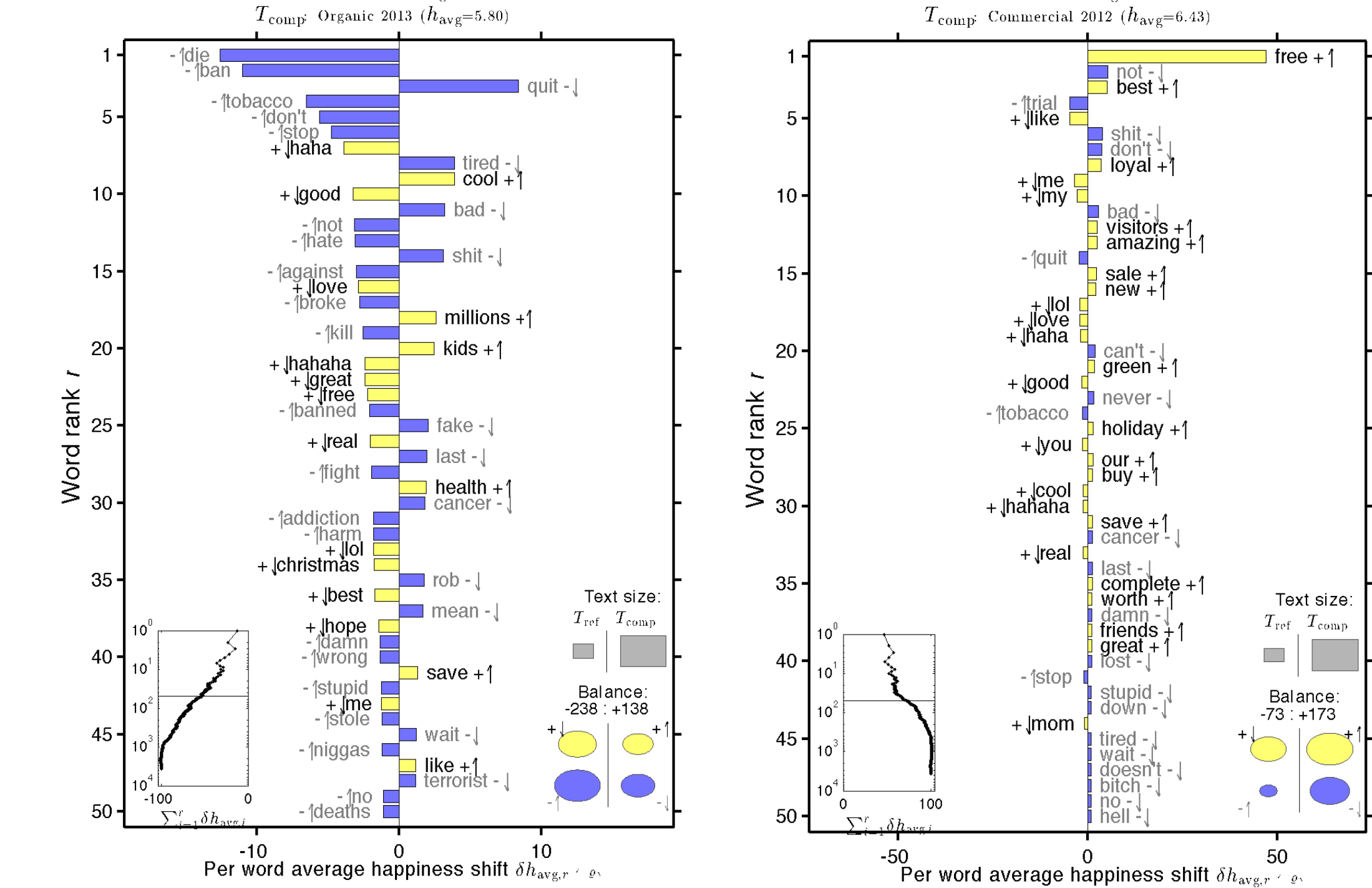


Electronic Cigarettes: Twitter Categorical Time-series Analysis

All E-cigarette mentions spanning January 2012 to July 2014 from the Twitter firehose, a 10% sample of all tweets, were collected and plotted as a function of time (below). The tweets were categorized into three classes: Organic, Commercial, and Infomercial. Marketing tweets use many overly positive words to advertise the product and vastly outnumber Organic Tweets. Since the use of Social Media as a marketing outlet for E-cigarettes is currently a hot political issue, it's important to isolate each of these categories and analyze each separately.



(left) The number of E-cigarette tweets from each user is plotted on log scale. (Below) Wordshift graphs illustrate the change in sentiments of organic users over time (left) as well as comparing the commercial and organic categories (right). On the left, 2012 is used as reference for 2013 where there is an increase in the negative words 'die', 'ban', 'hate', 'against', 'stop', and a decrease in positive words 'love', 'good', and 'hope'. On the right, commercial tweets are using more positive words related to marketing, 'free', 'best', 'sale', 'new'.



(left) A word shift binned by each hour of the day illustrates the emotionally charged words contributing to the change in happiness. Here, 2pm is compared to 8am. An increase in the negative words 'hate', 'disgusting', 'never', 'gross', and a decrease in the positive words 'coffee' and 'morning' (among others) are the culprits for the negative shift.

Acknowledgments

The authors wish to acknowledge the Vermont Advanced Computing Core, which is supported by NASA (NNX-08AO96G) at the University of Vermont which provided High Performance Computing resources that contributed to the research results reported within this poster. EMC was supported by the UVM Complex Systems Center, PSD was supported by NSF Career Award # 0846668. CMD and PSD were also supported by a grant from the MITRE Corporation. C.J, DG, RR, TJW, AK, and AS are supported in part by the National Institute of Health (NIH) Research wards R01DA014028 & R01HD075669, and by the Center of Biomedical Research Excellence Award P20GM103644 from the National Institute of General Medical Sciences.