CRIME ANALYSIS ACROSS MAJOR CITIES OF INDIA WITH TWITTER

Nisanth T. S
Student
REVA Academy for Corporate Excellence, REVA University
Mainframe Administrator, IBM India
Bangalore, India
NishanthTS.BA02@reva.edu.in

Sonali Sucharita
Student
REVA Academy for Corporate Excellence, REVA University
Technology Lead, Cognizant Technology Solutions
Bangalore, India
sonalis.ba02@reva.edu.in

ABSTRACT

Police traditionally use maps and tacit knowledge about a city to determine which public areas are to be observed to prevent or reduce crime rates. Most of the times, the accuracy of such predictions are very low. Crime prevention should be a proactive measure rather than a reactive one. Data from social networking sites like Twitter can enable the prediction of future crimes in major cities and is used extensively in countries like the USA.

The data used in this study are Twitter feeds from 5 major cities in India. Based on the latitude and longitude, this study attempts to collect the data for top 10 types of crimes across these major cities and then compare the results with the data from Government website, National Crime Records Bureau (NCRB).

Data collection is planned in three stages; 1. Collecting crime data from Twitter 2. Data from Govt. Crime records website and 3. Creating a combined data set with both twitter data and the Government data.

Text mining and sentiment analytics will be used to predict the crime rates in these cities. Our attempt is to come up with a better predictive model which can predict crimes that could occur in these cities for the next 3 months.

Keywords: *Text Mining, Crime prediction, Twitter, Density estimation*

1. INTRODUCTION

In this modern era, various social media sites are filled with so much of personal information, which can be instrumental in predicting human behaviours. The authors' interest area is Crime Analysis using such social media as demographic data and crime analysis have been found to be correlated (Xingan, et al., June 2015).

Nowadays social media is real time and information update is almost immediate. If crime related updates are channelized properly and hidden correlation could be unearthed, then actionable insights can be drawn. Demographic data can be used to enrich derived or extracted information from social media.

Twitter is widely used by people to tweet about their routine activities and to express opinions about a particular topic. The general tendency is to think that there is no correlation between Twitter use and crime. Because people will never tweet about what they intend to or have just committed a crime. But the interesting part is, what people do share are things like social events or outings that could lead to criminal activity. There are studies in this area and Twitter is the choice of data gathering because of its openness. However, there are challenges using the Twitter data like misspellings, usage of symbols and also the character limit imposed by Twitter. The results of such studies could help police in resource allocation, deciding when and where to deploy officers. Twitter has been very useful in predicting large-scale events like elections, infectious disease outbreaks and social upheavals (MIT Technology Review, 2014).

Currently, there are studies from the U.S (Wang, et.al., 2012) for predicting crimes in certain cities using Twitter data analysis. The motivation for this article is to conduct a similar study for the major cities in India, which will help the Indian Police to better allocate the resources and in a timely way so as to avoid crime. The intention of the research team is to predict crime at a ward or block level resolution or fine. The authors have collected Government data from Bruhat Bengaluru Mahanagara Palike (BBMP) website which has exhaustive data can also supplement the Twitter data that is being extracted. For example, Government data includes the presence or absence of street lights in a particular area which can sometimes lead to crime. So the authors are looking at this external data source from Government which can help to consider other factors apart from the content of tweets and geography. Also, the team would like to compare the analysis to the data provided by National Crime Records Bureau (NCRB).

National Crime Record Bureau (NCRB) was founded in 1986 with the intention of helping the investigating agencies by providing them with extensive information on criminal data at state and national level such as modus operandi, personal profile, fingerprint, photograph, criminal history and details of property which may be subject matter of crime (*Cited from Wikipedia*).

The software used by Delhi Police is CMAPS (Crime Mapping Analytics and Predictive System) which accesses real-time data from 100 helplines and, using ISRO's satellite imageries, spatially locates the calls and visualises them as cluster maps to identify crime spots. The models and algorithms used by the software can help police understand where the next crime is likely to occur. This concept called 'predictive policing', marks a paradigm shift in the way policing is done. CMAPS replaces the mechanical crime mapping which involves manually gathering data at an interval of 15 days. This software uses data generated through dial 100 to be plotted on the geospatial map of Delhi. Usage of such predictive policing has raised concerns that it might create prejudice against particular communities or areas (Karn Pratap Singh, 2017)

Here the model uses data from dial 100 calls. This research initiative will try and augment the current models by bringing in data from social media platforms like twitter which is very popular among the public when it comes to expressing opinions or sharing information on events, outings and what they are planning to do next etc. Data coming from social media platforms can have an indirect impact on the crime prediction.

2. LITERATURE REVIEW

Sentiment Detection approach for Twitter Messages (Barbosa, L & Feng, J, 2010) had used biased and noisy labels as input to build its models. The models were very robust because they had created an abstract representation of the Twitter messages, not raw word representation. They had also found that although noisy and biased, the data sources provide labels of reasonable quality and, since they have a different bias, combining them was beneficial.

On similar lines but extended to three class (positive, negative, neutral), Agarwal, et.al., (2011), had used unigram model as a baseline model and reported an overall gain of over 4% for classification (positive, negative, neutral). They tested with two models, tree kernel and feature-based models and found that these models surpass the unigram model in performance.

Another approach was based on the automatic semantic analysis and understanding of natural language tweets, combined with dimensionality reduction via latent Dirichlet allocation and prediction via linear modelling by Wang et.al., (2012) "Automatic Crime Prediction Using Events Extracted from Twitter Posts".

Evaluation results demonstrated the model's ability to forecast hit-and-run crimes using only the information contained in the training set of tweets. It outdid the prediction of baseline model which predicted hit-and-run incidents uniformly across all days.

Extending the work by Wang et al., and overcoming the three limitations (they had used tweets from hand selected news agencies, tweets were not associated with GPS Location information and the authors had investigated only two of the many crime types tracked by police organizations and did not compare their model with traditional hot-spot maps) Matthew Gerber's (2014) work uses Twitter-specific linguistic analysis and statistical topic modelling to automatically identify discussion topics across Chicago, then incorporates these topics into a crime prediction model and shows that, for 19 of the 25 crime types the author studied, the addition of Twitter data improves crime prediction performance versus a standard approach based on kernel density estimation.

Another work on Geo-Spatial Social Media Activity by Bendler et.al., (2014) threw light on their Zero-Inflated Poisson Regressions which indicated that analysing the spatiotemporal intensity of Social Media usage is valuable in explaining Burglary, Motor Vehicle Theft, Robbery, and Theft/Larceny. Authors could confirm that that Robbery and Theft/Larceny are more likely to occur near increased social activity, while Burglary and Motor Vehicle Theft are less frequent in such areas and times etc. by examining the coefficient estimate that corresponds to the Social Media variable for these four crime types. In order to account for the shortcomings of a Zero-Inflated Poisson Regression in analysing geospatial relations, they additionally executed Geographically Weighted Regressions.

To find the tweet origin location and its classification, Thom et.al., (2014) introduces two strategies that are suitable to assign probable locations of origin to social media messages of unknown locations. They are based on aggregated knowledge about the author and/or the textual content of the message.

Related Work to find Tweet Origin, Alsudais et.al., (2014) applied random forest classifier. These findings could be helpful to perform the crime analysis better if they are risk-prone areas or localities.

Another relevant work which would be helpful for analysing behaviour patterns. The question of who tweets with their location? This can be better understood with the help of this paper. Understanding the Relationship between Demographic Characteristics and the Use of Geo services and Geotagging on Twitter (Sloan, L & Morgan, J 2015).

SensePlace3 Framework by MacEachren et.al., (2011) to analyse place–time–attribute information in social media presented a design framework for a geo-visual analytics system that tackles the problem of understanding spatiotemporal dimensions in real-time, streaming social media. The design of SP3 provides a pattern for system architecture, computational processing, and web-based geo-visualization.

Albeit the Analysis and Predicting of Crime has been done by many authors, but scalability and uptake should also be discussed and put forward. Keeping this in mind quite a few authors have suggested their approaches. Kalampokis et.al., (2011) paper introduces a two-phased approach for supporting participatory decision-making and a Web data-driven architecture that enables the implementation of the proposed approach. This integrates Social and Government Data and also the social Data Analysis works for predicting future events. The architecture uses linked data paradigm as a layer which enables integration of data from heterogeneous sources.

3. METHODOLOGY

3.1 Research Problem Statement:

The problem that the authors are addressing here is to find out the crime spots in various parts of Bangalore and visually represent it with the help of a heat map. Twitter data is used to extract the crime related tweets. Demographic data from BBMP website data is being used to enrich the twitter data with more information like ward/assembly-constituency names where the crimes are reported. The final extension to this research would be to predict the crimes based on events being tweeted (Researchers used Twitter to Predict Crime, a report in business insider, 2014) and to extend the work to other cities in India.

The various stages of the research work are broken down into the following sections:

The research process thus include,

(1) Extract Twitter data specific to a major city's latitude and longitude.
(2) Segregate the extracted data based on the most popular crime related keywords.
(3) Enrich the segregated data with critical data from Government sources which can have an impact on crime occurrence.
(4) Use analytical approach to analyse the formatted data and report the results.

3.2 Data Collection & Cleaning

The team has collected live streaming tweets from October 16th until October 27^{th,} 2017 for the academic research that is being performed as part of this paper. Until October 27th, authors were able to extract around 22,000 tweets. The team used the twitter package 'tweepy' in python to perform this extraction. The filter criteria specified for the tweet extraction was the latitude and longitude of Bangalore city so as to collect maximum tweets from various parts of Bangalore. The polygon had a Southwest (bottom left) corner of 77.540558, 12.906525 and a Northeast (top right) corner of 77.5946, 12.9716. The tweets were extracted and saved in JSON format. The team then used the JSON loads function to

read each and every tweet. The tweets were then converted to a tabular format using the python pandas DataFrame object. Various attributes were selected when the team converted the tweets to tabular format. The attributes were, the creation time of the tweet, text content of the tweet, Location of the tweet, Country from where the tweet was made. As part of data cleaning, the data extracted was converted to DateTime object in python. The next step was to convert the DataFrame object to CSV format.

The CSV format tweet file was subsequently used for crime-related data analysis.

4. ANALYTICAL APPROACH

The first step in aggregating crime related tweets was to parse each and every tweet and to search for any crime related keywords. The team had chosen a list of crime-related keywords for applying the filter criteria. This parsing was performed on the text content of each tweet. The resultant records were stored in a DataFrame object.

4.1 Feature Engineering

Team performed topic labelling for each tweet based on the content which was being discussed in it. Labelling was based on the crime related keywords that were present in the tweet content. If the tweet was a crime that happened in the locality, then the corresponding tweet was marked as crime related tweet. As part of feature engineering, the team created two additional attributes called ‘Category’ and ‘Sub-Category’ to capture the information whether the tweet was crime related and if so, what is the type of crime being discussed. This DataFrame object with the crime-related records was then concatenated with the original DataFrame and duplicate records were removed as part of data cleaning.

4.2 Aggregating the crime reports Ward/Assembly Constituency wise

The next part of the data analysis included finding crime rate ward/assembly-constituency wise using the extracted tweets. 'Sub_Category' and 'Tweet_Location' were the attributes used to find out the crime rate. The team then grouped the 'Sub_Category' which included the crime type as per the location specified. A crosstab view of crime category and location was used to understand the area wise crime occurrences (Table 1).

Tweet_Location / Sub_Category	Andhra Pradesh	Anekal Bangaluru	Bengaluru	Bengaluru South	India	JP Nagar	Karnataka	Tamil Nadu
Accident	0	0	9	2	0	0	0	0
Assault	0	0	8	4	2	0	0	0
Blast	1	0	7	5	1	0	1	1
Explosion	0	0	5	0	0	0	0	0
Murder	0	0	14	5	3	0	0	0
Pothole	0	1	9	8	1	0	0	0
Snatch	0	0	1	0	0	1	0	0
Snatching	0	0	0	0	0	1	0	0
Terror	0	0	22	2	4	0	0	0
Theft	0	0	1	0	0	0	0	0

Table 1 Area wise tweets with sub-category – a sample

Mapping the crime spots (Crimes Tweets vs Location of Tweets):

The final part of the approach was to map the ward/assembly-constituency wise crime rates in a heat map of Bangalore city. Top crimes were identified location wise and were mapped. Tableau was used to map the number of crime-related tweets according to the location from where the tweet was made (Figure 1).

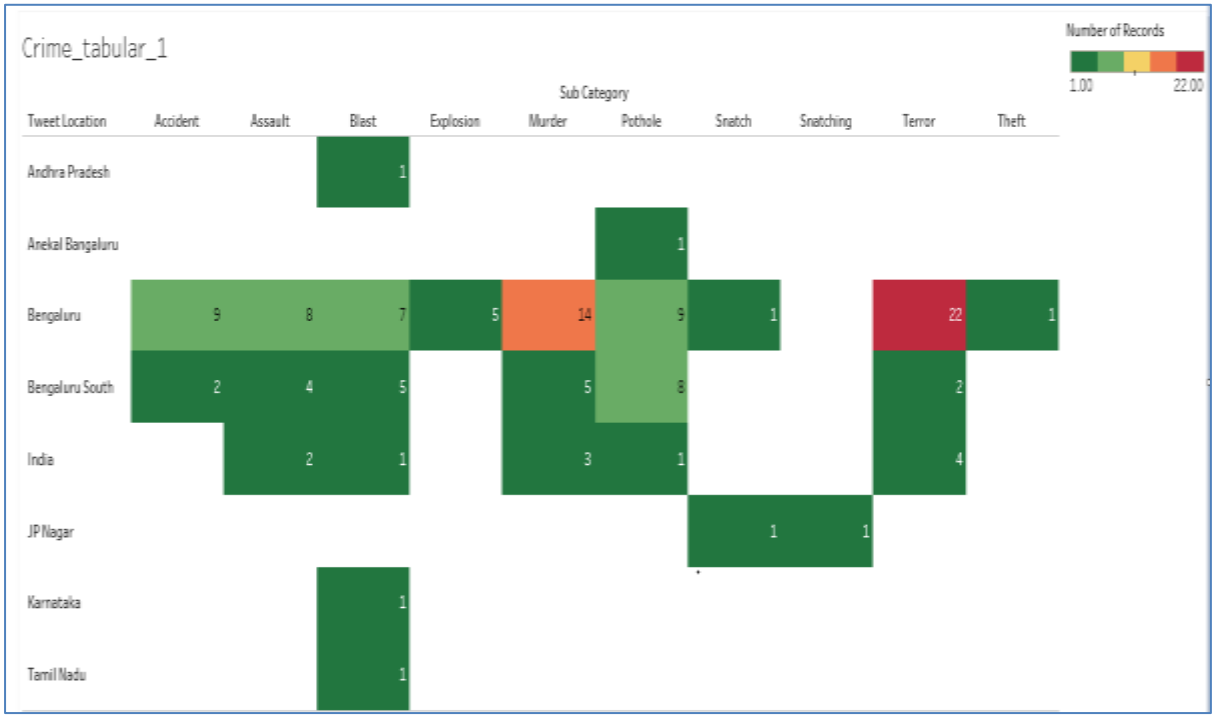

Figure 1 Location and Crime related Tweets using Tableau

Crime Spots after enriching the data with BBMP Assembly Constituency Information

Figure 2 Crime spots based on Assembly constituencies

5. FINDINGS/DISCUSSION

With the help of the analytical approach, we were able to locate the areas and a respective number of crimes reported in those areas. A very high-level view of ward/constituency wise crime rate is presented. For example, Bangalore south reported more incidents of assaults compared to other parts. We were able to identify the top crimes in various parts of Bangalore city. Also based on the nature of the crime, we were able to classify which place was more prone to a particular type of crime when compared to others. A heat map of crimes across the various assembly constituencies in Bangalore was created.

6. CONCLUSIONS AND FUTURE WORK

The work done as part of this paper helps in arriving at a very high-level view of crime reported location wise, top crimes and the crime-prone areas based on nature of the crime. At the next phase, the authors would like to add the demographic factors of Bangalore city. Further, the data from BBMP website would be mapped where there is ward level information which includes data on the number of police stations, number of street lights etc. Further studies will focus on exploring whether these demographic factors have any correlations between the crime rate and these factors. The current dataset would be enriched with these external data and would be trying to draw any new conclusions. Also, the plan is to extend the work to analyse other major cities in India once we complete our analysis of Bangalore city. One major extension of this work would be the ability to predict crime based on past data, demographic factors etc.

REFERENCES

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of Twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics.

Alsudais, A., Leroy, G., & Corso, A. (2014, June). We know where you are tweeting from: Assigning a type of place to tweets using natural language processing and random forests. In Big Data (BigData Congress), 2014 IEEE International Congress on (pp. 594-600). IEEE.

Barbosa, L., & Feng, J. (2010, August). Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 36-44). Association for Computational Linguistics.

Bendler, J., Ratku, A., & Neumann, D. (2014). Crime mapping through geo-spatial social media activity.

Corso, A.J., Leroy, G., Alsusdais, A. (2015). "Toward Predictive Crime Analysis via Social Media, Big Data, and GIS". In iConference 2015 Proceedings.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, 115-125.

Kalampokis, E., Hausenblas, M., & Tarabanis, K. (2011). Combining social and government open data for participatory decision-making. Electronic participation, 36-47.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., ... & Blanford, J. (2011, October). Senseplace2: Geotwitter analytics support for situational awareness. In Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on (pp. 181-190). IEEE.

Sloan, L., & Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geo-services and geotagging on Twitter. PloS one, 10(11), e0142209.

Thom, D., Bosch, H., Krueger, R., & Ertl, T. (2014, January). Using large-scale aggregated knowledge for social media location discovery. In System Sciences (HICSS), 2014 47th Hawaii international conference on (pp. 1464-1473). IEEE.

Wang, X., Gerber, M. S., and Brown, D. E. 2012. "Automatic Crime Prediction Using Events Extracted from Twitter Posts,"

X. Wang, D. Brown, M. Gerber, "Spatio-temporal modelling of criminal incidents using geographic, demographic, and Twitter-derived information", in Intelligence and Security Informatics, Lecture Notes in Computer Science, IEEE Press, 2012.

WEBLIOGRAPHY

Business Insider - http://www.businessinsider.com/twitter-crime-predict-2014-4?IR=T (Accessed on 10/10/2017)

MIT Technology Review - https://www.technologyreview.com/s/524871/can-twitter-predict-major-events-such-as-mass-protests/ (Accessed on 13/10/2017)

Karn Pratap Singh, (2017, February 27th) Hindustan Times, http://www.hindustantimes.com/delhi-news/delhi-police-is-using-precrime-data-analysis-to-send-its-men-to-likely-trouble-spots/story-hZcCRyWMVoNSsRhnBNgOHI.html (Last accessed on 26/11/2017)

Wikipedia - https://en.wikipedia.org/wiki/National_Crime_Records_Bureau (Accessed on 02/10/2017)