Crime Prediction Using Twitter Sentiment and Weather

Xinyu Chen, Youngwoon Cho, and Suk young Jang
University of Virginia, xc7xn, yc5ac, sj2rh@virginia.edu

Abstract - **Social networking services have the hidden potential to reveal valuable insights when statistical analysis is applied to their unstructured data. As shown by previous research, GPS-tagged Twitter data enables the prediction of future crimes in a major city, Chicago, Illinois, of the United States. However, existing crime prediction models that incorporate data from Twitter have limitations in describing criminal incidents due to the absence of sentiment polarity and weather factors. The addition of sentiment analysis and weather predictors to such models would deliver significant insight about how crime. Our aim is to predict the time and location in which a specific type of crime will occur. Our approach is based on sentiment analysis by applying lexicon-based methods and understanding of categorized weather data, combined with kernel density estimation based on historical crime incidents and prediction via linear modeling. By testing our model's ability to predict future crime on each area of the city, we observed that the model surpassed the benchmark model, which predicts crime incidents using kernel density estimation.**

Index Terms - Crime prediction, Kernel density estimation, Twitter sentiment analysis, Weather

INTRODUCTION

Automatic crime prediction is a key technique used to identify the optimal patrol strategies for police departments. In order to maximize the allocation of scarce resources, advancements in crime prediction are required. However, conventional crime prediction techniques have some drawbacks in capturing crime incidents due to the lack of criminal predictive factors in the statistical model. There are multiple factors that could affect future crime incidents other than crime hot spots. Nevertheless, a deeper investigation of criminology is necessary for selecting the possible contributing factors for crime incidents. The conventional crime predictive models are lacking in two specific areas of study. First, they ignore the sentiment of social media content. Secondly, the models do not consider the environmental conditions of incidents of interest. Social media content has rich contextual information about its users' daily activities based on posted textual data. Each textual post is compiled to the platform of service providers. These threads or posts generate an unstructured form of data. Therefore, data generated by social media can be considered as a strong tool for crime prediction. To develop the advanced crime prediction model, we select Twitter to supply data. Previous research has been established on using Twitter data and our work is the extension of these previous works.

In addition to Twitter data, we considered weather factors as the environmental factors that may affect the occurrence of criminal incidents. Weather factors, especially temperature, were discovered to be one of the significantly influential factors that lead a person to have aggressive behaviors. Anderson investigated crime incidents in two major cities to track the correlation between temperature and the frequency of violent crimes. In his paper, the incidence of violent crime has positive linear relationship with temperature of the day [1]. We collected Twitter data from official Twitter Streaming API. Along with the Twitter data, we also obtained weather data from the Weather Underground website to build a crime prediction model that has sentiment polarity and weather factors to make an accurate prediction on crime incidents using linear modeling.

PROBLEM STATEMENT

Automatic crime prediction is one of strongest tools for maximizing the allocation for scarce resources for preventing crime. However, conventional crime prediction models that employ Twitter data have limitations on describing the real time reflection of criminal incidents. Polarities of sentiment and possible weather factors have the ability to improve the accuracy and maximize the predictive power of crime models. In order to achieve our goal, we set four objectives: (1) to analyze textual content in Twitter data by using sentiment analysis to score the positivity/negativity of tweets and their trends in different neighborhoods. (2) to identify the weather factors that serve as significant indicators to predict certain crime incidents. (3) to employ kernel density estimation (KDE) to derive the distribution of crime incidents in the Chicago area. (4) to build and evaluate the predictive model by adding Twitter-derived data and weather factors to criminal incident data.

RELATED WORK

I. Crime Forecasting Using Weather Data

The past studies of aggressive behavior on an uncomfortable days demonstrated clear correlations between weather and criminal activities. The studies explained that it was

978-1-4799-1832-4/15/$31.00 (c) 2015, IEEE

predicted that violent crimes would be more prevalent during the hotter quarters of the year and in hotter years [2]. From the psychological aspect of the human being as a decision maker, the designated actor would not respond to certain situations logically. In fact, the actor would act irrationally, affected and controlled by his surroundings. The previous studies that discuss the relationship between crime incidents and weather factors only used statistical inference approach. The studies are more focused on how independent variables, such as temperature, humidity, and other weather factors, contribute to crime incidents, reporting how an increased or decreased range of each factor related to the type of crime that increased or decreased. Previous studies of weather and crime incidents relationship model only focused on singular weather factors, but rarely treated weather factors as a set of explanatory variables.

II. Sentiment Analysis

The rapid growth in the volume of users in social network services has provided the predictive ability in extensive fields, in which allow us to predict the reaction in selected public groups. Examples of predictive modeling based on social media contents are election results [3], the box office performance of movies [4], product sales [5], and stock market trends [6]. These researches primarily use the technique of sentiment analysis. Researchers employ semantic analysis on the contextual contents of each tweet and draw the predictive response of the selected group of people. However, the previous researches are deficient in the prediction model for a wild range of population. These studies only collected data from selected groups of people to predict their future response in certain circumstances. This limited information based on certain groups of people does not portray the overall response towards criminal incidents.

Since other research studies have addressed limitations on representing the response of the population to criminal incidents, we approached from another angle to reduce the restricted predictive information regarding the selectiveness of people. Other researchers developed criminal prediction models using topic modeling on Twitter data. Previous study of Wang et al only focused on the tweets from new agencies to find the correlation between topics, which used in the tweets, and specific types of criminal incidents [7]. Gerber et al, further developed the prediction model via topic modeling [8]. He combined historical crime incidents with GPS-tagged Twitter data, which was collected from all twitter users within the Chicago city area. However, these models only considered topic modeling, but did not apply sentiment analysis on Twitter data. Meanwhile, they also did not explore weather factors. These factors may have influence on crime in combination with Twitter messages.

In our study, we addressed these limitations. We applied sentiment analysis to evaluate the polarity of tweets. We also integrated weather factors into the model with sentiment polarity and historical crime record as explanatory variables. Taking full advantage of all these features, we are able to develop more accurate prediction on future crime incidents.

Data Collection

The primary data source that we used for making a crime predictive model was Twitter data. We utilized tweets with GPS coordinates, which were generated within the Chicago city boundary from January 1st, 2014 to Jan 31st, 2014 (n=1069804). The twitter posts we used came from the official Twitter streaming API, bounded with coordinates [-87.94, 41.64] (South-West limit) and [-87.52, 42.02] (North-East limit) [9]. Figure 1 shows a kernel density estimation plot for tweets generated within the Chicago city boundary during the time period. In addition to Twitter data, Chicago criminal incidents data shows the historical trends of theft incidents occurred in Chicago. This data originated from the Chicago data portal website, which was developed by the Chicago Police Department by tracking theft incidents committed on spatial points indicated with specific latitude and longitude, and the time of the theft incidents [10]. These historical theft records within the boundary of the Chicago are from December 25th, 2013 to January 31st, 2014. The data contains 5395 theft incidents points. The last data we utilized was weather data. We collected this data from Weather Underground, which is one of the web sites that provide the history of weather forecast and the future forecast for Chicago [11]. The weather data contains information including minimum, mean, and maximum temperatures in Fahrenheit; dew point in Fahrenheit; humidity in %; sea level pressure in Inches; visibility per Miles; wind speed in Mph; cloud cover indicated by 1 to 8; and events. Events give us information about sunny, rain, snow, and fog events on daily basis.

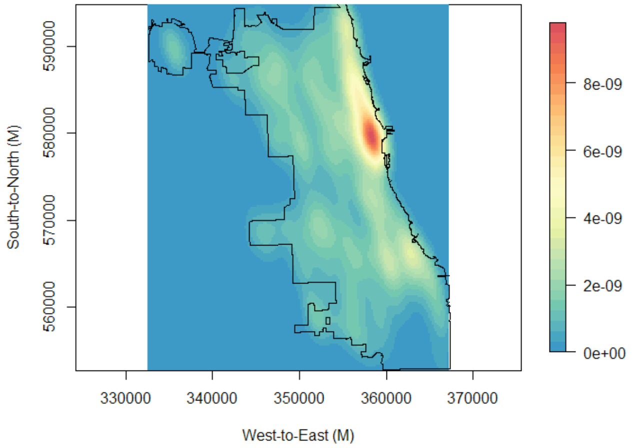

FIGURE 1
KDE For GPS-Tagged Tweets That Originated Within The City Limits Of Chicago Between January 1, 2014 And January 31, 2014

Modeling

To construct our crime prediction model, we first determine the training set of theft density, Twitter data, and weather data for 6-hour intervals in a day. This data contains the locations of occurrence in latitudes and longitude. From these geo-spatial points, we gained information on whether the designated regions have crime within 200-meter square. To begin sectioning the regional space across Chicago, we separate the Chicago area into multitude of smaller regional sectors that have length of 200 meter by 200 meter. Assigning all the crime points to the smaller sectors (neighborhoods) that we created a set of binary classifiers (crime or non-crime). After assigning binary classifiers to the designated regional sectors containing either crime points or no crime point, we computed polarity scores for each tweet within the boundary of Chicago. The scoring process is explained in the later work section below. Besides Twitter data, we modified weather data into an appropriate format. Each 6 hours, we assigned some of weather factors including temperatures and dew points, into 6-hour periods. For the time period between 12 a.m. and 6 a.m., we assigned minimum temperatures and dew points from our original raw data, and for the time period between 12 p.m. and 6 p.m., maximum temperatures and dew points were assigned. Except for these two specific time periods, we assigned mean values of temperature and dew points to the other time periods. Next, combining weather, twitter polarity score and its trend, we determined the explanatory variables in our training set as described below.

Response variable
$y_t(p)$ = (0 = non-crime points, 1= crime points)
Explanatory variable
$x_{1,t}(p)$ = Crime density
$x_{2,t}(p)$ = Polarity score of tweets (-1 to 1)
$x_{3,t}(p)$ = 3 –day Trend of polarity score
$x_{4,t}(p)$ = Temperature (Fahrenheit)
$x_{5,t}(p)$ = Dew points (Fahrenheit)
$x_{6,t}(p)$ = Mean humidity (%)
$x_{7,t}(p)$ = Mean sea level pressure (Inch)
$x_{8,t}(p)$ = Mean wind speed (MPH)
$x_{9,t}(p)$ = Precipitation (Inch)
$x_{10,t}(p)$ = Cloud cover (1 to 8)
$x_{11,t}(p)$ = Events (Sunny, Rain, Snow, Fog)

I. Kernel Density Estimation

Our crime densities for theft incidents were computed by KDE at a point p. The computational equation of KDE is following,

$$x_{1,t}(p) = k(p,h) = \frac{1}{Ph}\sum_{j=1}^{P} K\left(\frac{||p-p_j||}{h}\right) \tag{1}$$

In this equation, we calculate crime density at each point of p. In additional to the crime point of p, we set h as the bandwidth parameter that manages the smoothness of the KDE The capital letter P in our equation is the total number of crime incidents of theft that occurred in Chicago during the 6-hour periods in a day, j indicates a single crime point from the data of a time period, K is the standard normal density function, and ||•|| is the Euclidean norm. p_j is the actual crime point of theft in Chicago. We utilized the ks package in R statistical program to determine $k(p,h)$, and we also utilized the Hpi function for estimating bandwidth to acquire the value of h. This set of work is popularly utilized to estimate crime densities [12]. In our training theft density of KDE, we used 7 days (28 6-hour periods) of prior density to compute each KDE density for a specific date during the one-month period of data.

II. Information from Twitter Messages

The originality of this research lies in the combination of traditional kernel density estimate and weather features that describe the sentiment polarity and its trend of point p using Twitter content. We defined spatial sectors (neighborhoods) in Chicago by laying down evenly spaced cells 1000 meters on each side. Figure 2 below shows neighborhood sectors of 1000 meters by 1000 meters.

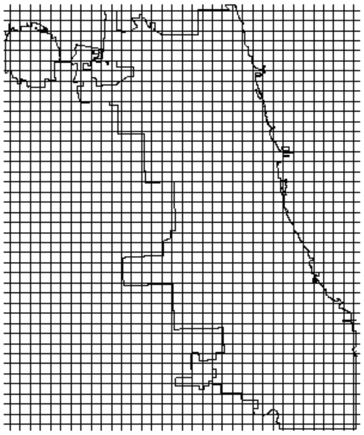

FIGURE 2
NEIGHBORHOOD BOUNDARIES FOR COMPUTING THE SENTIMENT POLARITY OF TWEETS

Given the neighborhood defined above, the problem becomes the estimation of the sentiment polarity values for each neighborhood given the tweets posted by users in each neighborhood during different 6-hour period of a day. We used a sentiment lexicon dictionary [13] and `polarity` function for this purpose.

In the polarity function, a context cluster of words is pulled from around a polarized word to be used as variant shifters. Each word in this context was labeled in four different categories including neutral, negator, amplifier, and de-amplifier. Each polarized word in each document was then calculated based on the scores of polarity. In order to compute the polarity, we assigned scores for words using the lexicon dictionary [13]. Words with negative connotations were assigned negative polarity scores, and words with positive connotations were scored as positive. After evaluating polarity scores of the documents, we then utilized amplifiers or de-amplifiers (default is .8, de-amplifier weight is constrained to -1 lower bound) to reweight the polarity score of each document. Finally, with the context cluster of words are cited, we divided them by the square root of the number of words in each document, which produced the polarity score for each document [14]. Furthermore, we paid attention to emoticons, since they might contain important information about the emotional states of users. In order to appropriately tokenize commonly used emoticons, we used the plyrDictionary package in R to replace emoticons with English words within the lexicon dictionary.

Besides sentiment polarity, we were also interested in the trend of sentiment within each neighborhood. Intuitively, consecutive periods of positive/negative sentiment might cause a higher risk of crime. In order to measure the trend, we created a trend index $x_{3,t}$(p). The inspiration of the algorithm comes from the evaluation of the trend of stock prices [15].

The index we created can measure the previews k periods trend of polarity. Firstly, let y_i be the percentage change of between previews k periods' polarity and today's polarity:

$$V_i = \left\{ \frac{P_{i-j} - P_i}{P_i} \right\}_{j=1}^{k} \tag{2}$$

Then, get the sum of the change over 10% a periods as an index $x_{3,t}$(p) for trend:

$$T_i = \sum_{v} \left\{ v \in V_i : v > 10\% \cup v < 10\% \right\} \tag{3}$$

Positive T signifies that the polarity of previews periods is bigger than today; while negative T means the polarity of previews periods is smaller than current period.

III. Logistic Regression

Our goal was to use the weather and sentiment feature we mentioned above to make predictions about future theft incidents. Specifically, we set a binary random variable y_{t+1}(p) to indicate whether crime incidents will occur in the next 6-hour period in the p neighborhood. Meanwhile, we derived a full model from logistic regression (4) by setting explanatory variables as we mentioned above.

$$\log\left(\frac{\Pr\left[y_{t+1}(p) = 1 \right]}{1 - \Pr\left[y_{t+1}(p) = 1 \right]} \right) = \beta_0 + \beta_1 x_{t,1}(p) + \cdots + \beta_{11} x_{t,11}(p) \tag{4}$$

Where parameter $\{\beta_0, \ldots, \beta_{11}\}$ can be estimated by maximum likelihood function using historical theft incidents.

For the prediction, we first calculated the estimated crime density of each neighborhood p on the period t' by using the KDE method we mentioned above. Then we processed tweets on the period t' to calculate the polarity score and trend index we mentioned above. The logistic regression model uses estimated density, twitter features and the forecasted weather factors to predict the likelihood of incidents occurring on period t'+1 in neighborhood p.

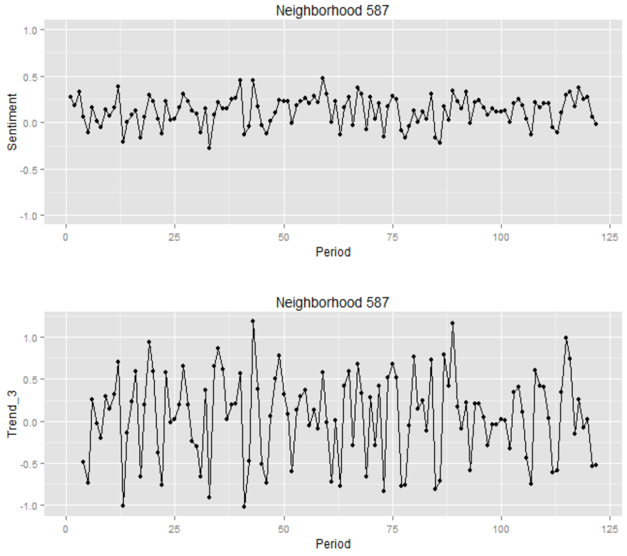

FIGURE 3
POLARITY SCORE AND 3-DAY TREND OF NEIGHBORHOOD 587 FROM JANUARY 1, 2014 TO JANUARY 31, 2014

EVALUATION

We evaluated our crime predictive model using Twitter sentiment analysis of January and effective weather factors by comparing its prediction to the actual theft incidents that occurred across Chicago, Illinois between December 25th, 2013 and January 30th, 2014. The Performance of logistic crime prediction model is measured by a surveillance curve.

With our training data set, we built a logistic regression to predict crime in the Chicago area. To avoid the influence of imbalanced response variable (0.07% of total response variable are crime points), we used ROSE package in R to under-sample the non-crime points [16]. For feature selection, the stepAIC backward elimination technique was used. The table below gives a summary of the information for our logistic regression model.

TABLE I
SUMMARY TABLE FOR LOGISTIC REGRESSION

Coefficients	Estimate	Std. Error	Probability (>|z|)
(Intercept)	-1.479e+01	5.870e+00	0.011724
Crime density	4.516e+08	2.070e+07	< 2e-16
Mean Humidity	-6.494e-02	1.170e-02	2.81e-08
Mean Sea level press.	5.782e-01	1.892e-01	0.002239
Mean Wind Speed	-4.515e-02	1.161e-02	0.000100
Cloud Cover	5.082e-02	3.094e-02	0.100519
Event Rain	1.085e+00	2.830e-01	0.000127
Event Snow	5.816e-01	1.748e-01	0.000879
Event Sunny	4.486e-01	1.886e-01	0.017353
Temperature F	3.828e-02	3.119e-03	< 2e-16
Trend_3	1.151e-01	4.956e-02	0.020217

Table I provides information on the variables that explain whether crime incidents occurred or not in every 200m*200m sector after feature selection. We can interpret our model based on the information. For instance, the coefficient of the 3-day trend indicates that for each unit increase of $x_{3,t}(p)$, the log-odds of theft crimes increases $1.151*10^{-1}$. Therefore, we can conclude that if there is an increasing trend of sentiment polarity, a higher crime rate could be expected. Meanwhile, in higher temperature, there is a higher chance of theft crimes, while low humidity lowers the chance of theft crimes.

In spite of the prediction model being significant based on our training data, it is necessary to evaluate the performance of our crime predictive model. In our crime prediction model, the performance of our model's prediction ability is measured by surveillance plot. The surveillance curve provides information that percentage about the actual crime captured per percentage of captured area [8]. The separated sectors in Chicago are sorted by their predicted crime density (p) from the highest density to its lowest. The function of the x coordinate and y coordinate could be written as:

$$\%\ Area\ Surveilled = x\% = \frac{\sum_{i=1}^{s} A_i}{\sum_{i=1}^{n} A_i} \quad (5)$$

$$\%\ Crime\ Captured = y\% = \frac{\sum_{i=1}^{s} c_i}{\sum_{i=1}^{n} c_i} \quad (6)$$

Where $\{A_n\}$ stands for the monitored area in Chicago sorted by p; $\{ci\}$ stands for actual crime with each monitored area; s stands for number of monitored area.

Meanwhile, a quantified measure of prediction performance can be derived from the area under the surveillance curve (AUC). The function of AUC can be written as:

$$AUC = \int_{-\infty}^{\infty} y(A) * x'(A) dA \quad (7)$$

We applied the model to predict theft incidents during the period of January 25th, 2014 to January 31st, 2014, which contained 964 incidents that occurred within the Chicago city area. Figure 4 shows the prediction performance via surveillance plot. Our prediction model (black line) has an AUC of 0.67, which means 0.67 of the crimes are covered under the monitored area. On the other hand, the benchmark logistic regression model (dashed line), that only contains crime density variable, has an AUC of 0.66. The difference of AUC between our model and benchmark model indicate only 1.5%. However, the difference of AUCs originated from the preceding portion of % Area Surveilled. In our plot the first 40% of Area Monitored exhibited significant difference between both AUCs of our model and the benchmark model. The above surveillance plot provided about 42% of crime captured by monitoring about 20% of area in the Chicago. The difference of AUCs is small, but the result is a satisfactory performance.

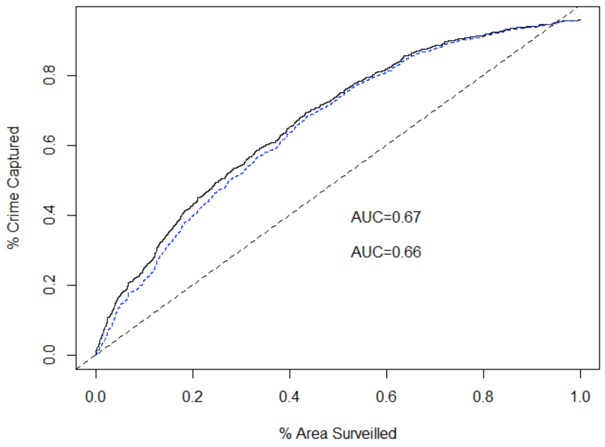

FIGURE 4
SURVEILLANCE PLOTS USING KDE-ONLY (DASHED-LINE) AND KDE + TWITTER + WEATHER MODEL (BLACK-LINE)

CONCLUSION/ FUTURE WORK

This paper has presented a preliminary investigation of the use of sentimental contents in social media and weather factors for criminal incident prediction. Though a few researchers have mentioned a possible relationship between crime and sentiment, the correlation has not been seriously verified. Moreover, very few of previous works investigate the combined effect of both sentiment and weather factors on crime. Our research has filled in the gap by proving the correlation between crime and predictors of weather and sentiment. Furthermore, our evaluation results demonstrated the model's ability to improve the forecast performance with the standard hot-spot (KDE) model on theft incidents. These results indicate not only potential benefits for the efficient allocation of policing resources but also a rewarding line for future research.

There are plenty of ways to extend this work. The most direct way to improve the predicting accuracy is to obtain weather forecast data in every 6-hour time period. For now, we can only retrieve daily weather forecasts from publicly accessible resources. Moreover, once we have access to spatially differentiated weather data for each different special sector, the predictive power of our model may improve. In addition, our hypothesis for this research focused on the correlation between Twitter sentiment and crime in general. However, we did not specify the connection between Twitter contents to specific kinds of crime. By detecting the topics that are highly correlated with positive and negative opinions, we might detect the correlation between those topics with sentiment polarity and the specific criminal activities [17]. Lastly, we used logistic regression for predictive modeling. Since the coefficient of sentiment polarity is insignificant in our logistic regression model, the better alternative might be to apply support vector machine or other advanced methods to verify the possible non-linear effect between polarity and crime incidents.

ACKNOWLEDGMENT

The authors would like to appreciate our advisor, Professor Mathew Gerber, for his kindly assist and practical feedback throughout the duration of our project. Additionally, the authors would like to thank Northrup Grumman Corporation for their support.

REFERENCES

[1] Anderson, C.A and Anderson, D.C. 1984. "Ambient temperature and violent crime:Tests of the linear and curvilinear hypotheses." *Journal of Personality and Social Psychology*, 46, pp. 91 – 97.

[2] Anderson, C.A. 1987. "Temperature and aggression: effects on quarterly, yearly, and city rates of violent and nonviolent crime." *Journal of Personality and Social Psychology,* pp. 1161 – 1173.

[3] Asur, S and Huberman, B. 2010. "Predicting the future with social media." *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, pp. 492 - 499

[4] Bermingham, A. and Smeaton, A. 2011. "On using twitter to monitor political sentiment and predict election results." *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology*, SAAIP, pp.2 – 10.

[5] Choi, H. and Varian, H. 2012. *"Predicting the present with Google Trends"*, Economic Record 88,pp. 1 - 8.

[6] Bollen, J. Mao, H. and Zeng, X. 2011. "Twitter mood predicts the stock market." *Journal of Computational Science.*

[7] Wang, Xiaofeng. Gerber, Matthew S. and Brown, Donald E. 2012. "Automatic Crime Prediction using Events Extracted from Twitter Posts." *Social computing, Behavioral – Cultural Modeling and Prediction Lecture Notes in Computer Science*, pp. 231 – 238.

[8] Gerber, Matthew S. May 2014. "Predicting Crime Using Twitter and Kernel Density Estimation" *Decision Support Systems*, Volume 61, pp. 115 – 125.

[9] "Twitter API" Twitter API. 2014. https://dev.twitter.com/rest/public. Accessed: January 31, 2015

[10] "Crime Data – 2001 to present" City of Chicago. 2015. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. Accessed: February 9, 2015.

[11] "Chicago, IL, weather data." 2015. http://www.wunderground.com/weather-forecast/US/IL/Chicago.html Accessed: February 11, 2015.

[12] Chainey, S and S. Uhlig, L. Tomposn. 2011. "The utility of hotspot mapping for predicting spatial patterns of crime." *Security Journal 21*, pp. 4 – 28.

[13] Hu, M and Liu, B. 2004 "Mining opinion features in customer reviews." *National Conference on Artificial Intelligence.*

[14] "Polarity Score (Sentiment Analysis)." 2015. http://www.inside-r.org/packages/cran/qdap/docs/polarity Accessed: March 20, 2015.

[15] Torgo, Luis. 2010. "Data Mining with R: Learning with Case Studies" Boca Raton: Chapman and Hall, pp. 107 – 109.

[16] G. King and L.Zeng, "Logistic Regression in Rare Events Data", University of Harward, Cambridge, MA, 2001, pp.137-163.

[17] Cai, Keke. Spangler, S. and Chen, Ying et al. 2008. "Leveraging Sentiment Analysis for Topic Detection" *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE.*

AUTHOR INFORMATION

Xinyu Chen, M.S. Student, Data Science Institute, University of Virginia

Youngwoon Cho, M.S. Student, Data Science Institute, University of Virginia

Suk young Jang, M.S. Student, Data Science Institute, University of Virginia

Mattew S. Gerber, Assistant Professor, Dept. of Systems and Information Engineering, University of Virginia