IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-04, Issue-04 E-ISSN: 2347-2693

Exploiting Social Network for Forensic Analysis to Predict Civil Unrest

Ruchika Ganar[1*] and Shrikant B. Ardhapurkar [2]

[1,2] *CT Department, YCCE, Nagpur India*

Available online at: www.ijcseonline.org

Received: Mar/22/2016 Revised: Apr /02/2016 Accepted: Apr/14/2016 Published: Apr/30/2016

***Abstract*—** Big Data analytics is new trending research area in IT industry and social media provides tremendous data for Big Data analysis. Social media analysis mostly includes mining people's opinion because mostly people share their views on social media platform (such as Twitter, Facebook, etc.). The opinions can easily flow in the society using Twitter. It is the easiest way to pass the information in the society. Crimes, riots, unrest, public movements and every activity is being planned or shared on Twitter and it is being delivered to individual within a short span of time. The opinions regarding every situation change as the individual change, so the people's reactions are also different. Sometimes the reaction can change hundreds of people to think the same and react on that which can lead towards civil unrest such as strikes, riots, March etc. Tweets can be analysed to understand the behaviour of the individual and groups. By predicting civil unrest the investigators will get the help to take certain action to prepare for the situation or to stop certain activities. The prediction can also help to find out the persons responsible for initiating certain activity. In this paper we have presented a system where tweets are processed and analysed to predict up to what rate the civil unrest will happen or not. Firstly, the real time Twitter data is being fetched by using flume service in hadoop. Then the tweets are pre-processed. The pre-processed tweets are filtered by using Content based filtering algorithm to filter out the tweets which are related to civil unrest. The filtered tweets are clustered according to the category to which the tweet belong such terrorism, politics and social using K-means algorithm. Then sentiment analysis is being performed followed with prediction of the civil unrest.

***Keywords*—** Big Data, Social Network Analytics, Hadoop, flume, Twitter, Sentiment Analysis, Prediction.

I. INTRODUCTION

Nowadays, people mostly use social media to stay connected with the outside world. People belonging to different field and different ages are now connected to each other by using Social media such as Twitter, Facebook etc. Social media is the easiest platform that has been used to connect individuals. Most of the people are addicted to social media and spend most of their time on social media. Social media is the best platform for transferring or circulating any information rapidly. Due to rapid flow of the information to every individual through social media makes them aware of each and everything whatever is happening. The social media platform also gives an individual to express their thoughts, present their views and also makes them to listen the other's views. If we look back the last few years we can notice that people tremendously use social media to mark their active presence in the society, and every individual try to take active participation in the discussions to put forward their opinion. Every community of the people use it, rather a politician or an actor or a common individual. Social media becomes a new trend of the society to express their thoughts.

Twitter is one of the social network and a platform to connect people and talk about the recent activities or to just flow the information. There is a rapid circulation of the information through Twitter. People share their opinions and discuss their thoughts publicly that thoughts may be positive or negative. Many of the people also use this platform to share negative opinions regarding any of the activity or movement and also can be used by people to discuss the issues. These opinions can influence many of the people to think the same as the crowd is talking. Such of the activity on Twitter can make the crowd to think negative and lead towards the civil unrest such as strikes, March, riots. People can also use this platform for planning any unrest. Individuals can share their views and ask the people to join it so that the activity can be planned. Mostly in India the civil unrest happen because of the reservation issues, religious issue, political issues and similar kind of things. Twitter can be a best platform to plan or to influence people for organizing any civil unrest. So there must be an application which can be used to analyse the tweets and predict the civil unrest before it happen. There are many issues which can cause tweets analysis challenging.

The issues that an investigator can face while analysing the Twitter data are

- The tweets are changing constantly and frequently updating every time.
- Every tweet contains special characters such as $, #, @, etc. for adding more information to that tweet. It

is difficult to find out that this special character is used for what purpose.
- Many of the people use numbers, abbreviations, etc. to their tweets to enhance it.
- Every individual uses its own language for the purpose of communication. So analysing is even more difficult.
- Some of other issues are large amount of data, lack of standard in writing a tweet, etc.[1]

Here we have presented a system where the tweets are collected and based on the result of analysis the prediction is performed. Twitter real time data is fetched by using flume service of hadoop and stored in HDFS (Hadoop Distributed File System).[15] These tweets are now filtered using the keywords related to civil unrest and by this more relevant tweets for analysis are identified. The tweets are now clustered into three different clusters namely social, terrorism and politics. The investigator then can review any particular domain based on searchable word. All the related tweets from different clusters are fetched, on that Sentiment analysis is performed to found out the tweet contents are positive, negative or neutral. Then finally the prediction is performed rating that what are the chances of civil unrest to happen labelling them low, medium and high according to chance of occurrence of unrest.

This paper is organized into eight sections. The Section I defines the introduction regarding the topic that what are the problems and what can be the solution. Section II describes the Forensic analysis and what is the need of forensic analysis in this research. Section III reviews the most cited research in this area. Section IV gives the theoretical knowledge regarding the design and implemented of the proposed framework. Section V presented the results of the proposed work. Section VI gives conclusion of the implemented work. The future enhancement of the proposed work is given Section VII. Information of the references used is given Section VIII.

II. FORENSIC ANALYSIS

Forensic analysis is the branch of the digital forensics. It examines structured data with regard of incidents of fraudulent activity. The aim of forensic analysis is to discover and analyse patterns of fraudulent activities.

Digital forensic, a new discipline was developed to deal with the problem of legal and lawful collection of digital evidence in official, unofficial or corporate internal investigations [11]. The goal of any investigator is to find evidences stored in devices. Digital forensic tools and techniques achieve the goals of locating data & capturing data and then analyze the data.

Why there is need of forensic analysis of social media? In this paper we have presented a model which will able to perform data mining on social media and find out the possibilities of occurrence of the civil unrest. Data mining for digital forensic analysis is a branch of Computer Science focused on pattern extraction from large scale data which has been used to support analysts. One of the most promising applications of data mining algorithms is to build recommendation systems, aiming to propose future directions to the investigation and to guide the analyst through the process.

Social network forensic analysis is use to find out the protestor's behavior and track there information by gathering and analyzing the status information of the protester's. A forensics investigation requires the use of disciplined investigative techniques to discover and analyze traces of unrest that has to be happen.

Fig 1 gives the flow that forensic involves three major steps that are gathering of information, analyzing and Result prediction. Gathering information means fetching the information from Twitter data source. Analyzing the data means performing certain operations to find out the results for prediction.

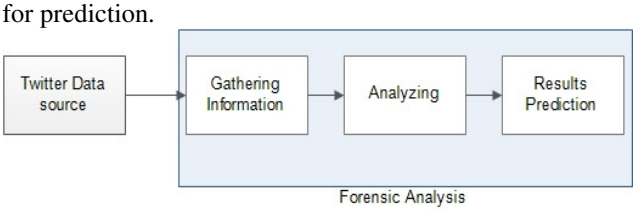


Fig.1 Forensic analysis

III. RELATED WORK

Several authors and researcher have analysed how social media plays an role in the civil unrest and how it is used during the planning or after any riot happen. Xiong Liu et al (2012) [1], in his paper he has discussed about a text cube architecture which was designed to organize social media data in multiple dimensions and hierarchies for efficient information query and visualization from multiple perspectives. Cube allows the analyst to examine public reaction (e.g., sadness, anger) to a range of social phenomena. Using that public reaction decision was made. But the issue in this work is that as the data size increase so it is not efficient.

Marc Cheong et al [2] discuss how social media plays a part in the London Riot 2011. They proposed a system that analysis social metadata to find out useful insights about major social events. A statistical correlation was found between the social media data and the real-world riot locations. They used Twitter metadata to investigate social dimensions of 2011 London riots. The analysis shows that the highest numbers of tweets were from mobile devices and suggesting a possibility of their catalyzing riots. From analysis FFRs and messages were from high-profile Twitter users, they detect other tweets focused on recovery and

commentary initiatives. Kohonens SOM was used for clustering the users and message properties and they obtained three clusters. Each of the clusters has unique and behavioural characteristics.

Ting Hua et al (2013) [3], proposes a system that uses for mining and analysing data from social media such as Twitter can reveal the causes of civil disturbances, including events and role of the politicians and different organizations in public opinion. In this system the Tweets are being classified or filtered on the basis of the keywords.

Ryan Compton et al (2013) [4], implemented a social media data mining system that can forecast events related to Latin American social unrest. The method extracts a small number of tweets from tweets that are publicly-available on Twitter, found similar tweets into coherent forecasts, and assemble detailed and easy interpretable audit which allows users to quickly collect information about the events that are going to be happen.

Elhadj Benkhelifa et al (2014) [5], presents framework which analyses the social media data and predict whether the social unrest will occur or not. In this paper the researcher uses the social snapshot framework (Huber's framework) which takes the snapshot of the social media at the given time. In this paper the framework is the combination of monitoring online social media and Digital Cloud Forensics.

Ryan Compton et al (2014) [6], present that in detail how it is now possible to examine social media and report on a large number of civil unrest events before their occurrence, while they are still in their planning stages. In this paper they restrict their attention to publicly visible data only. In this work they have provided a straightforward approach for the detection of upcoming civil unrest events in Latin America based on successive textual and geographic filters.

Nasser Alsaedi et al (2015) [7], presents an in-depth comparison an in-depth comparison of two types of feature that could be useful for identifying disruptive events: temporal and textual features. On the basis of these features, they investigate the dynamics of event/topic identification over time. They make several interesting observations: first, disruptive events are identifiable regardless of the "influence of the user" discussing them, and over a variety of topics. Second, temporal features play a central role in event detection and hence should not be disregarded or ignored. Third, textual features can be used to improve the overall performance of the event detection. We believe that these findings provide new insights for gathering information around real-world events, in particular for detecting disruptive events.

Qian Yu et al (2015) [8], analyzes real-time Sina Weibo tweet data streams and study volume correlations and temporal gaps between user searches and tweeting activities on hot topics. In addition, we examine the correlations between hot topic searches on social media and on search engines to understand hot topics and user behaviors across different platforms.

Harvinder Jeet Kaur et al (2015) [9], provides a brief overview of different techniques being developed for analyzing social media data, particularly twitter data. They developed a workflow for applying sentiment analysis to a comparatively new domain of natural disasters to detect public emotions in crisis. A base line model is designed and trained on unigram features using Naïve Bayes. The model is further tested on Kashmir floods dataset collected from twitter and an overall accuracy of 67% is achieved. The result provides valuable information which will assist the authorities to strategize their actions with due consideration to public sentiments and hence ameliorate the process of managing such situations

IV. DESIGN AND IMPLEMENTATION OF PROPOSED FRAMEWORK

We have implemented a framework for gathering and analysing the tweets and predicting the possible future civil unrest related activities. Currently the systems which are developed only focus on the prediction based on the content of the tweet. But there are many parameters that has to be considered while prediction such as tweet is send by which person, how many persons have replied to that tweet, is that person who initiate the movement is really genuine, and so many. So the system should be developed considering all the parameters while prediction. Here in the proposed system the prediction is done on different par we have implemented a framework for gathering and analysing the tweets and predicting the possible future civil unrest related activities.

Currently the systems which are developed only focus on the prediction based on the content of the tweet. But there are many parameters that has to be considered while prediction such as tweet is send by which person, how many persons have replied to that tweet, is that person who initiate the movement is really genuine, and so many. So the system should be developed considering all the parameters while prediction. Here in the proposed system the prediction is done on different parameters such as sentiment of the tweet, person who has initiated the movement, cluster to which it belong etc.

The proposed system is developed to predict the civil unrest collecting the tweets from Twitter. The fig. 2 gives the complete flow of the proposed system. The proposed system involves five steps for predicting the civil unrest.

That steps are Fetching data from Twitter, Keyword Filtering, Clustering, Sentiment Analysis and Prediction.

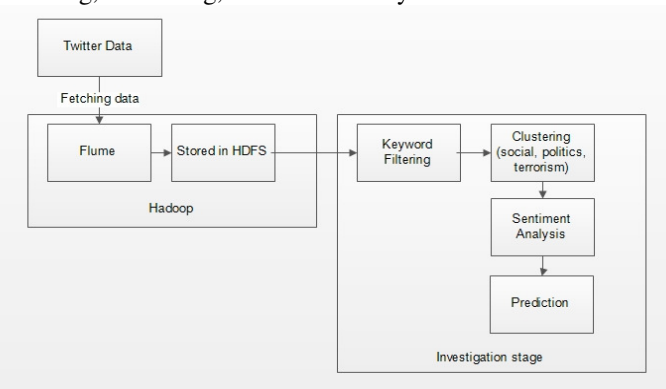


Fig 2.Proposed Framework for Prediction of civil unrest

A. *Fetching data from Twitter*

Data is collected from Twitter from the API to the HDFS using Apache Flume. Apache flume is a reliable and distributed system for effectively gathering and moving large amounts of data from various sources to a common storage area.[14] Major components of flume are source, memory channel and the sink. We have registered on Twitter as a developer [10]. Then creating a new application and Twitter will provide consumer key and secret key. We can use the consumer key and secret along with the access token and secret values by which we can access the Twitter and we can get the information that what we want exactly here we will get everything in JSON format and this is stored in the HDFS that we have given the location where to save all the data that comes from the Twitter.

The working how twitter data is fetched by using Apache Flume is given Fig 3. There is a Flume Twitter agent which is used to connect flume and Twitter data storage. In the proposed system MyTwitAgent is the Twitter agent. The data is then collected to the Data collector of Flume. Data collector is the Collecting Agent in the Flume. The data which is being fetched from Twitter is being collected by using Data collector. And now finally the data is being stored on HDFS(Hadoop Distributed File System). From there the data can be used for further analysis[12][13].

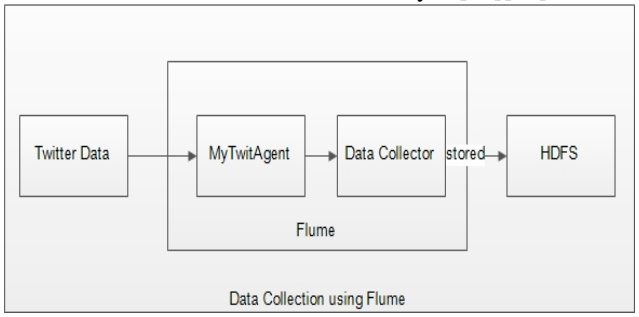


Fig 3.Working of flume to fetch real time data

B. *Keyword Filtering*

The posts and the tweets are made from combination of words that gives meaning to that post or tweet. There exits certain words in the posts that reflect the complete meaning of that post. Those words are referred as keywords which gives complete information about those posts/tweets. The tweets words are matched with the keyword dataset then the tweets are the filtered according to the tweets which are matched to the keywords in the dataset. The complete flow of keyword filtering is given in Fig 4.

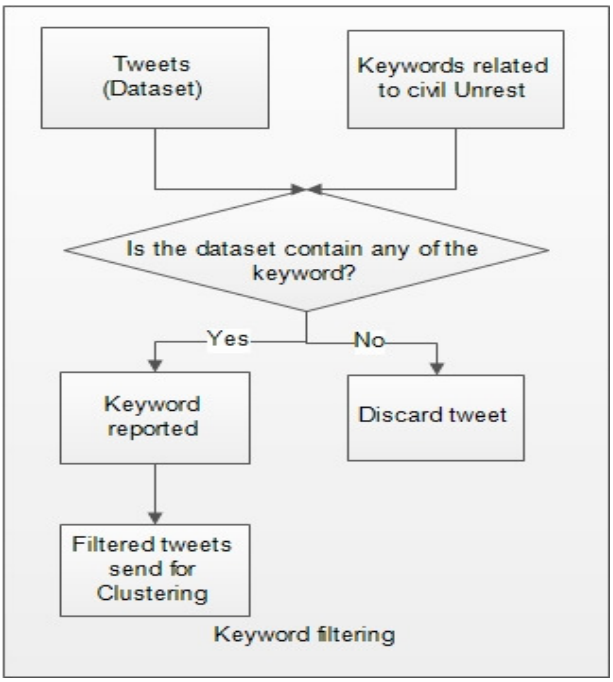


Fig 4 . Keyword Filtering

C. *Clustering*

Clustering is the technique of collecting the similar type of components in one cluster. There is collective behaviour of the tweets. Number of people tweet on Twitter. The user can be a politician, student, teacher, priests and other community's people. So clustering tweets on the basis of what type of tweet it is. In three generalized clusters the clustering is performed. The different types of clusters are Social, Terrorism and Politics. Mostly in India the unrest happen from the reason such as reservations, religious, political issues, terrorist attack and similar types of reason. So the clustering is done creating these three clusters. How the clusters are formed is given in Fig 5.

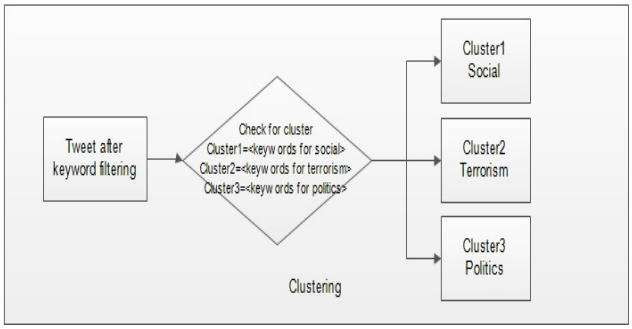


Fig 5. Clustering

D. *Sentiment Analysis*

Sentiment analysis would be used to see how the affected people are feeling and how they are reacting to any of the situation or movement. The people opinion and reaction only helps to know how the people are feeling. The result of Sentiment analysis is one of the parameter in prediction. For sentiment analysis two variables are assigned pos and neg. Certain negative terms are collected in dataset for checking whether terms in the tweets are negative. If the negative terms is their then neg = 1. In same positive terms are checked if positive term is there then it will assign pos = 1. If pos=1 and pos>neg then we can say that the tweet is positive. If neg=1 and neg>pos then tweet is negative. If the value of pos and neg are same then the tweet is considered as neutral. Fig 6 gives the complete working of how sentiment analysis is being performed.

The technique which is currently used in the proposed sysem is the Dictionary based Sentiment analysis. Here we have created two dictionary of the positive and negative words and according to that analysis is performed.

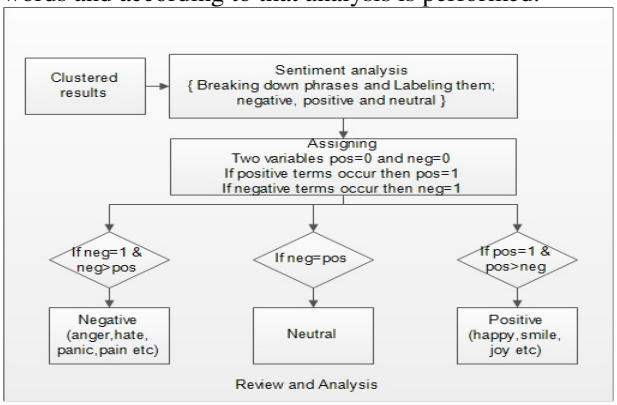


Fig 6. Sentiment analysis and review

E. *Prediction*

In Prediction the techniques are put together to predict events. There is the training dataset which contains certain tweets that has happen and on that predictions are being done. Based on the results from the training dataset the prediction on the real time data is made. The data after Sentiment analysis is compared with the prediction results of the training dataset. On the basis of certain parameters such as the result of sentiment analysis, cluster to which post belongs, who has twitted that tweet, etc. prediction is being performed.

Fig 7. Gives the complete flow that how prediction is being performed.

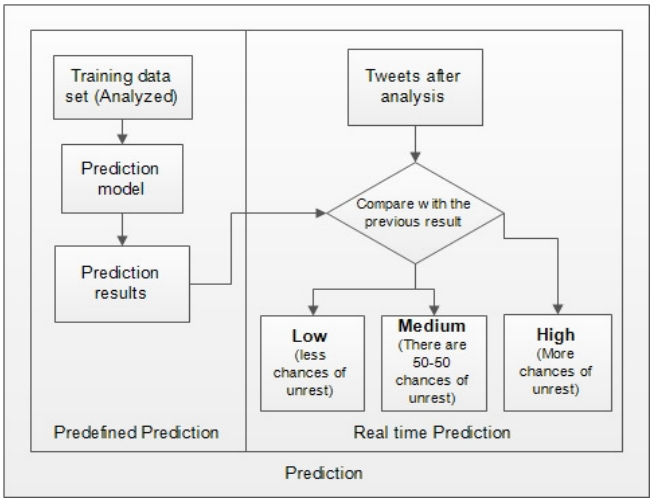


Fig 7. Prediction of civil unrest

V. EXPERIMENTAL RESULTS

A. *Fetching Real time data*

The data which is being fetched from Twitter is being plotted against the time to the data which is fetched. The Consistency of the fetching of data is being checked. The fetching data is plotted and the graph is shown in Fig 8.

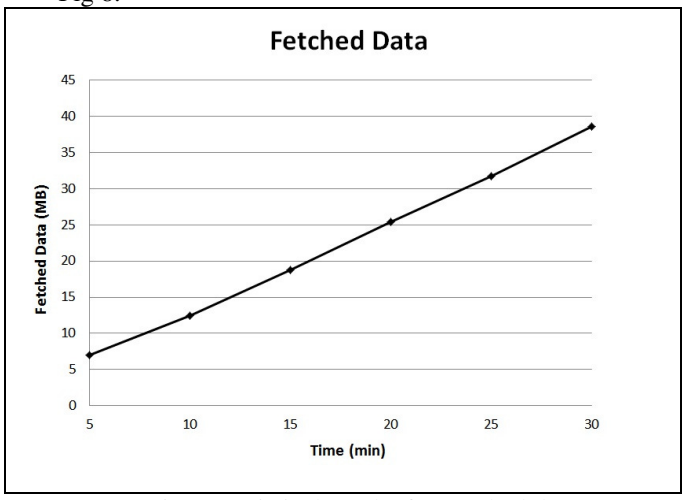


Fig 8. Statistical results for Fetched data

B. *Analysis Results*

The sentiment analysis is being performed on the data to check whether the post is Negative or Positive or neutral. The Data is being reviewed according to the domain the user needs to check. The analysis results after every review is being plotted to check how many of the positive or negative words are there in the analysed tweets. The result is given in the Fig 9.

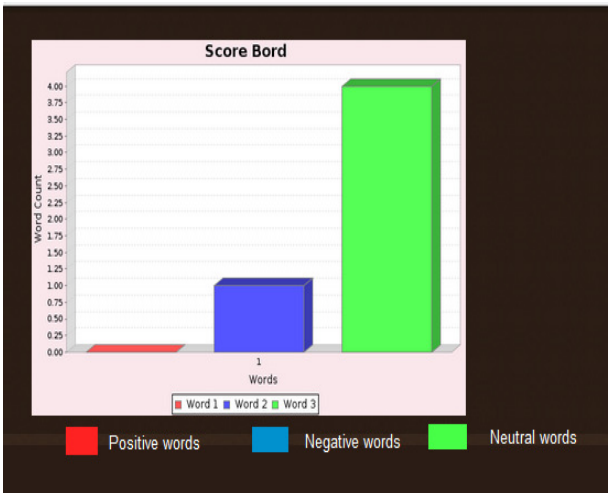
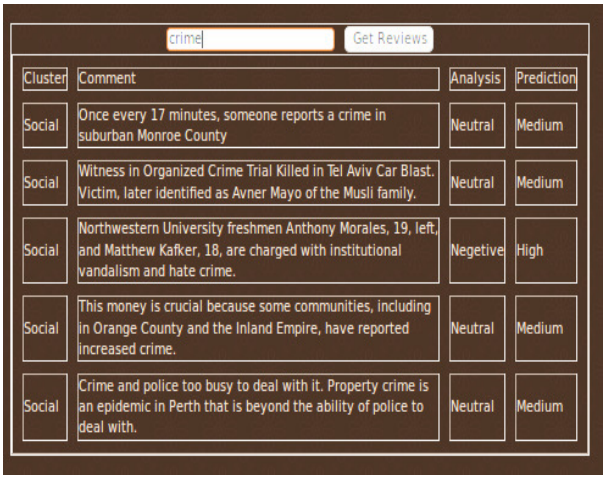


Fig 9. The Analysis results

C. *Prediction Results*

The prediction is being given in Fig 10. The final result of the system gives the details of the cluster to which the tweet belongs, along with the tweet, Analysis results and Prediction results.



Cluster	Comment	Analysis	Prediction
Social	Once every 17 minutes, someone reports a crime in suburban Monroe County	Neutral	Medium
Social	Witness in Organized Crime Trial Killed in Tel Aviv Car Blast. Victim, later identified as Avner Mayo of the Musli family.	Neutral	Medium
Social	Northwestern University freshmen Anthony Morales, 19, left, and Matthew Kafker, 18, are charged with institutional vandalism and hate crime.	Negetive	High
Social	This money is crucial because some communities, including in Orange County and the Inland Empire, have reported increased crime.	Neutral	Medium
Social	Crime and police too busy to deal with it. Property crime is an epidemic in Perth that is beyond the ability of police to deal with.	Neutral	Medium

Fig 9. The Analysis results

VI. CONCLUSION

As day by day the use of Twitter is increasing the people are more digitalized. They use more social network for communication. This paper examines the tweets and their role in civil unrest. The framework is developed to analyse the Twitter network and predict the incidents of civil unrest. For prediction of civil every parameter is being considered which can help the prediction to be strong. For time efficiency clustering is the performed.

VII. FUTURE SCOPE

In the current proposed system we have implemented the Dictionary based sentiment analysis i.e sentiments are drawn only on the basis of contents in the tweet compared to the dictionary. For increasing the accuracy of the system Corpus-based approach can be used along with the Dictionary based approach.

In Corpus-based approach a domain-specific sentiment lexicon is created to carry out the analysis. This can give more accurate results.

Current system only works on the Real time data from Twitter. So in future data from heterogeneous sources(such as Facebook, Google+ and many more) can be collected for analysis.

VIII. REFERENCES

[1] I1-Chul Moon, Alice H. Oh and Kathleen M. Carley, ”Analyzing Social Media in Escalating Crisis Situations”,IEEE,2011.

[2] Marc Cheong, Sid Ray and David Green,”Interpreting the 2011 London Riots from Twitter Metadata”,IEEE, 2012.

[3] Ting Hua, Chang-Tien Lu, Naren Ramakrishnan, Feng Chen, Jaime Arredondo, David Mares, San Diego and Kristen Summers,” Analyzing Civil Unrest through Social Media”,IEEE,2013.

[4] Ryan Compton, Craig Lee, Tsai-Ching Lu, Lalindra De Silva and Michael Macy, ”Detecting future social unrest in unprocessed Twitter data”, IEEE, 2013.

[5] Elhadj Benkhelifa; Elliott Rowe; Robert Kinmond; Oluwasegun A Adedugbe and Thomas Welsh, ”Exploiting Social Networks for the prediction of social and civil unrest”, IEEE,2014.

[6] Ryan Compton, Craig Lee, Jiejun Xu, Luis Artieda-Moncada1, Tsai-Ching Lu, Lalindra De Silva and Michael Macy, ”Using publicly visible social media to build detailed forecasts of civil unrest”, Security Informatics,springeropen journal, 2014.

[7] Nasser Alsaedi and Pete Burnap, “Feature Extraction and Analysis for Identifying Disruptive Events from Social Media”, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.

[8] Qian Yu, WeiTao Weng, Kai Zhang, Kai Lei and Kuai Xu, “Hot Topic Analysis and Content Mining in Social Media”, IEEE, 2015.

[9] Harvinder Jeet Kaur and Rajiv Kumar, "Sentiment Analysis from Social Media in Crisis Situations", International Conference on Computing, Communication and Automation, 2015.
[10] Twitter Developer[Online]. Available: http://dev.twiiter.com/.
[11] Forensic data analysis[Online]. Available: https://en.wikipedia.org/wiki/For-
ensic-data-analysis
[12] Flume 1.6.0 User Guide[Online].Available:https://
ume.apache.org/FlumeUserGuide.html/.
[13] Acquiring Big Data Using Apache Flume[Online]. Available:///E:/00%20M.Tech%20Project/Acquiring%20Big%20Data%20Using%20Apache%20Flume%20%20Dr%20Dobb's.html/.
[14]Hadoop[Online]. Available: https://hadoop.apache.org/.
[15] Flume[Online]. Available: https://flume.apache.org/.

