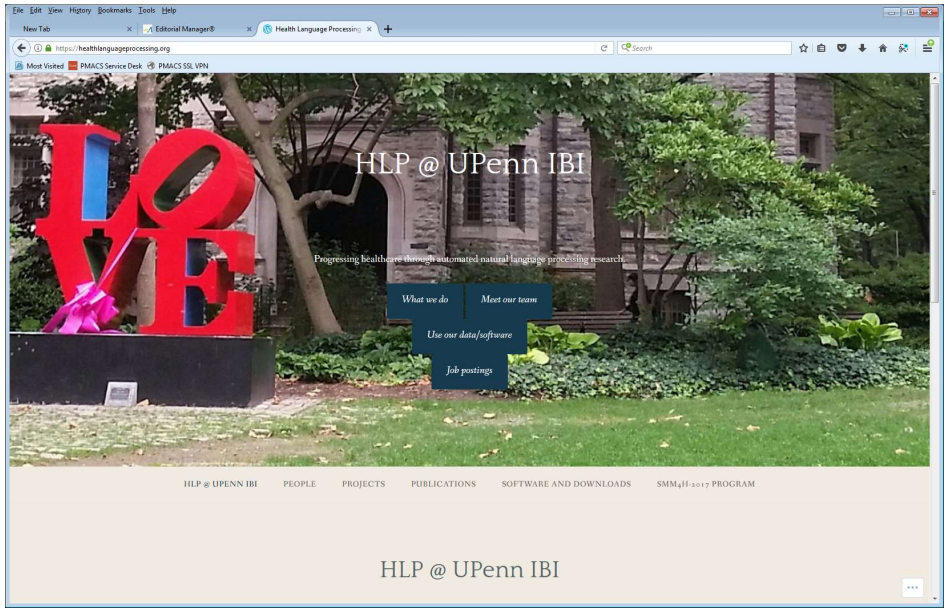
Natural Language Processing for Health (HLP)

November 2018

Tweet @UPennHLP #HLPMeeting



https://healthlanguageprocessing.org

Graciela Gonzalez-Hernandez

Contact: gragon@upenn.edu

Twitter: @gracielagon

Institute for Biomedical Informatics

Perelman School of Medicine
University of Pennsylvania

Program & Speaker List

- **Welcome/Introduction –** Abeed Sarker, Ph.D.
- **KEYNOTES** ***(15 minutes + 15 minutes for questions each)***

Twitter and Opioids: Correlations between Opioid-related discourse in Twitter and Regional Overdose Deaths
Rachel Graves, MD
HUP Emergency Medicine Resident, PGY1

Automatically Detecting Self-Reported Birth Defect Outcomes on Twitter for Large-scale Epidemiological Research
Ari Klein, Ph.D

Health Language Processing Lab – Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania

Opioids in the Twittersphere

Rachel Graves MD
Emergency Medicine Resident (PGY1)
University of Pennsylvania

The Team

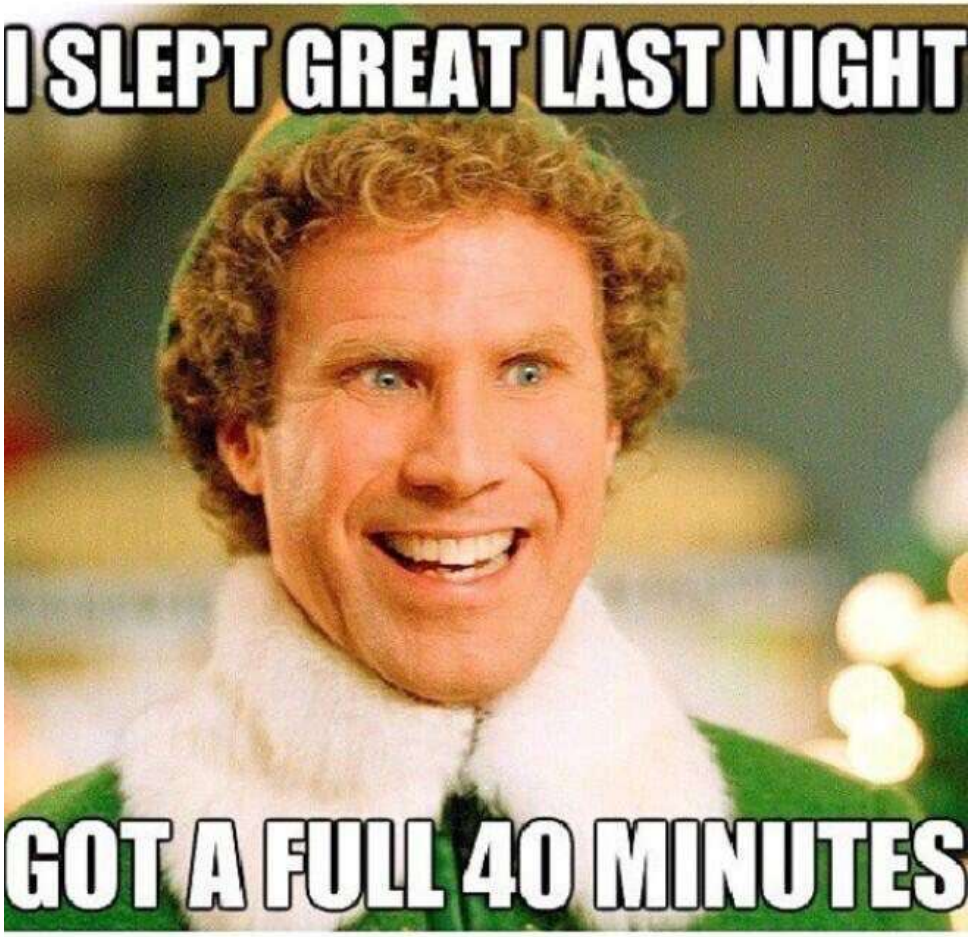
- **My supervisors and co-authors**—Raina Merchant MD MSHP FAHA, Zachary Meisel MD MPH MSHP, Dan Polsky PhD MPP

- **Computer and data scientist co-authors**—Christopher Tufts MS, Lyle Ungar MS PhD

- **Grant funding**—Philadelphia Department of Public Health

- **Special thanks** to Jeanmarie Perrone MD

Disclaimer



The Approach

- Goal: Determine whether Twitter data can be used to identify geographic differences in opioid-related discussion and whether opioid topics were significantly correlated with opioid overdose death rate

- Method summary:
  - Filter tweets (10 billion!) for "opioid keywords" (7/09-10/15)
  - Generate thematic topics (50) using *Latent Dirchlet Allocation (LDA)*, a machine learning analytic tool
  - Correlate topic distribution with census region, census division, and opioid overdose death rate

Key Findings

Unique opioid-related topics were significantly correlated with different Census Bureau divisions and with opioid overdose death rates at the state and county level.

- Drug-related crime, language of use, and online drug purchasing emerged as themes in various Census Bureau divisions.

- Drug-related crime, opioid-related news, and pop culture themes were significantly correlated with county-level opioid overdose death rates*, and online drug purchasing was significantly correlated with state-level opioid overdose death rates*.

  *CDC WONDER database based on ICD-9 and ICD-10 codes

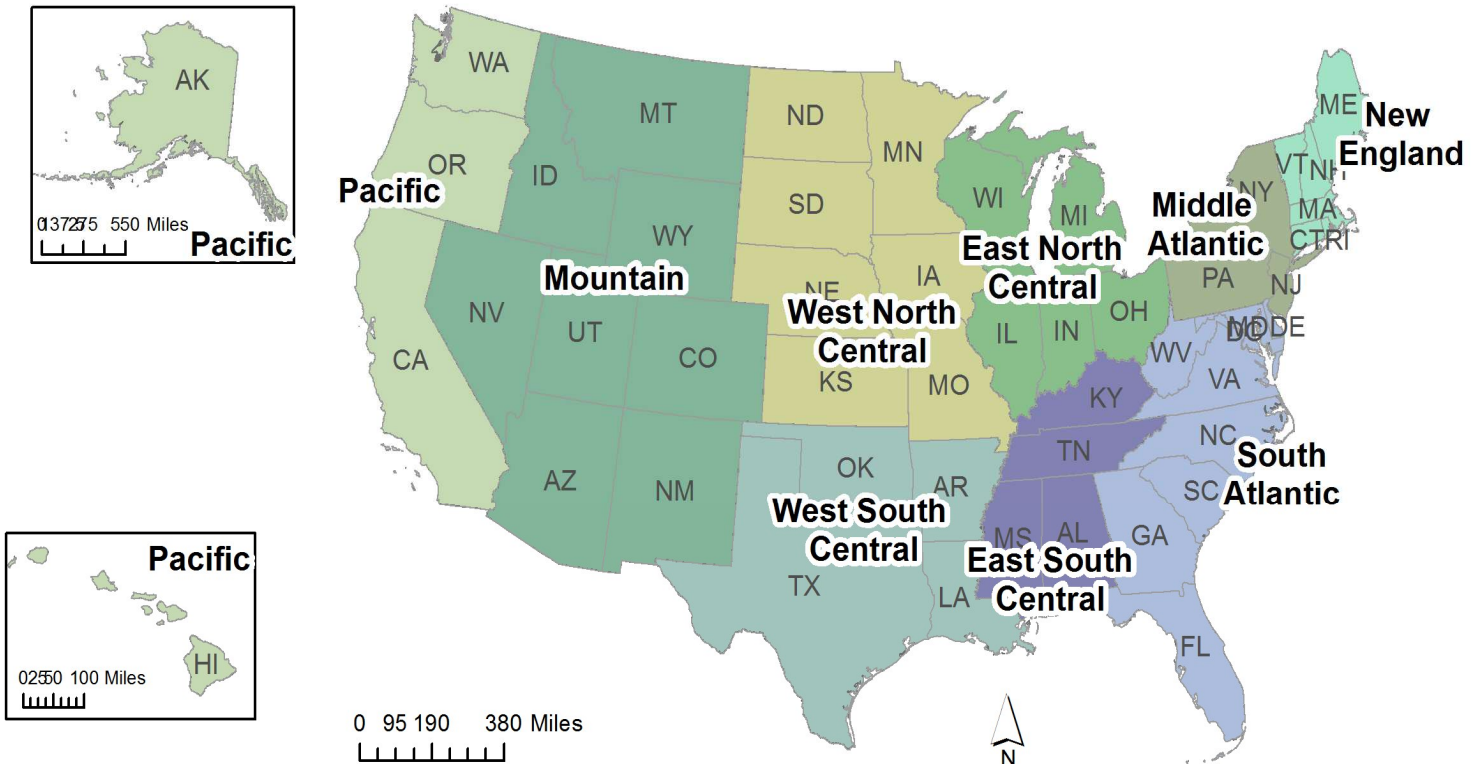
ASIDE: A FEW KEY TERMS

Latent Dirichlet Allocation (LDA)

- LDA: a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar

- For example, if observations are words (84k tweets) collected into documents it posits that each document is a mixture of a small number of topics (50) and that each word's presence is attributable to one of the document's topics.

Census Bureau Geographic Divisions

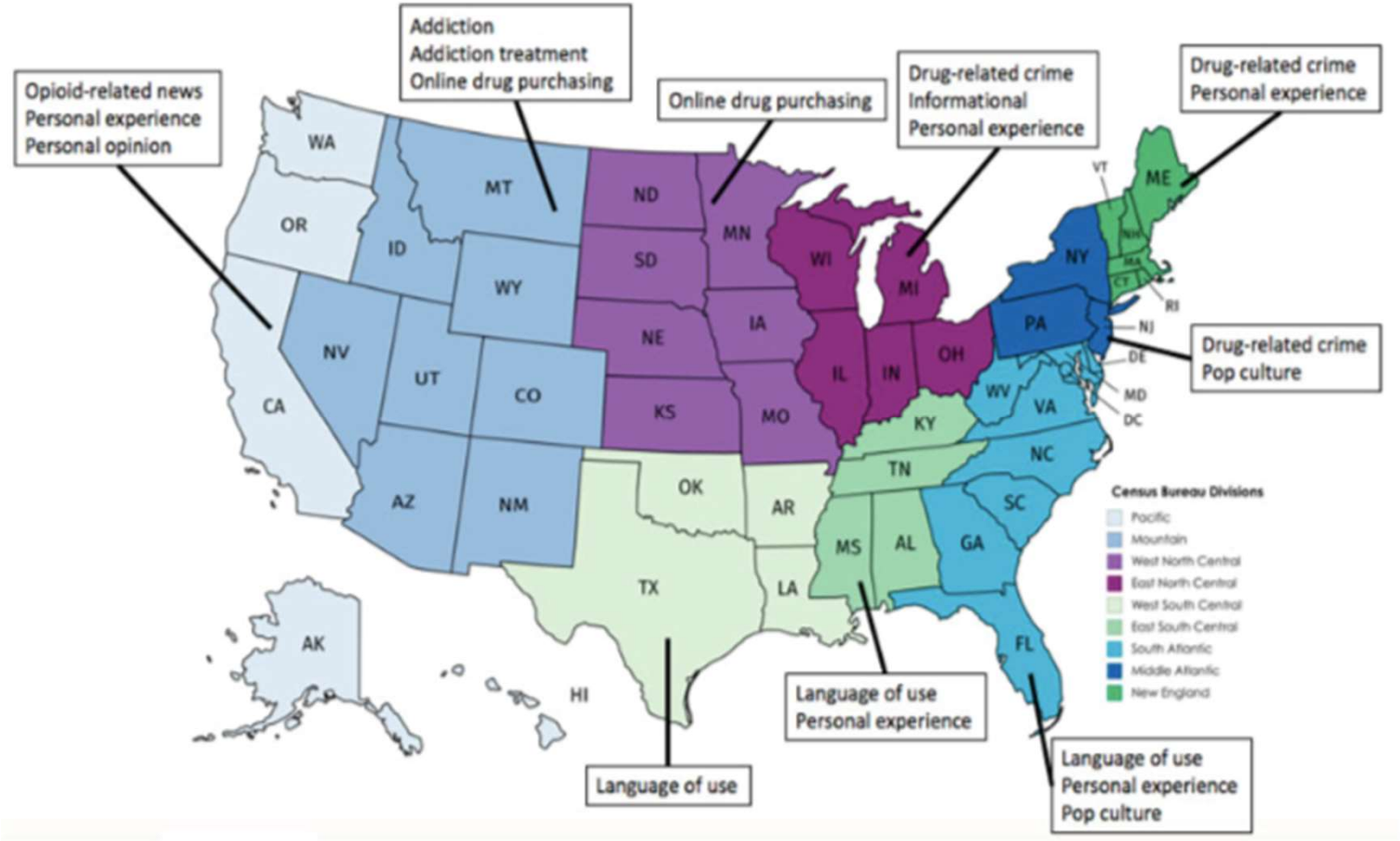


RESULTS

Generating Themes

Theme	Example topic
Addiction	
Addiction treatment	
Drug-related crime	

Correlating Themes with Census Bureau Divisions



Correlating Themes with Overdose Death Rates

Geographic division	Theme (r value)	Example topic
County	Drug-related crime ($r = 0.331$)	
State	Online drug purchasing ($r = 0.449$)	

Conclusion

Linguistic themes from Twitter are significantly correlated with Census Bureau divisions and with county- and state-level opioid overdose death rates. Content of opioid-related topics from Twitter may offer important insight into the drivers and consequences of opioid misuse in different areas.

Automatically Detecting Self-Reported Birth Defect Outcomes on Twitter for Large-scale Epidemiological Research

Ari Z. Klein, Abeed Sarker, Davy Weissenbacher, Graciela Gonzalez-Hernandez

Ari Z. Klein, Postdoctoral Researcher

Email: ariklein@pennmedicine.upenn.edu

Health Language Processing Lab (https://healthlanguageprocessing.org)

Department of Biostatistics, Epidemiology, and Informatics

Nov. 1, 2018 | HLP Special Interest Group

Penn Medicine

Background

- **Birth defects are the leading cause of infant mortality in the United States.**
- **The causes of the majority of birth defects remain unknown.**
- **Methods for observing pregnancies with birth defect outcomes remain limited.**
  - Pregnant women are largely excluded from clinical trials.
  - Data from animal reproductive studies may not translate to human risk factors.
  - Pregnancy exposure registries have a variety of limitations.
- **36% of Americans between ages 18-29 use Twitter.**

Related Work

- **Sarker A, et al. Discovering cohorts of pregnant women from social media for safety surveillance and analysis. *Journal of Medical Internet Research* 2017;19(10):e361.**
  - Developed an NLP and machine learning pipeline for automatically collecting and storing the publicly available tweets of users who have announced their pregnancy on Twitter.
- **Klein AZ, et al. Social media mining for birth defects research: A rule-based, bootstrapping approach to collecting data for rare health-related events on Twitter. *Journal of Biomedical Informatics* 2018;87:68-78.**
  - Identified 195 Twitter users whose pregnancies with birth defect outcomes could be observed via their publicly available tweets.
- **Golder S, et al. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. *Drug Safety* 2018;in press.**
  - Studied the identified cohort and a comparator group, and found evidence on Twitter that taking medication during pregnancy is associated with a higher risk of birth defect outcomes.

Objective

- **To develop methods for automatically detecting tweets by mothers reporting that their child has a birth defect**
  - The relatively small Twitter cohort initially identified prevented our epidemiological study from focusing on individual birth defects.
  - Additional Twitter users are being constantly added to our pregnancy database over time.
  - Exploiting social media on a scale potentially large enough for studying individual birth defects depends on methods capable of automatically identifying cohorts.

Corpus

- **22,999 annotated tweets that mention birth defects**
  - 1,192 (5.12%) “defect” tweets
    - Refer to a person who has a birth defect and identify that person as the user’s child
      - Ex.: “My little miracle, we are so blessed to have you **#hypoplasticleftheartsyndrome #hlhs**”
  - 1,196 (5.20%) “possible defect” tweets
    - Ambiguous about whether a person referred to has a birth defect and/or is the user’s child
      - Ex.: “He was born with **hypospadias** that fixed itself so he's going to get circumsized in 2 weeks. 😬😬😬”
  - 20,611 (89.67%) “non-defect” tweets
    - Do not refer to a person who has or may have a birth defect and is or may be the user’s child
      - Ex.: “Can’t watch this **cleft palette** infomercial, I can’t wake up depressed”
- **Inter-annotator agreement: κ = 0.86 (Cohen’s kappa)**

Supervised Classification

- **Trained and evaluated NB, SVM, and Bi-LSTM RNN classifiers**
- **Pre-processing and features**
  - NB and SVM
    - Lowercase, stem, remove non-alpha characters, normalize user names, URLs, birth defects, and people's names; word n-grams
  - Bi-LSTM
    - Word embeddings learned from GloVe word vectors trained on 2 billion tweets; tweets pre-processed the same as word vectors
- **Data-level approaches for class imbalance**
  - Under-sampling majority ("non-defect") class
    - Removed lexically similar "non-defect" tweets in the training set
    - Removed "non-defect" tweets lexically similar to false negative tweets in the validation set
    - Random under-sampling controls
  - Over-sampling minority ("defect" and "possible defect") classes
    - Over-sampling with replacement
    - Synthetic Minority Over-sampling Technique (SMOTE)

Results

Classifier	Training Set	F (+)	F (?)	F (-)
NB	original, imbalanced training set (14,716)	0.35	0.25	0.89
SVM	original, imbalanced training set (14,716)	**0.62**	**0.52**	0.96
SVM	under-sampling based on similar majority class tweets (5,551)	0.62	0.43	0.96
SVM	random under-sampling control set (5,551)	0.62	0.49	0.96
SVM	under-sampling based on similarity to false negative tweets (8,015)	0.58	0.51	0.95
SVM	random under-sampling control set (8,015)	0.62	0.50	0.96
SVM	over-sampling instances of minority classes with replacement (40,675)	0.62	0.46	0.95
SVM	SMOTE on original training set (39,148)	0.62	0.51	0.96
Bi-LSTM	original, imbalanced training set (14,716)	0.60	0.35	0.96
Bi-LSTM	under-sampling based on similar majority class tweets (5,551)	0.55	0.33	0.91
Bi-LSTM	random under-sampling control set (5,551)	0.54	0.37	0.92
Bi-LSTM	under-sampling based on similarity to false negative tweets (8,015)	0.48	0.36	0.90
Bi-LSTM	random under-sampling control (8,015)	0.59	0.45	0.95
Bi-LSTM	over-sampling instances of minority classes with replacement (40,675)	0.55	0.45	0.95

- **NB: Outperformed by SVM and Bi-LSTM for all three classes**
- **SVM: Under-/over-sampling did not improve performance for any of the classes.**
- **Bi-LSTM: Most under-/over-sampling improved performance for the "?" class, but similarity-based under-sampling methods did not outperform their controls. Overall, Bi-LSTM did not outperform SVM.**

Conclusions and Future Work

- Our automatic NLP and classification methods are the first step towards scaling social media for birth defects research.
- Deep neural network-based classifiers are outperformed for imbalanced social media data.
- Over-sampling methods for addressing class imbalance with CNNs may not generalize to RNNs.
- Future work will focus on automating user-level analyses for cohort inclusion.