**An Approach for Real Time Crime Prediction Using Twitter**

# Introduction:

Criminal activities happen more likely at the space-time gathering of the potential attackers, victims, and non-appearance of any predictive elements. Hence, it is very important to predict patterns that could help to prevent these crimes by effectively predicting crime and policing. The regular conventional crime-prediction models are lacking mainly in taking the sentiment of social media content into consideration. Social networking sites like Twitter and Facebook have a lot of unseen potentialities to reveal useful information when users post their daily activities on these platforms. Each textual post is compiled to generate an unstructured form of data and this data can be considered a very strong tool for analyzing the factors behind crime and future crime prediction. Social networking platforms like Twitter have a strong potentiality in predicting and describing scenarios like election results, natural disasters, and crimes. In order to source the crime prediction model, Twitter data is supplied to the built model. Hence, it is necessary to investigate the relationship between our text-enriched next-place predictions and the occurrence of crimes to come up with a more meaningful prediction model to predict the crimes more accurately (Chen et al., 2015). In this model, we will be getting all the GPS tagged tweets matching the keywords of the different crime categories using the latitude and longitude coordinates. The extracted tweets will be pre-processed so that all the garbage text is cleaned and once the pre-processing is done the model will be trained for predicting the crimes accurately. After training the model we will be giving the user a choice of selecting a particular state where the coordinates of the state are sent to the model and all the tweets related to the crime are extracted and categorized into different categories. Based on the total number of tweets extracted for each category, ranking will be done. These rankings can help the police authorities to focus more on the top criminal activity that is happening in that particular state.

# Problem

# 2.1. Problem statement:

Automated crime prediction models are some of the strongest tools for expanding the limited resources to prevent criminal activities. Nonetheless, regular conventional crime predicting models that use Twitter data have certain limitations in describing the up-to-date reflection of criminal activities. Hence the rate of violent crimes in the United States increases consecutively every year and making the people feel unsafe.

# 2.2. Motivation to study, significance of problem, and potential benefits

According to the recent report of the Federal Bureau of Investigation, it has been evident that violent crimes in the United States have been increased for the second consecutive year making the need to consider crime prediction with thoughtfulness. Hence, in this model, real-time tweets will be used for crime classification and their ranking which will provide real-time information on the trending crimes in different states of the United States of America. This information will then be used by the necessary officials to allocate limited resources and limit the crime rate effectively. For example, if a crime like a robbery has shown an unusual increase in the recent week the police department can concentrate more on patrolling in those areas to identify potential suspects who have already committed similar kind of crimes in the past and monitor them closely to control the future crimes.

# 2.2.1. Project Goals

# The primary goal of this project is to build a model to make predictions about possible crimes and rank them based on the frequency of each crime.

# The model will also help to alert the police officials to limit their resources and target more towards a particular crime based on the rankings generated by the model.

# The model should be accurate enough to rely on it and should be able to predict the crimes properly so that necessary actions can be taken to reduce crime.

# 2.2.2. Project Objectives

The main objectives of this project are to predict crimes accurately and rank different crimes based on their frequency and notify the law enforcements so that they can concentrate more on the areas where crime is more prevalent to prevent possible future crimes. The main intention of this project is to reduce crime and make people feel safe.

1. **Prior Work**

The accelerated increase in the total volume of users in social networking sites has added the predictive capability in broad fields allowing us to analyze and predict the reactions among the groups. Some of the forecasting models established on social media are mainly used for predicting the results of the election (Asur and Huberman, 2010), sentimental analysis of the movie performances (Bermingham and Smeaton, 2011), and the analysis of the trending stock market (Bollen et al., 2011). These models mainly use the techniques of sentiment analysis. Many researchers utilize the circumstantial data of the tweet and perform semantic analysis to forecast the response of certain groups of people. Nonetheless, earlier studies are lacking in prediction for a wide range of populations and these studies collected only limited data from a preferred group of people to analyze and make a prediction of their future response in similar circumstances. Due to this incomplete information collected from these models they failed to picture the total response against the criminal incidents. The limitations of the other research studies to represent people's responses related to criminal incidents, this model is built from approaching the situation from a different perspective to reduce the shortened information concerning people's selection. Other types of research developed some crime prediction models using the topic modeling on the extracted Twitter data. One of the earlier studies done by Wang et al mainly focused only on the tweets from the latest bureaus to analyze the correlation among the topics, which were mainly used in the tweets and different types of crimes like hit and run and burglary cases (Wang et al., 2012). After Gerber et al combined the historical crimes with the GPS-tagged data extracted from Twitter and built a prediction model. The data extracted for this model has been collected from the Chicago city area. But this model referred only topic modeling without applying the sentiment analysis and not considering the weather factors on the obtained Twitter data as these factors might influence crimes in combination with the Twitter messages (Chen et al., 2015).

Numerous efforts have been undertaken to utilize the content obtained from the micro-blogging data to analyze and forecast real-time updates, social rivalry, and many other public related risks. The obtained content contains patterns relating to various communities, points of view, and social behaviors. Patterns collected from the individuals are the result of employing various phases, suggesting various sentiments, or having differentiating language design among the users. Taking all these social topics, little scrutiny has been paid to the effect of the data generated and their relations with the crime-related incidents from the online users. Although most of the researches are defined to particular locations, types of crimes, users, societies, and specific events but, the built-in this model can be generalized to any selected location. The model proposed employs content without segregating keywords to establish forecasting signals rather than using the content of the previous incidents. As this is the first model that considers crime forecasting as a trend prediction it cannot be related or compared to the previous models. The assumption of this model is to investigate whether a relationship exists between the previously obtained content and future crimes. The interest in generating the trends related to the crimes is to make the policymakers, government officials, or the law enforcement agencies to see whether these crime patterns in the neighborhood are declining (Aghababaei and Makrehchi, 2016).

Many crime forecasting methods have employed different machine language techniques like regression analysis, support vector machine, and kernel density estimation using data like statistical data and social media data. (Liao et al, 2010) built a model based on Bayesian-based crime prediction employing geographical information. In this model, the crime site data has been segregated into private and public regions by using discrete decay functions to generate geographic profiles, which is equal to the probability distribution of crime occurrences. Lastly, to precisely measure the location of the next possible crime geographic profiles are combined with the Bayesian learning theory. (Shingleton, 2012) employed a regression-based approach to mainly predicting three types of crimes like violence, homicide, and assault in Salinas, California based on least squares, Poisson regression, and negative binomial regression models. The obtained results approve identical performances among the 3 models employed. Although, both negative binomial regression and Poisson regression models are based on the assumption that data should follow the Poisson distribution and if the data is unable to fit using Poisson distribution then ordinary least squares are used (Kang, 2019).

1. **Methodology** 
   1. **System Architecture**

**A close up of a sign

Description automatically generated**

**Application’s System Architecture**

* 1. **Approach**
     1. **Data Extraction**

As this approach mainly deals with the twitter data. The data set for prediction is extracted from twitter using the keywords for respective crimes. In order to make a connection with Twitter, there is a need to register for the Twitter’s developer account first to use the Twitter API. Once, the registration is complete twitter provides a set of keys like consumer\_key, consumer\_secret, access\_token, and access\_token\_secret which can be used inside the application to interact with Twitter API and extract tweets. But this API only allows limited requests per day and the maximum number of tweets we can fetch using this API is limited to 100. Due to this reason I have used a twitter library called “GetOldTweets” to fetch the historical crime data using the specific keywords. Since, more data is required for better prediction I have collected the entire world’s crime data rather than limiting it for just USA. I have collected more than twenty-five thousand tweets using “GetOldTweets” library. Multiple crimes have been categorized into single crime category and the keywords used for tweet extraction and the crime category are listed below.

# keywords used for extracting the tweets

Rape keywords = '#molested OR #rape OR #statutoryrape OR #sexualvoilence OR #molest OR #gangrape OR #girlabuse'

Assault keywords='#assault OR #childabuse OR #abuse OR #domesticabuse OR #kidnapping'

Theft keywords='#burglary OR #larceny OR #robbery OR #autotheft OR #shoplifting OR #theft'

Murder keywords = '#murdered OR #murder OR #killed'

Statutory keywords = '#drugcrime OR #trafficoffense OR #financialcrime OR #fraud OR #blackmail OR #drugtrafficking'

The tweets are extracted with the following columns

# columns of the csv file

COLS = ['id', 'created\_at', 'source', 'original\_text', 'clean\_text', 'sentiment', 'polarity', 'subjectivity', 'lang','favorite\_count', 'retweet\_count', 'original\_author', 'possibly\_sensitive', 'hashtags','user\_mentions', 'place', 'place\_coord\_boundaries']

Below are the sample tweets extracted from the Twitter using “GetOldTweets” library and stored in system storage in CSV format.

A screenshot of a cell phone

Description automatically generated

**Sample Tweets Collected**

* + 1. **Feature Selection**

In this stage I have removed all the unnecessary columns from the extracted data like created\_at, source, subjectivity, lang, favorite\_count, retweet\_count, original\_author, possibly sensitive, user\_mentions, place, and place\_coor\_boundaries. The main features that are selected for model building are original\_text and hashtags through which we can identify the crime hashtag and description of the crime in the tweet. Once the required features are selected the data is sent for cleaning and preprocessing.

* + 1. **Data Preprocessing**

As the extracted data contains missing values, null values, and garbage text there is a need to clean and preprocess the data. Based on the extracted tweets I have noticed that it contains a lot of non-ASCII words, repeated text or duplicate tweets, and some irrelevant symbols which needs to be removed to make the data perfect for building the model. I have used some preprocessing techniques like tokenization, stemming, and stop words removal.

**Stop Words:** Using this process all the stop words like “ ‘but’, ‘again’, ‘there’, ‘about’, ‘once’, ‘during’, ‘out’, ‘very’, ‘having’, ‘with’, ‘they’, ‘own’, ‘an’, ‘be’, ‘some’, ‘for’, ‘do’ ” have been removed from the tweets to make the analysis more accurate.

**Tokenization:** In this method all the words inside the tweets are converted into tokens and unwanted tokens which are not required for the analysis will be removed. Punctuation and macron text are also removed as they don’t play any major role for building the accuracy and might even alter the model’s accuracy prediction.

**Stemming & Lemmatization:** With the help of stemming and lemmatization we can convert all the words into their core or base form. This process will help to predict the words easily as it is easy to match the words in the base or original form. This process mainly converts all the words that are connected with each other into a central word if they have the same core meaning.

Few other methods like using regular expressions to remove the extra white spaces, numbers, and links from tweets to make the data perfect for prediction. Sample tweets after feature selection and preprocessing are shown below.

A screenshot of a newspaper

Description automatically generated

**Sample tweets after Preprocessing**

* + 1. **Data Exploration:**

After preprocessing I have explored the data collected using tableau to see the keywords and percentage of tweets extracted for each crime category and few other analyses.

A picture containing vector graphics

Description automatically generated

**Tweets Extracted for Each Crime Category**

Word cloud for the keywords extracted from the tweets is shown below. Using the below figure, we can identify that frequency of crime related keywords is high compared to other words.

A screenshot of a social media post

Description automatically generated

**Word Cloud for the Tweets Extracted**

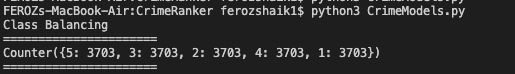
Top hashtags extracted for each class are shown in the below figure. Using the below figure, we can see that the categories are classified correctly as per our requirement.**A picture containing screenshot

Description automatically generated**

**Top Hashtags for Each Crime Category**

* + 1. **Balancing Classes Labels:**

As there are multiple classes involved in accuracy prediction it is very necessary to balance all the classes to train the model for accurate accuracy prediction. The classes have been balanced using a method called under sampling. From the below figure it is clear that all the classes are equally distribute after performing under sampling on the dataset.



**Class Balancing using Under Sampling**

* 1. **Model Building and Algorithm Selection:**

In this phase, a predictive model will be built using the pre-processed data to classify the extracted tweets into their respective categories and after training the model with preprocessed data a multiple core algorithm like Naïve Bayes, Logistic regression, and Linear SVC will be applied for model’s accuracy prediction. As it is difficult to identify the suitable algorithm for crime prediction it is very essential to run different models. Based on the accuracy and other evaluation identifiers such as F-measure, recall, and precision we can predict the best suitable algorithm for our application.

* 1. **Real-Time Prediction**

Once the model is selected then the model can be used to predict the live test dataset is fetched from twitter. Based on the keywords and state boundaries twitter will fetch the latest tweets posted on twitter. Once the tweets are extracted, they are saved inside a JSON file and then converted to a local csv file. The csv file is then used for data cleaning and preprocessing. The preprocessed data is then sent into the built model to predict the class labels for the tweets. Based on the total number of tweets retrieved and tweet frequency with respect to each crime category ranking is done. Real-Time crime prediction for Texas state is shown in the below figure.

A screenshot of a cell phone

Description automatically generated

**Real-Time Crime prediction**

* 1. **Building User Interface:**
     1. **Home Page**

In the final phase, a user interface will be built so that the user can make his/her choice of selection and check the crime prediction for a particular state. The homepage for the application is shown below figure.

**A close up of a map

Description automatically generated**

**Home Page for Crime Prediction**

* + 1. **Loading Screen:**

Once the user selects a state a spinner will be displayed on the right side indicating that the process has been started and results will be fetched soon. The below image shows the visual representation of loading state on the UI.

A close up of a map

Description automatically generated

**Loading Screen for Crime Prediction**

* + 1. **Results Screen:**

Once the processing is done the results will sent back to UI. Once the response is detected UI will stop the spinner icon and a results table will be shown on the screen. The table displayed contains the top crimes in that particular state. Below images show the visual representation of results page.

A close up of a map

Description automatically generated

**Results Screen Indicating the Top Crimes**

1. **Results and Analysis:**

**Naïve Bayes Model:**

**A close up of a logo

Description automatically generated**

**Accuracy Score for Naïve Bayes Classification Model 🡪 0.6521 (65.21%)**

**Logistic Regression:**

**A close up of a logo

Description automatically generated**

**Accuracy Score for Logistic Regression Model 🡪 0.7095 (70.95%)**

**Linear SVC:**

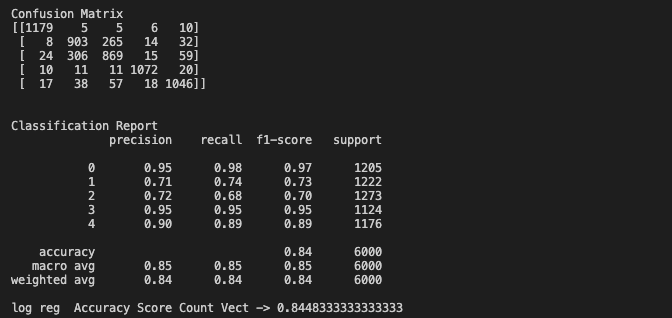
**A screenshot of a cell phone

Description automatically generated**

**Accuracy Score for Linear SVC Model 🡪 0.8546 (85.66%)**

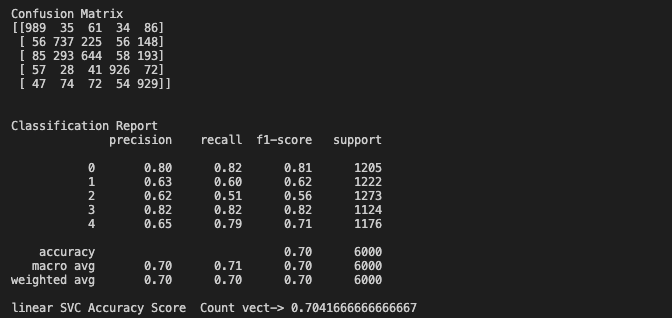
**Accuracy After performing Count Vectorization:**

**Logistic Regression:**

****

**Accuracy Score for Logistic Regression using Count Vectorization 🡪 0.8448 (84.48%)**

**Linear SVC:**

****

**Accuracy Score for Linear SVC using Count Vectorization Model 🡪 0.7041 (70.41%)**

**Naïve Bayes:**

**A screenshot of a cell phone

Description automatically generated**

**Accuracy Score for Naïve Bayes Score using Count Vectorization 🡪 0.8368 (83.68%)**

**Live Prediction Results:**

For predicting live results. I have used twitter stream API which accepts state boundaries as parameters. Using that I have extracted the live tweets and then wrote them to a JSON file and then converted it to CSV and then preprocessed those tweets. After preprocessing I have sent the tweets to the logistic regression model for predicting the class labels for each tweet. Once the tweets are obtained, I calculated the percentages based on their frequency and displayed on the UI.

**North Dakota:**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**Texas:**

**A screenshot of a cell phone

Description automatically generated**

**A close up of a map

Description automatically generated**

**New York:**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

1. **References**

Aghababaei, S., & Makrehchi, M. (2016). Mining social media content for crime prediction. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. <https://doi.org/10.1109/WI.2016.0089>

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, *1*, 492-499. <https://doi.org/10.1109/WI-IAT.2010.63>

Bermingham, A., & Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results. *Asian Federation of Natural Language Processing*, 2-10. Retrieved from <https://www.aclweb.org/anthology/W11-3702.pdf>

Bolla, R. A. (2014). Crime pattern detection using online social media. *Missouri University of Science and Technology*, 33-34. Retrieved from <https://scholarsmine.mst.edu/masters_theses/7321/>

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8. <https://doi.org/10.1016/j.jocs.2010.12.007>

Chen, X., Cho, Y., & Jang, S. Y. (2015). Crime prediction using twitter sentiment and weather. *2015 Systems and Information Engineering Design Symposium*, 63-68. <https://doi.org/10.1109/SIEDS.2015.7117012>

Jarrod S, S. (2012). Crime trend prediction using regression models for salinas, california. *Calhoun: The NPS Institutional Archive DSpace Repository*. Retrieved from <http://hdl.handle.net/10945/7416>

Kang H-W, Kang H-B (2017) Prediction of crime occurrence from multi-modal data using deep learning. PLoS ONE 12(4): e0176244. <https://doi.org/10.1371/journal.pone.0176244>

Liao, R., Wang, X., Li, L., & Qin, Z. (2010). A novel serial crime prediction model based on bayesian learning theory. *2010 International Conference on Machine Learning and Cybernetics*. <https://doi.org/10.1109/ICMLC.2010.5580971>

Wang X., Gerber M.S., Brown D.E. (2012) Automatic Crime Prediction Using Events Extracted from Twitter Posts. In: Yang S.J., Greenberg A.M., Endsley M. (eds) Social Computing, Behavioral - Cultural Modeling and Prediction. SBP 2012. Lecture Notes in Computer Science, vol 7227. Springer, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-642-29047-3_28>