CRIME MONITORING FROM NEWSPAPER DATA BASED ON SENTIMENT ANALYSIS

BY

MD. IQBAL HASAN
ID: 173-25-639

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

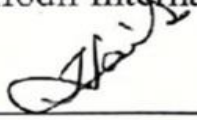
MAY 2019

APPROVAL

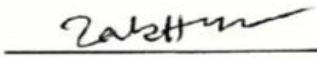
This thesis titled **"Crime Monitoring from Newspaper data based on Sentiment Analysis"**, submitted by Md. Iqbal Hasan, bearing ID No: 173-25-639 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on May 05, 2019.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



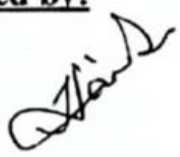
Md. Zahid Hasan — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head, Department of Computer Science and Engineering,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Dr. Sheak Rashed Haider Noori
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Submitted by:

Md. Iqbal Hasan
ID: 173-25-639
Department of Computer Science and Engineering
Faculty of Science & Information Technology
DaffodilInternationalUniversity

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty Allah for his divine blessing makes us possible to complete the final year project successfully.

We really grateful and wish our profound our indebtedness to **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head,** Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining*" to carry out this Research. His endless patience ,scholarly guidance ,continual encouragement, constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Syed Akhter Hossain** , Head of Department of Computer Science and Engineering, for giving me an opportunity to carry out the research work, without him I should not reached my goal and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

DEDICATION

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, specially to my let father **Md. Kamrul Hasan**, a strong and gentle soul who taught me to trust in Allah, believe in hard work and that so much could be done with little. Also, I want to show my gratitude to one of the most helping friend of mine *S.M Mazharul Islam* for his courage and great inspiration.

ABSTRACT

Crime is one of the major challenges of the world which is affecting the normal life and socio-economic development. Therefore, many governments are trying to use advanced technology to address or tackle such issues to maintain the peace of the country. So the analysis on Crime data has a great impact and value for the current scenario of the world. On the other hand, Sentiment Analysis is a growing effective method to assess written or spoken language to decide if the expression is favorable, unfavorable, or impartial, and to what degree. Nowadays, online newspapers are very popular among the people and contents varieties of crime news can be a great source to understand the types and occurrence of crime. The aim of this paper is to monitor the crime, based on the headlines of the online newspapers provided in. The approach is based on sentiment analysis by applying lexicon based methods and understands the crime categorized in a day, month, location and week. This piece of research work will help to deep understanding the pattern of the crime as well as the possibilities of occurrence of the crime in the specific time or day which will bear a great value to ensure the security purpose.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Big data has made things complex for general data analysis techniques. People around the world are working on different domains of data for vast verity, and certainly the crime data is one of the most used verities among them. Crimes of a certain locality or even a country can highly be reduced using data mining which is already a proven fact for the time being. Data mining can simply pin-point assorted crimes for different areas, even the reasons behind the crimes can be acquired. Online newspapers are the great source of news of different niche. In most cases, a headline tells the story and let people guess what the news is about. Therefore, the headlines can be the base in order to consider whether news is about crime or not. This research mainly focuses on crime monitoring based on newspaper headlines posted on Twitter. The lexical based analysis is used for crime prediction whereas a manually built real dataset is used to get the desired matches.

1.1 RESEARCH AIMS AND OBJECTIVES

The aim of this research is to monitor the crimes happens in a certain region based on weeks. Here, the Online Newspapers are used as the main data source. Almost each and every online news portal has their channel on many social websites like Facebook, Twitter, YouTube, etc. So, this research only grab the second option, meaning the twitter channels of the corresponding news portal will be used here as the data source. As the social channels update their news every minute, therefore it will be very easy to get updated news constantly and the crime data can be separated and monitored statistically in an arranged way so that it can be very easily understandable. Along with to this, this research wish to be extended like- to figure out the most crime affected area according to a certain region based on online newspaper data source. Besides, after researching the statistical research output, it can be assumed very easily that what types of crime happens in which place and even in which day in a week. So, necessary steps can be taken to prevent the crime or make the criminal plan in veins.

So, in aggregate this research does have the following objectives:

- To Study Data Analytics as well as Big Data.
- To Study and Learn Sentiment Analysis.
- To Develop and seek Knowledge on how to deal with Online Data Source.
- To Learn and implement How to deal with Twitter Data.
- To Develop Computational tool for Sentiment Analysis.
- To apply the Knowledge of Sentiment Analysis on the real life Examples.
- To monitor crime news weekly basis.
- Try to identify the most Crime Affected spot in a certain region.
- Try to predict the crime attempt in a certain area in a particular for weekly basis.

1.2 MOTIVATION

The revolutionary growth of the textual information on the web in the past few decades has brought a dramatic change in human life. In the social web, almost every newspaper share the up to date news and update them in every minute or so. So, it can be stated that social websites are the source of a large collection of news in the form of texts, which can be analysed to make this research fruitful. There may have some garbage data containing hyperlinks, images, videos which will not be needful in this regard and hence they are to be modified before the main process is to be run.

In current time, Twitter is a very popular social website and it’s a great source of engagement of people from all over the world. They connect with each other and communicate and express their thought through texts along with to this, they are not ready sometimes switch there an app to get news information. Therefore, people nowadays are habituated to read news even without leaving the twitter page and here Twitter has done a great work indeed. They grab all the news from all over the world and present in front of their audience so that they can hold their traffic and at the same time the traffic can be satisfied having all their needs from the same place. So, this is the main motivation to select Twitter as the main online data source for the research. On the other hand, it's very true that without having that level of motivation, it’s not possible to run research

or even any initiative at all. It's a proven statement that "In some cases, only motivation can complete a huge part of analysis or task".

In short, the motivations of this research are listed below:

- Current Revolution of Data Analysis along with Big Data.
- My Personal Interest on Data Mining and Sentiment Analysis.
- My Personal Interest to work with Crime based data source.
- Opportunity to learn many things that are very new to me.

1.3 EXPECTED OUTCOME

As this research uses Twitter as the source of news data, the app I wanted to create will be able to create interaction between tweet data and data mining. It will be able to collect news from the twitter channels of daily newspapers and analyze them. After the analysis, the application will provide a statistical overview of daily news from one or multiples newspaper at a time.

Then, the application, which was decided to be built using **R** for checking whether the news is negative or positive. This may happen using sentiment analysis strategies and once I get the decision about the news that it is positive or negative, it will be half done of the total expected outcome. The positive news will be discarded than hence there will only the negative news be available. Finally, after being compared the negative words with my Crime data Set (that contains thousands of crime-related words) I expect to have the desired output that means the statistics of which news are crimes and which not.

So, if I analyze the crime news based on a week, it can easily be guessed that in which day crimes occurs mostly. Along with that, most crime-affected spots can also be identified. Outputs will be viewed as Histogram Plots, Bar Diagram, Polarity Charts and Word Clouds.

That actually has the following advantages:

- The most crime occurred days in a week can be identified,
- The most crime affected areas can be identified,
- Necessary initiatives can be taken against the crime on those certain areas.

1.4 REPORT LAYOUT

There are five Chapters in this research paper. They are: Introduction, Literature Review, Proposed Method, Results and Output Visualization, Conclusion and Future Scopes.

Chapter One: Introduction; Aims and Objectives, Motivations, Expected Outcome, Report layout.

Chapter Two: Literature Review; Related Terms, Challenges.

Chapter Three: Proposed Method; Research Method, Data Collection, Creating Twitter App, R Setup and Preparation, Collection Data from Twitter, Data Pre-processing, Data Analysis, Data Flow Model.

Chapter Four: Result and Output Visualization; Output of Sentiment Analysis, Output of Crime Analysis.

Chapter Five: Conclusion and Future Scope; Conclusion and Future Scope.

CHAPTER 2
LITERATURE REVIEW

2.1 RELATED WORKS

In present world people are much more dependent on online newspaper or news channel and the amount is much higher than before because of availability of news and fast transmission. Therefore different kind of news is available and most of them are related to crime. Because of that researchers around the world are trying to analyze those data to predict crime before it happens.

In 2014 Researcher Xiaofeng Wang, Matthew S. Gerber and Donald E. Brown worked on crime prediction [1]. They used an automatic process to extract twitter post and used those data to predict crime. In their research they used a liner modeling for automatic crime prediction based on semantic analysis. Another team of researcher with Mr. Lei Zhang in 2015 worked on sentiment analysis for unique characteristics identification of twitter data [2]. In their research they used lexicon based analysis to perform sentiment analysis. They used bag of words for lexicon based analysis.

Researcher Mikhail Bautin, Lohit Vijayarenu and Steven Skiena worked on sentiment analysis and they used News and Blog posts as their data source [3]. They experimented on Lydia text analysis system to varying languages and news sources. They calculated entity based sentiment scores and compared it with two different translators for Spanish which is translator independent. On the other hand, in 2007 Namrata Godbole, Manjunath Srinivasaiah and Steven Skiena studied on the importance of sentiment scoring news and blog posts for large scale analysis of data [4]. They did sentiment composition of adjectives and also did comparison between news and blog posts.

Yanghui Rao and his team worked on emotion detection of social media data, associated with human emotions based on online dictionary used for sentiment analysis of online news in 2013 [5]. They used their custom algorithm and three types of strategy for build up a dictionary which has ability to predict the emotional state of news articles, news events and social media data.

Researcher Matthew S. Garber worked on crime analysis in 2014 and he built a model based on kernel density to get prediction on 25 types of crime [6]. In his research necessary data were collected from twitter and those data were relevant to crime. Those data were used to build the crime prediction model. In 2018 Hitesh Kumer Reddy and his co-researchers worked on area based crime prediction, which is a very important topic of research to reduce crime rate [7]. They developed a framework to predict crime using machine learning algorithms. In this research Google map and various data packages were used for crime monitoring and repost visualization.

Xinyu Chen, Youngwoon Cho and Suk Young Jang worked on location based crime monitoring and prediction in 2015 [8]. In their research data were collected from twitter and they used sentiment analysis for crime prediction and monitored the time and location for different types of crime. Using those data they tried to predict what kind of crime is going to happen in which area. They applied lexicon based method combined with kernel density estimation to predict crime destination and weather via linear modeling. Researcher H. Chen and his team developed a framework for crime data mining [9]. Their team used different case study for crime data mining framework, where the framework recognizes different types of crimes in local, national and international level.

2.2 RELATED TERMS

In this section, some mostly used terms used in this research will be described for better understanding. As the one of the aims of the research is to make it understandable by almost all categories of researchers or even the common educated people, it is crucial to make the whole content as understandable as possible. Therefore, this section of research holds some technical terms and their easy grabale definition along with the descriptions. They are given below one after another.

2.2.1 DATA SET

Most commonly a data set is refers to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. In short, a Data Set (or dataset) is

a collection of data[10]. In another words it can be said that, a data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity[11].

2.2.2 BIG DATA

Big Data refers to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data. It seldom to a particular size of data set[12].

2.2.3 DATA PREPROSESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

2.2.4 SENTIMENT ANALYSIS

Sentiment means feelings, it also be attitudes, emotions or opinions. And sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Using natural language processing, statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit. Sometimes it referred to as opinion mining.

The purpose of sentiment analysis is to determine the attitude of a communicator through the contextual polarity of their speaking or writing. Their attitude may be reflected their own judgment, emotional state of the subject, or the state of any emotional communication they are using to affect a reader or listener.

Sentiment analysis is a fast growing subject in the technical communication field. With the increase in social media, online retail, and personal blogs and publications knowing where public sentiment is leaning has translated into a rapid evolution in sentiment analysis that can become a valuable skill.

When we perform sentiment analysis on some content, the main searching point is the opinions in content and we picking the sentiment based on those opinions. An opinion is an expression that consists of two key components. One of them is a target or topic and another one is a sentiment on the topic. Consider a sentence, “I love this company“, here “this company” is the topic and the sentiment that is expressed by the verb “love” is positive.

Sentiment analysis is just not a feature in a social analytics tool – it’s a field of study. This field is still being studied, albeit not at great lengths due to the intricacy of this analysis, in the same way that some aspects of linguistics are still up to debate or not fully understood. Existing approaches to sentiment analysis can be grouped into three main categories: knowledge-based techniques, statistical methods, and hybrid approaches.

Knowledge-based techniques classify text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Some knowledge bases not only list obvious affect words, but also assign arbitrary words a probable "affinity" to particular emotions. Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and Semantic Orientation Point-wise Mutual Information. More sophisticated methods try to detect the holder of a sentiment (the person who maintains that affective state) and the target (the entity about which the affect is felt). To mine the opinion in context and get the feature which has been opinionated, the grammatical relationships of words are used. Grammatical dependency relations are obtained by deep parsing of the text.

Hybrid approaches leverage on both machine learning and elements from knowledge representation such as ontology and semantic networks in order to detect semantics that are

expressed in a subtle manner, e.g., through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so.
Open source software tools deploy machine learning, statistics, and natural language processing techniques to automate sentiment analysis on large collections of texts, including web pages, online news, internet discussion groups, online reviews, web blogs, and social media. Knowledge-based systems, on the other hand, make use of publicly available resources, to extract the semantic and affective information associated with natural language concepts. Sentiment analysis can also be performed on visual content, i.e., images and videos. One of the first approaches in this direction is SENTIBANK utilizing an adjective noun pair representation of visual content[13].

The applications for sentiment analysis are endless. More and more we're seeing it used in social media monitoring and VOC to track customer reviews, survey responses, competitors, etc. However, it is also practical for use in business analytics and situations in which text needs to be analysed.

2.3 CHALLENGES

As we know that sentiment analysis is the analysis of image, text, audio, video sentiment recognition. Sentiment analysis and evaluation process are facing several challenges. These challenges become obstacles in analysing the accurate meaning of sentiments and detecting the suitable sentiment polarity. To identify and extract subjective information from text the sentiment analysis is the practice of applying natural language processing and text analysis techniques [14].

The degree of accuracy issue is hard to answer, said Bing Liu, a University of Chicago computer science professor specializing in data mining. It depends on what are measuring, the level of analysing text, and the number of data sets across domains and the voice sound quality of videos, among other variables. Still, he thinks that progress is being made in this regard [13].

In sentiment analysis it is more challenging to detect a more in depth sentiment/emotion. Positive and negative is a very simple analysis but the challenging one is to emotions like how much hate there is inside the opinion, how much happiness, how much sadness, etc.

Emotion detection is really a difficult task because sometimes it happens that someone tells something that seems positive but in real it's not positive the sense was negative. So sometimes it is difficult to understand meaning of a sentence cause the emotions are too much complex.
Mainly sentiment analysis tries to detect the mental situation of a person. But sometimes it become tough to tell what the person meant. If we consider audio sentiment analysis, then noise or voice tune difference can create major error in output. Same for text analysis, because some texts word wise meaning is totally different from its actual meaning. That's why sentiment analysis is facing major challenges now a day.

Besides, the list of following challenges I faced at the time of working with this research-

- Identify the Negative News
- Building Crime Dataset
- Justifying the true and reliable news
- Crime Prediction

CHAPTER 3
PROPOSED METHOD

3.1 RESEARCH METHODOLOGY

To accomplish this research, the main thing it needs which is the data to work with. So, to collect data from online data source I blindly rely on Twitter for particular news channel. Everyday Twitter generates tons of terabytes of data by its users, where most of them are unstructured in some way. So before starting the research work and implementation, it is needed to consider this fact as my challenge for sentiment analysis.

I actually have found different platforms for data mining such as Orange, Weka, Rattle GUI, Apache Mahout, R, Hadoop, UIMA, SenticNet API, Natural Language Toolkit, Python Scripting etc. But, I have selected "R" as the Data Mining tool and platform. There are some reasons behind that. R has a very user friendly UI and bunch of strong libraries for Data Processing, Data mining and output visualization as well. Therefore, I firstly learnt how to work with R and then started learning R programming language that leaded me to successfully complete this research with ease. Also, I cannot but mention the website (https://www.datacamp.com)[16] from where I took a lot of help and its really a good kick-start for a new R learner as I personally feel. It helps people to learn R showing step by step process.

Having created a twitter app for data collection, I have created an application using R language for further process (sentiment scoring). This application can connect to twitter via internet, able to collect data and analyse them. Collected data then are stored in a CSV file.

3.2 DATA COLLECTION

There are thousands of newspapers all over the world currently. Most of them have their online version, where they share their daily news frequently. Along with to this, they also share the news in different social media like Facebook, Twitter against their corresponding page or account. Indeed it is much easier and reliable to collect data from the social Media rather than scrap their source websites. Therefore in this research data will be collected from Twitter, and different news channels like BBC, CNN, etc. will be the Media. Using Twitter API, necessary

keys will be collected and with the help of R language, necessary data will be collected and stored to a hard-drive location in the computer. The logic for collecting data is –

```
IF (KEY == match)
     Search_key = Keyword;
     Tweet_Number = Amount;
     Data_Date = Date;
     IF (DATA == Found)
          Store_to = Location;
     Else
          Return "error_message";
Else
     Enter "valid_key";
End;
```

According to the pseudo code, Twitter will verify access of the program using keys from API. If matches, the found program will be able to collect data. On the other hand, the program will search data- based on different keywords, the amount of data expected, and date. If the information provided for search finds a match, it will collect data and store it to an allocated location as a CSV file format. Otherwise, it will give an error message. In this research, only news headlines will be used which is the tweet of different news channels. Data will be collected in a continuous flow and daily basis.

3.2.1 CREATING TWITTER API

In order to connect the R compiler with Twitter for collecting data, the system needs two crucial keys (**`consumer key`** and **`consumer secret key`**) those can only be obtained from a Twitter App. The reason why, it's a must to create a Twitter App first and then sequentially comes to install the RStudio to connect them through Twitter API.

Creating the Twitter Application is absolutely easy and straightforward. Just need to fill up a form providing some required fields like name of the application, what is the purpose of the application (in at least 300 words) and a little more.

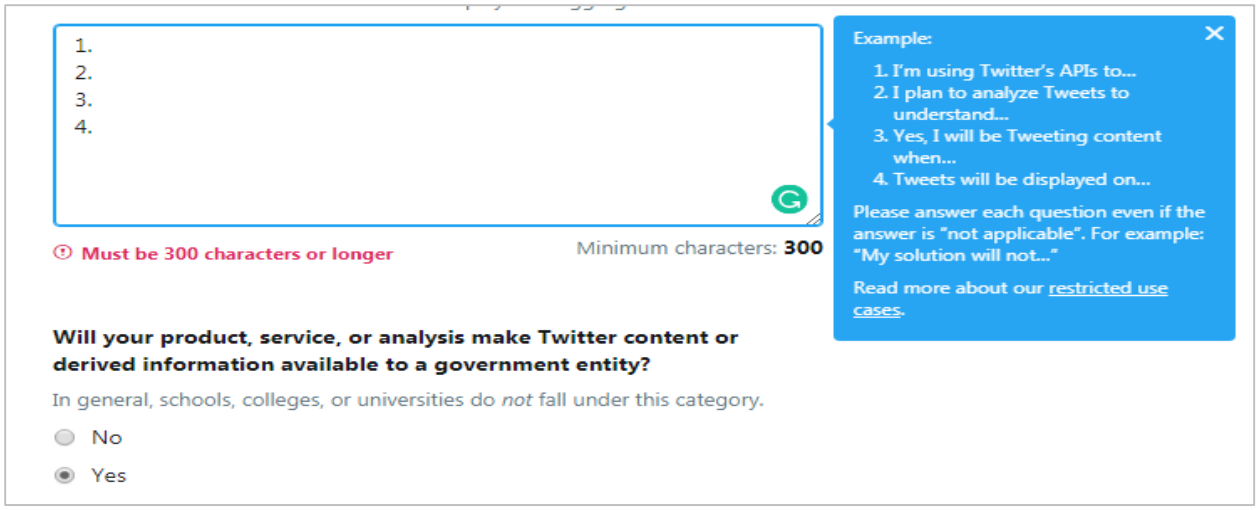


Fig.1: Twitter Application Creation

After providing all information, the application would be created and the two crucial keys (`consumer key, consumer secret key`) will be available under **Application Settings** to use for further work.

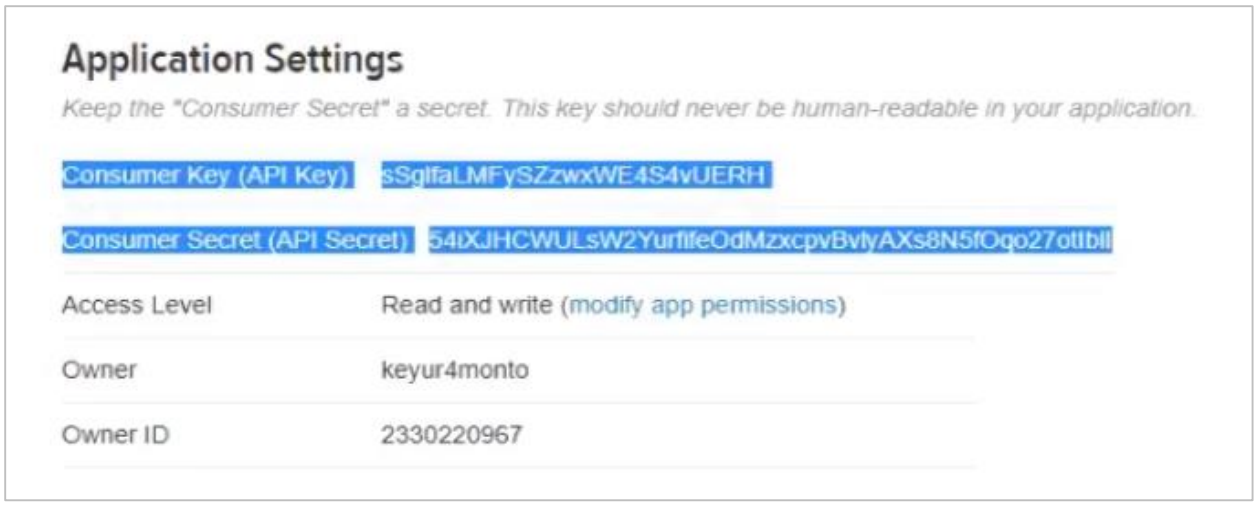


Fig.2: Consumer Key and Consumer Secret Key

3.2.2 R SETUP AND PREPARATION

Having created the Twitter application, it necessary to setup R compiler completely step by step and it's because- R[17] will be used to build the application to collect News Data from Twitter, accessing the Twitter application. To get R for free, firstly one needs to head over the following

site (https://www.r-project.org/) and download the latest version of it. After downloading, it needs to be successfully installed without interruption into the host PC /Laptop.

Fig.3: R Download

Having successfully installed R, the UI will be look like below.

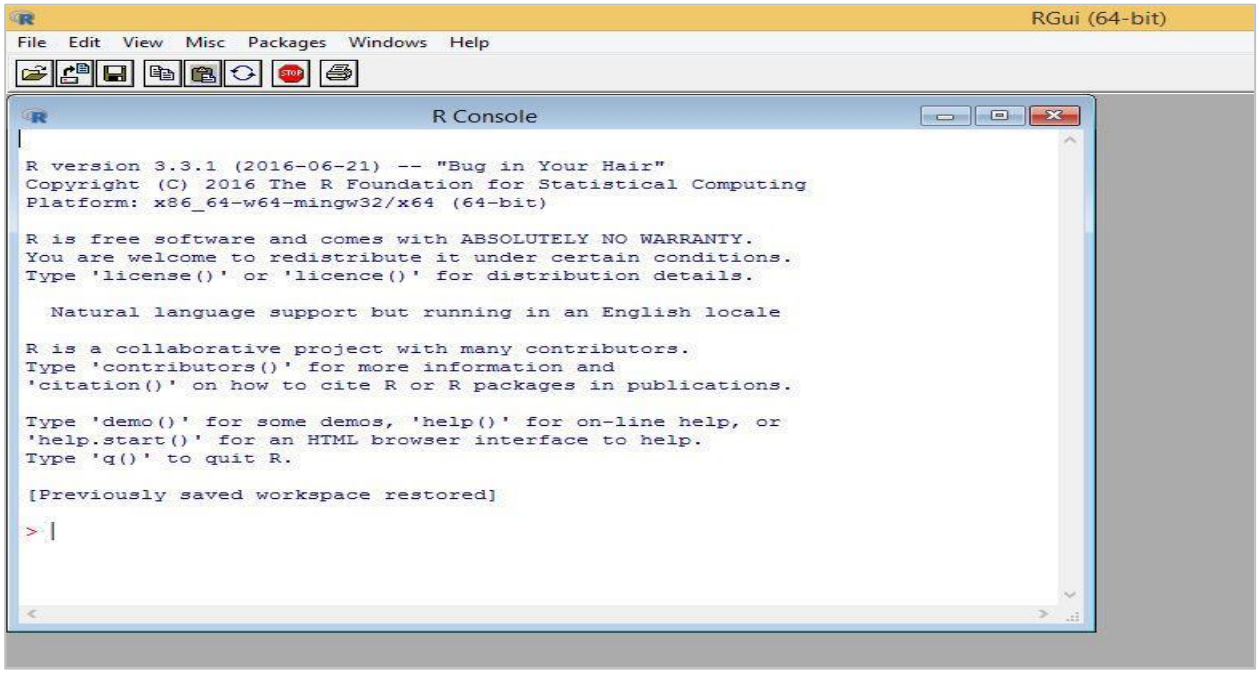


Fig. 4: R User Interface

Since the User Interface doesn't look much comfortable to work in, and there is no visual option, it is better to install another application named RStudio. RStudio gives R programmers a very user-friendly UI to comfortably work in. It is very easy and eases to visualize statistical result or output in an arranged way.

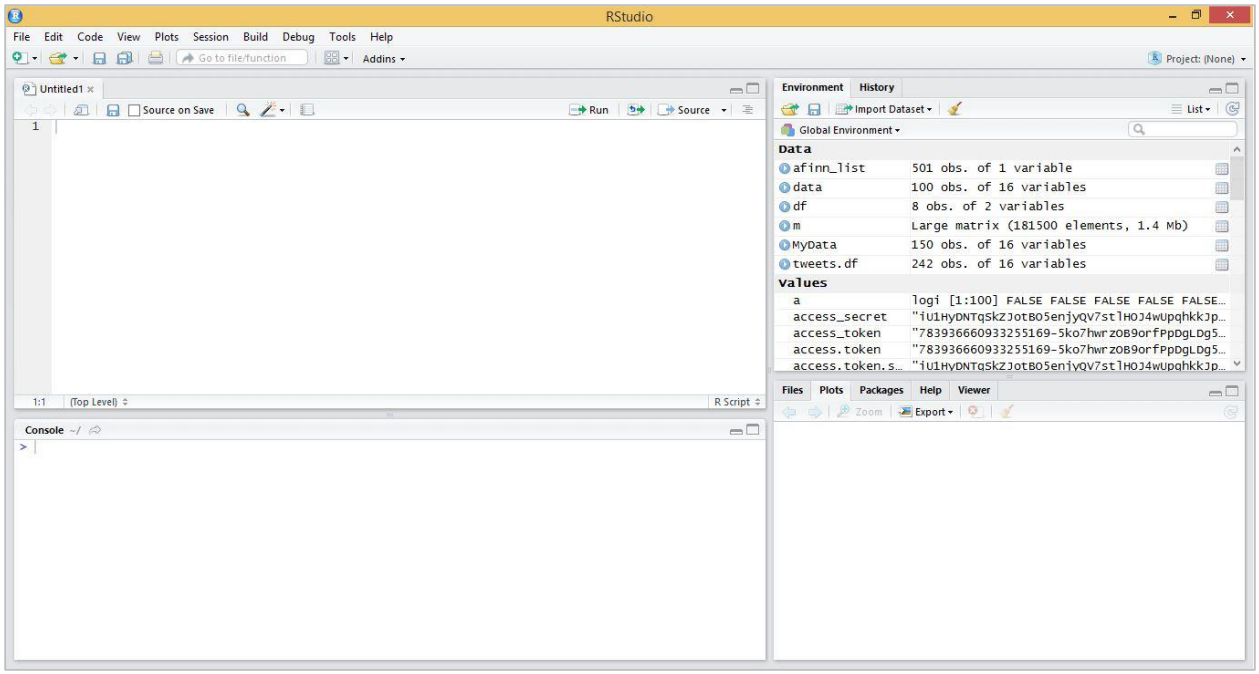


Fig. 5: Rstudio User Inteface

After installing all the above mentioned necessary components, it's ready to start the programing task for the research.

3.2.3 COLLECTING DATA FROM TWITTER

In order to collect data from Twitter, a Script to be written in Rstudio and running the script will connect the application through Twitter API and collect the tweets from twitter newspaper channels. The Script for Establish Connection between twitter with custom built app (using R) is given below:

```
library(twitteR)
library(ROAuth)

consumer.key <- "***Can be obtained from Twitter App***"
consumer.secret <- "***Can be obtained from Twitter App***"
access.token <- "***Can be obtained from Twitter App***"
access.token.secret <- "***Can be obtained from Twitter App***"
```

```
setup_twitter_oauth(consumer_key=consumer.key,
consumer_secret=consumer.secret, access_token=access.token,
access_secret=access.token.secret)

query <- 'From:CNN', 'From:BBC', 'From:ALJazira'
maxTweets <- 5000
startDate <- '2019-06-16'

tweets <- searchTwitter(query, n=maxTweets, lang='en',
locale='en', retryOnRateLimit=3)
tweets.df <- twListToDF(tweets)
write.csv(tweets.df, file='data.csv', row.names=FALSE,
fileEncoding='UTF-8')
```

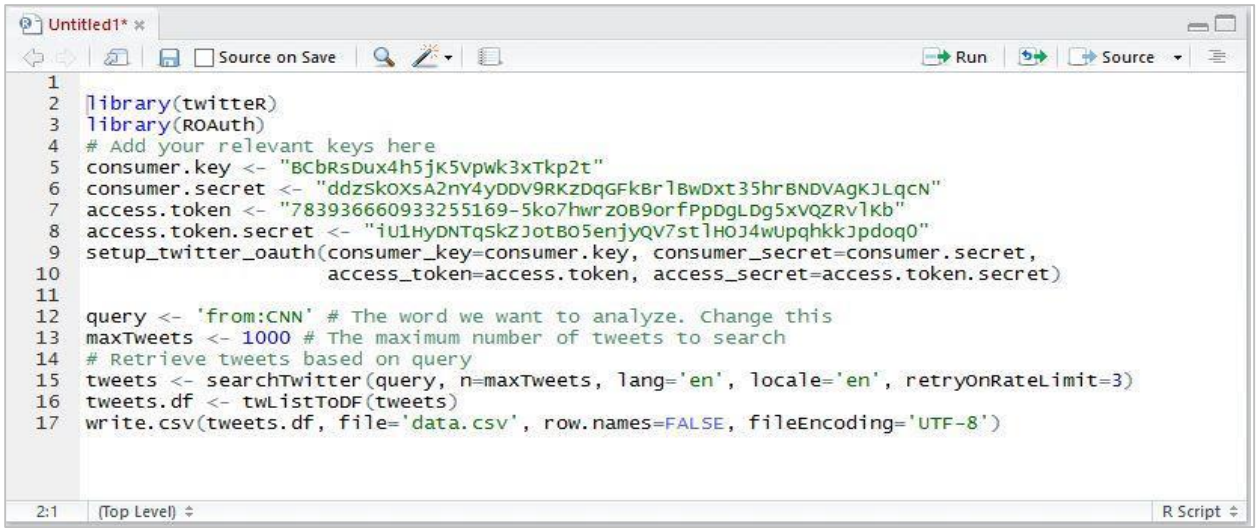


Fig. 6: Data Collection

3.3 DATA PREPROSESSING

After collecting the desired data, it is important to preprocess the data before it goes for analysis. Because, data with unwanted value can create difficulties and may highly affect the result. Hence, a data cleaning operation will analyze the collected data and remove the links, special characters, unnecessary columns, and almost all those data which does not pose a value to this analysis. When the preprocessing is completed, data file will be stored again with a different

name. This time it will be in a more normalized form and ready to be analyzed for the final result.

To cut off unexpected data, the following Script can be run using R:

```
sentence <- gsub("[^[:alunum:]///']", "", sentence)
sentence <- gsub('[[:punct:]]', '', sentence)
sentence <- gsub('[[:cntrl:]]', '', sentence)
sentence <- gsub('\\d+', '', sentence)
sentence <- tolower(sentence)
word.list <- str_split(sentence, '\\s+')
words <- unlist(word.list)
```

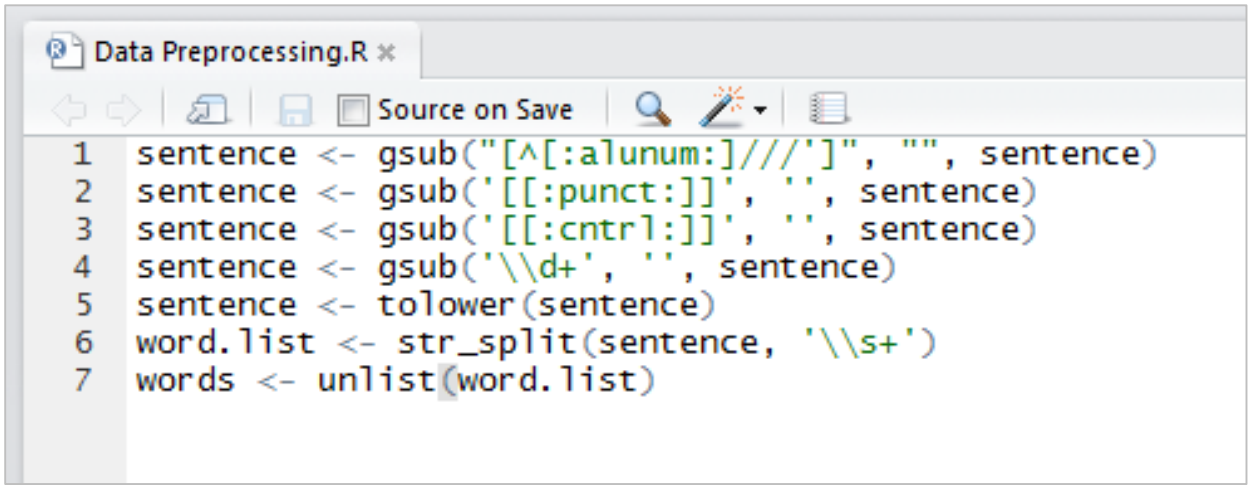


Fig. 7: Data Preprocessing

3.4 DATA ANALYSIS

After completing the prepossessing phase, now it's time to send the output .csv file through the final phase/ analysis. This phase is divided into two parts: Sentiment Analysis, and Crime Analysis.

3.4.1 SENTIMENT ANALYSIS

In this part, preprocessed data will be used for sentiment analysis. Using sentiment analysis, it is possible to predict positive news and negative news with ease. As this research focuses on crime analysis, determining news sentiment is crucial. The tweets having crime-related words will be

considered as crime news. But before considering crime related words for crime words, sentiment analysis needs to be done for better prediction.

Due to sentiment analysis, a pair of libraries containing positive and negative words individually will be used. A lexicon based analysis will be done to match words in preprocessed tweet data set with the bag of word, which will work as negative and positive word libraries. Here, each of the word (contains either positive or negative sentiment) will hold one point. For each tweet, the total number of positive score will be deducted from the total number of negative tweet.
If the total number of positive words are considered as P_S, the total number of negative word as N_S , and final score as S, the algorithm will be like below –

```
START
IF (WORD == POSITIVE)
{
P_S++;
GO TO NEXT WORD;
}
ELSE IF (WORD == NEGATIVE)
{
N_S++;
GO TO NEXT WORD;
}
ELSE
{
GO TO NEXT WORD;
}
S = P_S + N_S;
END;
```

Once the analysis is done, all the tweets will have positive, negative or neutral sentiments. At this part of analysis only positive and negative sentiment related tweets will of work. According to the analysis plan, system will then count only the tweets having value more or less than zero. Therefore it is now possible to make a pie chart that will be helpful to visualize ratio of the positive and negative sentiment related tweets amount considering those value. Let's consider P_S as total score of positive words and N_S as total score of negative sentiment. So, we get-

```
T_S= P_S + N_S.. .. .. .. .. .. .. .. .. ... (1)
```

```
S_P = (P_S/T_S)*100%..................... (2)
S_N = (N_S/T_S)*100%..................... (3)
```

Where S_P score of the total positive is value in percent and S_N is the score of total negative value in percent. On the other hand, T_S is the addition of positive and negative tweets or tweets that have value more than and less than 0. Using all those values it is possible to draw a pie chart and a bar chart for crime analysis.

3.4.2 CRIME ANALYSIS

After completing sentiment analysis crime data analysis process will start. A data set is under construction for crime analysis which is containing different types of crime-related word. As like as sentiment analysis also this can be done using lexicon based analysis. Where a bag of word is containing crime-related words and those will be matched with the preprocessed dataset. In this section for each crime related word system will count 1 and if there are multiple crime words in a single tweet, the system will add all their values. After processing, all the tweets system will store the data set with the value with a different name. The following pseudo-code is explaining the task-

```
START;
IF (word == crime_word)
{
Crime_value++;
CHECK NEXT WORD;
}
ELSE
CHECK NEXT WORD;
END;
```

Using the result of this analysis it is possible to draw a sequential pattern according to date and a bar diagram can be drawn comparing with sentiment analysis report, especially using negative sentiment and crime report.

3.5DATA FLOW MODEL

From the below **Fig. 8**, it can be seen that the process starts with collecting data which is clearly the required thing, and we comfortably rely on the online newspaper as the source since they contain tons of day to day data. Collected data may contain various unwanted symbols or special characters, which may affect the final output. Therefore after collecting, it is needed to go through a process in order to eliminate the unwanted pieces as long as the system gets it as desired. One thing to remember- the data is collected row-wise, meaning that each headline lies on a single row on the document.

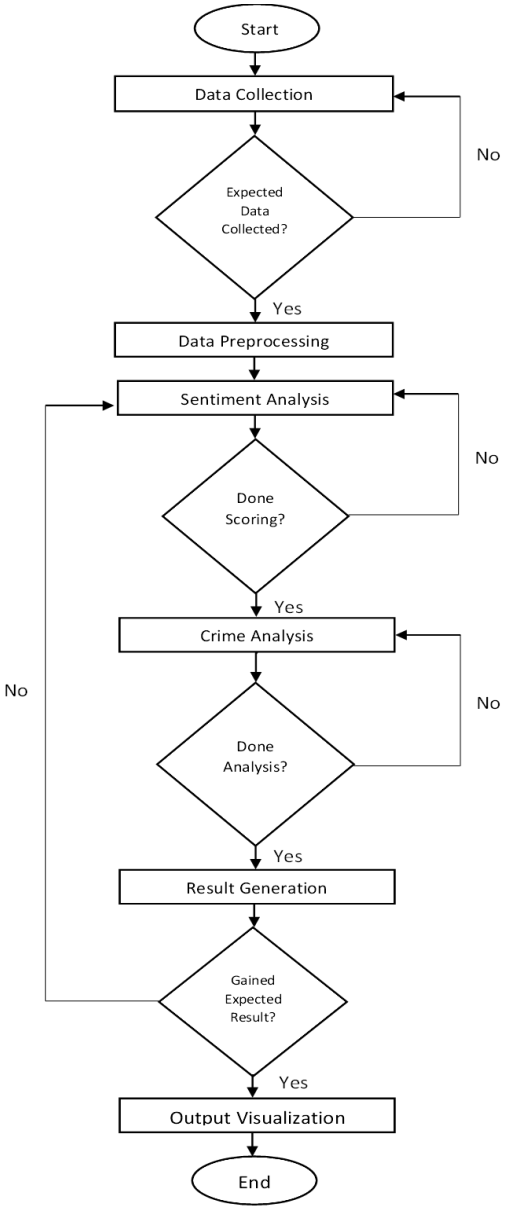


Fig. 8. Data Flow Model

It is required to process them which results the data, ready for further phases. At this phase, the whole collected data is used for acquiring sentiment analysis against each headline, and suppose this analysis is named as scoring. A custom algorithm will check each of the rows containing a single heading in each and will return an integer value against them individually. Typically the returned score is either a positive integer or a negative integer or even a zero.

The score is positive integer denotes that the headline is for positive news, and a negative score denotes that the headline is for negative news. The score zero denotes that neither the news is positive, nor negative. So they are named as neutral. This sentiment analysis phase will be run until getting the scores against all the data we passed through the SA (sentiment analysis) process.

Now the time comes for crime analysis, which is, in fact, the soul process of this research indeed. The output set of data along with the score against each headline will be then compared to the custom-made crime dataset and eventually will find whether a news is about a crime or not. In this case, the headlines having a positive score or even the score zero headlines will be ignored as there is no probability for a positive or neutral news to be a news describing crime incidents or identical.

This phase will be on action until obtaining the decision for all the negative news headlines. Next phase is of result generation. As soon as the decisions are made on the previous phase, the system will automatically generate a statistical report with the help of a bar chart, showing the percentage of crime headlines on daily basis. Here, an exception might occur in the result. If it seems not to be as expected, there must be something unwanted happened in the sentiment analysis phase, hence the phase sentiment analysis might have to be gone through again. The process ends with visualizing the desired output result having the amount of crime in each day of a week.

CHAPTER 4
RESULT AND OUTPUT VISUALIZATION

4.1 OUTPUT OF SENTIMENT ANLYSIS

By running sentiment analysis process on preprocessed data it is possible to get a day to day report on the sentimental state of news collected from different news pages. As data are collected separately for each day, so sentiment analysis will generate reports of each day accordingly. For example sample output for a few days are given below in figure 2, 3, 4 & 5 –

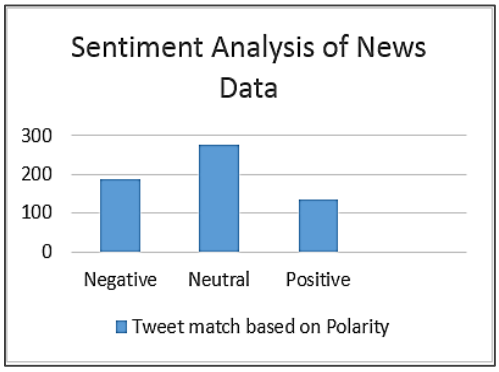


Fig. 9. Day 1 Sentiment Analysis

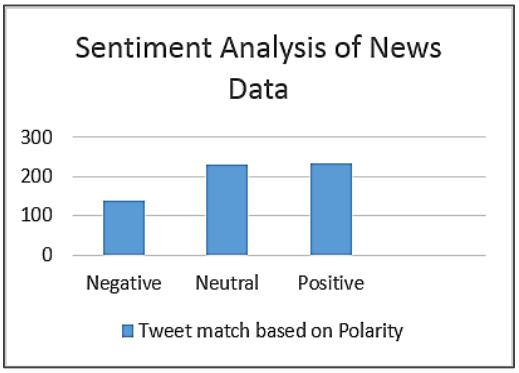


Fig. 10. Day 2 Sentiment Analysis

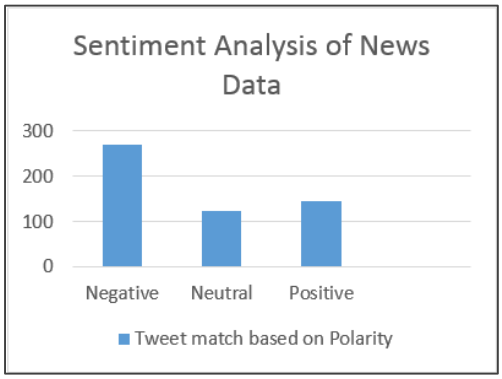


Fig. 11. Day 3 Sentiment Analysis

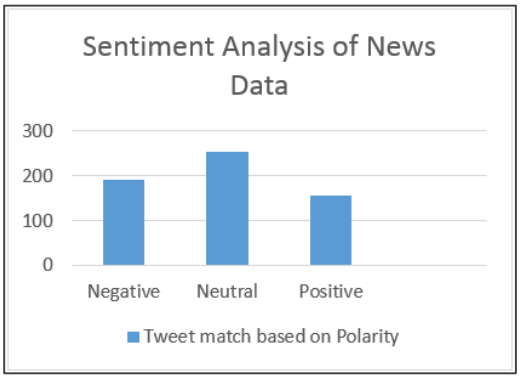


Fig. 12. Day n Sentiment Analysis

According to the given output by considering 30 days of a month it is possible to determine in which day of the week having more negative news than other days. Even it is also possible to determine in which part of the month negative news rate is high and which part is low. Even by dividing news reports into four weeks it is also possible to generate a weekly report.

4.2 OUTPUT OF CRIME ANALYSIS

On the other hand, Crime Analysis will generate another report based on the crime data which will analyze news and provide a pie diagram based on crime database match. Like sentiment analysis, crime analysis will be analyzed on daily basis. Sample output of the crime analysis is given below in figure 6, 7, 8 & 9 –

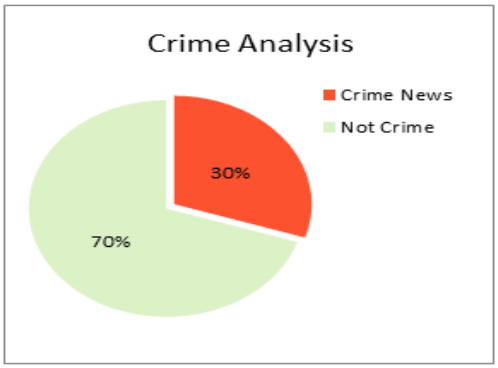


Fig. 13. Day 1 Crime Analysis

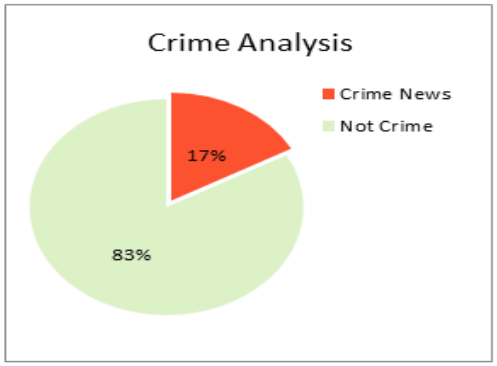


Fig. 14. Day 2 Crime Analysis

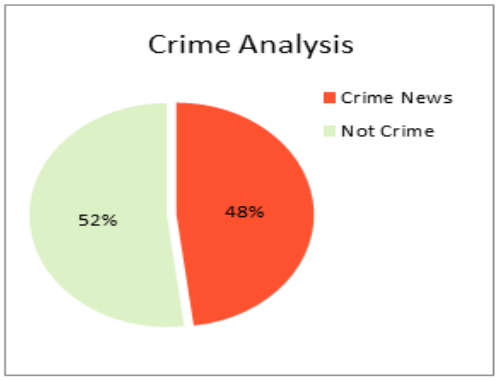


Fig. 15. Day 3 Crime Analysis

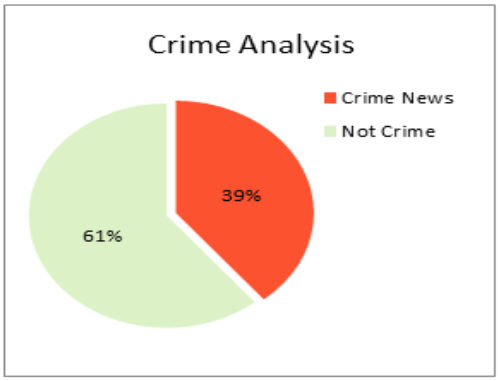


Fig. 16. Day n Crime Analysis

Based on both sentiment analysis and crime analysis now it is a possible to predict whether those negative news are actually crime or not. Moreover, this analysis will give a more valid and accurate report.

4.3 WORD CLOUD

There is a simple and straightforward approach to create a posh Word Cloud for any niche text data. This approach requires four R packages; **tm, SnowballC, wordcloud,** and **RcolorBrewer**. I will be using five simple phase for this instance to create my desired World Cloud of crime words using R. The steps are given below along with their proper description.

Step 1: Creating the World File

This is the first and foremost phase. In order to create a World Cloud, its needs to have all the words in the same file with .txt format. It's very simple, just to copy and paste all the words in a text file and save it with a meaningful name.

Step 2: Installing the Required Packages

The following script will be of installing the required packages needed for creating the Word Cloud. Once they are installed, they need to be loaded for further work phase.

```
install.packages("tm")
install.packages("SnowballC")
install.packages("wordcloud")
install.packages("RColorBrewer")
```

Step 3: Mining Text

This phase has three steps. Load Text, Load Data as Corpus, and Inspect the Content. The following three scripts will be able the do these work with ease.

```
text <- readLines(file.choose())
docs <- Corpus(VectorSource(text))
inspect(docs)
```

The frequencies of the words can be seen by building a Term-Document Matrix. The first column will be populated with distinct words and the second column will be populated with the frequency of that word. The script is given below-

```
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
```

```
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```

Step 4: Generating Word Cloud

This is the final phase to generate the Word Cloud. The following script will do it perfectly where **max.words** denotes the number of words will be taken randomly from the previously saved .txt file and **min.freq** denotes the words with minimum frequency will be taken as its minimum size. It can be set as user wish according to the need.

```
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
max.words=200, random.order=FALSE, rot.per=0.35,
colors=brewer.pal(7, "Blue"))
```

After doing all the previously mentioned phased successfully, the Word Cloud will be look like below.

Fig. 17: Word Cloud of Crime Data Set

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 CONCLUTION

Crime activities are growing parallel with the emerging technology. So the prediction of occurrence of crime will be very helpful to reduce crime. Monitoring crime from the newspaper based on sentiment analysis can play a vital role in future to predict the possible places and time to occur crime which will be a breakthrough in the security system in the general life. The major benefit of this research work is that types of crime can be predicted before happening including the day and week can also be identified. This framework can also be used to monitor different types of problems where sentiment analysis is a mandatory concern.

5.2 FUTURE SCOPE

Crime is such thing that perhaps can't be eradicated in a way that it never ever happens. The at most possible thing is to predict them before occurrence and take necessary steps. So accurate this prediction would be, the rate of eviction of Crimes would be super higher. We know that the evolvement of Sentiment analysis very fast and rapid and the demand of sentiment analysis is increasing in every sectors in our social life specially in Research.

So, this research work has potential to be used for-

- Detecting Crime even before it happens,
- Identifying the most Crime Affected Area for a certain region,
- Identifying the most Crime Occurrence Happening Day in terms of a week or a month,
- Necessary crime protection initiatives can be taken before it is too late,

In future I am planning to implement more algorithms in this research work to make the result more and most importantly, I want to contribute more in this research field by keeping carry on studies.

REFERRENCES

[1] Wang, X., Gerber, M. S., Brown, D. E.: Automatic Crime Prediction Using Events Extracted from Twitter Posts, In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Pages 115-125, DOI: 978-3-642-29047-3_28, USA (2014)

[2] Zhang, L., Ghosh, R., Dekhil, M., HSU, M., Liu, B.: Combining lexicon based and learning based methods for twitter sentiment analysis, In: National Conference on Advanced Technologies in Computing and network, page: 89-91, ISSN: 2277-9477, Maharashtra, India (2015)

[3] Bautin, M., Vijayarenu, L., Skiena, S.: International Sentiment Analysis for news and blogs, In: International Conference on Weblogs and social media, Page: 19-26, Washington, USA (2008)

[4] Godbole, N., Srinivasaiah, M., Skiena, S.: Large Scale Sentiment Analysis for news and blogs, In: International Conference on Weblogs and social media, DOI. 10.1.1.591.4890, New York, USA (2007)

[5] Rao, Y., Lei, J., Wenyin, L., Li, Q., Chen, M.: building Emotional Dictionary for sentiment analysis of online news, In: World Wide Web, DOI. 10.1007/s11280-013-0221-9, PP. 723–742 (2014)

[6] Gerber, M. S.: Predicting crime using Twitter and kernel density estimation, In: Decision Support Systems, Volume 61, Pages 115-125, DOI: http://doi.org/10.1016/j.dss.2014.02.003, USA (2014)

[7] Reddy, H. K., Reddy, T., Saini, B., Mahajan, G.: Crime Prediction & Monitoring Framework Based on Spatial Analysis, In: Procedia Computer Science, Volume 132, PP. 696-705, DOI. j.procs.2018.05.075, India (2018)

[8] Chen, X., Cho, Y., Jang, S. Y.: Crime prediction using Twitter sentiment and weather, In: 2015 Systems and Information Engineering Design Symposium, pp. 63-68, Charlottesville, VA (2015)

[9] Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., Chau, M.: Crime data mining: a general framework and some examples, In: Computer, PP. 50 – 56, Volume. 37, Issue. 4, DOI: 10.1109/MC.2004.1297301 (2004)

[10] Wikipidia, “Data set” Online available at << https://en.wikipedia.org/wiki/Data_set>>

[11] Whatls.com, “Data set” Online available at << https://whatis.techtarget.com/definition/data-set>>

[12] Wikipidia, “Big Data” Online available at << https://en.wikipedia.org/wiki/Big_data>>

[13] Wikipidia, “Sentiment Analysis” Online available at<<https://en.wikipedia.org/wiki/Sentiment_analysis >>

[14] Quora, “What are the applications of sentiment analysis? Why is it in so much discussion and demand?” Online available at <<https://www.quora.com/What-are-the-applications-of-sentiment-analysis-Why-is-it-in-so-much-discussion-and-demand>>

[15] ADWEEK, “5 Key Challenges in Sentiment Analysis” Online available at <<http://www.adweek.com/prnewser/5-key-challenges-in-sentiment-analysis/116604>>

[16] DataCamp, “Learn Data Science Online” Online avaialable at <<https://www.datacamp.com/>>

[17] R, “R Language” Online available at << https://www.r-project.org/>>