Fall 2014

Crime pattern detection using online social media

Raja Ashok Bolla

Department:

CRIME PATTERN DETECTION USING ONLINE SOCIAL MEDIA

by

RAJA ASHOK BOLLA

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN COMPUTER SCIENCE

2014

Approved by

Dr. Sriram Chellappan, Advisor

Dr. Wei Jiang

Dr. Zhaozheng Yin

ABSTRACT

In this research, we show online social networks can be used to study crime detection problems. Crime is defined as an act harmful not only to the individual involved, but also to the community as a whole. It is also a forbidden act that is punishable by law. Crimes are social nuisances that place heavy financial burdens on society. Here we look at use of data mining followed by sentiment analysis on online social networks, to help detect the crime patterns. Twitter is an online social networking and microblogging service that enables users to post brief text updates, also referred to as "tweets". These updates can convey important information about the author. A filter was designed to extract tweets from cities deemed to be either the most dangerous or the safest in the United States (US). A geographic analysis revealed a correlation between these tweets and the crimes that occurred in the corresponding cities. Over 100,000 crime-related tweets were collected over a period of 20 days. Sentiment analysis techniques were conducted on these tweets to analyze the crime intensity of a particular location. This type of study will help reveal the crime rate of a location in real-time. Although the results of this test helped in detecting crime patterns, the sentiment analysis techniques did not always guarantee the proper results. We conclude with applications of this type of study and how it can be improved by applying media to text processing techniques.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Dr. Sriram Chellappan, without whom I would not have been able to complete this thesis. Despite his busy schedule, Dr. Chellappan was very patient in answering all of my questions and offering invaluable assistance. His innovative thinking, determination and critical reviews helped me to work on various interesting problems. He trusted my skills during our first meeting and gave me the opportunity to work on various projects. He has been a great advisor, and I will never forget him.

I would like to thank Dr. Wei Jiang and Dr. Zhaozheng Yin for taking the time to serve on my thesis committee.

None of this work would have been possible without the constant love and support I received from my family and friends. I am truly indebted to my Parents, Raghu Bolla and Vimala Bolla, my Uncles Satyanaraya Vegulla, Suryanarayana Kommireddy, Satyanarayana Arumilli, and Sudhakar Nunna for the sacrifices they made in helping me pursue an advanced degree at Missouri University of Science & Technology. Finally, I would like to thank my lab mates with whom I have had several interesting conversations.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

1. INTRODUCTION

National security concern is the primary goal of any nation. Criminology studies focus on identifying criminal characteristics. The application of data mining techniques can help with this identification. Crime analysis, a part of criminology, is a law enforcement function that involves the systematic analysis of identifying and analyzing both patterns and trends in crime and disorder.

In the current world, the criminals are becoming technologically sophisticated, often expressing their emotions on the web. The World Wide Web's phenomenal growth has resulted in more users expressing their opinions online. Customers use these opinions to buy a product, conduct market analysis, and so forth.

Twitter is one of the most popular online social networks to date, where users post their opinions in short text called "tweets". These tweets are typically limited to 140 characters. Twitter has approximately 500 million users; approximately 340 million tweets are sent every day. Twitter is used, primarily, for the following four reasons [1].

- Daily Chatter (e.g., status messages on what the user is doing)
- Conversations (e.g., tweeting to either a user or a group of users within a community)
- Sharing information (e.g., posting links to web pages)
- Reporting new (e.g., status updates on current affairs).

According to Bollen, "a tweet is a microscopic, temporally-authentic instantiation of sentiment" [2]. Since tweets are crispy and brief, the public sentiment can be easily explored. Twitter also provides the feature of Retweet (RT), which allows users to share content posted by another user.

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials. It is used to determine an author's attitude, with respect to a particular topic or the overall contextual polarity in the text.

The rapid growth of Social media has spurred interest in sentiment. Various forms of online expressions (e.g., opinions-like reviews, ratings, and recommendations) have become major sources of information for businesses looking to market their products and manage their reputations.

The challenge of detecting crime patterns lies in geographically analyzing crime-related tweets and then performing sentiment analysis to identify crime prone zones in nearly real-time. Most of the studies that focused on crime pattern detection [8, 9] used data mining techniques to better understand historic data. This study used online social media to detect crime prone areas in almost real-time.

1.1 GOALS

This work was conducted in an attempt to accomplish the following:

- Conduct geographic analysis of tweets within selected cities.
- Analyze certain city intensity by applying sentiment analysis techniques to collected tweets.

- Identify the applications needed for this type of study.

1.2 CONTRIBUTIONS

This study was conducted to better understand the crime intensity of a particular location, in almost real-time, through the online social media. As stated earlier in this paper, the existing studies draw the crime patterns using the historic data which lacks the real-time feasibility. As technology is growing rapidly, the data exchange can be done at a glance. Using the power of online social media, we believe this approach could be very useful in drawing patterns for crime detection.

The approach used in this study began with the identification of the top ten crime prone cities and the top ten safest cities in the United States as determined by Forbes [3, 4]. The tweets generated within certain geographical area around these cities were then collected. The data collection process ran for nearly 21 days; which resulted in over 100,000 tweets in our database. Geographic analysis is performed using the density of population in the respective cities. The results drawn from this phase matches the pattern mentioned in the *Forbes* articles.

Sentiment analysis was applied over the collected crime-related tweets to measure the crime intensity of a particular location. Both Stanford's Recursive Deep model [5] and the dictionary-based approach, using Affective Norms for English words (ANEW) [6, 7], were used to conduct the sentiment analysis technique. The sentiment obtained from these techniques was used to identify a location's intensity in almost real-time.

1.3 THESIS STRUCTURE

The following discussion is divided into four sections. A brief background on the crime distribution in the United States is given in Section-II. This section also includes the existing techniques that tried to detect crime patterns using the historic data and a brief background on the existing sentiment analysis techniques available for use, including machine learning and lexicon-based approaches. Section-III contains information on the data collection and data processing used in this approach. The results collected are presented in Section-IV. The entire paper is concluded in Section-V.

2. BACKGROUND

2.1 CRIME DISTRIBUTION

The main challenge behind crime data mining is to understand patterns in criminal behavior in order to predict crime and prevention. Any research that can assist in solving crimes is preferred to protect individuals. A number of studies examined data obtained from either a sheriff's office [8] or a Crime Analysis Unit [9]. Clustering and Series Finder algorithms, respectively, were applied to the data in an effort to predict crime. Twitter, a powerful online social network, was used in this study to detect crime in almost real-time. The top ten most dangerous cities in the United States, as listed by Forbes magazine, were chosen for examination; the top ten safest cities were also examined for comparison. A complete list of cities analyzed is given in Table 2.1.

Table 2.1. List of Cities and Their Central Geolocation

Rank	**Crime Cities**	**Geolocation (lat, long)**	**Safe Cities**	**Geolocation(lat, long)**
1	Detroit	42.352711, -83.099205	Plano	33.061262, -96.7366254
2	East St. Louis	38.6106505, -90.1125948	Portland	45.5424364, -122.654422
3	Oakland	37.7919584, -122.2287941	Honolulu	21.3280681, -57.7989705

Table 2.1. List of Cities and Their Central Geolocation (cont.)

4	Memphis	35.129186, -89.9696395	San Jose	37.2970155, -121.8174129
5	Birmingham	33.5312374, -86.850137	Omaha	41.2918589, -96.0812485
6	Atlanta	33.7677129, -84.420604	New York	40.7056308, -73.9780035
7	Baltimore	39.2847064, -76.6204859	Santa Ana	33.7380535, -117.887414
8	Stockton	37.9730234, -121.3018775	Anaheim	33.7380535, -117.887414
9	Cleveland	41.4949426, -81.70586	San Diego	32.8245525, -117.0951632
10	Buffalo	42.8962389, -78.854702	Glendale	33.6030034, -112.3064516

2.2 SENTIMENT ANALYSIS

Sentiment analysis was used to determine a writer's/speaker's attitude with respect to either a topic or the overall contextual polarity of a text. Researchers use this analysis to measure emotions in online texts. The rise of social media fueled interest in using sentiment analysis to identify public opinions and interests. Several open source software tools utilize machine learning, statistics, and natural language processing techniques to automate sentiment analysis on a large collection of texts that have been gathered from various sources.

Sentiment analysis is a two-step process that includes both subjectivity classification and sentiment classification. The term subjectivity classification [7], is defined as distinguishing factual sentences from those used to present opinions, before analyzing sentiments. Paragraphs that present facts are typically removed so the researcher can focus on those paragraphs in which the author expresses opinions. Both naive Bayes classification [10] and Cut Based classification [11] are used for subjectivity classification.

The term sentiment classification [7], is defined as detecting sentiment polarity of the subjective sentences. This sentiment classification is also divided into two categories: binary sentiment classification and multi-class sentiment classification. Binary sentiment classification involves classifying sentiments either positive or negative. Multi-class sentiment classification involves classifying sentiments into one of five categories: strong positive, positive, neutral, negative and strong negative.

The most common machine learning techniques used for sentiment classification include naive Bayes, maximum entropy, and support vector machine [12]. Most sentiment analysis algorithms use simple terms to express sentiment. However the cultural factors, linguistic nuances, and differing contexts prevent researchers from drawing the sentiment accurately.

2.2.1 ANEW Based Approach. The ANEW is being developed to provide a set of normative emotional ratings for a large number of words in the English language [6]. This was developed to aid researchers when studying emotions; it is often used to determine a tweet's sentiment [7].

Siddharth and Dr. Healey [7] adopted a dictionary-based approach for determining the sentiment of tweets. They used the ANEW dictionary to provide pre-existing, normative emotional ratings for 1034 words along the three dimensions of valence, arousal and dominance. They used an independent matching technique to map all of the words in a tweet that were found in ANEW. They used two approaches (the arithmetic mean and normal distribution) to calculate both the mean valence and arousal.

These values were then plotted in a 2D emotional circumplex model. The tweet emotion was determined its position within the model. They used the model proposed by Russell and Barett [13] which is shown in Figure 2.1.

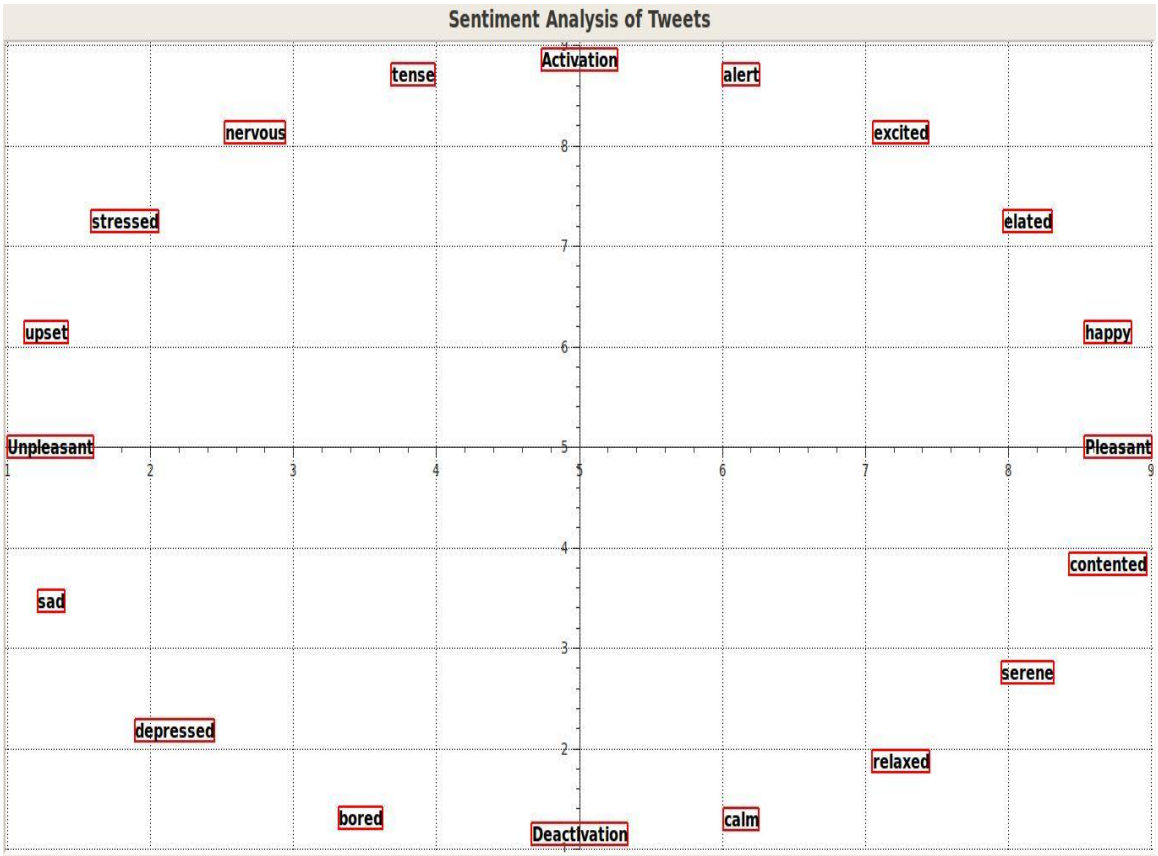

Figure 2.1. Semantic Structure of Emotional Model

2.2.2 Deep Learning for Sentiment Analysis. Richard, Alex, Jean, and Jason [5] introduced the Recursive model, a state-of-the-art in sentiment analysis. They also introduced both the Recursive Neural Tensor Networks (RNTN) and the Stanford Sentiment Treebank. The Treebank includes fine-grained sentiment labels for over 200,000 phrases in the parse trees of over 11,000 sentences. When the RNTN model is trained on the new Treebank, it outperformed all previous methods on several metrics. This approach follows the multi-class sentiment classification; it predicts five sentiment classes: very positive, positive, neutral, negative and very negative. The sentiment prediction's accuracy can reach 80.7%. An example of the RNTN accurately predicting

five sentiment classes at every node of a parse tree is given in Figure 2.2. It also captures the negation and its scope in the sentence.

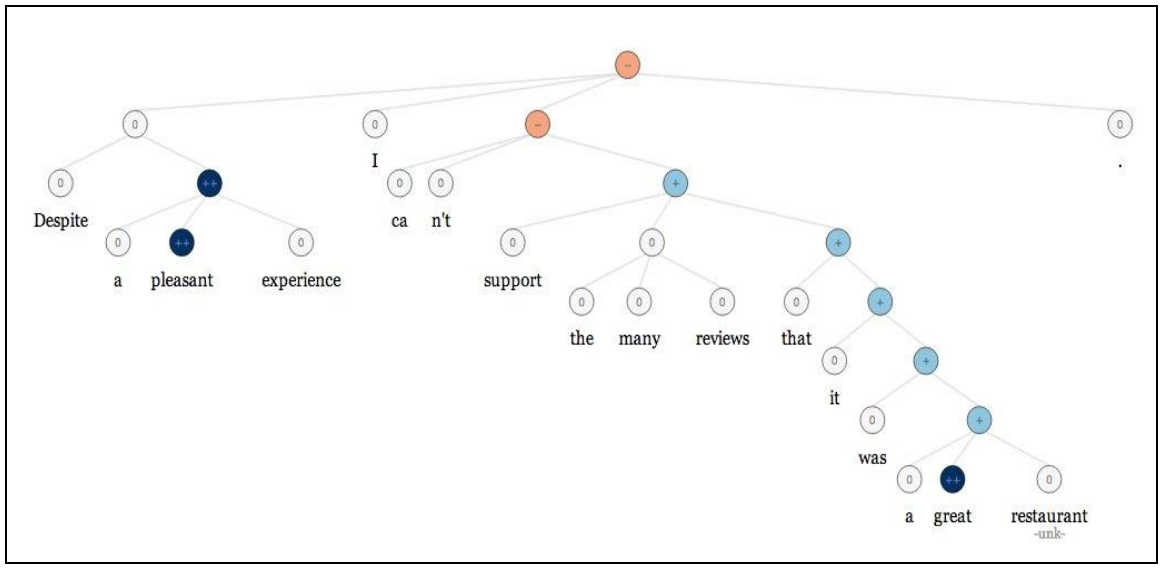

Figure 2.2. Example of the Recursive Neural Tensor Network

RNTN proposed here [5], takes the input phrases and represent through word vectors and a parse tree, then compute vectors for higher nodes in the tree using the same tensor based composition function. Unlike bag of words techniques, it also accurately captures the sentiment by giving negative results for the negation of positive phrases. This study was focused primarily on semantic vector spaces, compositionality in vector spaces, logical form, deep learning and sentiment analysis.

3. TWITTER DATA PROCESSING

In this section, a detailed explanation of our data collection from twitter, geographic analysis and sentiment analysis over the collected data including the implementation details of the proposed solution is provided. The tweet analysis used to detect crime consisted of the following steps:

1. Collect tweets

2. Clean and parse the data

3. Conduct geographic analysis on the extracted tweets

4. Conduct sentiment analysis on the extracted tweets

3.1 COLLECT TWEETS

The top ten dangerous and safe cities in the United States, with their geolocation details were identified before the experiments were begun. The list of the cities and the geolocation details are provided in Table 2.1. Tweets were collected within a 50km radius around the city's central geolocation. This radius was kept constant for all the cities examined.

Twitter allows developers to explore its platform. Twitter4J, an unofficial Java library for the Twitter Application Program Interface (Twitter API), was used to automate the application so that it could be integrated into Twitter. Tweets were then collected according to crime-related topics within a certain constant area of each city. A keyword search strategy was adopted for collection purposes. Keywords used to identify crime-related tweets included "gun," "crime," "sinister," "kill," and so forth.

Unfortunately, Twitter places limits on the developer's account for reliability purposes. The default rate limit for calls to the API varies according to the authorization method being used and whether or not the method itself requires authentication [18]. The limitations are:

- Unauthenticated calls are permitted 150 requests per hour. Unauthenticated calls are measured against the public facing IP of the server or device making the request.
- OAuth calls are permitted 350 requests per hour and are measured against the oauth_token used in the request.

Because of these Twitter Developer limitations over its database, research devices were able to pass 150 requests in semi-hourly intervals. Our research used around 20 distinct developers, students at Missouri S&T, keys and managed to pass over 3000 requests every 15 minutes. In order to make the dataset short, research was able to extract tweets that contain the keyword "gun". Sample tweets collected during this process are listed in Table 3.1

Table 3.1. Sample Tweets Collected

S. No	**Location**	**Date**	**Tweet**
1	Detroit	Sat Jul 26 22:25:35 CDT 2014	So I'm walking home some dude pulls up and puts a gun in my face... Great fucking night..

Table 3.1. Sample Tweets Collected (cont.)

2	New York	Thu Jul 24 20:53:22 CDT 2014	Good guy with a gun takes out bad guy with a gun. Injuries reported after shots are fired at Pennsylvania hospital
3	East St. Louis	Wed Jul 16 23:50:31 CDT 2014	gun shots are so pleasant to hear #uhm...

Tweets were continuously collected in these cities for a period of 21 days, between July 7, 2014 and July 21, 2014 which ultimately ended up with over 100,000 tweets in the database. Although research obtained a variety of information from the tweets, relevant and irrelevant, particular interest was in specific content from the tweets.

3.2 CLEAN AND PARSE THE DATA

Each tweet was parsed before sentiment analysis was conducted. These parsing steps included the following:

1. Separate the individual terms in a tweet according to the white-space boundaries
2. Convert the tweet into lower case letters
3. Remove all non-alphanumeric characters from tweets (e.g., hash signs and dashes)

These steps helped with identifying the individual set of string tokens needed to word match as part of the ANEW-based sentiment analysis approach. While the Deep Learning for sentiment analysis approach requires only non-alphanumeric characters to be excluded. The study can be further extended by Gingering [14] all the terms in the

tweet which can help to increase the number of terms mapped to corresponding stemmed equivalents in ANEW.

3.3 CONDUCT GEOGRAPHIC ANALYSIS

Geographic analysis is a component of data analytics that involves both collecting and scrutinizing each data sample in a set of items from which the samples are drawn. The primary goal is to identify the trends. The datasets in this study that were extracted from twitter needed to be filtered before trends could be drawn. The parameters defined for this analysis phase included the following:

1. **TweetMean:** The TweetMean was defined as the average number of tweets per day in a particular city. The average tweet count of all days in a particular city were computed to estimate TweetMean μ_t :

$$\mu_t = \frac{\sum_{i=1}^{n} C_{i,t}}{n} \tag{3.1}$$

Where $C_{i,t}$ is the tweet count in a city for day i.

2. **SearchArea:** The SearchArea was defined as the area within the city from which the tweets were extracted:

$$SearchArea = \pi r^2 \tag{3.2}$$

Where the radius (r) was fixed to 50km.

3. **PopulationCount:** The PopulationCount was defined as the count of people in a particular SearchArea:

$$PC = Density \times SearchArea \tag{3.3}$$

Where $Density$ is the population per unit area.

4. **TweetRatio:** The TweetRatio was defined as a PopulationCount per TweetMean:

$$TweetRatio = \frac{PopulationCount}{TweetMean} \quad (3.3)$$

5. **PeopleTR:** The PeopleTR was defined as a reciprocal of the TweetRatio; it was simply the probability that a single person would commit a crime.

$$PeopleTR = \frac{TweetMean}{PopulationCount} \quad (3.3)$$

The steps used to design a filter for geographic analysis on data collected included the following:

1. Read the database and cluster the tweets according to the city
2. Cluster the city-specific data according to the date the tweet the tweet was posted in Twitter
3. Count the number of crime-related tweets in each city. For example, CityName#TweetCount_Day1#TweetCount_Day2#.....#TweetCount_Day20
4. Load the population density defined as the population per unit area of each city
5. Calculate the predefined parameters (e.g., TweetMean, SearchArea, TweetRatio, PopulationCount, and PeopleTR)

3.4 CONDUCT SENTIMENT ANALYSIS

Sentiment analysis is used to determine an author's attitude with respect to either a particular topic or a document's overall contextual polarity. As stated earlier, effort was given to evaluate the sentiment which existed in our tweets using two sentiment analysis techniques, namely ANEW based technique and Deep Learning model.

In ANEW based technique, effort was made in mapping every term from ANEW to its equivalent in the tweet. This was then applied by Porter Stemming [15] to improve the mapping. Selection of matching words that existed in both ANEW and the tweet was given further consideration. The mean valence μ_v and mean arousal μ_a for each selected ANEW term is fetched to compute average of the valence and arousal of all the ANEW terms. For example, the tweet “Good guy with a gun takes out bad guy with a gun. Injuries reported after shots are fired at Pennsylvania hospital " has four words 'good', 'hospital', 'fire' and 'gun' which map to ANEW. The valence and arousal scores for the terms 'good', 'hospital', 'fire' and 'gun' are [7.47, 5.43], [5.04, 5.98], [3.22, 7.17] and [3.47, 4.02] respectively. The overall mean valence and mean arousal score of the tweet is calculated using the formula (3.1) as

$$\mu_v = \frac{7.47 + 5.04 + 3.22 + 3.47}{4} = 4.8$$

$$\mu_a = \frac{5.43 + 5.98 + 7.17 + 4.02}{4} = 6.4$$

These final mean scores were plotted in Figure 2.1 so that the sentiment analysis could be determined.

Most sentiment prediction tools work by matching certain keywords, giving positive points to positive words and negative points for negative words and summing up these points to give the sentiment of the text. The order of words is not considered which results in lack of important information. In order to overcome the drawbacks in keyword matching techniques, new deep learning model that actually works on sentence structure was proposed in [5]. By passing the tweets accordingly, we can extract the sentiment using deep learning models. As stated earlier this method is based on multi-class classification, where the outcomes are one among the following list [very positive, positive, neutral, negative, and very negative]. For example, the tweet “Good guy with a gun takes out bad guy with a gun. Injuries reported after shots are fired at Pennsylvania hospital “falls under negative class out of five classes. The tweet “gun shots are so pleasant to hear #uhm..." falls under positive class.

4. RESULTS AND DISCUSSIONS

In this section, we analyze the crime related tweets in the selected cities and demonstrate the correlation between the crime intensity based on Forbes article and trends observed on Twitter. Sentiment analysis techniques were then used to examine the crime intensity within each city's twitter dataset.

4.1 SCENARIO I

An exact correlation was drawn between the crime trends described in a Forbes article and sentiments identified in Twitter. The tweets were run through a geographic analysis phase after both the data collection and cleaning phase were complete. The geographic analysis results are included in Scenario I. Stepped Area Charts, a step-wise graphical representation of quantitive data, were used to visualize the datasets. The number of tweets collected from the crime cities are graphed in Figure 4.1.

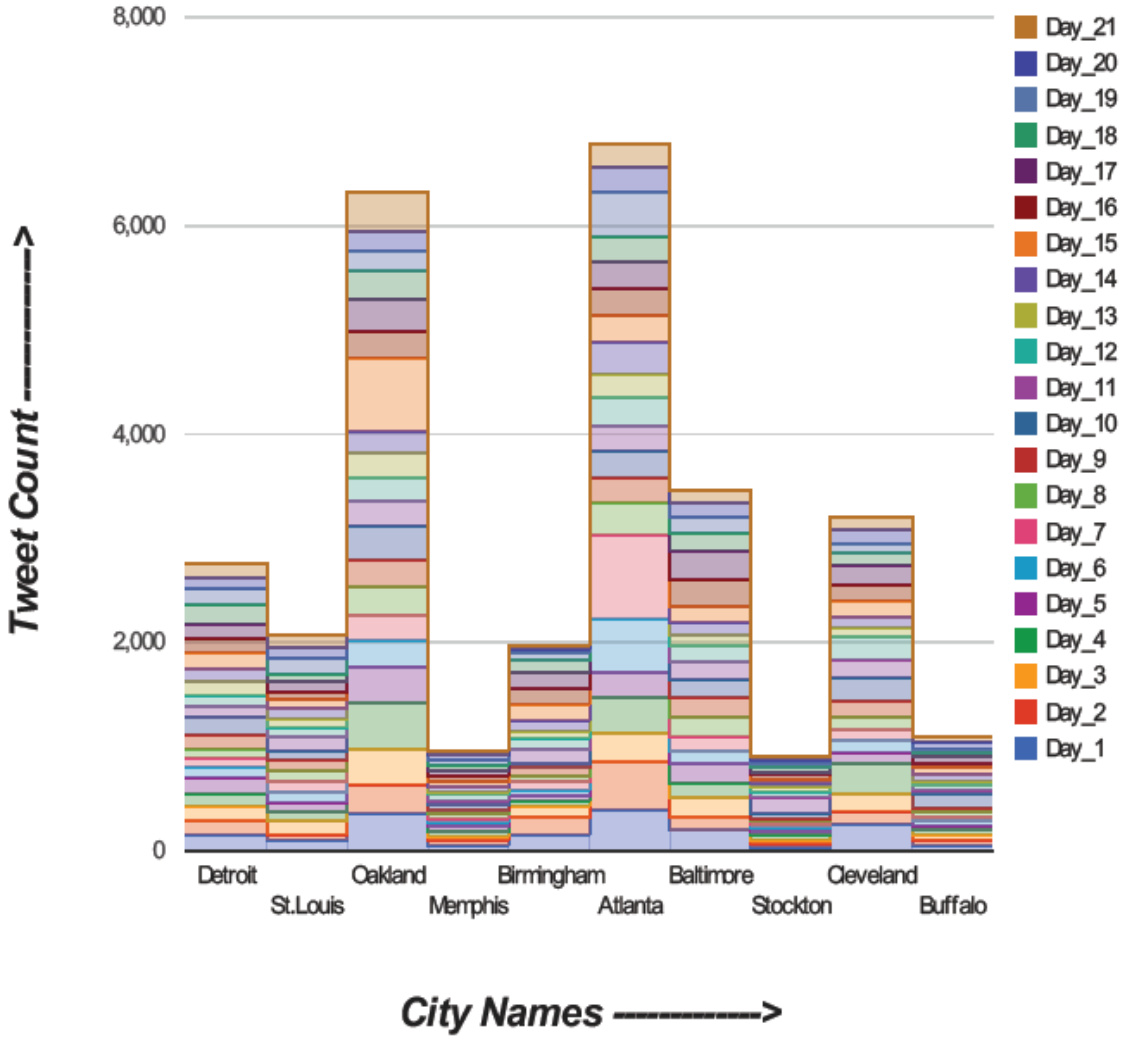

Figure 4.1. Distribution of Tweets in Cities Identified as Crime

The top ten crime cities are listed sequentially along the horizontal axis. The number of tweets collected each day is listed in a step manner along the vertical axis. Those cities with largest amount of crime (e.g., Detroit, Oakland, and Atlanta) had a larger number of tweets than any other cities. The number of tweets collected from select cities identified as safe is illustrated in Figure 4.2.

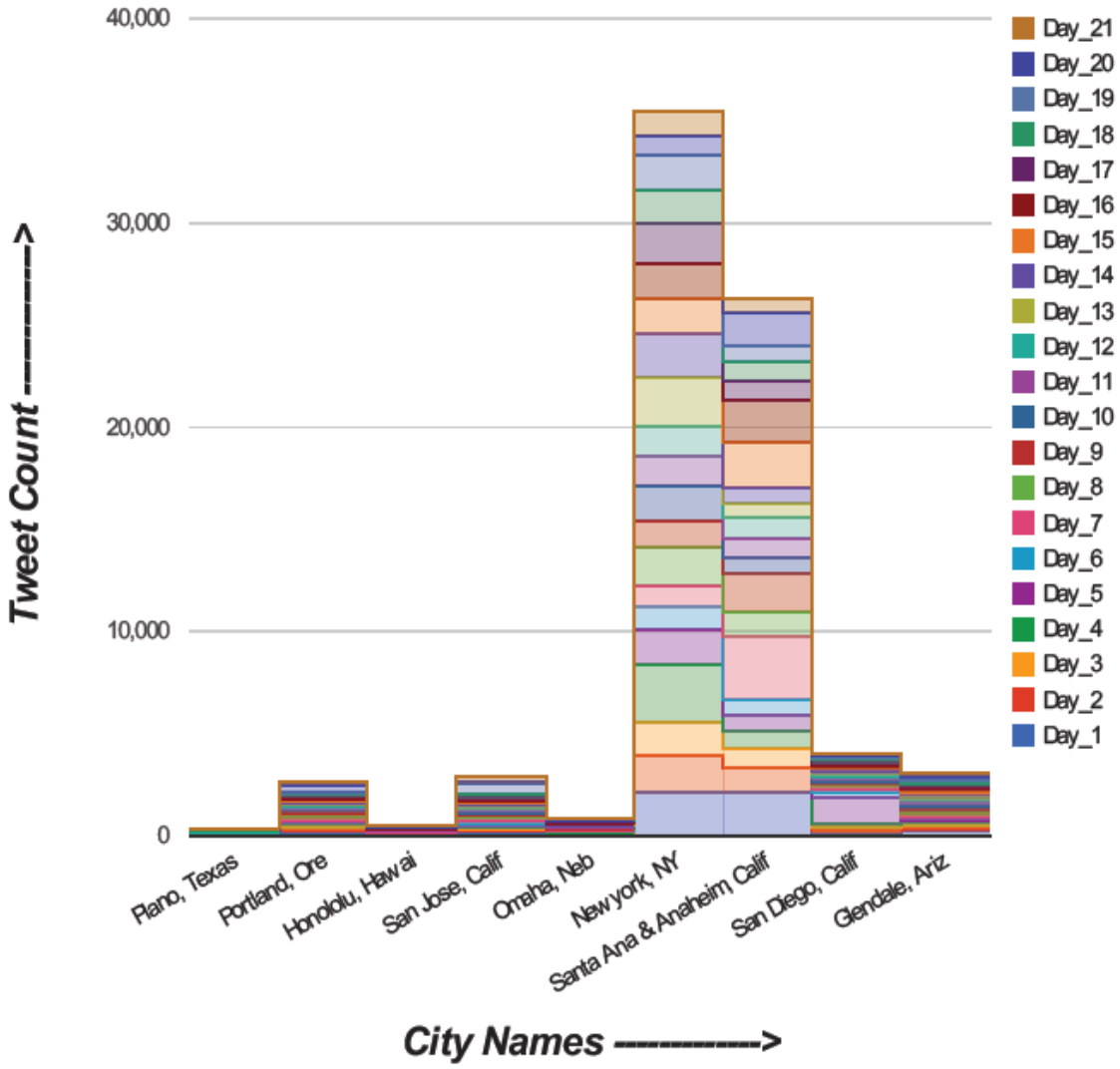

Figure 4.2. Distribution of Tweets in Cities Identified as Safe

The top ten safe cities are listed sequentially along the horizontal axis. The number of tweets collected each day is listed in a step manner along the vertical axis. The top safe cities (e.g., Plano, Honolulu, and Omaha) had a smaller number of tweets than any other city. The number of tweets collected from random cities, from list of collected cities, is illustrated in Figure 4.3.

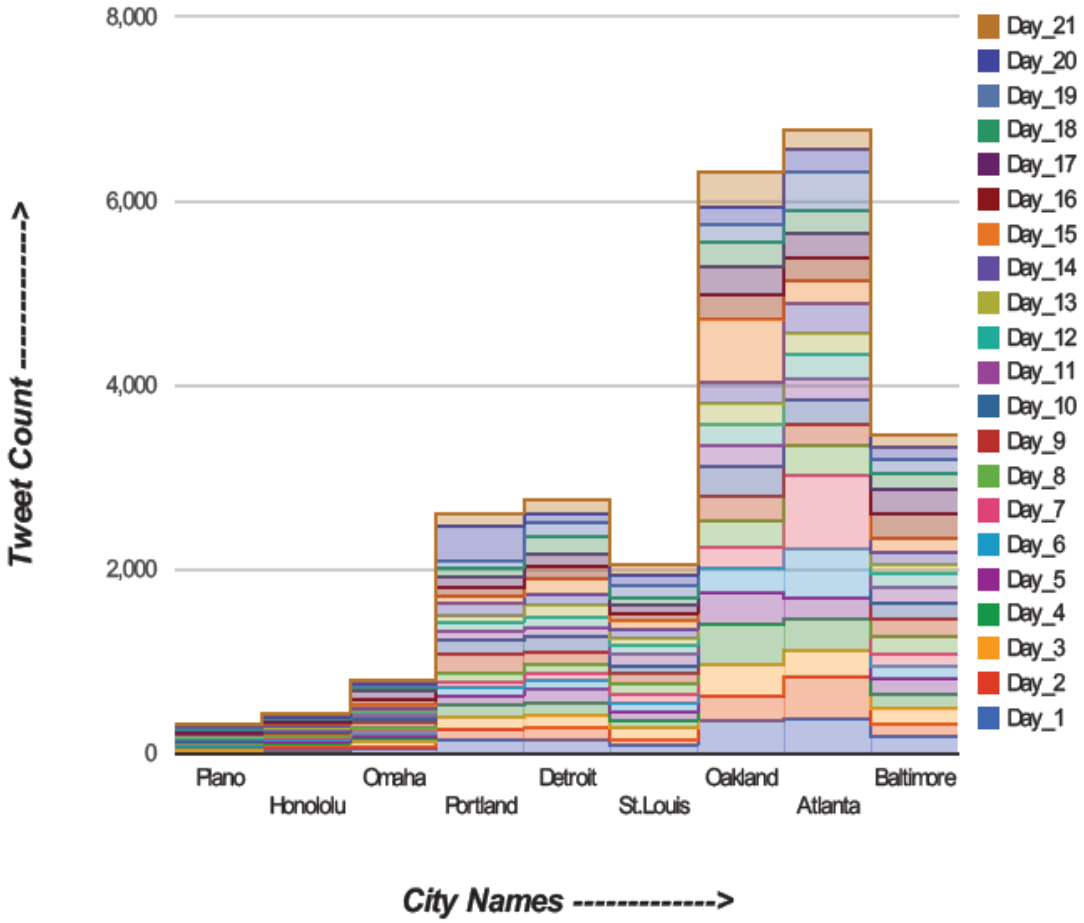

Figure 4.3. Distribution of Tweets in Select Cities

Several of the cities from each list (both dangerous and safe) were combined into one list so that any trends present in the data could be identified. The cities in Figure 4.3 are distributed along the X-axis. The safe cities (e.g., Plano, Honolulu, and Omaha) had significantly fewer tweets than the crime cities (e.g., Detroit, Atlanta, and Oakland).

The results gathered from the geographic analysis are represented five dimensionally in Figure 4.4. Each bubble has a set of five parameters (e.g., abbreviated city name, crime intensity, tweet count, TweetRatio, and PeopleTR). The number of tweets appears along the x-axis, and the TweetRatio appears along the y-axis. Each city's

abbreviated name appears on its corresponding bubble. Both the crime intensity and the PeopleTR can be observed in the colour variance (as per the scale at the top of Figure 4.4) and bubble's size, respectively.

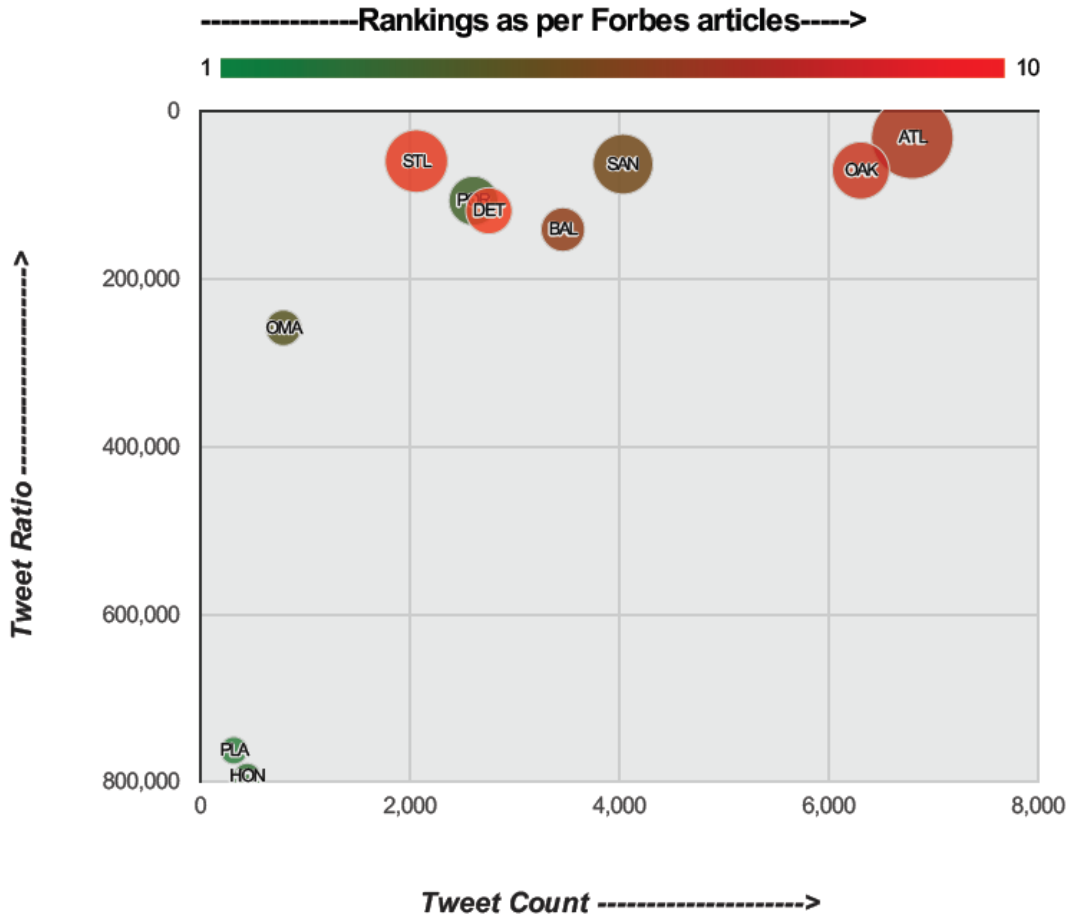

Figure 4.4. Five-dimensional View of Geographic Analysis

Each of the safe cities is located at the bottom of the graph. They are safe based on the following:

- These bubbles registered a smaller tweet count on the x-axis and a larger tweet ratio on the y-axis. Thus, although these cities have large populations, their crime patterns are small.

- The bubbles are nearly green in colour, indicating they have been reported as safe in Forbes.
- The bubbles are comparatively small, suggesting that the probability of a person in that area committing a crime is small.

Cities at the top of the graph are considered crime cities based on the following:

- These bubbles registered a larger tweet count on the x-axis and a smaller tweet ratio on the y-axis. Thus, a larger population has expressed crime patterns than those in other cities.
- The bubbles are nearly red in colour, indicating they have been reported as crime-prone in Forbes.
- The bubbles are large, suggesting that the probability of a person in that area committing a crime is large.

A number of cities examined in this study that had been reported as safe were found here to be prone to criminal activity. Thus, a location's crime intensity should be measured in real-time rather than from previous statistics.

4.2 SCENARIO II

As detailed in Section 3, we first tried to measure the sentiment involved in the tweet using dictionary based approach. All of the tweets were clustered according to city, and each cluster was passed through the sentiment analyzer. This analyzer was implemented in such a way that it chose the tweet, cleaned and parsed the data, and provided sentiments according to the ANEW based approach. The output from this multi-class sentiment analyzer was as follows:

- Very Positive
- Positive
- Neutral
- Negative and
- Very Negative

Tweets that were categorized as either Negative or Very Negative were identified as contributing to the crime intensity. Crime trends exist in every city, even those identified as safe in Forbes magazine.

The probability of the tweets that contribute to crime intensity within a particular cluster was calculated so that the crime intensity of that cluster could be identified. Tweets that contribute to crime intensity are that classified as either very Negative or Negative. A heat map, in which the findings of this study are plotted, is illustrated in Figure 4.5. It is detailed as a day-wise crime intensity plot of the cities. The crime intensity is represented as a colour scale. The crime pattern in this map is not limited to crime-prone cities. Safe cities (e.g., Plano, Portland, and San Hose) each recorded

several days of high crime intensities. A number of crime cities (e.g., Stockton and Cleveland) also recorded high intensities.

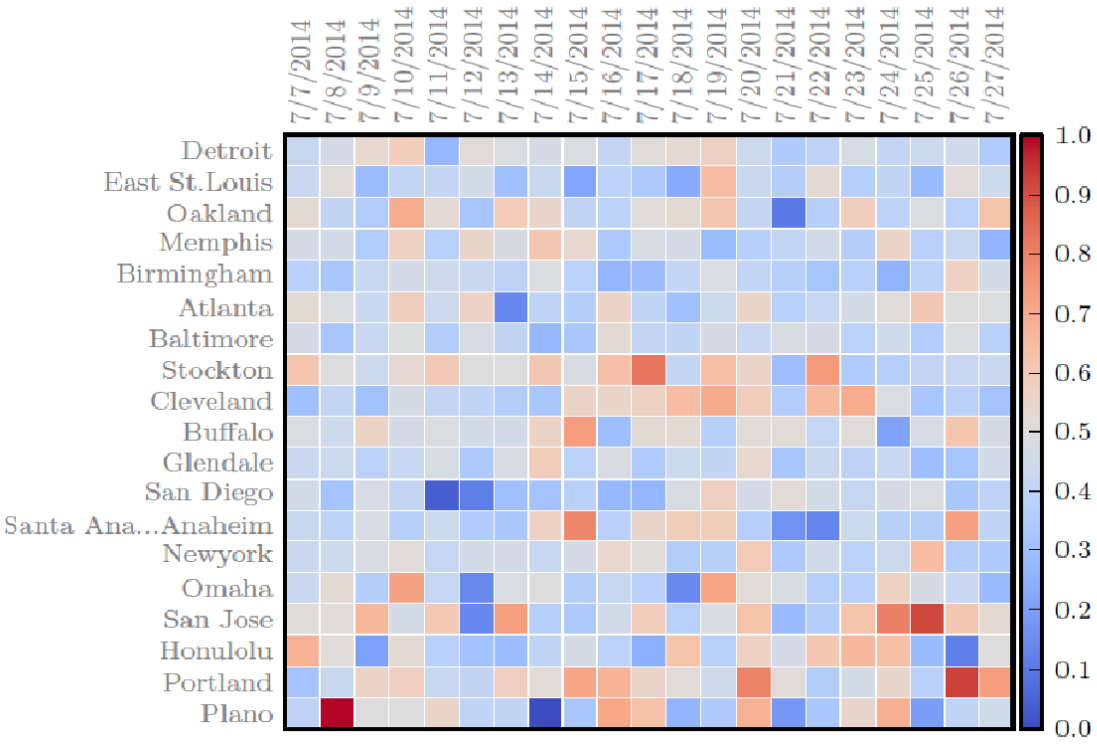

Figure 4.5. Results from Dictionary-Based Sentiment Analysis Conducted Over the Tweets

Recursive deep models were used for sentiment analysis to overcome the limitations of a dictionary-based approach. The recursive deep models accurately capture not only the effects of negation but also its scope at various tree levels for both positive and negative phrases. Crime intensity is measured in the following steps:

- Cluster the extracted tweets according to city
- Pass the tweets through both the data cleaning and parsing phase

- Use an analyzer to measure the sentiment involved in the tweets
- Cluster the tweets according to day, thereby narrowing the intensity variance

This model also follows a multi-class sentiment analysis scheme. Again, the output includes the following:

- Very Positive
- Positive
- Neutral
- Negative and
- Very Negative

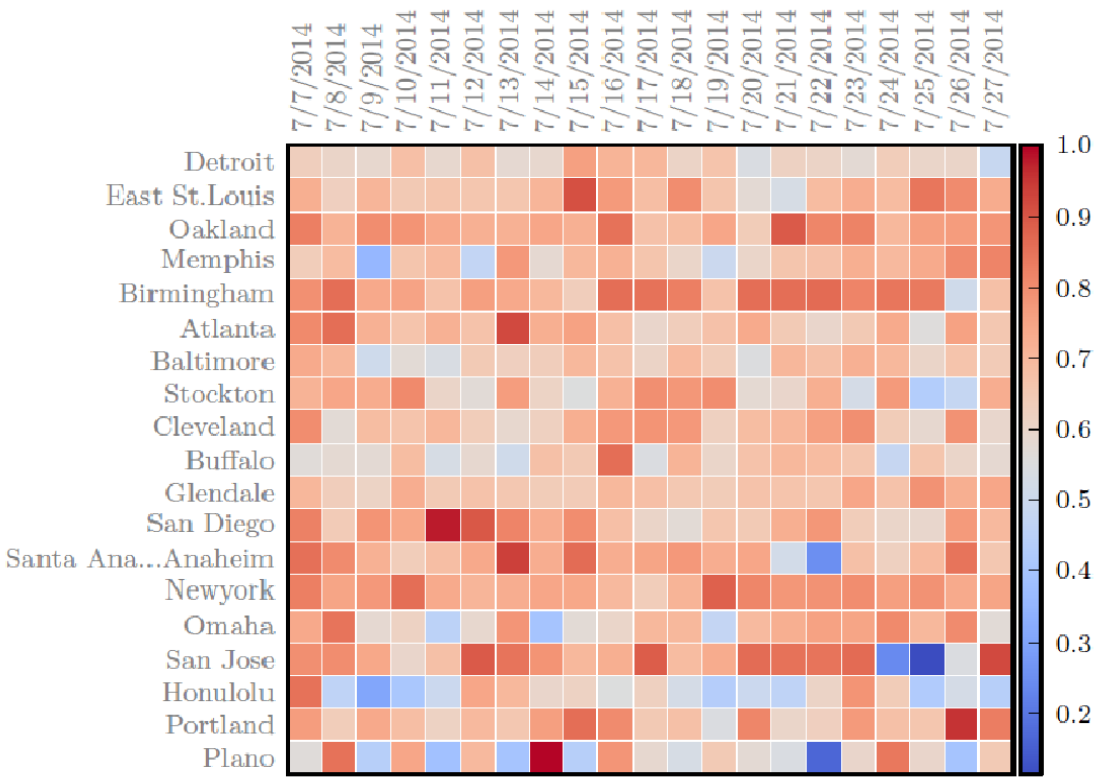

Figure 4.6. Results from Deep Learning Models Over the Tweets

A heat map was again used to plot these results (see Figure 4.6). This day-wise crime intensity plot indicates that the crime prone cities had a greater crime intensity than safe cities. The crime trends observed in safe cities, however, must still be considered.

A city's average crime intensity per thousand people was calculated next. In Figure 4.7, each block represents a city's crime intensity per thousand people on a specific day.

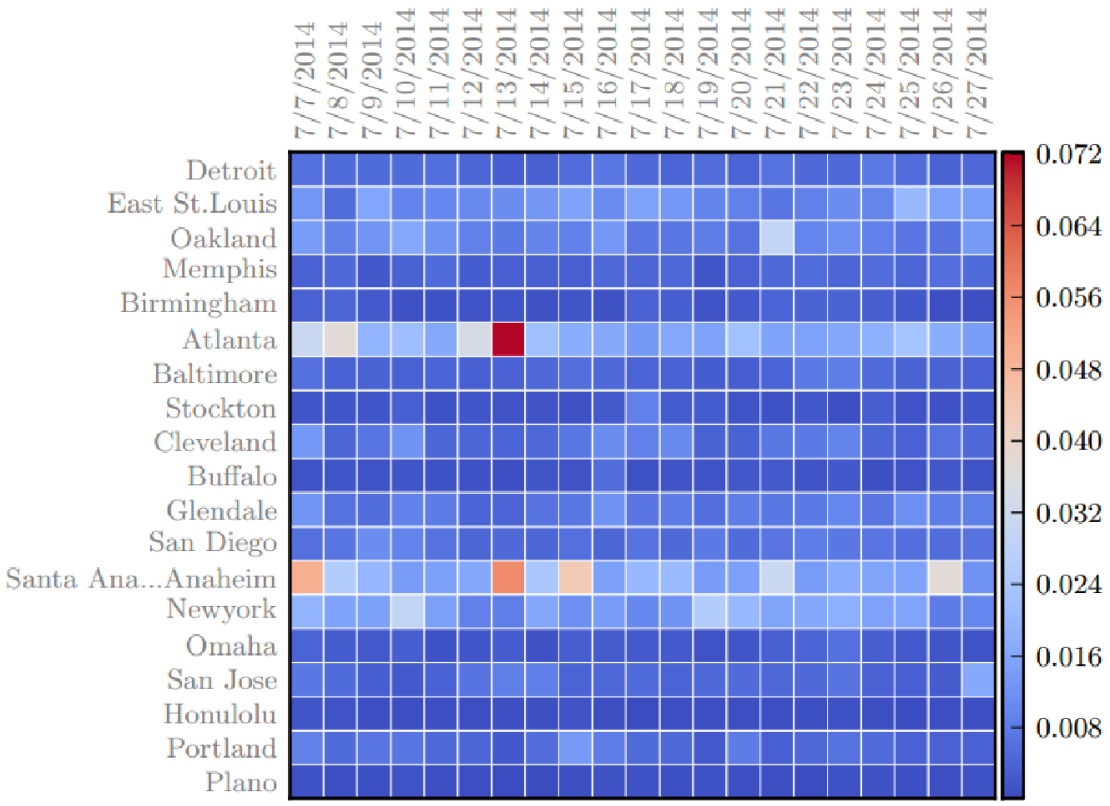

Figure 4.7. Refined Results Gathered From Sentiment Analysis

The safe cities at the bottom of Figure 4.7 appear to be safer according to the colour scale. The crime-prone cities appear to be dangerous according to the colour scale. Crime trends in various safe cities (e.g., New York and Santa Ana) reveal that crime patterns do exist in these areas.

4.3 SCENARIO III

The shootings that recently took place in Ferguson, St. Louis were analyzed as part of the third scenario. The situation began when an unarmed, 18-year-old African-American male, Michele Brown, was fatally shot by Darren Wilson, a white Ferguson police officer, on August 9, 2014 [17]. The incident sparked protests and vandalism. The street violence continued for two weeks, creating terror throughout the community. The crime trend divergence observed in Twitter for St .Louis, both before and after the Ferguson shooting is depicted in Figure 4.8.

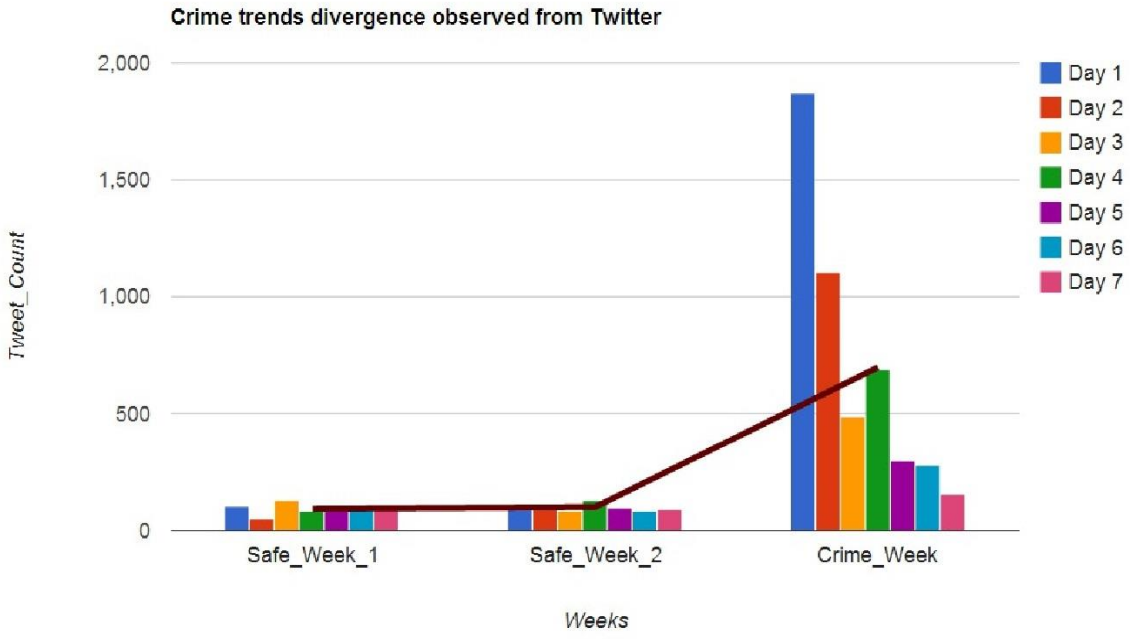

Figure 4.8. Crime Trend Divergence in St. Louis County

From Figure 4.8, two safe weeks had an average of 93 and 101 tweets per day, as per the crime pattern search strategy utilized. This average rose sharply the week the street violence occurred. An average of 697 tweets per day indicates the city was

terrified as a result of the continuing crime. Studies previously conducted on crime pattern detection held the crime intensity of a geolocation. The strategy utilized in this study drew the crime patterns in almost real-time.

In depth, the tweets like "Yeah applying for my gun license!" and "I'm not familiar with the gun laws in Missouri but aren't you allowed to carry a gun?" can warn the audience that the city has become violent, and residents are trying to protect themselves. Tweets that help to raise awareness, that people in the city are going crazy are:

- “Forget cameras, let's disarm the police. Gun control for the 5-0. #Ferguson"
- “@johnnybHEAT3 suspended? If I pointed a gun in a cop's face and threaten to kill him, where would I be?" and
- " @staceyhopkinsga @kroger @citygear shut the fuck up bitch , ima carry my gun in there regardless of their policy"

.

4.4 SCENARIO IV

Tweets were next extracted from the city-specific crime identification groups (e.g., city police, the city crime page, and the local media). These pages rely on information provided by sheriffs. A number of tweets were extracted from these accounts to improve the scheme. For example, Atlanta's Twitter account "Atlanta_Crime" is monitored by Spot Crime [19]. This page maintains tweets that report all of the crime scenes registered by the sheriff. The data collected from this page is pictorially represented in Figure 4.9.

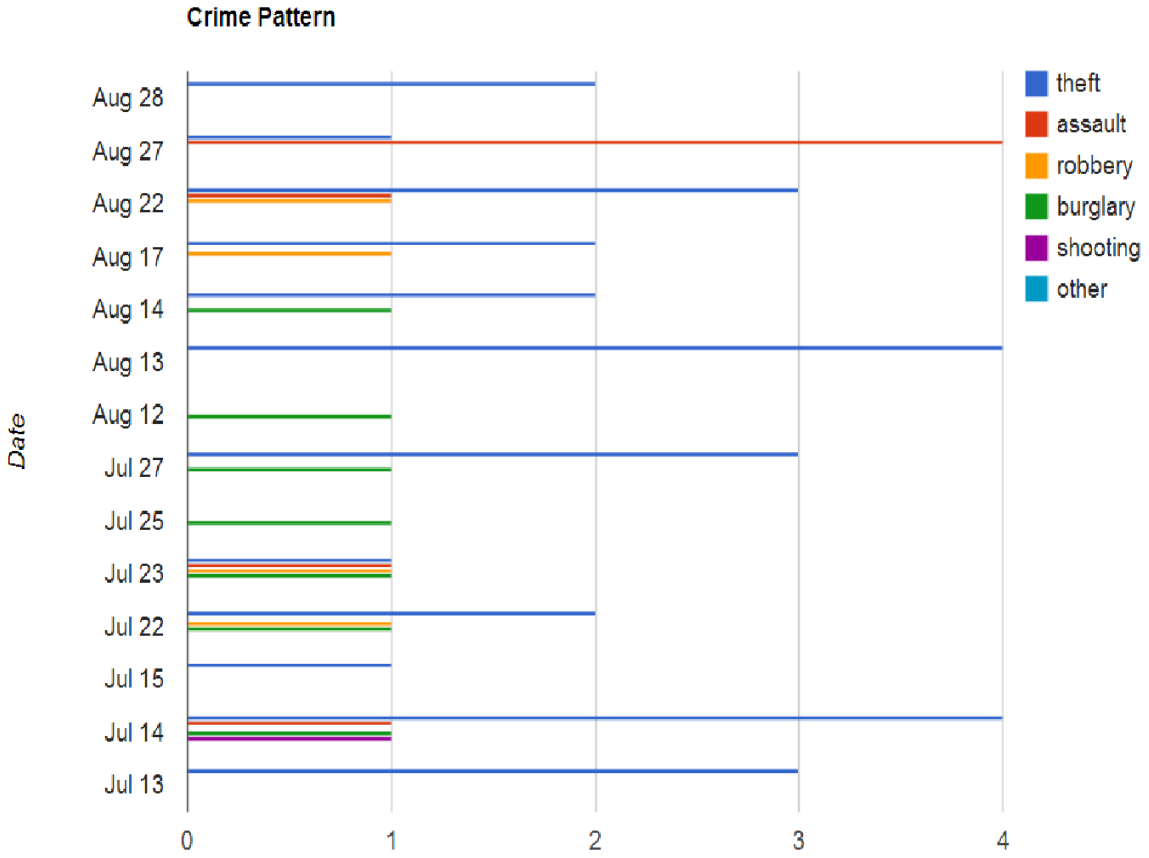

Figure 4.9. Crime Trends Drawn From @Atlanta_Crime Twitter Account

The crimes in this image that were registered on July 14th include one shooting, four thefts, one assault, and one burglary. Also in the Figure 4.7, these days are reported as experiencing more crime than usual. This proves the fact that, monitoring online social media will help with identifying crime trends in advance. In future, we can always design some algorithms on top of our study, to give priorities to different twitter pages [City Police, Spot Crime, Media, etc] which ultimately improve the accuracy in measuring city's crime intensity.

5. CONCLUSION

A crime pattern can be detected, nearly in real-time, when online social media is monitored. Crime can occur anywhere at anytime. Previous statistics do not accurately identify the crime intensity of a specific location. More accurate results can be drawn from social media. Results from geographic data analysis conducted on various tweets provided a clear picture of the criminal trends in several different cities. The crime intensity day-wise positively correlated with crime statistics from cops, which ultimately prove the hypothesis. The Ferguson shooting case study clearly differentiates the city's safe and dangerous pattern. To be more precise, we analyzed the specific twitter accounts which tweet only about the crime scenarios happened in the city based on sheriff data and visualized.

The results gathered from this study were positive. An advanced sentiment analysis algorithm will aid in differentiating a sinister murderer from tweets within a specific location. Video-to-text processing, image-to-text processing, and data from various online sources would also help improve accuracy. This type of study would help with informing others of the crime pattern both within and around their location, ultimately assisting them with staying in a safe zone. Monitoring various social media outlets (e.g., Facebook, Google+, Tumblr, and Myspace) would improve accuracy.

BIBLIOGRAPHY

[1] A. Java, X.Song, T.Finn, and B.Tsang. Why We Twitter; Understanding Microblogging Usage and Communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop 2007 on Web Mining and Social Network Analysis, pages 56-65, 2007.

[2] J. Bollen, A.pepe, and H.Mao. Modelling public mood and emotion: Twitter sentiment and socio-economic phenomena. arxiv:0911.1583v0911 [cs.CY], pages 09-19, 2010.

[3] Daniel Fisher. "The 10 Most Dangerous U.S. Cities". Forbes Article, 2012. [Online]. Available: http://www.forbes.com/sites/danielfisher/2012/10/18/detroit-tops-the-2012-list-of-americas-most-dangerous-cities/

[4] Francesca Levy. "America's Safest Cities". Forbes Article, 2010. [Online]. Available: http://www.forbes.com/2010/10/11/safest-cities-america-crime-accidents-lifestyle-real-estate-danger_slide_11.html

[5] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.

[6] Bradley, M.M., Lang, P.J.: Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida (1999).

[7] Siddarth Shankar, Ramaswamy. 2011. Visualization of the Sentiment of the Tweets. Master's Thesis, North Carolina State University, Raleigh, NC.

[8] Shyam Varan Nath, Crime Pattern Detection Using Data Mining, Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, p.41-44, December 18-22, 2006.

[9] T. Wang, C. Rudin, D.Wagner, and R.Sevieri. Learning to detect patterns of crime. In ECML, PKDD, 2013.

[10] L. Pang, B. Lee and S. Vaithyanathan. Thumbs Up? Sentiment Classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 10:79-86, 2002.

[11] B. Pang and L. Lee. A Sentimental education: Sentiment Analysis using subjectivity summarization based on minimum cuts. In Proceedings 42nd ACL, pages 271-278, 2004.

[12] H. Tang, S. Tan and X. Cheng. A survey on sentiment detection of reviews. Expert Systems with Applications, 36(7):10760-1073, 2009.

[13] J. A. Russell and L. Feldman Barerett. Independence and bipolarity in the structure of current affect. Journal of Personality and Social Psychology, 74:976-984, 1998.

[14] Miki Feldman-Simon and Edina Fitzpatrick. A New Generation Writing Solution. [Online]. Available: www.gingersoftware.com/files/writing.pdf

[15] M.F. Porter. An algorithm for suffix stripping. Program-Automated Library and Information system, 14(3):130-137, 1980.

[16] 2014 Isla Vista Killings. (2014, August 26). In Wikipedia, The Free Encyclopedia, The Free Encyclopedia. Retrieved 17:20, August 26,2014, from http://en.wikipedia.org/w/index.php?title=2014_Isla_Vista_Killings&oldid=622895353

[17] Shooting of Michael Brown. (2014, August 26). In Wikipedia, The Free Encyclopedia, The Free Encyclopedia. Retrieved 18:48, August 26,2014, from http://en.wikipedia.org/w/index.php?title=Shooting_of_Michael_Brown&oldid=622919817

[18] Twitter Rest API Rate Limiting in v1.1. In Twitter Developers, The Free Service for Developers to extract data from Twitter. [Online]. Available: https://dev.twitter.com/docs/rate-limiting/1.1

[19] SpotCrime.com. (2014, April 1). In Wikipedia, The Free Encyclopedia. Retrieved 21:36, August 31, 2014, from http://en.wikipedia.ord/w/index.php?title=SpotCrime.com&oldid=602208308

VITA

Raja Ashok Bolla was born on 27th August, 1990 in Razole, India. He received a Bachelor of Engineering degree in Computer Science from GITAM University, Visakhapatnam, India. Immediately upon receiving his Bachelors degree he joined in Tata Consultancy Services and worked as Systems Engineer for a period of two years. He subsequently joined Missouri University of Science and Technology in fall 2013 and received his Master's degree in Computer Science in December 2014. During his course of study, he worked as Graduate Research Assistant under Dr. Sriram Chellappan. His areas of interest include Algorithms, Natural Language Processing, Software Engineering, Psychology, and Distributed Systems.