INFORMS 2016: Text Mining in Health and Security Analytics
- Control Number 5677

Key Conversation Trends and Patterns about Electronic Cigarettes on Social Media

Wenli Zhang, Sudha Ram

INSITE: Center for Business Intelligence and Analytics,
Eller College of Management,
University of Arizona

{wenlizhang, sram}@email.arizona.edu

https://www.insiteua.org/

E-cigarette

- Electronic cigarette (e-cigarette): handheld electronic device that vaporizes flavored liquid
- Ingredients: nicotine, propylene glycol, glycerine, and flavorings

- Since the introduction to the market in 2004, global usage of e-cigarettes has risen exponentially
- Use of e-cigarettes greatly increased in a relatively short period of time
- By 2013, there were several million users globally

- Growth in the US and UK had reportedly slowed in 2015, lowering market forecasts for 2016

Hot Debate Over e-cigarettes

- Motivation
  - Recreation
  - Quitting smoking
  - Healthier than smoking
  - Circumvent smoke-free laws and policies
- Health effects
  - Expose users to fewer toxicants than tobacco
  - May have a role in smoking cessation, but others disagree
  - The safety of electronic cigarettes is uncertain
  - Addiction

Related Literature

- Goniewicz, Maciej L., Elena O. Lingas, and Peter Hajek. "Patterns of electronic cigarette use and user beliefs about their safety and benefits: **an internet survey**." *Drug and alcohol review* 32.2 (2013): 133-140.
- Pearson, Jennifer L., et al. "e-Cigarette awareness, use, and harm perceptions in US adults." *American journal of public health* 102.9 (2012): 1758-1766. **[2 surveys]**
- Regan, Annette K., et al. "Electronic nicotine delivery systems: adult use and awareness of the 'e-cigarette'in the USA." *Tobacco control* 22.1 (2013): 19-23. **[consumer-based mail-in survey]**
- Polosa, Riccardo, et al. "Effect of an electronic nicotine delivery device (e-Cigarette) on smoking reduction and cessation: a prospective 6-month **pilot study**." *BMC public health* 11.1 (2011): 1.

Research objective

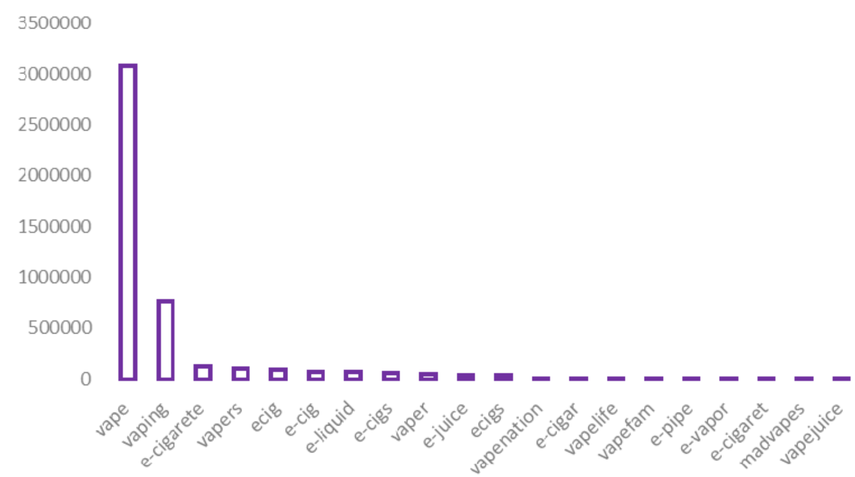
- Traditional survey is not only expensive but not timely, nor adequate to understand
  - Public health impact of e-cigarette
  - Better understanding of population-wise use patterns
  - Perceptions regarding the use and abuse liability of e-cigarette

- Research objective
  - Explore using social media data to identify key conversations, trends, patterns about the usage of e-cigarette
  - Natural language processing, word embedding, topic modeling, content and sentiment analysis, and social network analysis

Data set

	# of tweets	Collection period	Geographical area	# of keywords	Keywords examples
E-cigarette dataset	9,644,416	3/12/2015 -- 4/27/2016	All over the word	50	e-cigarette, e-juice, e-vapor

- Twitter Streaming API
- Total: 9,644,416 (57.926 G)
- Tweets with coordinates information (latitude, longitude): 60,987 (**0.06%**)
- Tweets have location information (user.location): 6,127,426 (**63.5%**)



Selected keywords and frequency of tweets

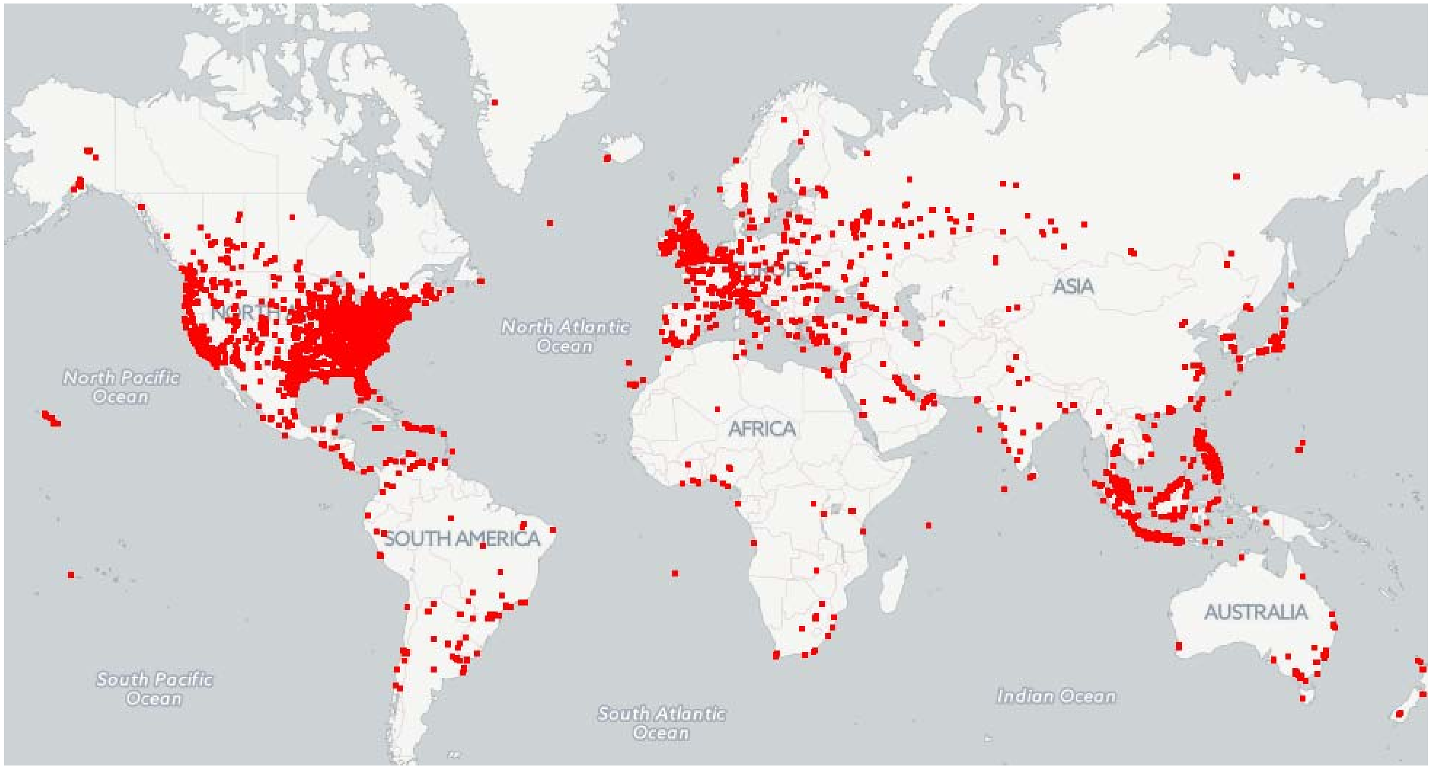
1. Geographic Analysis

- **Research questions**:
  - What are the prevalence and characteristics of e-cigarette users in the US and all over the world?

Geographic Analysis

- Geographic location for each tweet:
  - **Location coordinates** (latitude/longitude) (**0.06%**)
  - **User-specified location** (**63.5%**)
- Address to coordinates (address -> latitude/longitude)
  - Combination of three popular geocoding web services
    - Nominatim (no rate limit, 1 request per second)
    - Bing (50,000 rate limit per day)
    - Googlev3 (2,500 rate limit per day)
- User-specified location (address -> name of US states)
  - Regular expressions analysis for the location field, i.e., Matching state names or postal abbreviations on the US states, followed by matching city names
  - 1,258,878 (13.05%) as one of us states

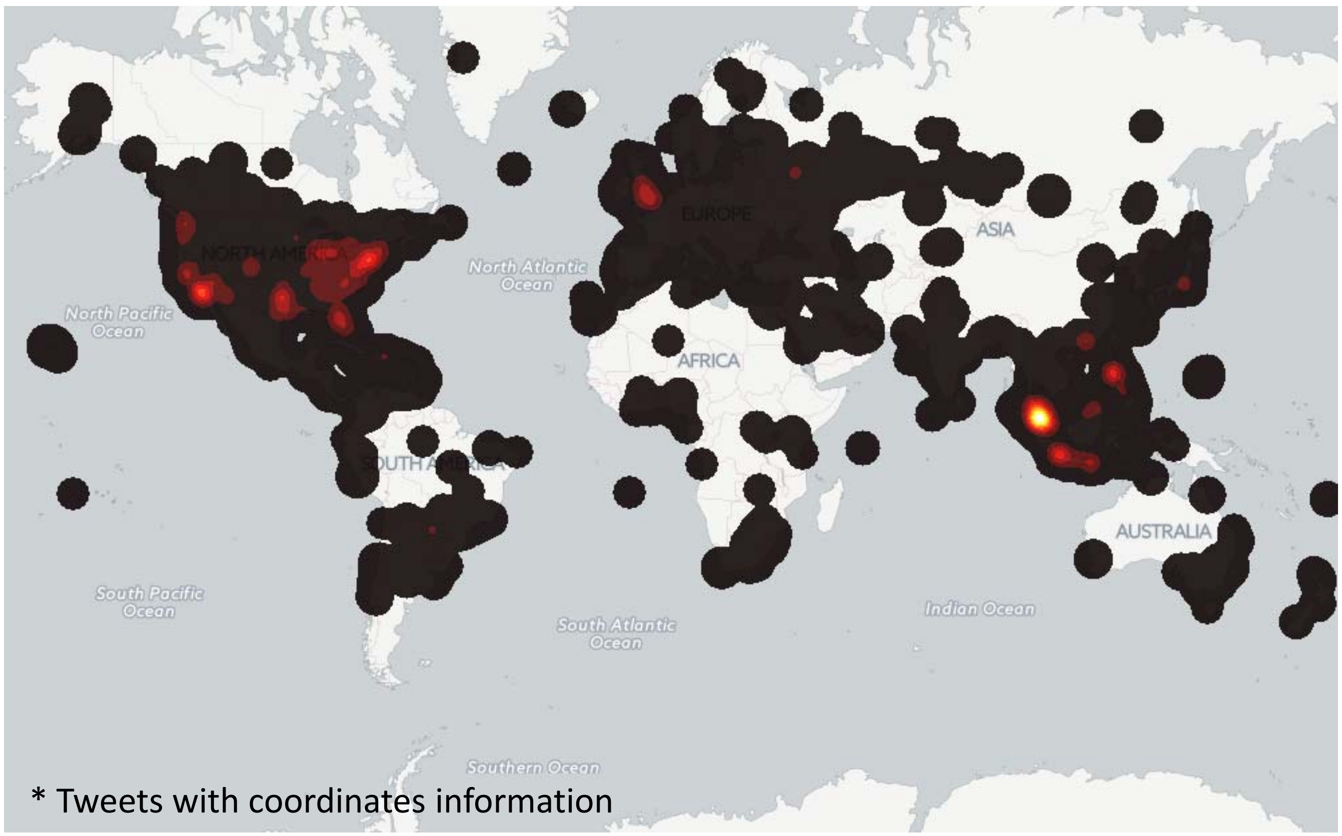
Geographic Mapping



* Tweets with coordinates information

Hot Spot Mapping Using Kernel Density Estimation

ARIZONA
Eller

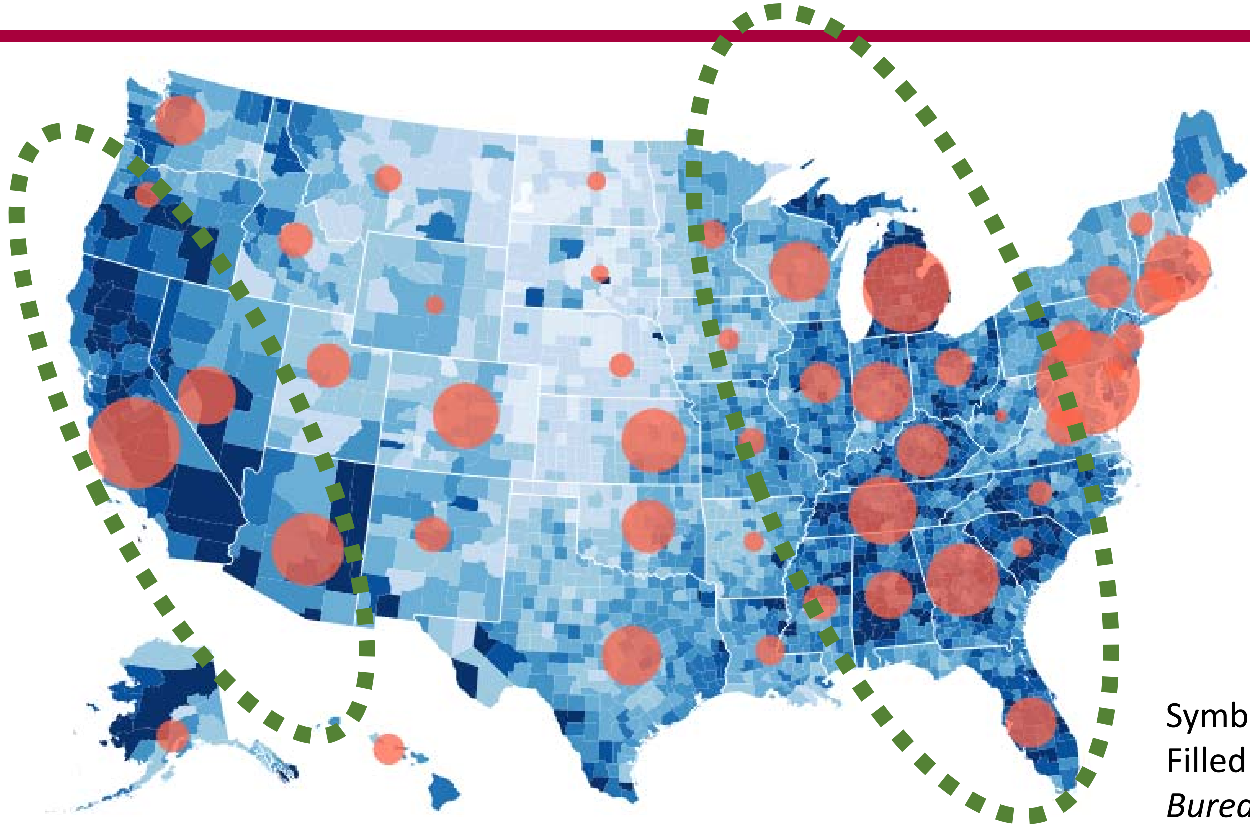


* Tweets with coordinates information

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}}$$

where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between feature i and j.

Number of e-cigarette related tweets aggregated by US States



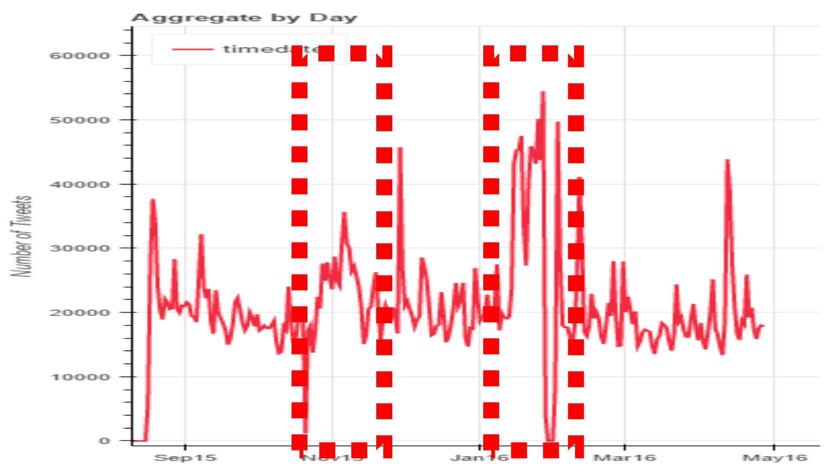
User-specified location (address -> name of US states)
1,258,878 (13.05%) as one of us states

Symbol map: number of tweets
Filled map: unemployment rate 2015
Bureau of Labor Statistics, Census Bureau

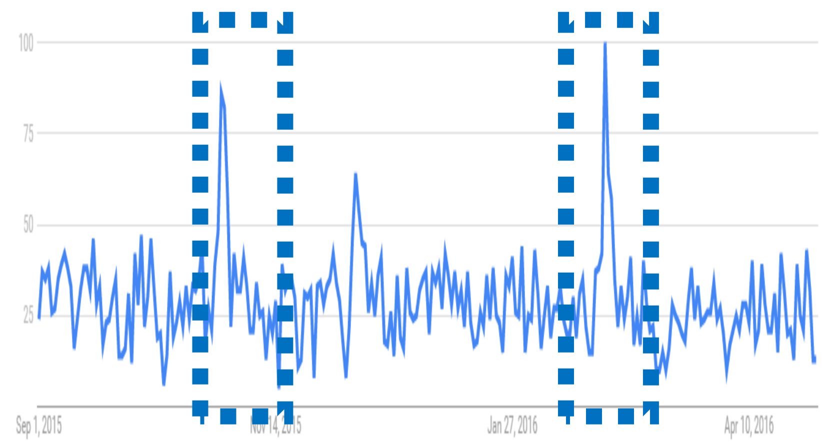
2. Time Series Analysis

- **Research questions**:
  - What are the patterns of the number of e-cigarette related tweets at successive time intervals?
  - * Can meaningful characteristics of the data be extracted and predict future values based on previously observed patterns?

Time Series Analysis - Number of Tweets per Day (US)



User-specified location (address -> name of US states)
1,258,878 (13.05%) as one of us states
2015/9 – 2016/4



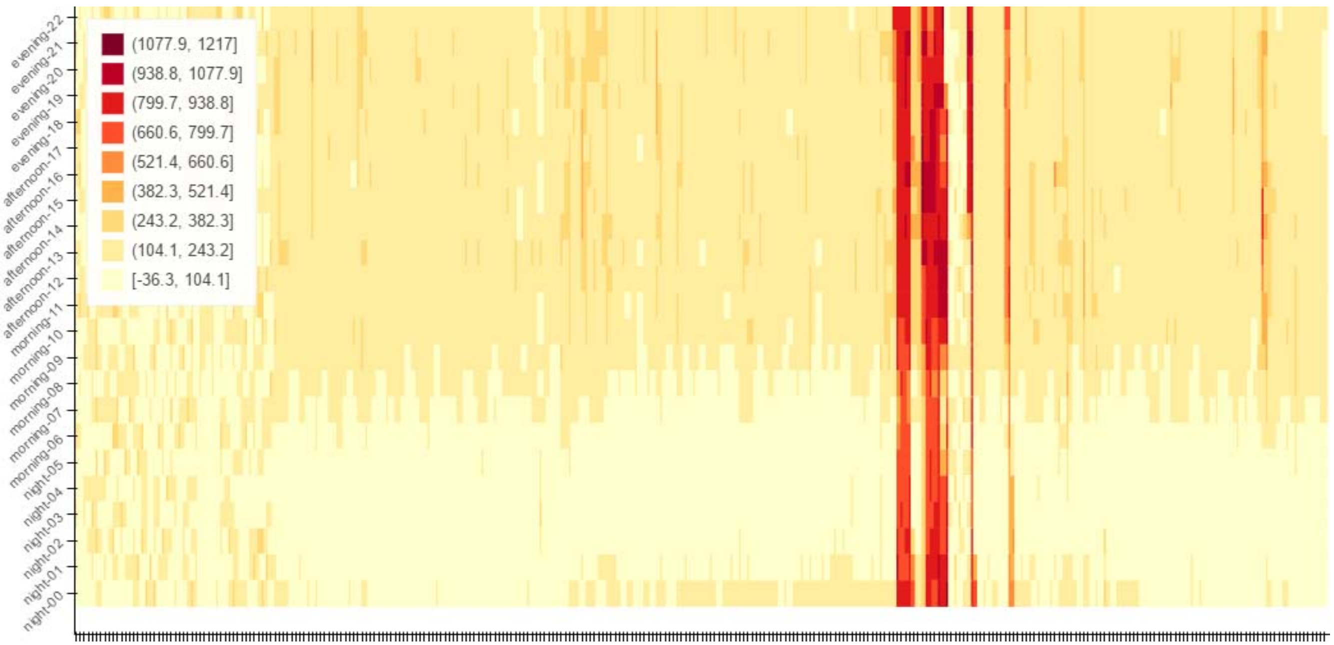
e-cigarette
Search term

Google Trends

United States ▾ 9/1/15 - 5/1/16 ▾ All categories ▾ Web Search

https://www.google.com/trends/explore?date=2015-09-01%202016-05-01&geo=US&q=e-cigarette

Time Series Analysis - Number of Tweets per Hour (US)

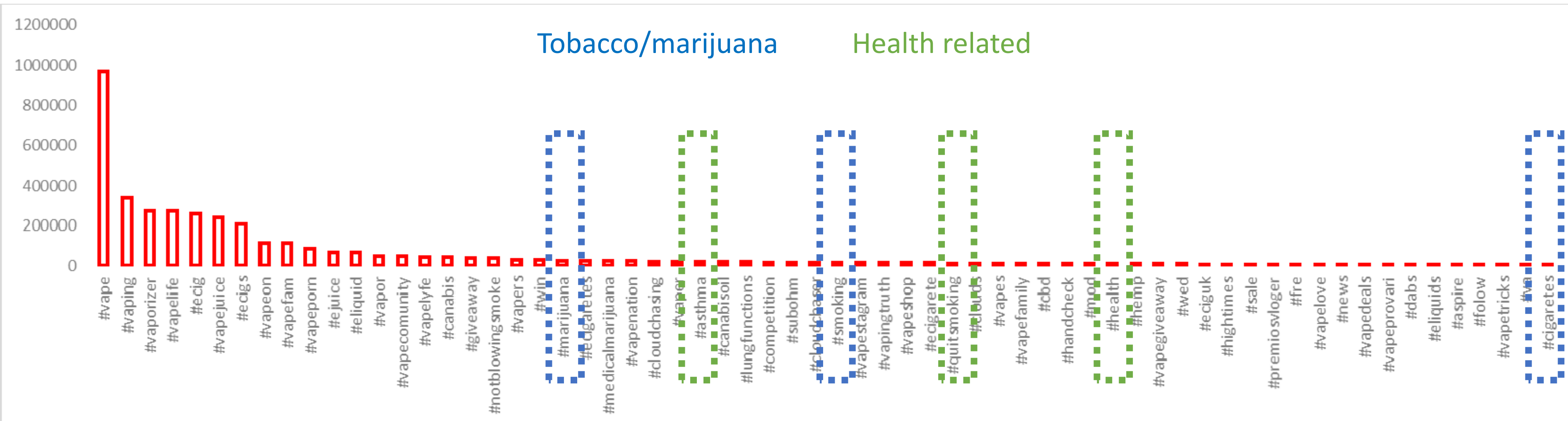


User-specified location (address -> name of US states)
1,258,878 (13.05%) as one of us states
2015/9 – 2016/4

3. Key Conversations and Trends

- **Research questions**:
  - What are the key conversations (topics) and trends about e-cigarettes on social media?
  - *Are e-cigarettes a replacement for tobacco/marijuana (or a new market)?

Key Conversations and Trends
Categorize Tweets by Hash-tagged(#) words

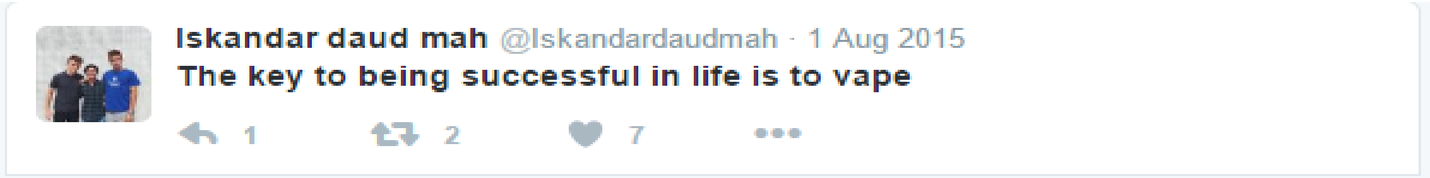


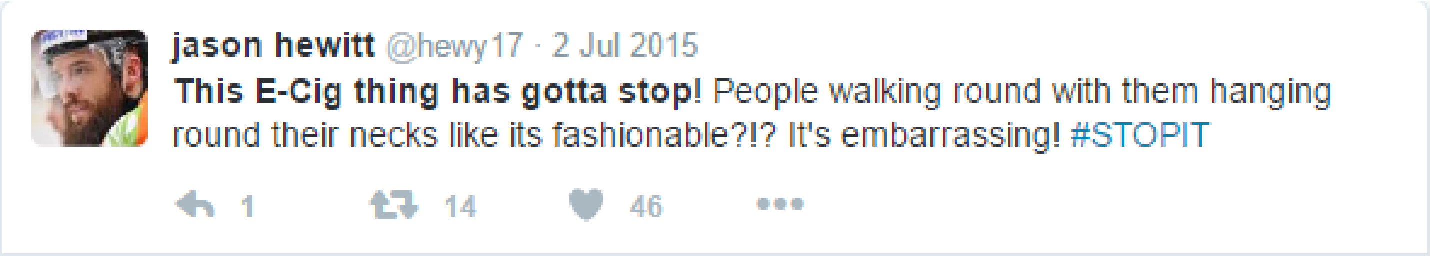
Key Conversations and Trends
Topic modeling

- Topic modeling: identifying patterns in a dataset.
- Latent dirichlet allocation (LDA): un-supervision learning methods

Topic 1: people’s feeling about e-cigarette

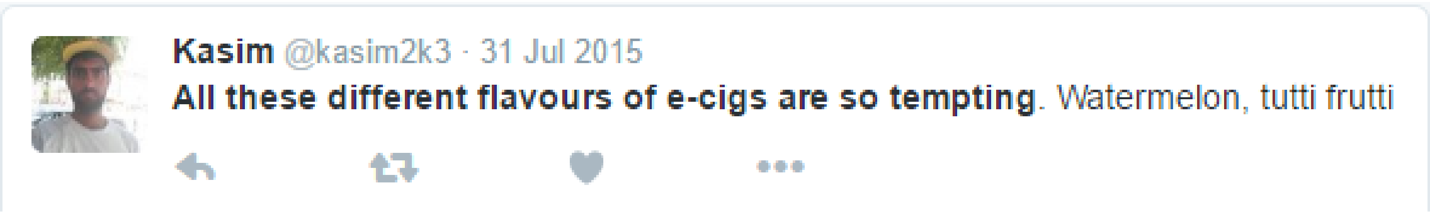
holy, chance, successful, active, liking, fightback, glorified, stout, tidy, fashionable, authentication, fans

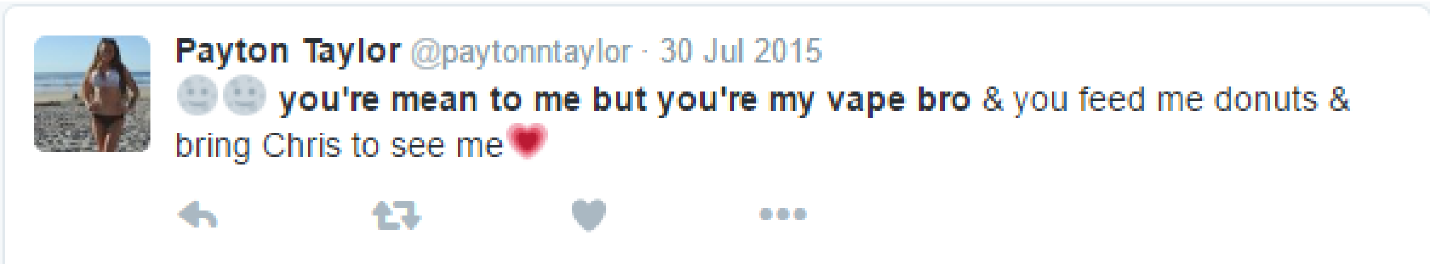




Topic 2:
-- drinks and foods that people have when they are using e-cigarettes
-- flavor of e-cigarette juice

milk, iced, pan, liquids, cigar, melon, donuts, shrimps, cafe, oil, tea





Topic 3:
phenomenon that shows when people are using e-cigarettes

air, beam, steam-punk-mods, fired, vaporization, bright, heat

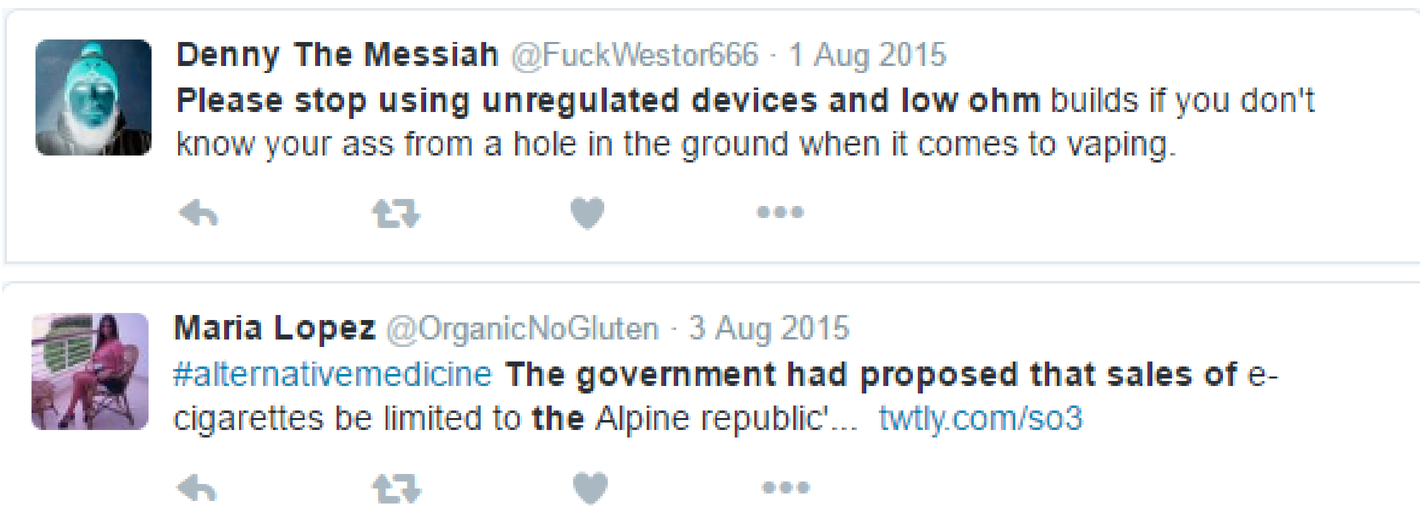
vapingbunny @vapingbunny · 30 Jul 2015
quad beam vape clouds trick! pinterest.com/pin/4324160016...

In reply to David Jones
Trial_Watcher @Trial_Watcher1 · 30 Jul 2015
@davidjones720 @GunnetteP making pot vape is very simple as well, simply soak an amount in vegetable glycerine and **heat at a low temp 225f**

1

Topic 4: e-cigarette legal regulations

complicated, unregulated, web, punish, launches, planed, exploding, proposed, demand, , smoking, quit-smoking, quick, mutation, girls-who-build



4. Content Analysis

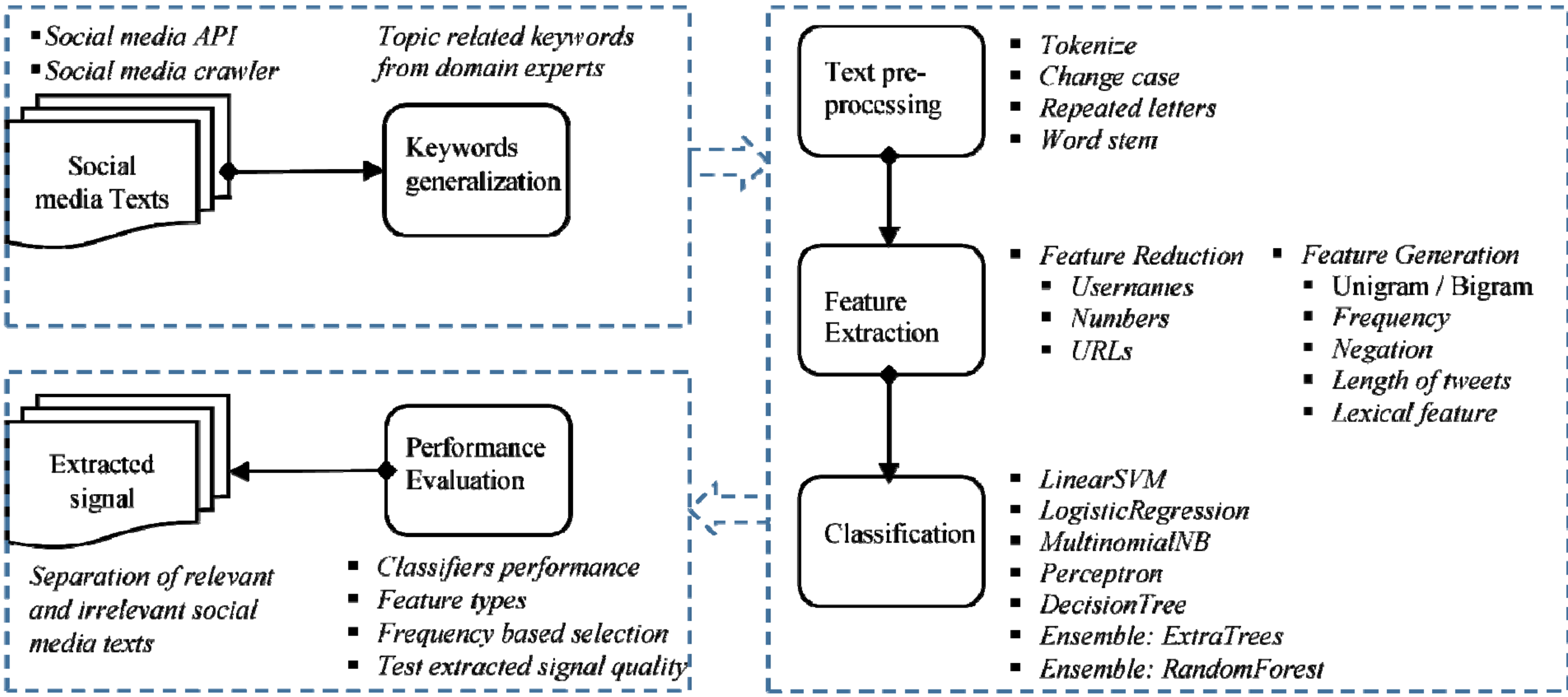
- **Research questions**:
  - In what percentage the e-cigarettes related tweets is about first-person experiences and opinions?
  - In what percentage these tweets is about news, marketing messages, and policy and government themes?

Content analysis

- Classify e-cigarette related tweets into Relevant, Irrelevant, News and Ads
- The training dataset is a collection of tweets that are labelled into categories manually
- Two e-cigarette researchers have manually classified 3,149 tweets

	# of tweets	Category	# of tweets / category
Content analysis training dataset	3,149	• E-cigarette relevant • E-cigarette irrelevant • News • Ads	1,396 (44%) 558 (18%) 311 (10%) 884 (28%)

Content Analysis - Twitter Textual Data Preprocess



Reference: Zhang, W., Ram, S. 2015. A Comprehensive Methodology for Extracting Signal from Social Media Text Using Natural Language Processing and Machine Learning. 25th Workshop on Information Technologies and Systems (WITS).

Content analysis – performance & results

Classifiers	Accuracy	Relevant			Irrelevant			News			Ads		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Logisticregression	0.844	0.821	0.898	0.858	0.800	0.784	0.792	0.905	0.679	0.776	0.885	0.852	0.868
Linearsvc	0.838	0.830	0.875	0.852	0.800	0.784	0.792	0.870	0.714	0.784	0.860	0.852	0.856
Multinomialnb	0.838	0.856	0.883	0.869	0.854	0.686	0.761	0.808	0.750	0.778	0.819	0.880	0.848
Perceptron	0.848	0.905	0.820	0.861	0.792	0.824	0.808	0.880	0.786	0.830	0.831	0.907	0.867
Decisiontree	0.811	0.859	0.801	0.829	0.683	0.778	0.727	0.764	0.750	0.757	0.839	0.855	0.847
Extratrees	0.836	0.836	0.871	0.853	0.752	0.782	0.767	0.853	0.743	0.794	0.874	0.843	0.858
Randomforest	0.823	0.812	0.881	0.845	0.754	0.755	0.754	0.853	0.679	0.755	0.866	0.823	0.844

- **Training dataset; 10 fold cross validation**

Relevant	Irrelevant	News	Ads
71.67%	3.70%	3.47%	21.16%

- Dataset: User-specified location (address -> name of US states)
  1,258,878 (13.05%) as one of us states
- Majority tweets are e-cigarette relevant (consider we use 50 e-cigarette related keywords to collect this dataset)

5. Sentiment analysis

- **Research questions**:
  - What is the attitude about e-cigarettes on social media (* and why)?
  - *How is that different from people's attitude towards tobacco and marijuana?

Domain adaptation for sentiment analysis

- Adaptation by **feature augmentation**
  - Take each feature in the original problem and make three versions of it: a general version, a source-specific version and a target-specific version
  - The augmented source data will contain only general and source-specific versions
  - The augmented target data contains general and target-specific versions

$$\Phi^s(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{x}, \mathbf{0} \rangle, \quad \Phi^t(\boldsymbol{x}) = \langle \boldsymbol{x}, \mathbf{0}, \boldsymbol{x} \rangle$$

- Reference: daumé III, hal. "Frustratingly easy domain adaptation." *Arxiv preprint arxiv:0907.1815* (2009).

Sentiment analysis training datasets

	# of tweets	Category	# of tweets / category
Target domain: e-cigarette	1,086	positive negative	737 (68%) 349 (32%)

15 junior and senior students from University of Arizona were invited to label **1,086 tweets** (randomly sampled from the dataset) as *"positive", "negative"*.

	# of tweets	Category	# of tweets / category
Source domain: Election debate	5,282	positive negative	2,418 (45%) 2,864 (55%)

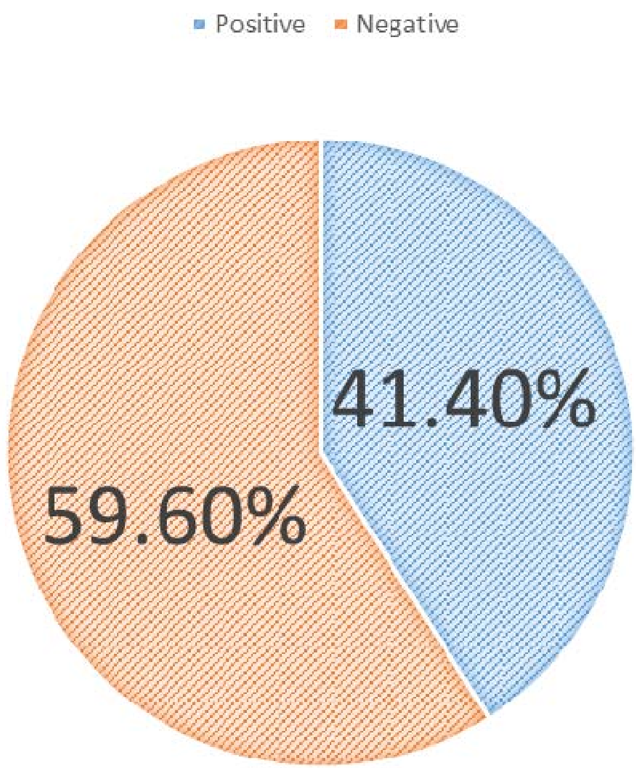
- Twitter sentiment dataset
  - 2008 US Election debate (http://www.ayman-naaman.net/2010/11/21/twitter-sentiment-dataset-online/)
  - Twitter sentiment corpus by Niek Sanders (http://www.sananalytics.com/lab /twitter-sentiment/)
- Only positive and negative records were kept

Sentiment analysis
Domain adaptation by feature augmentation

Classifiers	Accuracy	Positive			Negative		
		Precision	Recall	F1	Precision	Recall	F1
LogisticRegression	0.736	0.714	0.680	0.697	0.751	0.780	0.765
LinearSVC	0.736	0.690	0.741	0.715	0.778	0.731	0.754
MultinomialNB	0.717	0.721	0.599	0.654	0.715	0.813	0.761
Perceptron	0.708	0.734	0.544	0.625	0.695	0.841	0.761
DecisionTree	0.613	0.567	0.574	0.570	0.652	0.645	0.649
ExtraTrees	0.706	0.706	0.586	0.640	0.706	0.802	0.751
RandomForest	0.686	0.691	0.542	0.607	0.685	0.803	0.739

10 fold cross validation

Sentiment analysis results



- Dataset: User-specified location (address -> name of US states)
  1,258,878 (13.05%) as one of us states
- More tweets are showing negative sentiments

Twitter users' sentiment in each content category

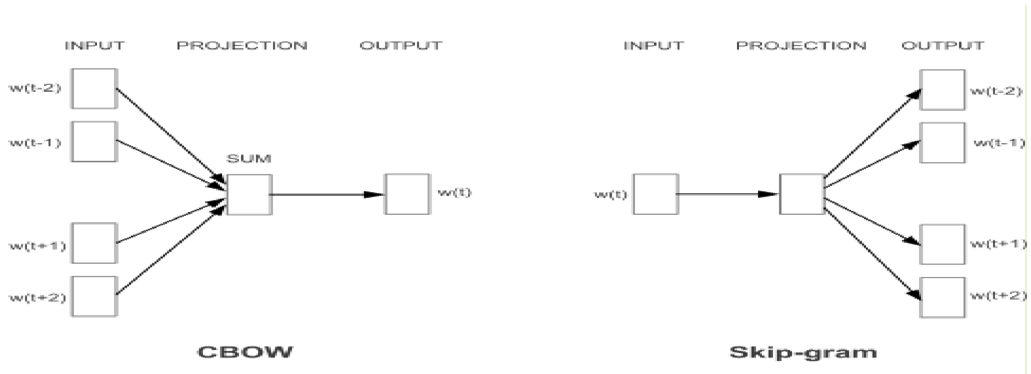
Content	Sentiment	%
Relevant	positive	39.55%
	negative	60.45%
Irrelevant	positive	40.62%
	negative	59.38%
News	positive	40.21%
	negative	59.79%
Ads	positive	43.17%
	negative	56.83%

6. Language Patterns on e-cigarette related social media data

- **Research questions**:
  - What are the language patterns on e-cigarette related social media data?

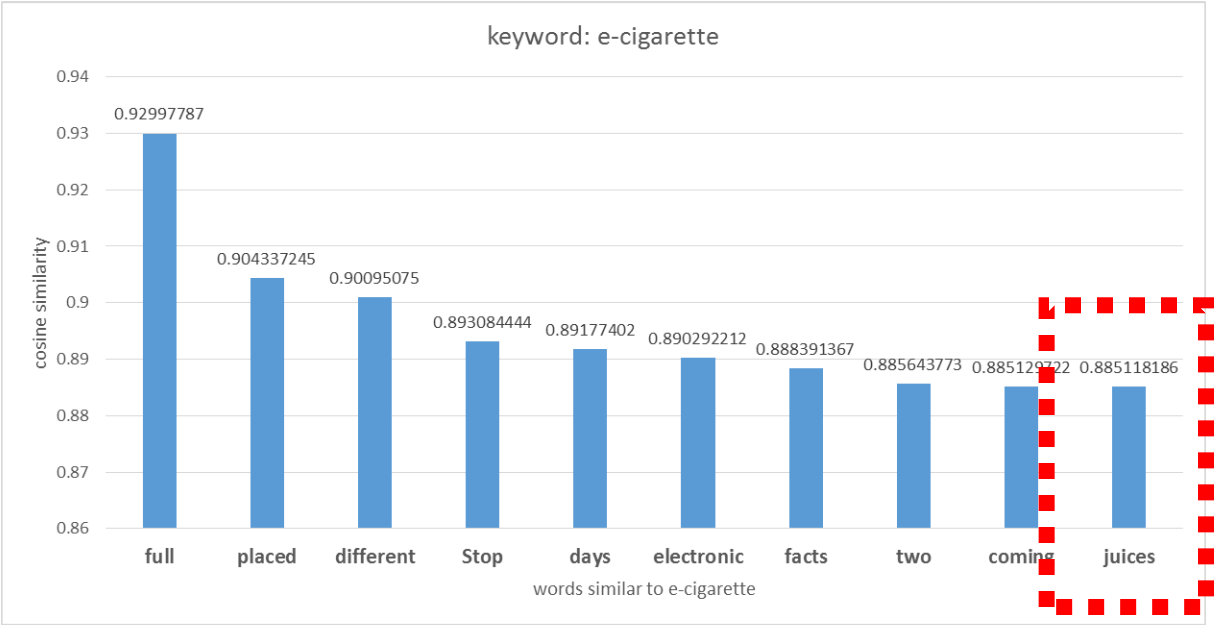
Language Patterns on e-cigarette related social media data

- Goal: learning high-quality word vectors
  - Continuous Bag-of-Words Model
    - Uses continuous distributed representation of the context
  - Continuous Skip-gram Model
    - Maximize classification of a word based on other words in the same sentence.



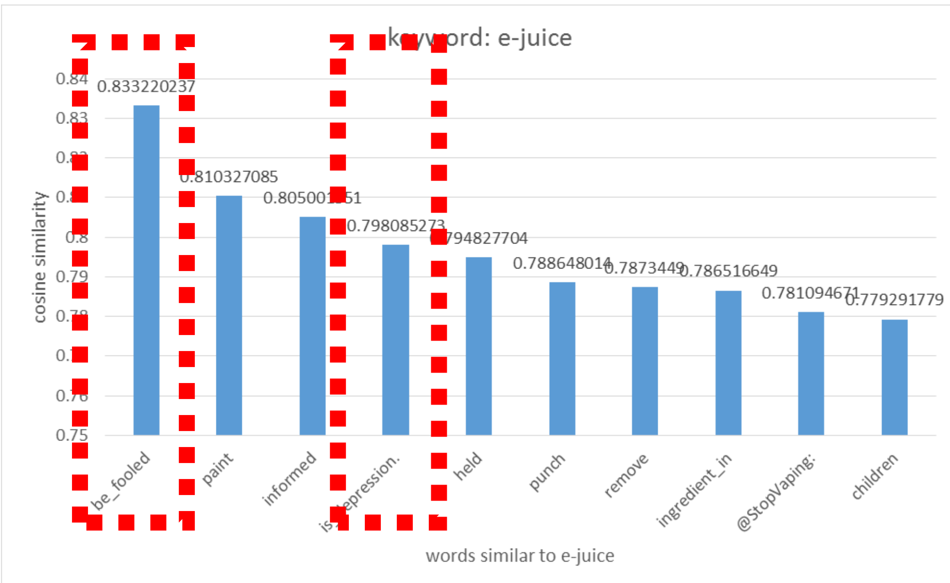
Reference: Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Words similarity: e-cigarette



The word “juices” has high cosine similarity with this keyword, as “juices” in this content normally means “nicotine juice”, we may consider these two words are synonyms in e-cigarette related social media text

Words similarity: e-juice



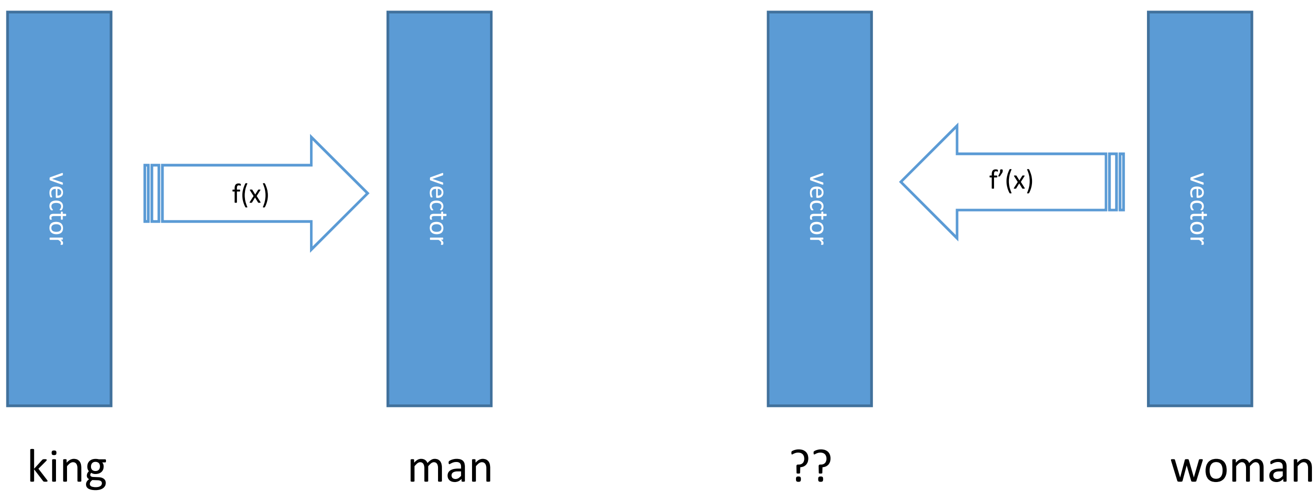
"fool" and "depression" tend to occur in the same context of the word "e-juice".

Literature: Many people have complained of lack of concentration, mood disorders, **depression**, anxiety, greater appetite and other symptoms which can last for months and are electronic cigarette side effects.

Bullen, Chris, et al. "Effect of an electronic nicotine delivery device (e cigarette) on desire to smoke and withdrawal, user preferences and nicotine delivery: randomised cross-over trial." Tobacco control 19.2 (2010): 98-103.

Analogies of the keyword

- Analogy: word that is comparable to the keyword in significant respects
- e.g., King – man → woman = w: queen
- **argmax** *cos(w, king) - cos(w, man) + cos(w, woman).*

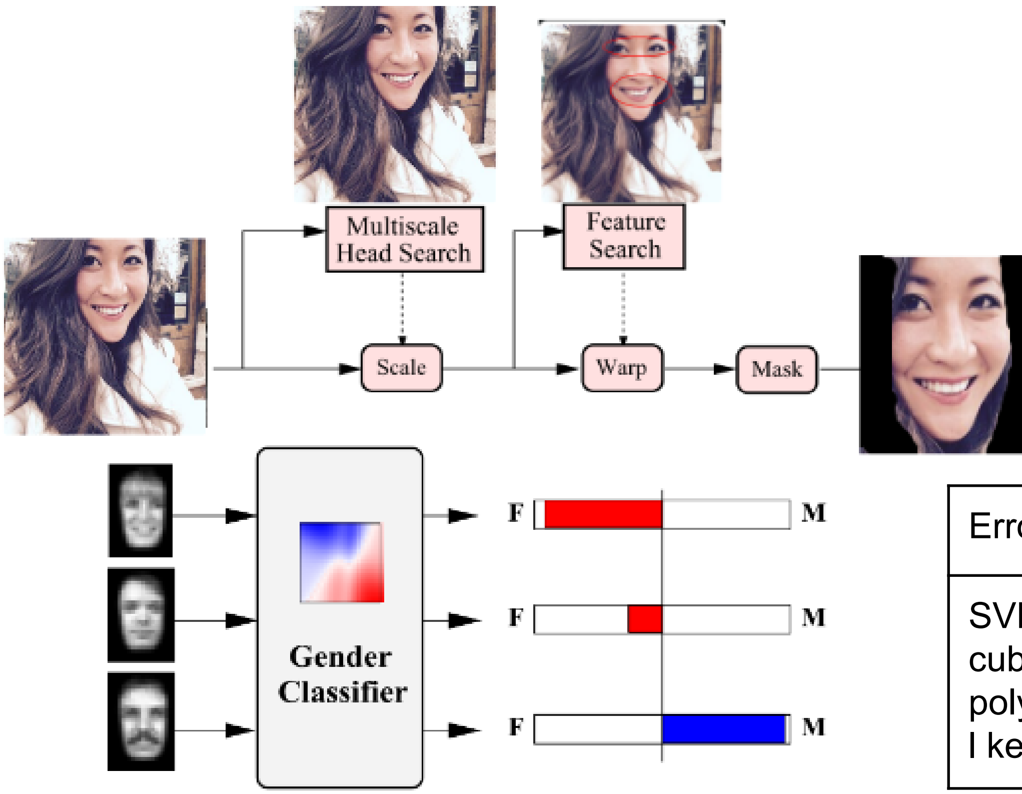


Reference: Levy, Omer, Yoav Goldberg, and Israel Ramat-Gan. "Linguistic Regularities in Sparse and Explicit Word Representations." *CoNLL*. 2014.

Gender Classification

- Distant supervision: classifier is learned given a weakly labeled training set
  - https://www.ssa.gov/oact/babynames/
  - The most popular given names for **male and female** babies born during 1970- 2000
- Twitter user profile
  - Screen name (e.g., jsmith92, kingofpittsburgh)
  - Full name (e.g., John Smith, King of Pittsburgh)
- Profile image URL
  - Dimension: 48 x 48

Gender Classification



B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-19(7):696–710, July 1997.

Error Rate	Overall	Male	Female
SVM with cubic polynomial kernel	27.16%	26.53%	28.04%

Analogies of *e-cigarette*

e-cigarette – _MAN_ → _WOMAN_ = ?

Words	Similarity	Words	Similarity
health	0.352852	danger	0.346252
ecigs	0.351051	tonight	0.344528
be	0.3469	county	0.340723
significant	0.34662	e-cigarette	0.340395

To some woman, e-cigarette means "health", however, to other woman, e-cigarette means "danger"

e-cigarette – _WOMAN_ → _MAN_ = ?

Words	Cosine similarity	Words	Cosine similarity
nicotine	0.35991	home	0.348752
back	0.358842	work	0.348257
ecig	0.356447	putting	0.347857
be	0.352272	vape	0.343035

To some man, e-cigarette means "home", however, to other man, e-cigarette means "work".

Thank you.

On going: Social network analysis

- News & Ads tweets: tweet – retweet network
- Relevant tweets: user – followers network