Predicting Opioid Overdose Crude Rates with Text-Based Twitter Features (Student Abstract)

Nupoor Gandhi,[1] **Alex Morales,**[1] **Sally Man-Pui Chan,**[2] **Dolores Albarracin,**[2] **ChengXiang Zhai**[1]
[1]Dept. of CS, University of Illinois at Urbana-Champaign, IL, USA
[2]Dept. of Psychology, University of Illinois at Urbana-Champaign, IL, USA
{nupoorg2, amorale4, sallycmp, dalbarra, czhai}@illinois.org

Abstract

Drug use reporting is often a bottleneck for modern public health surveillance; social media data provides a real-time signal which allows for tracking and monitoring opioid overdoses. In this work we focus on text-based feature construction for the prediction task of opioid overdose rates at the county level. More specifically, using a Twitter dataset with over 3.4 billion tweets, we explore semantic features, such as topic features, to show that social media could be a good indicator for forecasting opioid overdose crude rates in public health monitoring systems. Specifically, combining topic and TF-IDF features in conjunction with demographic features can predict opioid overdose rates at the county level.

Introduction

With the rise of opioid abuse in the US, there has been a growth of overlapping hotspots for opioid overdose-related deaths in Springfield, Boston, Fall River, New Bedford, and parts of Cape Cod (Conway and O'Connor 2016). A large part of population, including rural communities, is active on social media, and social media presents a common informal resource for at-risk individuals to express their struggles. Twitter has been proven adequate for prediction of not only suicide, influenza rates , and human immunodeficiency virus (HIV) rates (Young, Rivers, and Lewis 2014), but also prescription opioid abuse at the individual level (Sarker et al. 2016). We explore the predictive power of Twitter more generally with respect to opioid overdose crudes rate per county.

We develop a model with primarily text-based Twitter features to predict opioid overdose crude rates at the county level. Though our preliminary results are mixed, we speculate that this work could serve as a tool for municipal governments to detect changes in opioid overdose crude rates using social media as a real-time signal.

Feature Construction

For our baseline text-based feature set we used Term Frequency Inverse Document Frequency (TF-IDF) and a Bag of Words (BoW) representation of the Twitter data.

Our primary feature set were topic-based features. We use an unsupervised Latent Dirichlet Allocation (LDA) model (Blei, Ng, and Jordan 2003). The granularity of document representation will ultimately need to match that of the prediction task, so the most intuitive approach would be to pool all of the tweets for a given county into a single document, which would likely produce incoherent topics. In this work, we exploit more detailed information about the tweets similar to Morales et al (2018). We partition the tweets for a single county into "attribute documents", where attributes are at the county and author level. For example, in the author document representation, we pool multiple tweets for a given Twitter user to compose a single document. This prevents very vocal Twitter users from dominating the topic distribution for a given county. Considering that White et al. (2009) found factors associated with risk for opioid abuse included depression and socioeconomic factors, we also add a feature for the crude rate for gender and age scaled in the range $[0, 1]$.

Experiments

Dataset

We obtained county-level crude rates for opioid overdose crude rates from the Centers for Disease Control and Prevention (CDC) per 100,000 including only people aged 13 and older. Data from regions with less than 10 cases per year or less than 100 inhabitants are routinely suppressed by the CDC, and this suppression criteria were also applicable for the present analysis.

Our Twitter corpus ranges from 2014 - 2016 with more than 3.4 billion tweets, including re-tweets. However in order to use this dataset at the spatial granularity of the overdose rates we geotagged our Twitter corpus to Federal Information Processing Standard Publication (county) codes, which uniquely identifies counties in the US. This resulted in CDC data and Twitter data for 352 counties in 2014, 402 counties in 2015, and 466 counties in 2016.

County level Prediction

Generally, in text-based prediction, data is leveraged to make a prediction of an interesting variable, which further helps support and optimize decision making. In our case, we

Feature Type	Classifiers											
	Linear SVM		MLP		MultinomialNB		Logistic Regression		Random Forest		Gradient Boost	
	Author	County	Author	County	Author	County	Author	County	Author	County	Author	County
LDA	0.472	0.522	0.583	0.416	0.416	0.416	0.554	0.548	0.609	0.606	0.662	**0.652**
BoW	0.563	0.554	0.562	0.48	0.421	**0.647**	0.540	0.599	0.565	0.582	0.601	0.601
TF-IDF	0.540	0.632	0.598	**0.677**	0.573	0.596	0.527	0.618	0.578	0.377	0.644	0.521
TF-IDF + LDA	0.541	0.41	0.576	0.571	0.419	0.509	0.527	0.534	0.611	0.483	**0.675**	0.329
BoW + LDA	0.563	0.566	0.584	0.571	0.426	0.428	0.541	0.538	0.555	0.538	**0.665**	0.637

Table 1: Prediction F-1 scores for author and county document representations, for six classifiers with our proposed text-based feature construction methods.

Feature Type	Classifiers		
	MultinomialNB	Random Forest	BernouilliNB
Age + Gender	0.517	0.602	0.318
TF-IDF + Gender	0.679	0.674	0.681
TF-IDF + Age	0.672	0.694	0.681
TF-IDF + Age + Gender	0.681	0.687	0.681
BoW + Age + Gender	0.681	0.692	0.676
LDA + Age + Gender	0.735	0.474	0.630

Table 2: Prediction F-1 scores for county document representations, for highest performing classifiers with our supplementary demographic features.

use a combination of text-based features derived from the Twitter data and the overdose rates from 2014-2015 to train and 2016 to test. Our baseline feature set is Term Frequency-Inverse Document Frequency (TF-IDF) and bag-of-words (BoW) matrix for the corpus. Using this baseline, we were able to compare how the addition of the topic features improved the predictive power of the feature-set. To select the best number of topics for the topic model, we experimented with the number of topics ranging from 50 to 300 by training and validating on the training corpus.

Considering that policy-makers would be interested in ranking counties in order of opioid overdose crude rates, we develop two class labels for each county based on whether the opioid overdose crude rate was lower or higher than the median rate in the training data. Then, we used several classifiers (Multinomial Naive Bayes Method (MultinomialNB), Neural network Multi-layer Perceptron classifer (MLP), Logistic Regression Linear Model, Random Forest Classifier, Linear Support Vector Machine (Linear SVM)) to predict labels for the counties in 2016.

Results

We use precision, recall, and F1 scores for the two labels to evaluate the prediction model. We saw that combining multiple text-based feature types did not improve performance necessarily (ex. LDA, TF-IDF). However, combining text-based features in conjunction with age and gender increased the F1 score.

One limitation is the lack of CDC data for opioid abuse rates, as for 2,630 counties, the overdose rate is missing for all of the years from 2014-2016. We observed that counties with fewer than the median number of authors were more likely to be predicted correctly with the author document representation as opposed to the county document representation. In the future, we plan to compare these counties with those that have a high number of authors to determine which document attribute scheme is best for this task. More generally, these results are subject to population-bias, in that counties with a high message count corresponded to stronger predictions.

We found a clear pattern across most classifiers that mental disease information improves the opioid overdose prediction task. Additionally, when we apply the same experiment methodology for opioid overdose crude rate prediction to mental disease prediction, we found that there was significant overlap between counties where our classifier predicted correctly. The following tweet, for example, would present a document that both prediction models performed well on.

> Side effects may include major mood swings and some cases severe depression -I really shouldnt have take that, Too bad I dont give a fu Keisha anymore, What a trip.

We defined "predictive" topics as topic features with higher weight for counties classified correctly, and upon closer inspection, we found that some of the top terms contained drug-related words, though the specific connotation of the words were ambiguous in context. In the future, the topics could be improved with semi-supervised models. We also found that demographic features dramatically improved our results, so we will expand the use of known indicators of opioid abuse rates for this task (ex. unemployment, access to healthcare).

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*.

Conway, M., and O'Connor, D. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology*.

Sarker, A.; O'Connor, K.; Ginn, R.; Scotch, M.; Smith, K.; Malone, D.; and Gonzalez, G. 2016. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter. *Drug safety*.

White, A. G.; Birnbaum, H. G.; Schiller, M.; Tang, J.; and Katz, N. P. 2009. Analytic models to identify patients at risk for prescription opioid abuse. *The American journal of managed care*.

Young, S. D.; Rivers, C.; and Lewis, B. 2014. Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes. *Preventive medicine*.