International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

# SENTIMENT ANALYSIS ON TWITTER POSTS USING HADOOP

**Soumy Sharma[1], Sandeep Kumar[2]**

[1] *Bachelor of Technology, Department of Computer Science and Engineering, I.T.S Engineering College, Greater Noida, India*
[2] *Assistant Professor, Department of Computer Science and Engineering, I.T.S Engineering College, Greater Noida, India*

---

**Abstract -** *In today's world, every keep themselves updated with world using the various social media platforms. Twitter, is the one of the most popular platforms for sharing or viewing of someone's opinions or views as well as post about their perspective and queries regarding it.*

*In this application, people share their perspective using 'tweets'. These tweets are visible to the public and accessible to everyone. Twitter API helps to retrieve these posts or tweets and creates a database with raw data gathered from tweets. In this research, we take into account the event of upcoming event of elections. People posting their views regarding the various political parties, their work and achievements, future plans or promises etc. and this data is being collected and used for the purpose of research about people's opinions and views on the kind of political image that persists in the minds of people and can further be used for the prediction of exit poll.*

***Key Words*: Hadoop, Opinion mining, twitter, tokenization, unstructured data, sentiment analysis, tweet.**

## 1. INTRODUCTION

Due to the sudden exponential increase in the number of users having access to the internet services, there is also in huge increase in the amount of data. The users using various platforms to share or access information add data to the service providers' database servers. This results in large accumulation of raw data. This data can be accessed using the data mining technologies and can further be used for business or research purposes. Twitter, is among the most popular social media platform for expressing one's opinion or seeing others point of view on various issues, events or incidents that occur in our lives.

Twitter being among top social media platforms, also becomes an important source for research of public opinions on various foci. Thus, it provides an ideal environment for the purpose of analysis.

There are millions of tweets that are been posted or viewed by people across the world, every single day. These tweets are stored on the computational machine. After this process, the data is collected in the unprocessed form and is first made into tabular form.

Using the Ni-Fi technology, we move the data from local storage to Hadoop clusters.

During this transfer process, 'hive' in used to classify and arrange the unstructured data into usable information in specific format under various categories. Finally, we have the data required for the analysis of tweets. The information is analyzed based upon the criteria that we assign for the same.

## 2. PROBLEM STATEMENT

Tweet examination can be significant in understanding the sentiments behind a tweet, which are in far reaching numbers. In evident various tweets, might be phony and can be ambiguous. With the use of Apache Hadoop, we can quicken the strategy in separating the tweets. The traditional system can't predict and separate these tweets from useful ones. Hadoop has couple of drawbacks such as: The structure does not make extraordinary usage of the Twitter API. The system is customer dependent, for example, the customer needs to research twitter autonomous from any other person which isn't possible. It can't help in sentiment examination of tweets. It was not capable of using big data.

## 3. SENTIMENT ANALYSIS

Social platforms are a rich stage to find out about an individuals' conclusion and sentiment seeing diverse themes as they can impart their insight effectively on social medias including Facebook and Twitter. There is distinctive sentiment situated data gathering frameworks which intend to remove individuals' assessment with respect to various points. The sentiment-mindful frameworks nowadays have numerous applications from business to sociologies.

Sentiment Analysis, otherwise called Opinion Mining is a field inside Natural Language Processing (NLP) that fabricates frameworks that endeavours to recognize and separate feelings inside the content. Normally, other than recognizing the conclusion, these frameworks remove traits of the articulation, example:

Polarity: if the speaker express a positive or negative sentiment,
Subject: what is being discussed,

Supposition holder: the individual, or substance that communicates the conclusion.

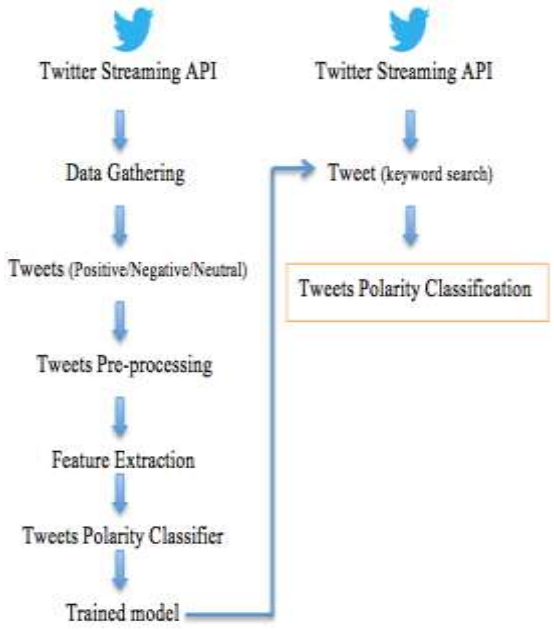


**Fig -1**: Tweet Processing

At present, opinion mining is a point of extraordinary intrigue and advancement since it has numerous pragmatic applications. Since freely and secretly accessible data over Internet is continually growing, countless communicating conclusions are accessible in audit destinations, blogs, web journals, and online media.

With the assistance of sentiment analysis frameworks, this unstructured data could be naturally changed into organized information of general conclusions about items, administrations, brands, legislative issues, or any theme that individuals can express suppositions about. This information can be helpful for business applications like investigation, advertising, item audits, customer feedback, market analysis, and client administration.

Since informal organizations, particularly Twitter, contains small texts (called tweets) and individuals may utilize diverse words and shortened forms which are hard to extricate effectively, in this manner a few researchers have utilized Hadoop ecosystems systems to concentrate and mine the extremity or polarity of the tweets. Some of the top and frequently used abbreviations are EC for election commission, EVM for Electronic voting machine, etc. Thus, due to use of abbreviations and internet acronyms, the process of sentimental analysis for short messages like Twitter's posts is challenging.

## 4. TOOLS & TECHNOLOGY

### 4.1 HADOOP

Hadoop is an open-source framework used for the purpose of storing data as well as running various applications on the clusters of computers. It provides with large room for storage of any kind of data, along with massive processing capabilities and the ability to effetely handle concurrent tasks.

Hadoop runs applications utilizing the MapReduce calculation, where the information is prepared in parallel with others. To put it plainly, Hadoop is utilized to create applications that could perform total factual investigation on immense measures of information.

### 4.2 HDFS

The Hadoop Distributed File System (HDFS) provides a distributed file system that is intended to execute on cluster of computers. It has high fault tolerance and is designed to run even on less expensive hardware. It gives quicker access to app information and is also practical for applications which consists huge datasets.

### 4.3 MapReduce

MapReduce is a programming model used for composing circulated applications for efficient processing of large datasets, on huge clusters (a huge number of computers) of hardware in a fault tolerant way. The MapReduce program keeps running on Hadoop.

The MapReduce calculation contains two imperative processes, to be specific, Map and Reduce. The Map task takes a lot of data and changes over it into another arrangement of data, where singular components are separated into tuples (key value pair). The Reduce task takes the yield from the Map as an information and joins those data tuples (key value pair) into a little arrangement of tuples.

### 4.4 APACHE HIVE

Apache Hive is used for the analysis and performing operations on the data using query language. It is regularly a piece of good instruments conveyed as a major aspect of the product biological system dependent on the Hadoop structure for taking care of huge informational collections in a disseminated processing condition.

Like Hadoop, Hive was created to deliver the need to deal with petabytes of information collecting through the various internet usage.

### 4.5 TABLEAU

Tableau is an important data visualization tool that is widely being utilized by the Business Intelligence Industry. It is used for transforming raw data into comprehensible format. This format helps to make it possible for anyone to easily decipher the processed data.

### 4.6 APACHE AMBARI

The Apache Ambari is a tool that is designed for the purpose of management of Hadoop clusters. It runs on the top of Hadoop clusters to keep record and manages the ongoing processes or tasks. It provides a user interface for the administrator to manage these running applications on the Hadoop clusters. The main aim of Apache Ambari is to ease the management of Hadoop by providing easy-to-use interface.

### 4.7 NIFI

Apache NiFi is a platform used for movement of data among dissimilar frameworks. It gives continuous control that makes it simple to deal with the transfer of information between any source and destination systems. Apache NiFi enables real time tracking of information.

## 5. METHOLOGY

The described methodology is used to defeat the issues looked before in the proposed framework. That is:-

1. A twitter application is made utilizing a twitter streaming API for getting the twitter information.
2. After that information is transferred in the local HDFS, utilizing a tool called Apache Flume. In Apache Flume, using the twitter API, all the desired tweets are extracted straight from the twitter website.
3. These tweets are stored in the Hadoop Distributed File System.
4. The collected data from twitter in the amorphous form and requires to be organized.
5. For organizing the unstructured data, Hive is used. It is used to transform the un-organized complex information into a meaningful structure.
6. The structured data is then preprocessed to evacuate the NULL values and tautologies from the information.

### 5.1 TWITTER API CREATION

A twitter application is made utilizing a twitter streaming API for extracting the twitter information in real time.

The following are the steps to make the twitter API:

1. Open web browser and go to https://dev.twitter.com/apps/new
2. Enter your Application Name, Description as well as the site address. Callback URL can be left empty.
3. Accept the TOS, and fill the CAPTCHA.
4. Submit by tapping the Create your Twitter Application.
5. Copy and use the customer key or the API key and buyer secret from the screen into your application.

## 6. RESULTS AND CONCLUSION

Apache Hadoop provides abetment in twitter post analysis. Using the tools like, FLUME and HIVE, helps to fetch a diverse range of results by simply changing the keywords in input query.

The tweets are extracted from different countries to examine about the views and opinions of the citizens of that country regarding elections. The complete analysis was conducted on real time data, so is more valuable. The analysis could be beneficial in decomposing individuals' sentiment. The results reveal about the polarity of tweets, that is, either they are positive, negative or neutral. This researched data could be beneficial for candidates who are contesting elections for developing new strategies as well as modifying the existing campaigning strategies for the elections or can be used for further research purposes.

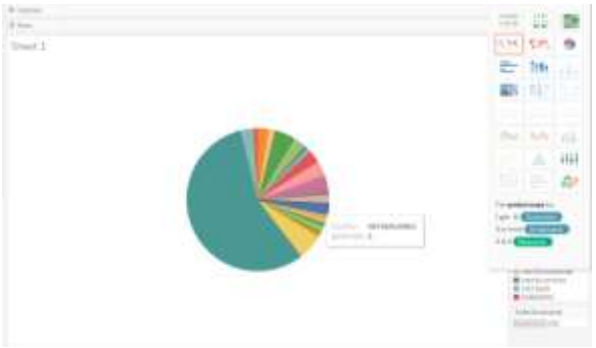

**Fig -2**: Output Analysis Chart

## 7. REFERENCES

1. Harshal Kapase, Kalyani Galande, Tanmay Sonna, Deepali Pawar, Dipmala Salunke “Sentiment polarity analysis on Twitter data from different Events” published on 03 March 2018
2. Md. Nowraj Farhan, Md. Ahsan Habib and Md. Arshad Ali “A study and Performance Comparison of MapReduce and Apache Spark on Twitter Data on Hadoop Cluster”. Published online on 08 | July-2018.
3. Harshal Kapase, Kalyani Galande, Tanmay Sonna, Deepali Pawar, Dipmala Salunke “A review on:

Sentiment polarity analysis on Twitter data from different Events" Volume: 05 Issue:03 | Mar-2018.

4. Riya Bhatia, Prachi Garg, Rahul Joharic "Corpus based Twitter Sentiment Analysis" 3rdInternational Conference on Internet of Things and Connected Technologies (ICIoTCT), Jaipur(India) on 26 | March, 2018.
5. Brinda Hegde, NagashreeH, Madhura Prakash "Sentiment analysis of Twitter data: A machine learning approach to analyse demonetization tweets" International Research Journal of Engineering and Technology (IRJET) ,Volume: 05 Issue: 06 | June-2018
6. Priyanshu Jadon, Miss Rupali Dave "A big data approach for Sentiment Analysis of Twitter Data using Naïve Bayes" International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 7, Issue: 04 | April 2018.
7. Mika V. Mäntylä, Daniel Graziotin, Miikka Kuutila "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers" M3S, ITEE, University of Oulu, Finland,Institute of Software Technology, University of Stuttgart, Germany, Issue : 21 | Nov-2017.
8. Amit Palve, Rohini D.Sonawane , Amol D. Potgantwar "Sentiment Analysis of Twitter Streaming Data for Recommendation using Apache Spark" Volume-5, Issue-3 | June 2017
9. Priya Gupta "Analysis of User Behaviour for Twitter Posts on Hadoop" International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 05 | May -2017