INFORMS 2016: Business Analytics and Text Mining - Control Number 5228

Extracting Signals from Social Media Text with Natural Language Processing, Machine Learning and Domain Adaptation

Wenli Zhang, Sudha Ram

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

https://www.insiteua.org/

{wenlizhang, sram}@email.arizona.edu

An extension of our previous work in Zhang, W., Ram, S. WITS 2015.

- **Background**
- Methodology
  - Text preprocessing
  - Feature extraction
    - Feature reduction
    - Feature generation
- Classification
- Domain Adaptation
- Experiments & results
- Implications & contributions

Social Media & Predictive Analytics

- Social media are widely used
- Using social media data for predictive analytics
  - Disease surveillance
  - Targeted marketing
  - Political campaigns
- Great potential for revealing latent population characteristics

Accuracy of These Systems

- Commonly used techniques:
  - Keyword matching
  - Linear regression

- Many of the predictions and analyses produced misrepresent the real world.

Misrepresent the Real World

- Flu surveillance
  - Not been correlated with CDC infection data in recent seasons
- Google's flu-tracking service
  - Wildly overestimated

Noise from Social Media Data (1)

Bias machine learning techniques toward misclassification of text

(A) loosely structured informal language:

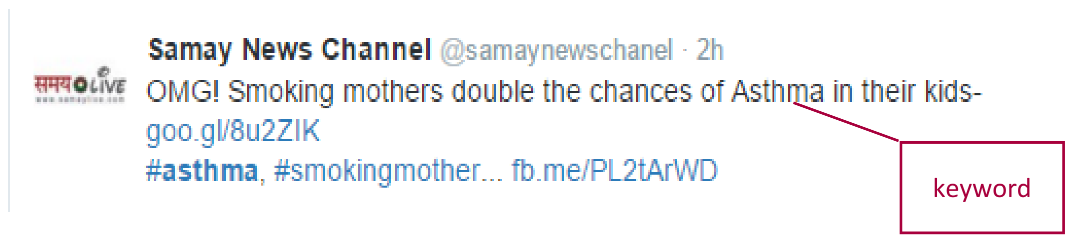
- Misspellings / abbreviations / urban slangs / emoticons

Noise from Social Media Data (2)

Overestimate population characteristics

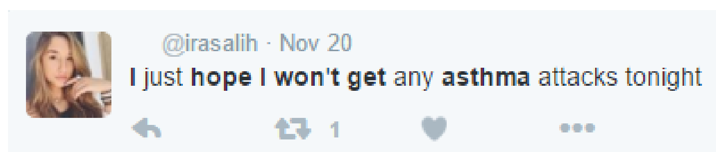
(B) Anomalous media spikes:

- Retweet asthma news stories
- Do not necessarily reflect actual disease affliction



(C) Use of misleading terms and phrases:

- Tweets indicating awareness of disease; clearly about the disease but not about an infection.

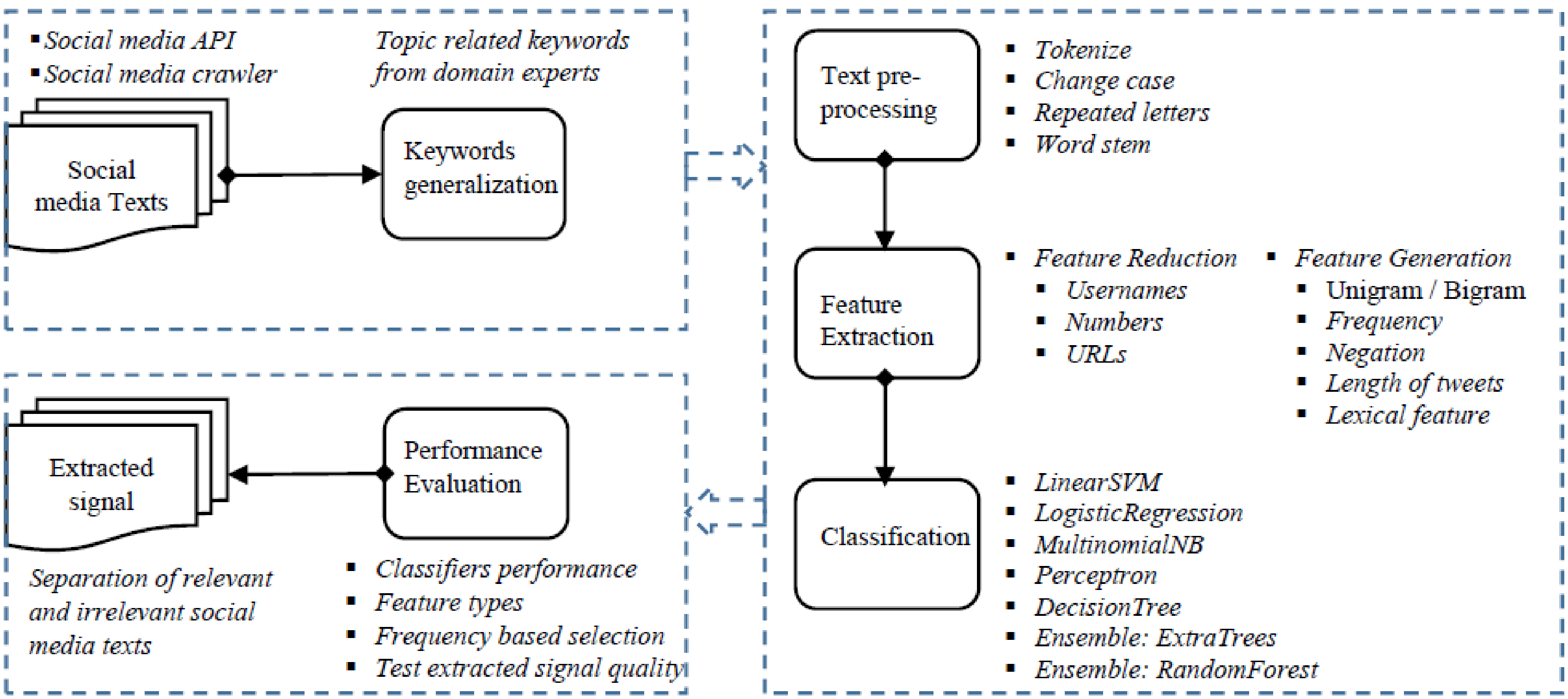


Research Objective

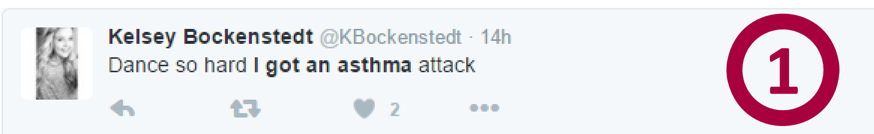
- Effective methodology to extract signal from social media text
- Clearly distinguish relevant text on a specific topic
    - Accurate
    - Timely
    - Economical

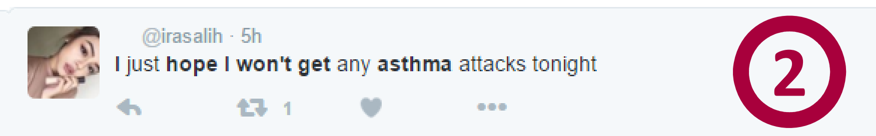
- Background
- **Methodology**
  - Text preprocessing
  - Feature extraction
    - Feature reduction
    - Feature generation
- Classification
- Domain Adaptation
- Experiments & results
- Implications & contributions

Methodology for Signal Extraction from Social Media Text



Feature Vector



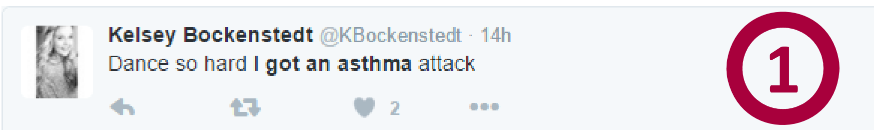


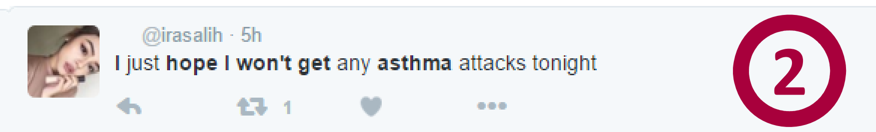
	dance	***so***	***hard***	***i***	***get***	***an***	***asthma***	***attack***	***just***	***hope***	***will***	***not***	***tonight***
tweet1	1	1	1	1	1	1	1	1					
tweet2				1	1		1	1	1	1	1	1	1
……													

- Directly determines how successful the signals could be extracted from social media text.

- Background
- Methodology
  - **Text preprocessing**
  - Feature extraction
    - Feature reduction
    - Feature generation
- Classification
- Domain Adaptation
- Experiments & results
- Implications & contributions

Preprocess (1)





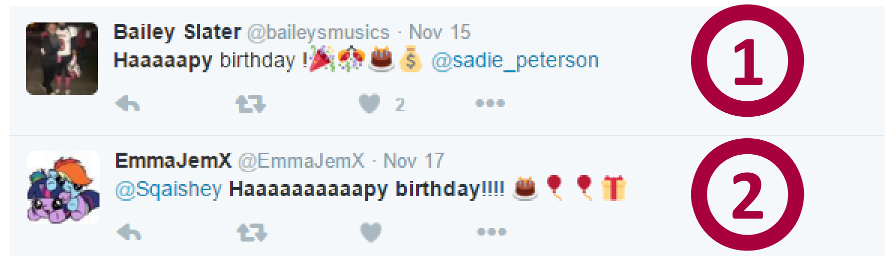
	dance	*so*	*hard*	*i*	*get*	*an*	*asthma*	*attack*	*just*	*hope*	*will*	*not*	*tonight*
tweet1	1	1	1	1	1	1	1	1					
tweet2				1	1		1	1	1	1	1	1	1
……													

- **Tokenize:** e.g., Hewlett-Packard / San Francisco
- **Change case**: lowercase.
- **Additional white spaces:** multiple whitespaces → single whitespace

Preprocess (2)

- **Repeated letters**: Any letter occurring more than two times in a row is replaced with two occurrences: haaaaappy → haappy.
- **Stem word**: Porter's algorithm.

Pre-processing can effectively reduce lexical noise.



	haappy	birthday
tweet1	1	1
tweet2	1	1

Word stem examples:

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

- Background
- Methodology
  - Text preprocessing
  - Feature extraction
    - **Feature reduction**
    - Feature generation
- Classification
- Domain Adaptation
- Experiments & results
- Implications & contributions

Feature Reductions

Original	happy	20	birthday	@thegob70!	#CowboysNation	like	us	http://fb.me/2iE7MvMin
Feature Reduction	happy	***NUMBER***	birthday	***USERNAME***	CowboysNation	like	us	***URL***

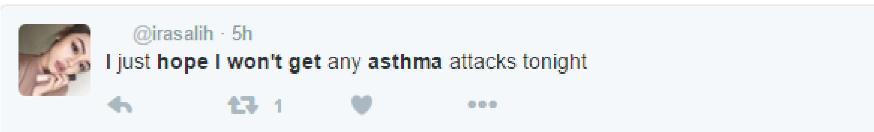
- **Usernames**: equivalence class token (USERNAME) replaced all words that start with the @
- **Numbers**: all the numbers were replaced with the token (NUMBER).
- **URLs**: equivalence class was used for all URLs, token (URL).

Effect of feature reductions: Shrink the feature set down to 45% of its original size.

Hugely improve the efficiency of machine learning algorithms.

- Background
- Methodology
  - Text preprocessing
  - Feature extraction
    - Feature reduction
    - **Feature generation**
- Classification
- Domain Adaptation
- Experiments & results
- Implications & contributions

Feature Generation (1)

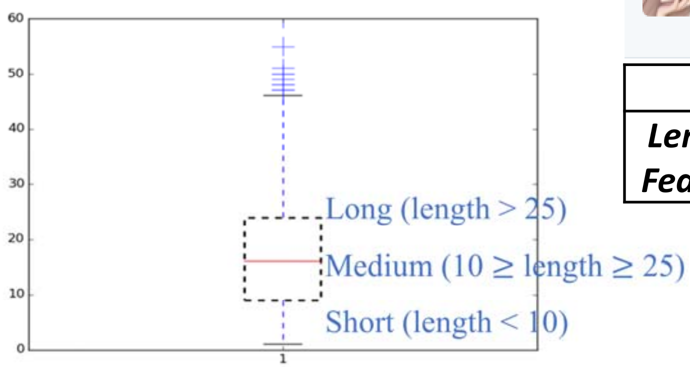


	i	just	hope	will	not	***not_get***	***not_asthma***	***not_attack***	***not_tonight***
Negation	1	1	1	1	1	***1***	***1***	***1***	***1***
Bigram	i_just	just_hope	hope_will	will_not	…	…	…	…	…

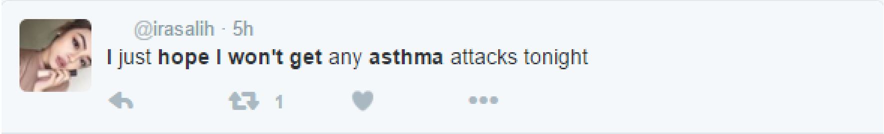
- **Unigram**
- **Bigram:** every sequence of two adjacent elements in a string of tokens
- **Negation**: Prefix all words between a negation word and a punctuation sign with (NOT).

Feature Generation (2)

Text Length Analysis



Asthma training dataset



	i	...	attack	tonight	***SHORT***	***MEDIUM***	***LONG***
Length Feature	1	...	1	1	***1***		

Feature Generation (3)

Text: I got an asthma attack.

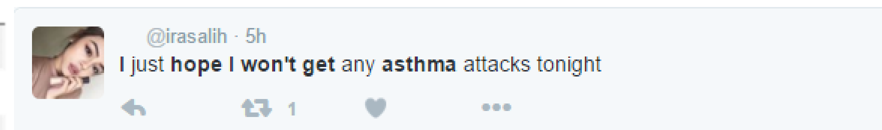
Part-of-Speech tag:

Tokens	*Part-of-speech*	*Tags*
i	List item marker	LS
got	Verb, past tense	VBD
an	Determiner	DT
asthma	Noun, singular or mass	NN
attack	Noun, singular or mass	NN

Feature extracted:

an_DT \ asthma_NN \ attack_NN

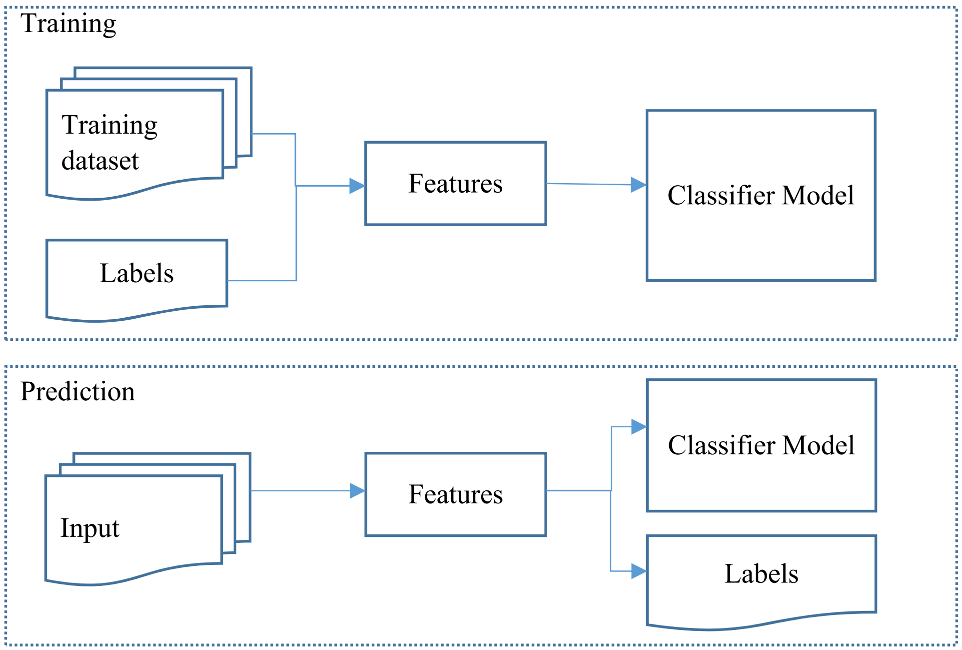
Part-of-Speech Tag



	i	...	*get_VBD*	*asthma_NN*	*attack_NN*
Lexical Feature	1	...	*1*	*1*	*1*

- Background
- Methodology
  - Text preprocessing
  - Feature extraction
    - Feature reduction
    - Feature generation
- **Classification**
- Domain Adaptation
- Experiments & results
- Implications & contributions

Classification:
Extracting Signal from Noisy Dataset



- Identifying categories a new observation belongs
- Training set of data

- Background
- Methodology
  - Text preprocessing
  - Feature extraction
    - Feature reduction
    - Feature generation
- Classification
- **Domain Adaptation**
- Experiments & results
- Implications & contributions

Domain Adaptation by Feature Augmentation

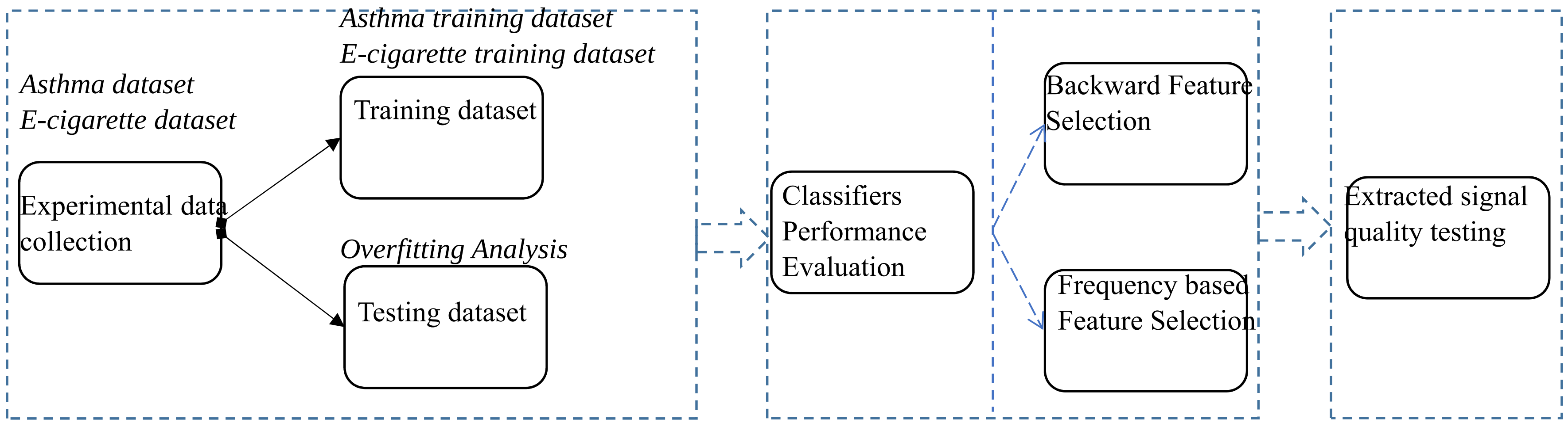
- Domain Adaptation by **feature augmentation**
  - Take each feature in the original problem and make three versions of it: a general version, a source-specific version and a target-specific version
  - The augmented source data will contain only general and source-specific versions
  - The augmented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

- Background
- Methodology
  - Text preprocessing
  - Feature extraction
    - Feature reduction
    - Feature generation
- Classification
- Domain Adaptation
- **Experiments & results**
- Implications & contributions

Experiments and Results



Dataset Description

	# of tweets	Collection period	Geographical area	# of keywords	Keywords examples
Asthma dataset	5,513,368	11/1/2013-6/30/2014	All over the word	18	asthma, inhaler, wheezing
E-cigarette dataset	921,173	5/1/2014-5/31/2014		50	e-cigarette, e-juice, e-vapor

Not used during classifier development

Training Datasets

	# of tweets	Category	# of relevant	# of irrelevant	
Asthma training dataset	4,500	• asthma relevant • asthma irrelevant	814 (18%)	3,686 (82%)	Unbalanced dataset
E-cigarette training dataset	3,149	• e-cigarette relevant • e-cigarette irrelevant	1,396 (44%)	1,753 (56%)	Balanced dataset

Performance of Baseline Method

	accuracy	asthma relevant		asthma irrelevant	
		precision	recall	precision	recall
ANN	0.86	0.67	0.20	0.87	0.98

ANN: artificial neural network

Classifier Performance Evaluation

	# of features		*asthma relevant*		*asthma irrelevant*		*# of features*		*e-cigarette relevant*		*e-cigarette irrelevant*	
	Unigram	a	p	r	p	r	Unigram	a	p	r	p	r
LinearSVM		0.88	0.61	0.63	0.93	0.92		0.88	0.84	0.86	0.90	0.89
LogisticRegression		**0.89**	**0.67**	0.60	0.92	0.94		0.87	0.82	0.86	0.90	0.87
MultinomialNB		0.82	0.44	0.34	0.88	0.91		0.89	0.87	0.86	**0.90**	0.91
Perceptron	5564	0.86	0.63	0.43	0.91	0.94	4212	0.87	0.82	0.86	0.90	0.87
DecisionTree		0.87	0.62	**0.68**	**0.94**	0.92		0.87	0.85	0.82	0.88	0.90
Ensemble: ExtraTrees		0.87	0.64	0.47	0.90	**0.95**		**0.89**	**0.87**	0.86	0.90	**0.91**
Ensemble: RandomForest		0.87	0.62	0.47	0.90	0.94		0.88	0.86	0.86	0.90	0.90

(a) Asthma training data set

(b) E-cigarette training data set

a: accuracy *p: precision* *r: recall*

10 Fold Cross Validation
Training data set

Overfitting Analysis

	# of features		500_tweets relevant		500_tweets irrelevant	
	Unigram	a	p	r	p	r
LinearSVC		0.88	0.67	0.78	0.94	0.90
LogisticRegression		0.88	0.70	0.68	0.92	0.93
MultinomialNB		0.82	0.63	0.32	0.85	0.95
Perceptron	5564	0.87	0.66	0.76	0.94	0.90
DecisionTree		0.78	0.48	0.53	0.88	0.85
Ensemble: ExtraTrees		0.85	0.69	0.47	0.88	0.95
Ensemble: RandomForest		0.85	0.65	0.59	0.90	0.92

a: accuracy *p: precision* *r: recall*

Training-Asthma Training Dataset; Testing-500 New Tweets

Not used during classifier development
Not used in feature generation

Backward Feature Selection

features	classifier	# of features	a	asthma relevant p	asthma relevant r	asthma irrelevant p	asthma irrelevant r	classifier	# of features	a	e-cigarette relevant p	e-cigarette relevant r	e-cigarette irrelevant p	e-cigarette irrelevant r
	LR		0.88	0.67	0.55	0.92	0.95	ET		0.89	0.86	0.86	0.91	0.90
U + N + L + P	LC	6789	0.87	0.59	0.56	0.92	0.93	NB	4913	0.89	0.87	0.86	0.91	0.91
	LR		0.87	0.63	0.56	0.92	0.94	ET		0.88	0.85	0.86	0.90	0.90
U + N + L	LC	5941	0.87	0.59	0.62	0.92	0.92	NB	4357	0.89	0.87	0.86	0.91	0.91
	LR		0.88	0.64	0.58	0.92	0.94	ET		0.89	0.87	0.87	0.91	0.91
U + N + P	LC	6774	0.87	0.60	0.60	0.92	0.92	NB	4902	0.89	0.87	0.86	0.91	0.91
	LR		**0.89**	**0.69**	0.55	0.92	0.95	ET		0.88	0.86	0.85	0.90	0.90
U + L + P	LC	6423	0.87	0.62	0.60	0.92	0.93	NB	4775	0.89	0.87	0.86	0.91	0.91
	LR		0.88	0.65	0.60	0.92	0.94	ET		0.89	0.87	0.86	0.90	0.91
U + N	LC	5938	0.85	0.55	0.56	0.91	0.91	NB	4354	0.89	0.87	0.86	0.91	0.91
	LR		0.88	0.64	0.58	0.92	0.94	ET		0.88	0.86	0.86	0.90	0.90
U + L	LC	5567	0.87	0.61	0.63	0.93	0.92	NB	4215	**0.89**	0.87	0.86	0.91	0.91
	LR		0.88	0.66	0.58	0.92	0.94	ET		0.89	0.86	0.86	0.90	0.91
U + P	LC	6408	0.87	0.61	0.62	0.93	0.92	NB	4763	0.87	**0.90**	0.87	0.91	0.91
	LR		**0.89**	**0.69**	0.58	0.92	0.95	ET		0.89	0.87	0.85	0.90	**0.92**
U + B	LC	26497	0.87	0.60	0.60	0.92	0.92	NB	17301	**0.90**	0.86	**0.90**	**0.93**	0.90
	LR		0.87	0.69	0.43	0.90	**0.96**	ET		0.87	0.85	0.83	0.89	0.90
B	LC	20933	0.87	0.64	0.51	0.91	0.94	NB	13089	0.88	0.84	0.88	0.91	0.89
	LR		0.89	0.67	0.60	0.92	0.94	ET		0.89	0.87	0.85	0.90	0.91
U	LC	5564	0.87	0.61	**0.63**	**0.93**	0.92	NB	4212	0.89	0.87	0.86	0.90	0.91

(a) Asthma training data set (b) E-cigarette training data set

U: unigram *B: bigram* *N: negation* *L: length of tweets* *P: lexical feature, POS tag*
LR: LogisticRegression *LS: LinearSVM* *ET: ExtraTrees* *NB: MultinomialNB*
a: accuracy *p: precision* *r: recall*

10 Fold Cross Validation
Training datasets

Excluding Terms with Document Frequency Lower than Threshold

	# of features	time (sec.)	asthma relevant			asthma irrelevant		# of features	time (sec.)	e-cigarette relevant			e-cigarette irrelevant	
	Unigram		a	p	r	p	r	Unigram		a	p	r	p	r
min_df: 0%	5564	3.16	0.89	0.67	0.60	0.92	0.94	4212	18.92	0.89	0.87	0.85	0.90	0.91
min_df: 3%	**60**	**0.50**	**0.85**	0.56	0.43	0.89	0.94	**54**	**0.79**	**0.86**	0.85	0.81	0.87	0.90
min_df: 6%	**29**	**0.38**	**0.86**	0.60	0.49	0.90	0.94	**23**	**0.49**	**0.83**	0.84	0.72	0.83	0.91
min_df: 9%	19	0.35	0.86	0.59	0.48	0.90	0.94	16	0.45	0.80	0.79	0.69	0.81	0.87
min_df: 12%	14	0.31	0.84	0.54	0.32	0.88	0.95	11	0.38	0.78	0.76	0.69	0.80	0.85
min_df: 15%	11	0.29	0.84	0.52	0.32	0.88	0.94	9	0.35	0.79	0.76	0.70	0.81	0.85
min_df: 18%	8	0.28	0.85	0.61	0.27	0.87	0.97	7	0.30	0.77	0.76	0.64	0.78	0.86
min_df: 21%	7	0.25	0.85	0.59	0.27	0.87	0.96	6	0.28	0.75	0.70	0.65	0.77	0.81

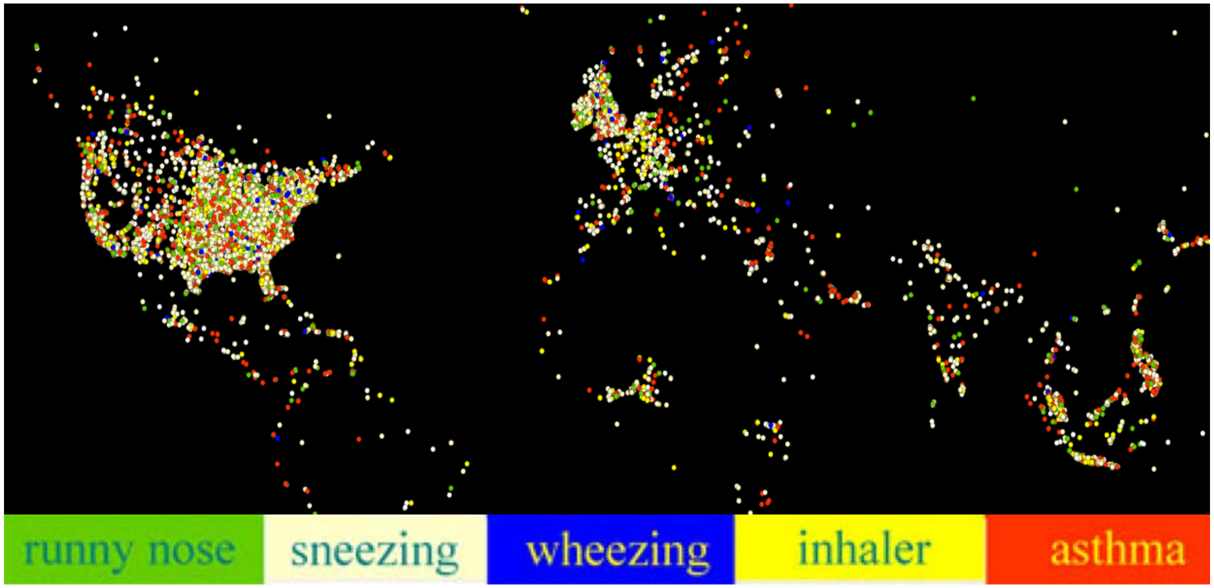
(a) Asthma training data set (LogisticRegression) (b) E-cigarette training data set (ExtraTrees)

a: accuracy *p: precision* *r: recall*

10 Fold Cross Validation / Training data set

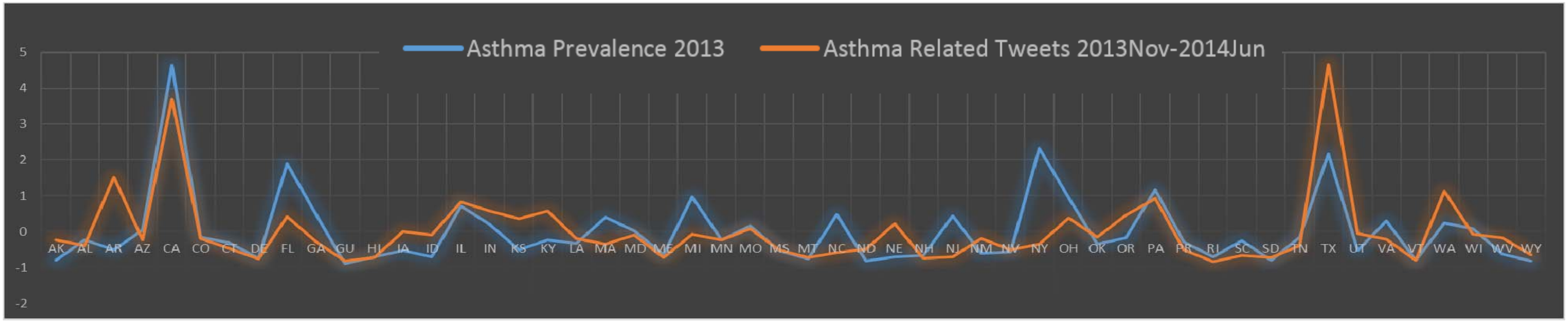
Ground Truth Based Evaluation: Geo-location Extraction

- 3.10% (171,165 / 5,513,368) of the tweets contained geographic coordinates
- 91.03% (5,019,319 / 5,513,368) tweets containing location information
- Identify 18.85% (63,093 /517,342) tweets as one of 50 US state names



8 months dataset

Ground Truth Based Evaluation: Asthma Prevalence Correlation



		After signal extraction	Before signal extraction
Asthma Prevalence 2013	Pearson Correlation	0.692**	0.303*
	N	50	50

**. Correlation is significant at the 0.01 level   *. Correlation is significant at the 0.05 level

8 months dataset

Domain Adaptation by Feature Augmentation

Source dataset	# of tweets	Category
Training dataset	1,850	• **News (1190 64%)** • **Ads (660 36%)**

Target dataset	# of tweets	Category

E-cigarette training dataset	3,149	• e-cigarette relevant
		  • First-person opinion (1,396 44%)
		• e-cigarette irrelevant
		  • **News (320 10%)**
		  • **Ads (1057 34%)**
		  • Other (376 12%)

Domain Adaptation by Feature Augmentation

Classifiers	First-person opinion			Other			News			Ads		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Logisticregression	0.719	0.716	0.718	0.318	0.665	0.431	0.557	0.397	0.463	0.552	0.562	0.557
Linearsvc	0.771	0.860	0.813	0.536	0.798	0.641	0.518	0.470	0.493	0.661	0.582	0.619
Multinomialnb	0.774	0.824	0.798	0.507	0.745	0.603	0.505	0.457	0.480	0.626	0.582	0.603
Perceptron	0.719	0.797	0.756	0.354	0.612	0.448	0.453	0.384	0.416	0.575	0.509	0.540
Decisiontree	0.717	0.796	0.754	0.347	0.625	0.446	0.468	0.368	0.412	0.585	0.540	0.561
Extratrees	0.770	0.857	0.811	0.526	0.798	0.634	0.516	0.464	0.488	0.651	0.577	0.612
Randomforest	0.710	0.802	0.753	0.502	0.654	0.568	0.421	0.362	0.389	0.603	0.574	0.588

Without domain adaptation; target dataset; 10 fold cross validation

Classifiers	First-person opinion			Other			News			Ads		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Logisticregression	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	0.830	0.875	0.852	0.800	0.784	0.792	0.870	0.714	0.784	0.860	0.852	0.856
Multinomialnb	0.856	0.883	0.869	0.854	0.686	0.761	0.808	0.750	0.778	0.819	0.880	0.848
Perceptron	0.905	0.820	0.861	0.792	0.824	0.808	0.880	0.786	0.830	0.831	0.907	0.867
Decisiontree	0.859	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	0.871	0.853	0.752	0.782	0.767	0.853	0.743	0.794	0.874	0.843	0.858
Randomforest	0.812	0.881	0.845	0.754	0.755	0.754	0.853	0.679	0.755	0.866	0.823	0.844

With domain adaptation; 10 fold cross validation

- Background
- Methodology
  - Text preprocessing
  - Feature extraction
    - Feature reduction
    - Feature generation
- Classification
- Domain Adaptation
- Experiments & results
- **Implications & contributions**

Contributions & Implications

- Contributions
  - New framework to extract signal from social media text
  - Accurate / timely / economical
  - Robust to overfitting
  - Applied in different domains

- Implications
  - Generating robust social media datasets for a variety of purposes
  - Development of various types of predictive models.

Future Work

- Population biases vary across different social media platforms
  - Teenagers and young adults
  - Gender bias

- Topic embedding

Thank you.