Predicting Localized, Fine-Grained Crime Types using Twitter

PI: Deepak Venugopal

Assistant Professor
317 Dunn Hall
Department of Computer Science
University of Memphis
Memphis TN 38152
dvngopal@memphis.edu
Phone: 901.678.1539

1 Abstract

Online social media is a massive information resource in the Big data era. IBM estimates that we generate 2.5 quintillion bytes of data each day, a large part of which is data generated through online social media. Twitter is one of the most popular forms of social media used by more than 100 million users, with more than 300 million messages being exchanged each day. Thus, *tweets* serve as a rich information resource, and due to the generality of the topics that are tweeted, it can be leveraged for several different types of predictive analytics. For instance, Twitter data has been successfully used to predict disease outbreaks such as the spread of flu within a community, predict election trends, etc. In this project, I propose to connect crime prediction to tweet patterns. Specifically, the topics in tweets corresponding to a certain geo-location can be leveraged to determine if that geo-location has a higher probability for specific types of crime. For instance, certain types of hate-crime can be analyzed through sentiments expressed in tweets, news agency tweets can be analyzed for crime reports, destructive or violent behavior can be analyzed through emotions expressed in tweets, etc. I will build an information extraction system to analyze text specified in tweets and automatically infer events that can help in fine-grained prediction of crime. Further, I will seek to combine this information with data from local law enforcement agencies to build a adaptive model that predicts the likelihood of crime in Memphis neighborhoods. This in turn will help the city of Memphis in allocating expensive law enforcement resources to those areas that are prone to specific types of criminal activity. Further, this will also allow the city of Memphis to take localized corrective/preventative actions based on the type of crime that has a higher likelihood in a particular area.

2 Methods and Goals

The purpose of this project is to build a framework that can automatically predict fine-grained crime types for various locations in and around Memphis. The main idea is to use text analytics on Twitter to continuously augment statistical models that predict the likelihood of different crime types such as robbery, assault, disruptive behavior, etc., in specific areas of the city.

Processing tweet data is well-known to be a huge challenge in the natural language processing community, for a multitude of reasons. Tweets are short, noisy, can be incomplete in terms of information and often use several abbreviations and jargon. The main technical challenge is to extract useful information from such noisy, short tweets that can be used to predict specific varieties of crime in a neighborhood. I propose to derive this information using Latent Dirichlet Allocation (LDA). Specifically, LDA is a probabilistic graphical model based approach that infers latent topics from documents. The main attractiveness of LDA is that it is an unsupervised topic modeling method, i.e., it does not require access to labeled topics, which is particularly useful in our case since it is hard to obtain annotated topics/events in tweets. Here, I will model clusters of tweets as defining a mixture model over our topics-of-interest, and infer the topics in these clusters using Bayesian inference methods. I will then develop a novel approach to combine the inferred topics with other features of the neighborhood including historical data, to come up with a joint probabilistic model that predicts crime in a neighborhood. Specifically, I will use inferred topics to re-calibrate prior distributions in the crime prediction model, and compute the overall likelihood for specific types of crime in neighborhoods around Memphis. Note that, since tweets are continuously updated, this predictive model is adaptive as compared to static models that are typically derived from purely historical data.

The main research questions that this project will address are (1) how can we extract reliable information from Twitter in a form that is useful for us in analyzing crime?, (2) how can we integrate this information with other features (e.g., historical data) to develop a joint predictive model for crime, and update this model dynamically?, and (3) how can we effectively deploy our system to communicate predictive results effectively (visualizations, alerts, etc.)?

To achieve the proposed project goals, I will build utilize my experience with fundamental inference and learning algorithms for probabilistic graphical models [7, 8, 9, 10, 11] , as well as my experience in utilizing probabilistic graphical models for core natural language processing applications including event extraction and multilingual event coreference resolution [3, 12].

3 Related Work

Utilizing social media data for analytics has been gaining widespread attention in several varied domains such as public health, disaster management, politics, etc. Sadilek et al. [5] used Twitter data to predict food-poisoning breakouts; Sakakil et al. [6] predicted earthquakes using time series information from Twitter. Culotta [2] used tweets to detect influenza epidemics; Ringsquandl et al. [4] used tweets for political analysis, etc. More recently, there have been attempts to adapt this in crime prediction/prevention analytics. Most notably, Wang et al. [13] proposed improving crime prediction through the use of Twitter feeds, with promising empirical results in the Chicago metropolitan area. In this project, I will build on their work, and integrate Twitter information within probabilistic models to obtain adaptive models that can predict crime at a fine-grained level in local neighborhoods.

4 Tasks

The main tasks that I will perform in this project are as follows.

Data Collection Using public Twitter API's, I will collect tweets, and pre-process them with the help of spatial and temporal information specified in the tweets.

Information Extraction I plan to use LDA to infer latent topics from noisy tweets. To minimize the effects of noise in tweets, I plan to use smoothing techniques by clustering users with similar profiles together and then infer topics on aggregated tweets.

Predictive Models I will use the inferred latent topics as priors in a probabilistic model for predicting crime. I will seek to work with the city of Memphis to obtain statistical or historical data on crimes. I will aggregate this data with the extracted information from tweets to construct a joint probabilistic model (using methods such as MaxEnt) that predicts the likelihood of a certain crime-type for a certain neighborhood in or around Memphis. I will also develop methods to adapt this model as we update our Twitter data.

Task	**Nov-Jan**	**Jan-April**	**May-June**
Data collection and pre-processing	X		
Information Extraction from Tweets	X	X	
Building Predictive Models		X	
Stacking predictive models with twitter information		X	
Evaluation			X
Deployment and visualization			X

Table 1: Timeline for completing the proposed research

Utilizing Predictions I will build visualization tools on top of the predictive model to display the predicted crime *hot spots* [1] for the city. This will allow uses to visualize crime-related information dynamically, since we can use new tweets to update the crime likelihood information in the map automatically.

A brief timeline for the activities of the proposed project is shown in Table 1.

References

[1] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, 21(1-2):4–28, 2008.

[2] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, pages 115–122, 2010.

[3] J. Lu, D. Venugopal, V. Gogate, and V. Ng. Joint inference for event coreference reolution. In *COLING*, 2016.

[4] Martin Ringsquandl and Dusan Petkovic. Analyzing political sentiment on twitter. In *AAAI Spring Symposium: Analyzing Microtext*, 2013.

[5] Adam Sadilek, Henry A. Kautz, Lauren DiPrete, Brian Labus, Eric Portman, Jack Teitel, and Vincent Silenzio. Deploying nemesis: Preventing foodborne illness by data mining social media. In *AAAI*, pages 3982–3990, 2016.

[6] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *In Proceedings of the Nineteenth International WWW Conference (WWW2010). ACM*, 2010.

[7] S. Sarkhel, D. Venugopal, T. Pham, P. Singla, and V. Gogate. Scalable training of markov logic networks using approximate counting. In *AAAI*, pages 1067–1073, 2016.

[8] D. Venugopal, S. Sarkhel, and K. Cherry. Non-parametric domain approximation for scalable gibbs sampling in mlns. In *UAI*, pages 745–755, 2016.

[9] Deepak Venugopal and Vibhav Gogate. On lifting the gibbs sampling algorithm. In *NIPS*, pages 1664–1672, 2012.

[10] Deepak Venugopal and Vibhav Gogate. Dynamic blocking and collapsing for gibbs sampling. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

[11] Deepak Venugopal and Vibhav Gogate. Giss: Combining gibbs sampling and samplesearch for inference in mixed probabilistic and deterministic graphical models. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 897–904, 2013.

[12] Deepak Venugopal, Chen Chen, Vibhav Gogate, and Vincent Ng. Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features. In *Empirical Methods in Natural Language Processing*, pages 831–843, 2014.

[13] Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown. Automatic crime prediction using events extracted from twitter posts. In *Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238, 2012.