Detection of Crime Prone Areas using Twitter Feed

Irfan Siddavtam
Information Technology, KJSCE
irfansiddavatam@somaiya.edu
Mumbai, India

Dhruv Doshi
Information Technology, KJSCE
dhruv.doshi@somaiya.edu
Mumbai, India
1814002

Shubham Bhakuni
Information Technology, KJSCE
shubham.bhakuni@somaiya.edu
Mumbai, India
1814006

Labdhi Jain
Information Technology, KJSCE
labdhi.jain@somaiya.edu
Mumbai, India
1814015

Kunj Gala
Information Technology, KJSCE
kunj.gala@somaiya.edu
Mumbai, India
1814021

***Abstract*—This paper sought to identify areas that are prone to crime in New York City using the latest data from Twitter feed. A pipeline is suggested in the paper which takes tweets from Twitter handles as input and gives geo-coordinates from the location details mentioned in the tweet after selecting the tweets which are related to criminal activity. Data is then plotted on a New York City map for better visualization.**

Keywords—crime prone areas, New York, NYC crimes, twitter feed, Boolean model, Named Entity Recognition, data analysis.

I. Introduction

Crime rate is increasing day by day and even though crimes can occur anywhere, they occur more in some places than others. When selecting a neighborhood to move to, crime rate is one of the important factors people consider. For authorities too, knowing the high crime rate areas is crucial to ensure the community's safety. Hence, the identification of crime prone areas will not only benefit the citizens but also the authorities to make better and informed decisions.

But, only analyzing the already available historic data will limit the scope of the results. To get fresh and updated data, such a source of data is required where the data gets constantly refreshed and updated. What better source of data than Twitter? Twitter being a microblogging and social media application, where people use short messages called 'tweets' to communicate, provides us with latest data. If verified accounts are selected for data extraction then the data obtained will be reliable. Latest news updates related to crime in New York City can be obtained from the feed of such Twitter handles.

This solution will help the authorities continually identify areas prone to crimes and policing-related incidents, making deployments and patrolling in those areas more effective and thereby preempt and prevent any untoward incidents that could result in an emergency

II. Literature Survey

(Refer Table 1)

III. Proposed Methedology

The following pipeline was followed in sequence, starting with extracting the raw tweets to classifying each tweet as related to crime (crime tweet) or not related to crime (non-crime tweet). Tweets from any twitter handle can be passed through this pipeline and the tweets classified into 2 categories will be given as output.

A. Selection of Twitter Handles:

Twitter handles that regularly update their crime news on their feed were required. Two Twitter handles were identified that regularly posted updates on crimes in New York City. First, is the New York Police Department's official Twitter handle named NYPD News (@NYPDnews). Latest crime updates and wanted criminals' news is posted on this account. Second, an account handled by NYPD named NYPD Crime Stoppers (@NYPDTips). The public can send anonymous tips to these accounts and in return they may get up to $3500 reward for a lead.

B. Extract Tweets using Twitter API:

The Tweepy library is Python is used to extract tweets using the Twitter API. This API allows the user to

1) *Get tweets from a timeline*
2) *Follow and unfollow accounts*
3) *Creating and deleting tweets*

As per requirements, only the first functionality of the Twitter API was useful. 200 tweets from each handle were extracted (200 was the limit per handle). Hence, these 400 extracted tweets were converted to Pandas dataframes by appending data from both sources.

Sr. No.	Name	Publication Type	Publication Year	Publication Agency	Aim	Conclusion
1.	Detection of crime and non-crime tweets using Twitter	Journal	2018	IJARCCE	Crime Detection system (CDS) detects if the post is related to crime or not. If the post is related to crime, then it is further classified into the subtype of crime: Crime against the person, property, country, etc.	The system designed in the paper achieved 93% accuracy in classifying tweets into crime and non-crime tweets.
2.	Analysis and Classification of Crime Tweets	Conference	2019	ICCIDS	The purpose of authors was to classify tweets into crime and non-crime. Here, the tweets that were classified as crime needed police action whereas others were general tweets. The authors have implemented approach based on text mining techniques for classification of tweets.	In this paper Naive Bayesian, J48, Random Forest, and ZeroR classifiers were used for classifying 369 tweets. Among all the algorithms, Random Forest outperformed all classifiers with an accuracy of 98.1% while classification of tweets into crime and non-crime.
3.	A Literature Review on Twitter Data Analysis	Journal	2016	IJCEE	This paper gives a detailed review about techniques used in analysis of twitter; it includes analysis of twitter's network-topology, spread of event over the network, hashtags, influence identification and lastly, sentiment analysis.	They were able to calculate the life cycle of topics using amount of tweets in a given time along with the sentiments of the users using NLP and ML techniques.
4.	A Basic Approach for Extracting and Analyzing Data from Twitter		2020	Springer	This paper presents different approaches to collecting tweets information. The paper covers a detailed study of the method used in extraction of tweets such as using twitter API and python Twint library. NLP and ML techniques were implemented to perform classification and extraction on data.	The authors were able to give detailed study about the approaches and methods to collect and analyze semantic information from tweets.

5.	Keyword extraction from Tweets using NLP tools for collecting relevant news	Conference	2020	IEEE	In this paper, a method was introduced to extract the prime keywords from the corpus of tweets and then based on those keywords the relevant news was retrieved which was uploaded on Twitter.	In order to extract keywords, the authors make use of NER, NLP, POS tagging and TF-IDF. With a total accuracy of 67.6%, 107 were identical records and 231 were positive records.
6.	Twitter Based Information Extraction	Conference	2017	IJNTR	presents a brief study of the work which shows that extracting useful information from Twitter and other social media platforms is indeed feasible	The authors provided detailed information about the extraction techniques applied which were based on this subject like the tasks involved in extraction of tweets, inputs, the types of methods of extraction used and the output produced.
7.	Automatic Crime Prediction Using Events Extracted from Twitter Posts	Conference	2012	Springer	In this paper, the authors presented an initial investigation related to criminal incident prediction based on twitter.	The paper concluded with results that were able to show ability of model to forecast cases of hit-and-run crime by using details provided by the training set.
8.	Automated Crime Tweets Classification and Geo-location Prediction using Big Data Framework	Journal	2021	TURCOMAT	The authors focused on a very complicated problem because of is heterogeneity, nature and the amount of data – automatic classification of tweets,	used Multiclass, Multi-level NB classifier; the results obtained from these classifiers were checked against and report floated by NCRB which showed that the data ws 82% correctly plotted.

Table 1: Literature Review

C. *Pre-processing tweets:*

Pre-processing the tweets forms a necessary step since tweets consists of a lot of information which is of no use for further processing. Hence, a procedure must be followed to remove unwanted information. This includes steps like- Removing unwanted characters (like @, # and $) since Twitter users use these symbols to tag other users. Next, tokenization is the process of separating out words from sentences. Using NLTK library in Python, tokenization can be easily carried out. Next, stop words like ‘and’, ‘at’, ‘or’, etc. need to be removed since they provide no significant impact on the end results. Once this is done, all the tweets need to be converted to lowercase since lowercase and uppercase words hold the same meaning here. For example, the words *‘murder’* and *‘MURDER’* hold the same importance. Hence, they must be converted to any one case, i.e., lowercase. The tweets have been cleaned and now ready for further processing.

D. *Dictionary building:*

The next objective is to decide if a tweet has information related to a crime or not. If it does, then it is called a ‘crime tweet’ and if it doesn’t then it is called a ‘non-crime tweet’. To help classify the tweets, a crime dictionary is required which contains words indicative of crime, which was not readily available. Hence, a dictionary-building approach was selected. To build a crime words dictionary, an invertedindex was created using the tweet corpus to help identify which crime related words are used frequently in the tweets. Manual labelling was carried out to mark the frequently occurring words which were related to crime. After the completion of this step, a crime dictionary consisting of 66 crime indicative words, was ready to be used for classification.

E. *Identification of crime tweets using Boolean model:*

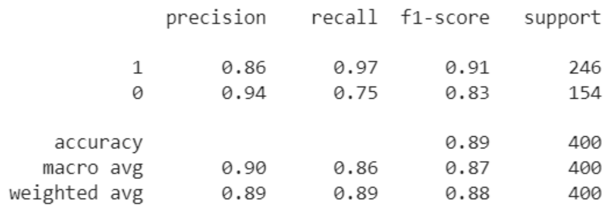
A simple yet efficient information retrieval model- Boolean Model was used to classify tweets as crime and non-crime. The Boolean model requires a query and corpus as input and gives a labelled corpus as output (label ‘1’ for relevant documents and ‘0’ for not relevant documents). Initially, a data structure called term-document incidence matrix is created. In this, each row corresponds to a term and each column corresponds to a document. Hence, the size of matrix will be 400 x 66. Each row in the matrix is a term incidence vector (or simply term vector). Each column is equivalent to a document vector. The values in the matrix are kept in sorted order for processing efficiency. By using this matrix,a particular term can be easily found. In the Boolean model, a document can be categorized relevant or non-relevant to a query but there is no degree of relevance. The model is established on Boolean logic and classical set theory. Terms are viewed as Boolean variables and the value of a term is either true(1) if the term is present in the document or false(0) in the other case. Just as documents are interpreted as document vectors, Boolean expressions are used to demonstrate queries. Boolean IR model doesn’t rank documents. This model produces high recall and lower precision. Using this model only the crime tweets are retrieved and the non-crime tweets are ignored. Other more complex models like TF-IDF could also be used but that level of complexity was uncalled for in this retrieval.

F. *Verifying classification results*

To verify the classification results, confusion matrix was calculated along with scoring parameters like precision, recall and accuracy. For the ground truth, manual labelling was performed. Following are the results:

	Crime	Non-crime
Identified	239	39
Not identified	7	115

Table 1: Confusion matrix



```
              precision    recall  f1-score   support

           1       0.86      0.97      0.91       246
           0       0.94      0.75      0.83       154

    accuracy                           0.89       400
   macro avg       0.90      0.86      0.87       400
weighted avg       0.89      0.89      0.88       400
```

Fig 1. Classification scoring metrics

G. *Extracting location information:*

After selecting the crime tweets, the location mentioned in the tweet has to be extracted. For example, a tweets states, “*WANTED for a ROBBERY: On 3/5/22, at 4 AM, near W. 23 St &amp; 7th Ave in Manhattan, the suspect approached a 31-year-old man…*”. From this entire tweet, only the location related keywords, i.e., *7th Ave in Manhattan* need to be extracted. This is done using Named Entity Recognition algorithm (with the help of spaCy, a free-open source library for Natural Language Processing in Python).

H. *Obtaining geo-coordintes:*

Finally, the extracted location keywords are converted to geo-coordinates so that it is possible to plot the data points on a New York City map and visualize where are crime prone areas are. This is done using the Geocoding API in Python. For example, the geo-coordinates for 7th Ave in Manhattan are 40°45′39″N 73°59′02″W. It is possible to plot these values on a map rather than the location name.

I. *Borough-wise grouping:*

Now that geo-coordinates and location keywords are associated with the tweets, it is possible to group the crime locations according to their boroughs. Borough is an administrative district/ town in New York City. Hence, by using groupby functions in Python or by simply running a loop over the location keywords, the tweets can be associated with a borough.

With the completion of this step, we have identified the crime tweets and have the location and geo-coordinates of the location where the crime took place, as mentioned in the tweet.

III. RESULTS

Out of the 400 tweets extracted, 278 tweets were identified as crime out of which 239 were actually crime tweets whereas 39 were falsely identified as crime tweets. On the other hand, 122 tweets were identified as non-crime tweets out of which 7 were falsely rejected since they were crime tweets. From the 278 crime tweets, the location keywords of 119 tweets were detected. Further, the location of 88 tweets was detected correctly, the rest of the locations were out of bound and hence were ignored.

Finally, these 88 crime locations were plotted on a New York City map using Folium library in Python. Folium maps in Python provide us with FastMarkerCluster module which helps to cluster the data points visually. Blue markers on the map indicate the location of crime.

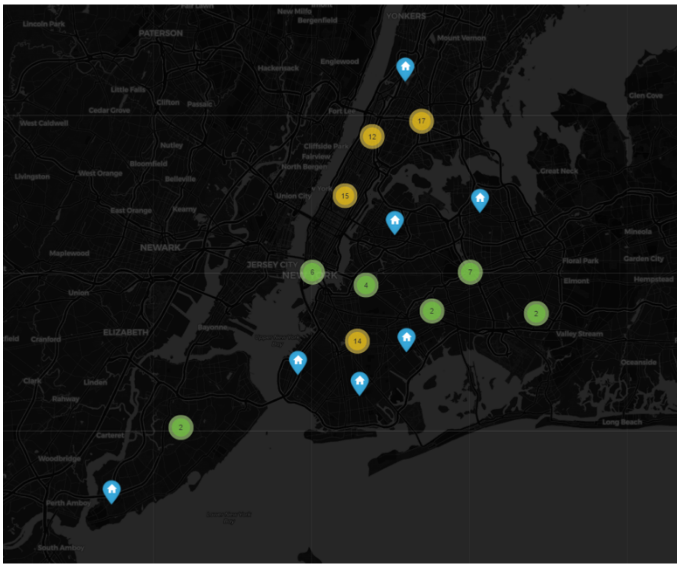


Fig 2. Plotting the crime locations on NYC map

Borough-wise count is also obtained from the pipeline created. The following table shows borough-wise distribution of the crimes:

Borough	Count
Brooklyn	22
Bronx	22
Manhattan	19
Queens	10
Staten Island	3

Table 2. Borough-wise crime distribution

IV. Conclusion

To conclude, a pipeline was created to facilitate the identification of crime location from the raw data pulled from any Twitter handle which provides data on criminal activity. This can actually be useful to authorities who want to analyze the patterns related to crime locations or want to identify places that are more prone to crime. Using this information, they can take necessary steps like increasing the number of CCTV cameras in a particular street, increasing patrolling frequency or facilitating better and faster ways to report a crime in crime prone areas.

The scope of this project can be increased in multiple ways. First, the number of tweets per handle was restricted to 200 in this project. But it is possible to increase this number to thousands and this would in turn make the results richer. Second, if it is not possible to increase the number of tweets extracted per handle, the number of Twitter handles taken under consideration can be increased. In this project, 2 handles related to crime in New York City were considered, increasing this would improve the results. Third, the scope can be taken beyond the bounds of New York City as well. Twitter feeds reporting criminal activity in other cities or countries can also be used to get more diversified results.

Hence, the pipeline is flexible and can also be manipulated to work according to the user's needs.

References

[1] N. Sharma, A. Dhamne, N. more, V. rathi and A supekar, "Detection of crime and non-crime tweets using Twitter", in IJARCCE, vol. 7, pp. 229-232, 2018.

[2] S. Tiwari, R. Ranjan, A. Verma, N. Sardana and R. Mourya, "Analysis and Classification of Crime Tweets", in ICCIDS, vol. 167, pp. 1911-1919, 2020, doi: 10.1016/j.procs.2020.03.211.

[3] H. Anber, A. Salah and A. El-Aziz, "A Literature Review on Twitter Data Analysis", in IJCEE, vol. 8, pp. 241-249, 2016, doi: 10.17706/IJCEE.

[4] Xavier, C.C., Souza, M., "A Basic Approach for Extracting and Analyzing Data from Twitter", in Special Topics in Multimedia, IoT and Web Technologies. Springer, 2020, doi: 10.1007/978-3-030-35102-1_7

[5] T. D. Jayasiriwardene and G. U. Ganegoda, "Keyword extraction from Tweets using NLP tools for collecting relevant news," International Research Conference on Smart Computing and Systems Engineering (SCSE), 2020, pp. 129-135, doi: 10.1109/SCSE49731.2020.9313024.

[6] M. Kumar, A. Garg, A. Munjal, A. Tanwar, "Twitter Based Information Extraction", International Journal of New Technology and Research (IJNTR), vol. 3, pp. 52-55, 2017.

[7] Wang, X., Gerber, M.S., Brown, D.E., "Automatic Crime Prediction Using Events Extracted from Twitter Posts in Social Computing, Behavioral - Cultural Modeling and Prediction, Springer, vol. 7227, pp. 231–238, 2012, doi: 10.1007/978-3-642-29047-3_28.

[8] Dr.K. Santhiya, Dr.V. Bhuvaneswari, V.Murugesh, "Automated Crime Tweets Classification and Geo-location Prediction using Big Data Framework", in Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, 2021.

[9] Heidi Cohen, "How reliable is Twitter?", 2013 Accessed on: 01/12/2022 [Online] Available: https://heidicohen.com/reliable-twitter-research/

[10] N Schwitter, "GOING DIGITAL Web data collection using Twitter as an example", 2020, Accessed on: 01/12/2022 [Online] Available: www.oxfam.org

[11] "Twitter API documentation" 2022, Twitter Developer Platform, Accessed on: 01/15/2022 [Online] Available: https://developer.twitter.com/en/docs/developerportal/overview

[12] Zhang AJ, Albrecht L, Scott SD, "Using Twitter for Data Collection With Health-Care Consumers: A Scoping Review", 2017 International Journal of Qualitative Methods, 2017, doi: https://doi.org/10.1177/1609406917750782

[13] W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in IEEE Access, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

[14] S. Sathyadevan, M. S. Devan and S. S. Gangadharan, "Crime analysis and prediction using data mining," 2014 First International Conference on Networks Soft Computing (ICNSC2014), 2014, pp. 406-412, doi: 10.1109/CNSC.2014.6906719.

[15] M. A. Boni and M. S. Gerber, "Area-Specific Crime Prediction Models," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 671-676, doi: 10.1109/ICMLA.2016.0118.