IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Detection of crime and non-crime tweets using Twitter

Nakul Sharma[1], Anupama dhamne[2], Nisha more[3], Vikash rathi[4], Ankita supekar[5]

Dept of Information Technology, Sinhgad Academy of Engineering, Pune[1-5]

Abstract: As the speedy growth of Twitter in social media, researchers get attracted towards the use of social media data for analysis. Twitter is one of the widely used social media platform to express thoughts. This paper approaches for analyzing tweets using sentiment mining in which classify highly unstructured data on Twitter and second is data mining. There are 500 million users of Twitter. The limit of character in twitter is 140 character because of this user uses shorthand notation example "ok" can be used as "kk". Analysis of tweets contains misspellings and grammatical error.

Keywords: Social network, natural language processing, Machine Learning, sentiment analysis and data mining.

I. INTRODUCTION

Now a day's social media plays important role in modern life. Online Social media such as Twitter, Facebook, and many enterprise social media, have become very much popular in the last few years. People are spending a large amount of time on social media to interact with people. The number of people who use social media increasing day by day. People tweets their opinion, view, thought or event on social media and also share their post. User posts their comment and others can follow them.

Twitter now has become the most popular online microblogging service. Twitter the user to send image and text-based posts up to 140 characters.
Social media are a medium of analytics on a huge amount of user data for many companies based on which many prediction models are built .this prediction models help to step-up new business or new ideas. But this is the positive side of social media. On the other side, people share their personal information on social media site and their information is misused .the social media is the easy target for distributing fuck and wrong information. people comment related to the crime and the post is increase the violence in public .the crime Detection system(CDS) Detect, the post is related to crime or not. If the post is related to crime the further classify into the subtype of crime:

1) Crime against the person
2) crime against the property
3) crime against the country
4) Other

Two approaches are used in the CDS .one is sentiment mining for detecting the crime directly from the post or comment and other is data mining for structured data and history data to find out the intensity of the crime.Sentiment mining is used for unstructured data and real-time data. Data mining is the practice of examining large preexisting databases in order to generate new information. CDS system help to reduce the crime. The remaining of the paper proceeds as follows: Section 2 shows the related works in the crime detection research area; Section 3 shows the architecture and proposed system; Section 4 describes the observation and result; Finally, Section 5 represents the conclusions.

II. RELATED WORK

The previous work [1] focuses on the streaming data on twitter which is classified as the tweet is Malicious or not. Crime prediction had been a trending research field using social posts and data analysis. The tweets are fetched using the Twitter API and then analyses it using machine learning. Then the data pre-processing is done[2] .the stop word removing is done with the help of Stanford NLP Libraries [5].the twitter comment contains misspellings, elisions, and grammatical errors[3], to make the sense of the twitter comment it transforms them into a canonical form, consistent with the dictionary or grammar.
Analyzing data from these social media sites is one of the new buzzwords for many business strategies, Technical concepts, World health issues, Election campaigns, inventions, Entertainment; all can be handled by using sentimental analysis. The sentiment mining system identification of tweet without knowing the previous background. Sentiment mining uses the negation algorithm to an identification of comment is positive or negative. Text mining aims to

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

accurately extract, identify and analyze information from unstructured data sources. The past studies of aggressive behavior on an uncomfortable day show clear correlations between location, day, time and criminal activities extracting specific tweet attributes[6] like username, location, time, re-tweet count etc. using the attributes find the intensity of crime. for data mining, the Naive Bayes algorithm is used.

III. PROPOSED SYSTEM

The proposed system Collection data from the Twitter social networking site and processes data using NLP techniques. We are using two approach one is sentiment mining and other is data mining.Sentiment mining is used for unstructured data and real-time data. As data mining is used for structured data and history data. The system consists of the following modules

1) Data collection module
2) Sentiment mining
3) data mining
4) output Classification

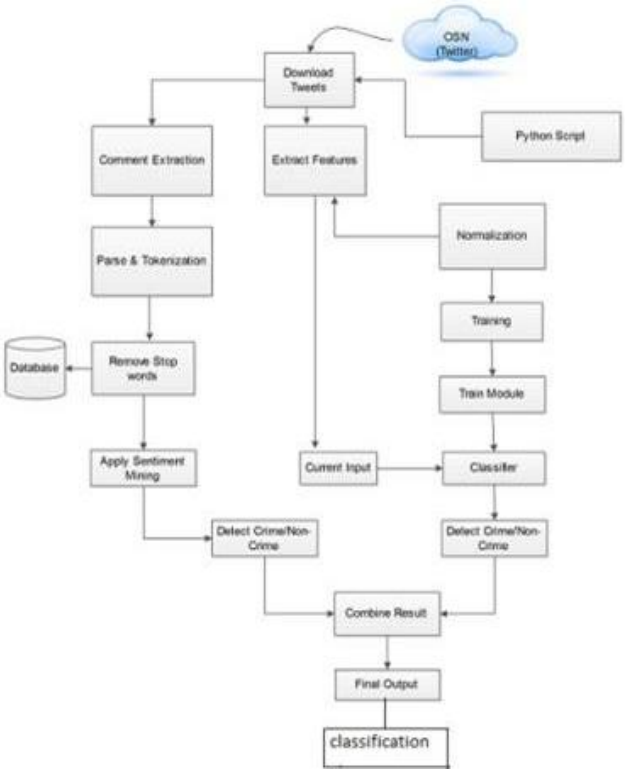

Fig.2. architecture diagram

1) Data collection module:-
The tweets are fetched using the Twitter API. The API provides a user-friendly programming interface through which download the tweet object in tweet Object format. This object format helps in extracting specific tweet attributes like username, location, time, re-tweet count etc. Once the data is fetched pre-processing of the gathered data is done to extract features.

2) Sentiment mining:-
Sentiment mining system identification of tweet without knowing the previous background. Before applying the algorithm data pre-processing is required. The data undergo the following processes.
1) Stop Word removal
2) Repeated letters removal
3) Noise data removal
4) Parsing and tokenization

1. Stop word removal:-
Stop words are those words which generally do not carry any useful information but are added to get the grammar of the sentence. For example prepositions like on, in, to, above etc., articles like a, an, the, question words like who, what,

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

where, how etc., generally do not add any information to the content. But they are always found in large amount in a sentence. So, these words are to be removed from a sentence before applying the algorithm.

2. Repeated letters removal: -
People tend to show their emotional state by repeating the letters of words in the tweets like 'happppppyyyyy'. In English, any word contains letters repeated twice to the maximum. If a letter is repeated more than twice consecutively, the number of its occurrence is reduced to two. Thus'happppppy' becomes'happy'.

3. Noise data removal:-
By noise data we mean the unwanted data in the tweets like URLs, hashed words, names etc. The URLs present in the tweets are removed.

4. Parsing and tokenization:-
Once the data are cleansed, Parsing and tokenization are done. Tokenization helps in part of sentence part of the word in a sentence. tokenizationbreaks a stream of text into tokens, usually by looking for whitespace. A parser takes the stream of tokens.

3) Data mining:-
data mining is used to find the intensity of crime using the Naive Bayes algorithm. before applying the algorithm do the data pre-processing.
1) Feature extraction
2) Normalization
3) Data training and train module

1. Feature extraction: -
After downloading the tweet, extracting specific tweet attributes like username, location, time, re-tweet count etc. All the extracted feature are stored in the database.

2. Normalization: -
The extracted tweet convert into the normal form for easy use and access .in normalization all the attributes give the index.the attributes are further use as an index.

3. Data training and train module:-
Algorithms learn from data. They find some pattern and behavior in given data set and learn from the data.Data training apply to the data set. New data is input to the trained module and predict the output Intensity of data.

4) Output Classification
If the tweet is related to crime then its divide into the type of crime. Mainly we use the system 4 type of crime
1) crime against the person
2) crime against the property
3) crime against the country
4) other

IV. RESULTS AND DISCUSSION

We have chosen using negation algorithm as our main classifier; the results are based on those experiments. For determining the accuracy of the system we worked on a random set of sample 1000 tweets, of which 60% was no crime and the rest were the crime. Classes for these users have known already, out of those 1000 tweets 93-95% were classified without mistake.

Confusion Matrix

	Crime	No crime
Crime	37%	4%
No crime	3%	56%

Accuracy	Recall	Precision
93%	90.24%	92.50%

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

V. CONCLUSION

This system collects particular types of tweets from Twitter social networking sites and does NLP technique to extract feature out from the tweets. Various databases are added to increase the accuracy of prediction. After those various methods of classification are applied sentiment mining of data as positive and negative. Hence this system will categorize tweets and let us know the possibility of particular types of tweets which will exist in which particular area by using Machine Learning(semantic analysis and data mining), thus helping us to find crime through users tweets on social media developed by us.

REFERENCES

[1] SagarGharge, Mr.ManikChavan, "An Integrated approach for Malicious Tweets detection using NLP", International Conference on Inventive Communication and Computational Technologies (ICICCT 2017)
[2] MonishaKanakaraj, Ram Mohana, Reddy Guddeti," NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers", 3rd International Conference on Signal Processing, Communication and Networking (ICSCN 2015)
[3] Bilal Ahmed, "Lexical Normalization of Twitter Data", IEEE Transactions on Computational social systems, July 2015
4] Marcello Trovati, Philip Hodgson's, Charlotte Hargreaves." A Preliminary Investigation of a Semi- Criminology Intelligence Extraction Method: A BigData Approach", International Conference on Intelligent Networking and Collaborative Systems 2015
[5] HaseSudeepKisan, HaseAnandKisan, AherPriyanka Suresh," Collective Intelligence & Sentimental Analysis Automatic of Twitter Data By Using Stanford NLP Libraries with Software as a Service (SaaS)", 978-1-5090-0612-0/16/$31.00 ©2016 IEEE
[6] Xinyu Chen, Youngwoon Cho, and Suk young Jang," Crime Prediction Using Twitter
Sentiment and Weather", IEEE Systems and Information Engineering Design Symposium 2015
[7] JazeemAzeez, D. John Aravindhar," Hybrid Approach to Crime Prediction using Deep learning", International Conference on Advances in Computing, Communications and Informatics (ICACCI 2015)