American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-7, Issue-8, pp-10--15
www.ajer.org
Research Paper Open Access

Tweet Analysis: Extractionand Sentimental Analysis On Twitter Data

A.Shreya Shetty[1]

[1] *(Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, India.)*

ABSTRACT:*Nowadays companies have realizedthe need for Big Data to make decisionsaboutcomposite problems.Twitter being one of the top used social media site receives millions of tweets every day, on a diversity of issues. This large amount of raw data can be accustomed for the industrial, social, economic or business purpose for their development. Hadoop is one of the best tools for twitter data analysis as it works on a variety of data like big data, streaming data etc. This paper discusses how to use FLUME and apply the concept of sentimental analysis on it. Flume is used to extract the real-time data from Twitter and Sentimental analysis is an analytics tool.*

KEYWORDS: *Twitter, Sentimental analysis, Hadoop, Flume, HDFS,Analysis*

Date of Submission: 20-07-2018 Date of acceptance: 04-08-2018

I. INTRODUCTION

Social mediaplatforms like Twitter have become popular in today's world. Companies and organizations have invented ways to analyze data from Twitter for information about how people feel and think about their services and products. The analysis of this data can be applied for decision making in various areas. Twitter data is usuallysemi-structured which makes the analysis trickier. For the purpose of development, Twitter provides us with an API which permits the designerto accessfew of tweets tweeted which contains the keyword. Twitter alsopermits the use of emoticons that are the pipe indicators of the author's view on the content. Tweets also consist of a time, date and the addict's name. The timestamp can be used for guessing the unborn trend application of the project. User location, if accessible, can also help to calculate the drifts in different geographical regions. For performing twitter data analysis,initially,the datais collected intothe local HDFS, using flume. Then the tweets are preprocessed for eliminating noise and pointlesslogos. Lastly,the sentimental analysiswill be used for tweets analysis.

II. HADOOP FOR DATA ANALYTICS

The important challenge of the IT world is to analyze and store thetremendous amount of datagenerated from various fields. Applications which are highly expandable and trustable for storage are used for analyzing thetremendous amount of data for understanding their customers. To meet these requirements Google had developed Google File System(GFS) and MapReduce programming model. GFS could handle the tremendous amount of data for storing, processing, analyzing and retrieving.Butthe main issue with GFS was that this filesystem was owned, so the researcher crew of Yahoo evolved an open source implementation of GFS and later this open-source project was named as Apache Hadoop. Hadoop mainly consists of two components: HDFS and MapReduce for storing and analyzing data respectively. Hadoop framework comprehendsvarious modules for variousfunctions as shownbelow.

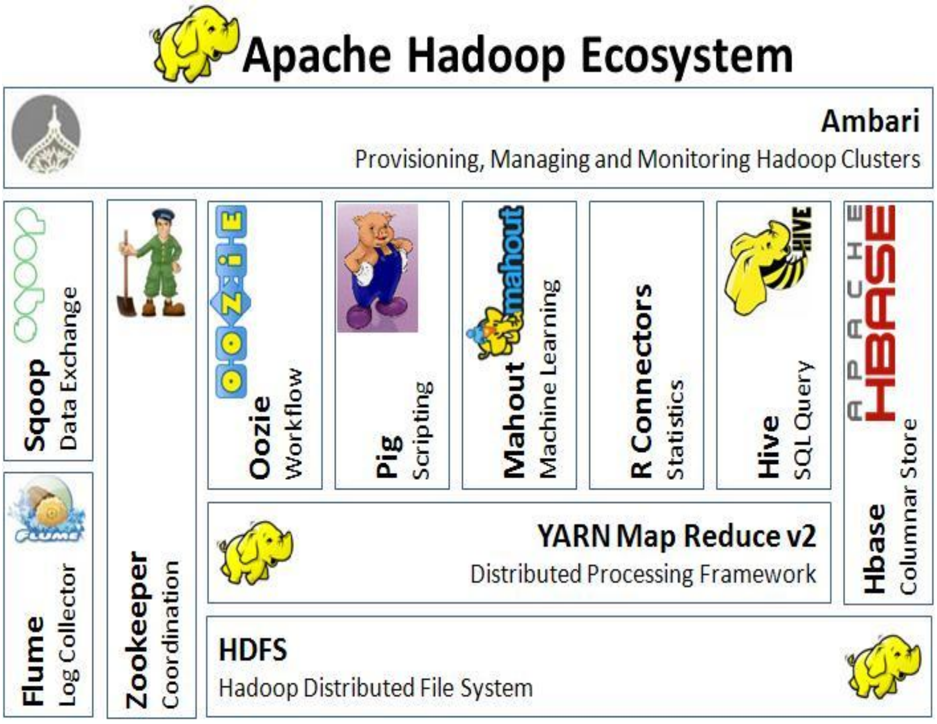


Fig. 1. Hadoop Ecosystem

2.1 HDFS

Hadoop uses HDFS (Hadoop Distributed File System) for storing the data. The HDFS is grounded on the Google File System (GFS) and deliversanetwork that aims to run on large blocks of small computer machines in a safe, fault-tolerant manner. HDFS uses a master-slave architecture wherein the master accords a single ‘Name Node’ that addresses the file system metadata and more than one slave Data Nodes that store the actual data.

Reason for using Hadoop HDFS for tweet storing:

Distributed storage and processing, Security, Reliability, Speed, Efficiency, Availability, Scalability

2.2 Apache Flume

Apache Flume is an application built on Hadoop which is distributed, dependable and accessible service for efficiently gathering, amounting, and moving huge amounts of streaming data into theHDFS. It has a simple and adaptableframe based on streaming data flowand is well-conditioned and fault resistant with tunable, dependable mechanisms for failover and recovery.

Source: This entity is used for extracting and receiving data from the web server.

Sink: This entity is used for delivering the data to the point of disembarkation.

Channel: It is the medium between source and sink

Agent: It is the physical Java virtual machine controlling flume.

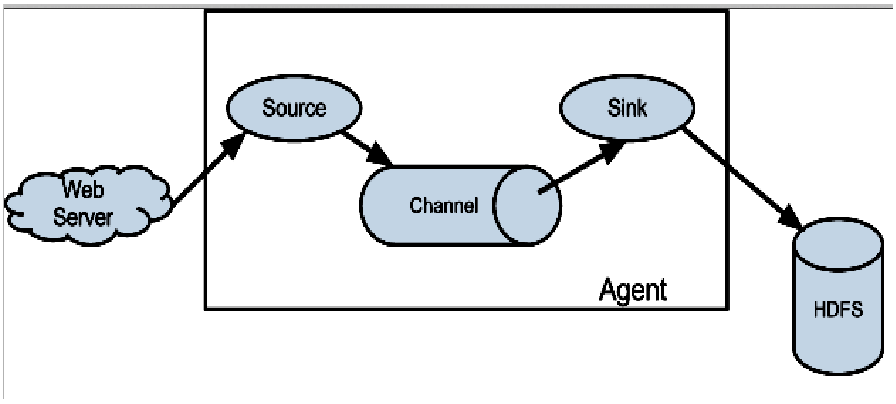
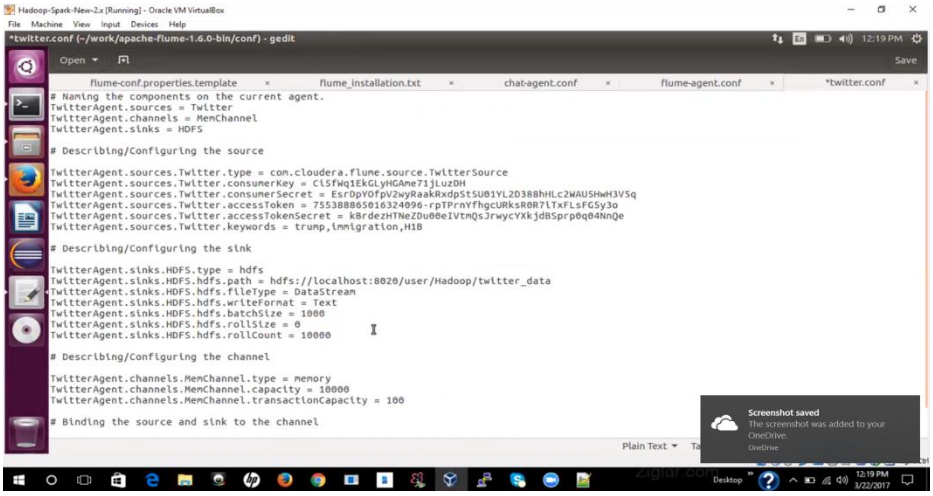


Fig. 2. Apache Flume

III. EXTRACTION OF DATA USING FLUME

1) To extract data from Twitter firstly, a Twitter Application must be created at “apps.twitter.com”. The developer has to sign-in using his existing account, click on the “create new app” to fill the required details and click on the “create your Twitter application” button.
2) Once the application is generated it is deadly important for the developer to generate the keys and access tokens. The consumer key and consumer secret key are pre-generated but the access tokens must be created. These are the major keys for retrieving data from Twitter.
3) Replace the keys and tokens in the configuration file with the newly generated ones. This is the step where Apache Flume gets access to the Twitter account. The snapshot below shows the configuration file.

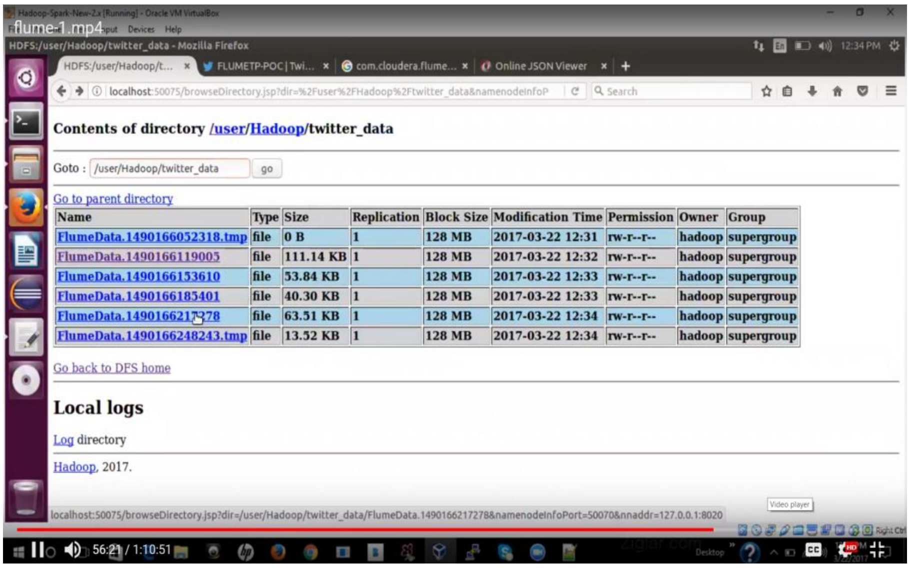

Hadoop-Spark-New-2.x [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
*twitter.conf (~/work/apache-flume-1.6.0-bin/conf) - gedit
Open
Save
flume-conf.properties.template
flume_installation.txt
chat-agent.conf
flume-agent.conf
*twitter.conf
Plain Text
Screenshot saved
The screenshot was added to your OneDrive.

```
# Naming the components on the current agent.
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.consumerKey = Ci5fWq1EkGLyHGAme71jLuzDH
TwitterAgent.sources.Twitter.consumerSecret = EsrDpYOfpV2wyRaakRxdpStSU01YL2D388hHLc2WAU5HwH3V5q
TwitterAgent.sources.Twitter.accessToken = 755388865016324096-rpTPrnYfhgcURksR0R7iTxFLsFG5y3o
TwitterAgent.sources.Twitter.accessTokenSecret = kBrdezHTNeZDu00eIVtmQsJrwycYXkjdBSprp0q04NnQe
TwitterAgent.sources.Twitter.keywords = trump,immigration,H1B

# Describing/Configuring the sink

TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:8020/user/Hadoop/twitter_data
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

# Describing/Configuring the channel

TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100

# Binding the source and sink to the channel
```

Fig. 3.Flume configuration file

Twitter data will be extracted based on a few specific words which must also be mentioned in the configuration file. In the “TwitterAgent.sources.Twitter.keywords” add the desired words for data extraction.
4)$<location of Flume> -ng agent -n <name of the agent> -f $<location of the agent>-Dflume.root.logger-DEBUG,console

The above command must be executed in the Hadoop terminal to establish a connection between the twitter application and HDFS and to receive the stream data from Twitter based on the given keywords in the configuration file. The extracted data will be stored in the location specified at “TwitterAgent.sinks.HDFS.hdfs.path” in the configuration file. The below-attached snapshot shows the location of the files which stores the extracted data.


Hadoop-Spark-New-2.x [Running] - Oracle VM VirtualBox
flume-1.mp4
Devices Help
HDFS:/user/Hadoop/twitter_data - Mozilla Firefox
HDFS:/user/Hadoop/t...
FLUMETP-POC | Twi...
com.cloudera.flume...
Online JSON Viewer
localhost:50075/browseDirectory.jsp?dir=%2Fuser%2FHadoop%2Ftwitter_data&namenodeInfoP
Search
localhost:50075/browseDirectory.jsp?dir=/user/Hadoop/twitter_data/FlumeData.1490166217278&namenodeInfoPort=50070&nnaddr=127.0.0.1:8020
Video player
56:21 / 1:10:51

Contents of directory /user/Hadoop/twitter_data

Goto : /user/Hadoop/twitter_data go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FlumeData.1490166052318.tmp	file	0 B	1	128 MB	2017-03-22 12:31	rw-r--r--	hadoop	supergroup
FlumeData.1490166119005	file	111.14 KB	1	128 MB	2017-03-22 12:32	rw-r--r--	hadoop	supergroup
FlumeData.1490166153610	file	53.84 KB	1	128 MB	2017-03-22 12:33	rw-r--r--	hadoop	supergroup
FlumeData.1490166185401	file	40.30 KB	1	128 MB	2017-03-22 12:33	rw-r--r--	hadoop	supergroup
FlumeData.1490166217278	file	63.51 KB	1	128 MB	2017-03-22 12:34	rw-r--r--	hadoop	supergroup
FlumeData.1490166248243.tmp	file	13.52 KB	1	128 MB	2017-03-22 12:34	rw-r--r--	hadoop	supergroup

Go back to DFS home

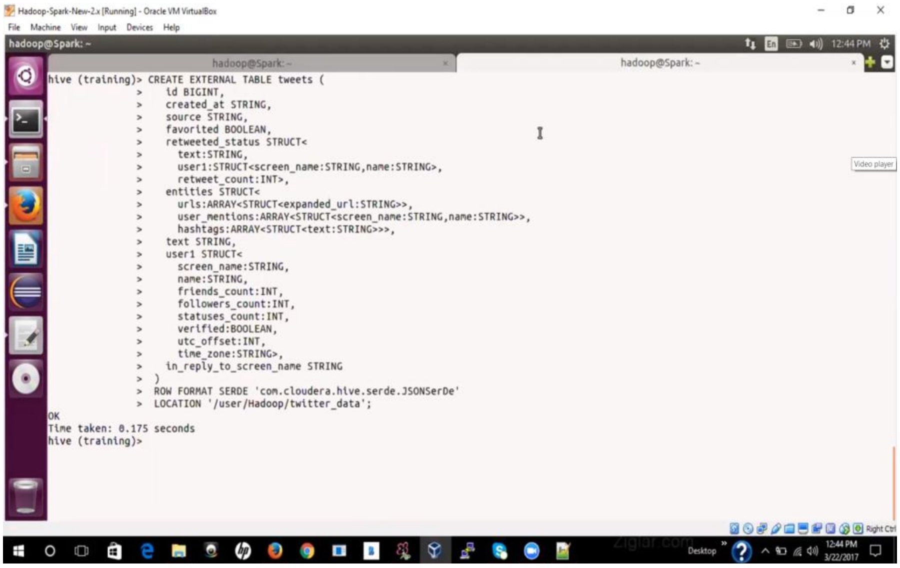
Local logs

Log directory

Hadoop, 2017.

Fig. 4. Extracted files

5)The contents of the files will be in 'JSON format. To view the actual data, we can either use online JSON viewers which displays the given json formatted data into an understandable format or we can use Apache Hive by creating tables and adding jar files to the Hive terminal. The attached snapshot shows the format for creating the table in Hive.



```
hive (training)> CREATE EXTERNAL TABLE tweets (
               >   id BIGINT,
               >   created_at STRING,
               >   source STRING,
               >   favorited BOOLEAN,
               >   retweeted_status STRUCT<
               >     text:STRING,
               >     user1:STRUCT<screen_name:STRING,name:STRING>,
               >     retweet_count:INT>,
               >   entities STRUCT<
               >     urls:ARRAY<STRUCT<expanded_url:STRING>>,
               >     user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
               >     hashtags:ARRAY<STRUCT<text:STRING>>>,
               >   text STRING,
               >   user1 STRUCT<
               >     screen_name:STRING,
               >     name:STRING,
               >     friends_count:INT,
               >     followers_count:INT,
               >     statuses_count:INT,
               >     verified:BOOLEAN,
               >     utc_offset:INT,
               >     time_zone:STRING>,
               >   in_reply_to_screen_name STRING
               > )
               > ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
               > LOCATION '/user/Hadoop/twitter_data';
OK
Time taken: 0.175 seconds
hive (training)>
```

Fig. 5. Table creation in HIVE

On performing this step, all the json data is converted into a tabular format. We can extract the data from these tables using the command, "select * from <table name> limit <number of records to be displayed>;"

Extraction of data from twitter ends here and is available on the HDFS.

Data analysis can be done in diverse methods. If we utilize the components of the Hadoop ecosystem, the data from HDFS can be used directly. In this document, we shall be using Sentimental analysis, so data has to be copied from HDFS to local FS. This can be completed by running the following 'HDFS get' command in the terminal.

```
bin/hadoop fs -get /<location of the file in HDFS>/opt/<location in the local fs>
```

IV. APPROACH FOR SENTIMENTAL ANALYSIS

Sentimental Analysis (SA) is a study of people's emotions, opinions and attitude on an event or an issue. SA can also be defined as a process for the classification of words. It targets towards finding people's opinions and classifying them based on their polarity.

4.1 TRIPOD OF SENTIMENTAL ANALYSIS

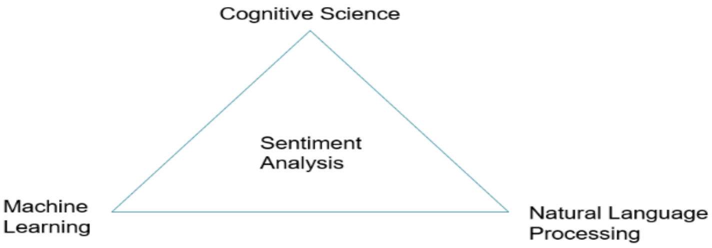


Fig. 6. Tripod of sentimental analysis

1.Cognitive science is an interdisciplinary study of an idea, learning, and study of mind.
2.Machine learning is the application of artificial intelligence, in the discipline of computer science, which gives the computer the power to grasp and improve without external programming. The aim is to construct algorithms to make the computers grasp automatically by receiving input, performing statistics and predicting outputs.
3.Natural language processing (NLP) is the concept of making the computer understand the general human language.

4.2 Analysis approach

The Sentimental analysis is divided into 5 major steps:

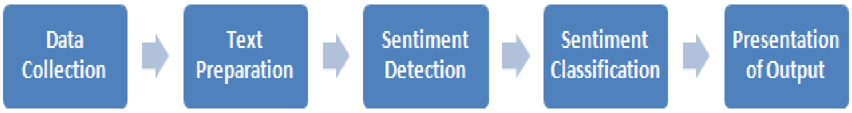


Fig. 7. Steps in the analysis

4.2.1 Data Collection

Consumers usually express their feelings or reviews on social media platforms like Facebook, Twitter, Blogs, Discussion boards, etc. In this project, the data is collected from Twitter using Flume as discussed above in section III.

4.2.2 Text preparation

Text preparation is filtering the extracted data before analysis.

i) It includes distinguishing and debarring the STOP WORDS like a, an, the which are irrelevant for the aim of analysis. These are called STOP WORDS and have no significance during analysis.

ii) Unstructured data to structured data: Twitter comments are usually unstructured, for example,'aswm' is written 'awesome'. Dynamic data records are used for these types of conversions.

iii) Emoticons: This is the commonly used method to express opinions or feelings. It is a symbolic representation which will be converted into words in this stage. Ex: ":)" denotes a smiling face which will be converted into the word "happy".

4.2.3 Sentiment detection

At this stage, the sentences are examined to identify the subjective and objective expressions. The expressions which are subjective are retained and the rest are discarded. The words of subjective expressions are rated depending on its nature. The rating is usually done on a scale of 0 to 1.

4.2.4 Sentiment classification

By using the concept of Machine Learning, the system is pre-trained about three main group of words i.e. positive, negative and nonpartisan(neutral) words. The detected subjective words are classified into the before mentioned groups and further rated. The most positive words like 'Excellent', 'Eloquence' and 'Magnanimous' is given a rating of 1 and the most negative words like 'Apathy', 'Cold-hearted' and 'Atrocious' is given 0. In this way, the trained system can rate the subjective words by comparing it to the stored ratings and the nature of the word.

4.2.5 Presentation of Output

Before the final output is displayed, all the ratings are summed up to get the final result of the analysis. The centralnotion of sentiment analysis is to convert semi-structured and unstructured text into meaningful analyzed information. After the completion of the analysis, the analysis results can be displayedusing text or anygraphical representation.

V. OVERVIEW OF THE APPROACH

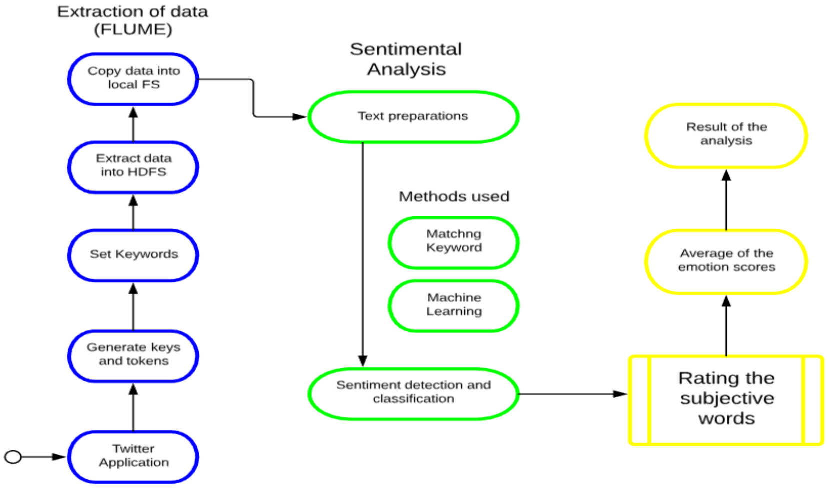


Fig. 8. Overview picture

VI. ACCURACY

The implementation technique has been tested on the standard dataset available at the following link:
http://www.cs.tau.ac.il/~kfirbar/mlproject/twitter.data
The accuracy of the project based on the above dataset is as following:

ACCURACY RESULTS

Sentiment	**Count**	**Correct count**	**Percentage(%)**
Positive	729	542	74.3
Negative	665	458	68.8
Neutral	72	53	73.6

So, the overall accuracy came up to 72.2%.

VII. CONCLUSION

Twitter isa very important source of opinion on different issues. It can give us a keen insight into anyissue and can be a good source for analysis. Analysis helps us in decision making in various areas. Apache Hadoop is one of the primeadd-ons for Twitter dataextraction, as it extracts real-time data also. Once the system is configured using FLUME, it helps in extracting avariety of subjects by setting the keywords in the configuration file. The sentimental analysis is the most felicitous method for data analysis as its accuracy is almost equal to 73%.

REFERENCES

[1]. Sunil B. Mane, Sunil B. Mane, Yashwant Sawant, SaifKazi, Vaibhav Shinde, “Real Time Sentiment Analysis of Twitter Data Using Hadoop”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3098 – 3100, ISSN:0975-9646.

[2]. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251

[3]. Sunil B. Mane, Yashwant Sawant, SaifKazi, Vaibhav Shinde ... (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3098 - 3100 http://ijcsit.com/docs/Volume%205/vol5issue03/ijcsit2014050393.pdf

[4]. L.M Patnaik, "Big Data Analytics: An Approach using Hadoop Distributed File System", International Journal of Engineering and Innovative Technology (IJEIT), vol. 3, pp. 239-243, May 2014.

[5]. Big data emerging technologies: A case study with analyzing Twitter data using apache hive

a. https://ieeexplore.ieee.org/document/7453400/references

A.Shreya Shetty "Tweet Analysis: Extractionand Sentimental Analysis Ontwitter Data..” American Journal of Engineering Research (AJER), vol. 7, no. 08, 2018, pp. 10-15