A MINI PROJECT REPORT

ON

"CRIME PREDICTION BASED ON TWITTER SENTIMENTS AND WEATHER"

B.Tech 5th Semester Information Technology

A Project By:

Anushree Goswami (LIT2016005)

Manisha Meena (LIT2016012)

Anjali Kumari (LIT2016033)

Shubhi Agarwal (LIT2016036)

Under the Guidance Of:

Dr. Vishal Krishna Singh

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, LUCKNOW

CONTENTS

- **Introduction**
- **Problem Statement**
- **Objective**
- **Related Work**
    a. **Crime Forecasting Using Weather Data**
    b. **Sentimental Analysis Using Twitter Data**
- **Data Collection**
- **Application**
- **Acknowledgment**
- **Reference**

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.

1.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, 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.

2. Problem Statement

a. 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.

b.There are many challenges to using Twitter as an information source for crime prediction. Tweets are notorious for (un)intentional misspellings, on-the-fly word invention, symbol use, and

syntactic structures that often defy even the simplest computational treatments (e.g., word boundary identification). To make matters worse, Twitter imposes a 140-character limit on the length of each tweet, encouraging the use of these and other message shortening devices. Lastly, we are interested in predicting crime at a city-block resolution or finer, and it is not clear how tweets should be aggregated to support such analyses (prior work has investigated broader resolutions, for example, at the city or country levels). These factors conspire to produce a data source that is not only attractive owing to its real time, personalized content but also difficult to process. Thus, despite recent advances in all stages of the automatic text processing pipeline (e.g., word boundary identification through semantic analysis) as well as advances in crime prediction techniques (e.g., hot-spot mapping).

c. Hot-spot maps are a traditional method of analyzing and visualizing the distribution of crimes across space and time. Relevant techniques include kernel density estimation (KDE), which fits a two-dimensional spatial probability density function to a historical crime record. These techniques are useful but carry specific limitations. First, they are locally descriptive, meaning that a hot-spot model for one geographic area cannot be used to characterize a different geographic area. Second, they require historical crime data for the area of interest, meaning they cannot be constructed for areas that lack such data. Third, they do not consider the rich social media landscape of an area when analyzing crime patterns.

d. 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.

3. Objectives

(1) Quantify the crime prediction gains achieved by adding Twitter-derived information to a standard crime prediction approach based on kernel density estimation (KDE).
(2) Identify existing text processing tools and associated parameterizations that can be employed effectively in the analysis of tweets for the purpose of crime prediction.
(3) Identify performance bottlenecks that most affect the Twitter-based crime prediction approach.
(4) Hopefully, Our research will fill in the gap by proving the correlation between crime and predictors of weather and sentiment and demonstrate the model's ability to improve the forecast performance with the standard hot-spot (KDE) model on crime incidents.

4. Related Work

a. Crime Forecasting Using Weather Data

The past studies of aggressive behavior on an uncomfortable days demonstrated clear correlations between weather and criminal activities [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.

b. Sentimental Analysis Using Twitter Data

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.

5. 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. The twitter posts we used came from the official Twitter streaming API. In addition to Twitter data, Chicago criminal incidents data shows the historical trends of theft incidents occurred in Chicago [7]. 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 [8]. 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. 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.

6. Applications

To predict various large-scale events like:
a. elections

b. infectious disease outbreaks
c. national revolutions
d. the essential hypothesis is that the location, timing, and content of tweets are informative with regard to future events

7. Progress

a. We have collected twitter data from the official twitter streaming API.
b. Chicago Criminal Incidence historical data is collected from Chicago data portal website.
c. History of Weather forecast for Chicago is gathered from weather underground.
d. Study of Implemented Algorithms:
    - Kernel Density Estimation (KDE)
    - Latent Dirichlet Allocation (LDA)
    - Natural Language Processing (NLP)
    - TF / IDF Matrix

8. Acknowledgment

We would like to thank our project mentor Dr. Vishal Krishna Singh for helping and guiding us throughout the process.

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] *Social Big Data: Recent achievements and new challenges Gema Bello-Orgaz, Jason J. Jung, David Camacho*
[8] *Predicting crime using Twitter and kernel density estimate Matthew S. Gerber*