Online Suspect Verification System

[1]Sujit Kumar, [2]Himanshu Ranjan, [3]Aman Deep, [4]Prema S.
[1,2,3]UG Students,[4]Assistant Professor
Department of Computer Science and Engineering
Siddaganga Institute of Technology, Tumakuru, Karnataka, India

***Abstract*—At present crime investigating departments use some of the old outdated file systems to view and analyze the crime details. There are many situations in which data about criminals should be shared with other police stations. When one police station requires any criminal information during their investigations, they need to call at that police station and get it, which is time consuming and require manpower. Also, citizens cannot get any information about criminals and the status of any crime investigation case. Our web based application provides an easy to use platform for the investigating departments concerned with the case in finding the prime suspect. It allows individual investigation teams in different locations to keep track and coordinate on cases. Also, it provides an easy to use interface for general public to submit evidences with respect to specific cases and check the status of any ongoing investigations. It extracts information from huge volume of data entered by the officers/general public in the social networking websites such as Twitter as well as our system. The data extracted is unstructured data and is converted into structured data which is analyzed and frequency of negative words with respect to each suspect is generated. Based on this, the probability of suspects in a crime is generated and the suspect who is having highest frequency of negative words is predicted as the prime suspect.**

***Index Terms*—Crime, suspect, predict, Twitter, web-based.**

I. INTRODUCTION

Nowadays crimes are increasing in our society. But due to inefficient existing system, we are not able to track the real culprit on time. Due to this, victims usually find difficulty in getting the right judgment at right time. The demands of forensics could prove costly as laws and investigations become more complex. Courts are becoming insistent on the need for systems to gather and preserve evidences. Nonetheless, the courts have a right to expect that litigants and counsel will take the necessary steps to ensure that relevant records are preserved when litigation is reasonably anticipated and such records are collected, reviewed and produced. In traditional forensics, the media seized at the crime scene are preserved and analyzed. However, after the collection of evidences, preservation and validation could be manipulated and presented in the court. In the literature we have found only the analysis of existing crime data and predicting the future trend of various crimes using different data mining techniques. None of the work focused on predicting the culprit/suspect in any crime. Therefore, our work focuses on predicting the prime suspect of any crime by using various data and information stored on social networking sites such as twitter and the data relative to suspects stored in the database. This system provides an easy way of collecting evidences with respect to crime scenes described. It also maintains the database of each and every previous criminal.

The basic objective of our work is to provide an easy to use platform for the investigating departments concerned with the case in finding the real suspect and in return provide a ray of hope for the victims, to allow individual investigation teams in different locations to keep track and coordinate on cases, to provide an easy to use interface for general public to provide evidences with respect to specific cases and also check the status of any ongoing investigations and to identify the prime suspect quickly, thereby establish a kind of belief and trust on the administration as well as on the law system in the eyes of victims.

The rest of the paper is organized as follows. We briefly discuss the related work (Section 2) and then present our proposed system (Section 3). Finally, we conclude our work followed by future work that can be carried out (Section 4).

II. LITERATURE SURVEY

Crime analysis [1] is defined as analytical processes which provide relevant information relative to crime patterns and trend correlations to assist personnel in planning the deployment of resources for the prevention and suppression of criminal activities. Authors have analyzed homicide crime and have concluded that it has seen downward trend during 1990 to 2011. They have used the k-means clustering technique for extracting useful information from the crime dataset using Rapid Miner tool as it is solid and complete package with flexible support options. K-means clustering algorithm divides n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Crime analysis is done on the crime dataset recorded by the police in England and Wales.

Linear regression model is used for predicting the occurrence of crimes in the city of Delhi, India [2]. The prediction of future crime trends involves tracking crime rate changes from one year to the next year and use of data mining concept to project those changes into the future. A dataset of the last 59 years (from 1953 to 2012) was reviewed to predict the occurrence of crimes including murder, burglary, robbery, kidnapping and abduction, theft and riot etc. 15 years ahead of time. The system was trained by applying linear regression using crime data taken from National Crime Records Bureau (NCRB), India for different crime types. This work will be helpful for the local police stations in decision making and crime supervision.

Data mining concept is used for crime analysis and prediction in [4]. Using it previously unknown and useful information is extracted from unstructured data. An approach between computer science and criminal justice to develop a data mining procedure which can help in solving crime investigation cases has been proposed. Authors have focused on crime factors of each day instead

of focusing on causes of crime occurrence like criminal background of offender, political enmity etc. Their method includes data collection, classification, pattern identification, prediction and visualization.

The growing availability of information technologies has enabled law enforcement agencies to collect detailed data about various crimes. Classification is the process of finding a model (or function) that depicts and distinguishes data classes or notions, with the end goal of having the ability to utilize the model to predict the crime labels. In [3] an improved method of classification algorithm for crime prediction is proposed by authors. They have compared Naïve Bayesian and Back Propagation (BP) classification algorithms for predicting crime category for distinctive state in USA. In the first phase, the model is built on the training and in the second phase the model is applied. The performance measurements such as Accuracy, Precision and Recall are used for comparing of the classification algorithms. The precision and recall remain the same when BP is used as a classifier. The analysis result shows that the Naïve Bayesian classification algorithm is better than Back Propagation algorithm

Classification is the process of dividing the data into number of groups which are either dependent or independent of each other and each group acts as a class [5]. The classification can be done by using several methods using different types of classifiers. However, it cannot be done so easily when it is to be applied on text documents i.e. document classification. This paper focuses on the multi-class document classification and to achieve high classification accuracy with respect to text documents. Naïve Bayes approach is used to deal with the problem of document classification via a deceptively simplistic model: assume all features are independent of one another, and compute the class of a document based on maximal probability. The Naïve Bayes approach is applied in linear as well as hierarchical manner for improving the efficiency of classification model. It has been observed that Naïve Bayes hierarchical classification technique is more effective than the linear classification technique. It performs better even in case of multi-label document classification.

III. Proposed System

Our proposed system is a web-based application having the architecture as shown in Fig.1. As and when crimes are committed, those details are entered into the database by the administrator, who later assigns officers to investigate the case. Investigating officer can perform various operations as shown in Fig.2. after getting authenticated by providing proper login id and password. During the course of investigation, officers can tweet about the findings in any social networking sites such as Twitter. This account may be followed by other investigating officers/public and provide relevant data, if any. Apart from this, officer can store the details of any suspects and evidences found during investigation into the database. One important feature of this system is that any person can provide evidence after registering in the system. Officer can analyze the twitter data / data in the database to predict the prime suspect in a specific case. Since the data is in unstructured form, it is converted into structured form using pre-processing and sentimental identifier algorithm. Then the frequency generator and suspect prediction algorithms are used to predict the prime suspect in a specific case.

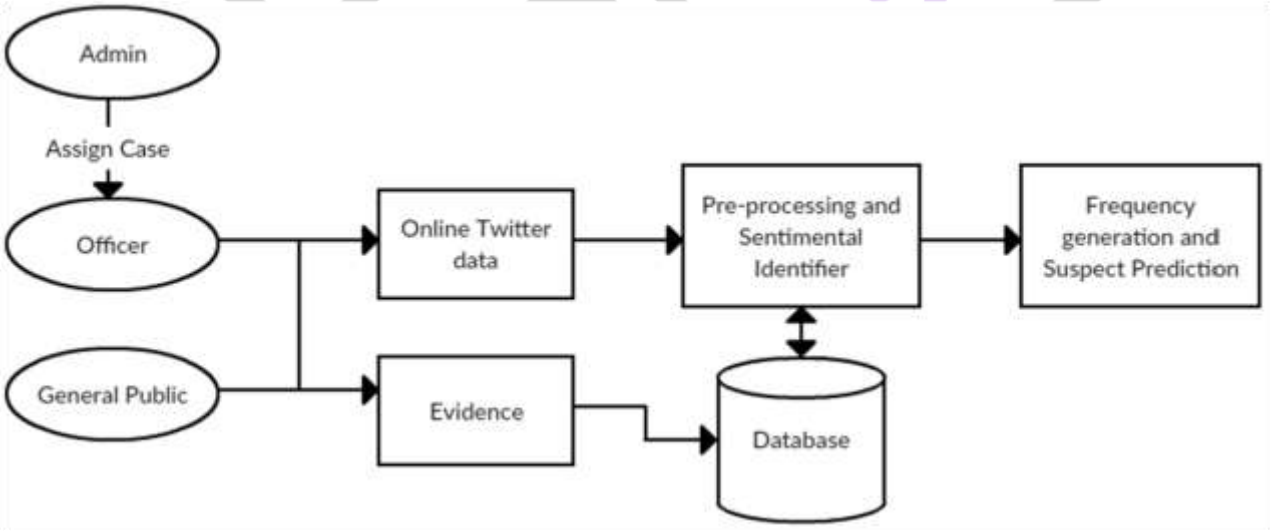

Fig. 1. System Architecture Diagram

Pre-processing

Data extracted from the twitter account as well as data stored in the database is unstructured data and contains words which have no scope in sentiment analysis. The data need to be refined before performing any sort of analytics. Text pre-processing is filtering of the extracted datasets before analysis as shown in the Fig. 3. It includes identifying and eliminating non-textual content and content that is irrelevant to the area of study. Tokenization is the task of chopping the data up into pieces, called tokens, and at the same time throwing away certain characters, such as punctuation. The pre-processor performs tokenization of unstructured data. Dictionaries of suspects, positive and negative words will be there for extracting the data sets. The dictionary words are extracted using a process called stemming. It involves mapping of two or more subsequent words and removing the duplicates to improve the performance of methods. Later, filtering is done to extract the appropriate data sets.

Sentimental Identifier

Sentiment identifier takes the pre-processed list as input and compares each element in the list with the dictionary of words. The dictionary of words contains many pre-defined words with a sentiment value attached with each word. If the element in the pre-processed list matches with the dictionary word, the corresponding sentiment value is stored in sentiment array.

Frequency Generator and Suspect Predictor

The stored sentiment value is processed and frequency for those particular negative words for a specific suspect is generated. Based on the frequency, the attributes such as suspect name, sentiment value and negative words are stored in a sentiment table for finding the most probable suspect.

Suspect predictor predicts the most probable suspect among the suspects in a crime. The final crime frequency is obtained by taking the average of the words to the total number of negative words taken from the dictionary. A suspect with the highest number of negative words i.e. highest crime frequency will be predicted as the prime suspect in a crime as depicted in Fig. 4. It also shows the frequency of each of the negative word with respect to suspects.

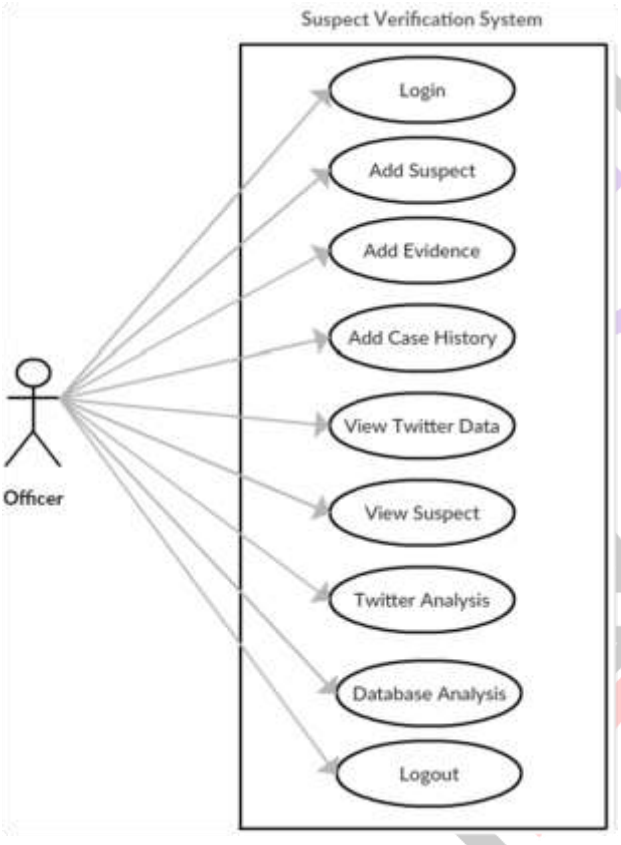

Fig. 2. Usecase diagram for Investigating Officer

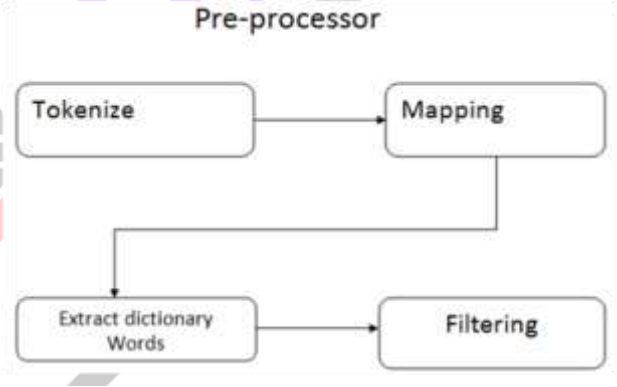

Fig. 3. Pre-processing of Unstructured Data

IV. Conclusion and Future Work

Crime analysis and suspect prediction is an emerging field. We have attempted to extract information from huge volume of data entered by the officers/general public in the social networking websites such as Twitter as well as our web application. The data extracted is unstructured data and is converted into structured data for further analysis. The structured data is analyzed and frequency of negative words with respect to each suspect is generated. Based on this, the probability of suspects involved in crime is generated and the prime suspect is predicted. Our web-based Suspect Verification System fast tracks the process of finding prime suspect of a crime by eliminating the communication delay between investigating organizations by fostering cooperation among them and also, by active participation of general public by way of providing evidence of crimes.

Our work can be further enhanced by incorporating machine learning approach to detect the prime suspect, including image analysis for utilizing more real time data and use of cloud for storing huge volume of data pertaining to a crime.

Fig. 4. Prediction of Prime Suspect

REFERENCES

[1] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime analysis using k-means clustering", International Journal of Computer Applications, Vol. 83, No.4, pp. 1-4, December 2013.

[2] P. Gera, and R. Vohra, "Predicting Future Trends in City Crime Using Linear Regression", International Journal of Computer Science & Management Studies, Vol. 14, Issue 07, pp. 58-64, July 2014.

[3] A. Babakura, N. Sulaiman, and M. Yusuf, "Improved method of classification algorithms for crime prediction", International Symposium on Biometrics and Security Technologies (ISBAST), Kuala Lumpur, Malaysia, 26-27 Aug. 2014.

[4] S. Sathyadevan, Devan M.S., and S. Gangadharan, "Crime analysis and prediction using data mining", IEEE 2014, First International Conference on Networks & Soft Computing (ICNSC2014), Guntur, India, pp. 406-412, 19-20 Aug. 2014.

[5] S. Joshi, and B. Nigam, "Categorizing the document using multi class classification in data mining", International Conference on Computational Intelligence and Communication Systems, Gwalior, India, 7-9 Oct., 2011.