A Study on Sentimental Analysis of Twitter Through Big Data Using Hadoop

K. Uma Maheswari[1], G. Hemaa Nandhini[2], S. Keerthna[3]
[1, 2, 3] Dept of Mathematics
[1, 2, 3] Sri Krishna Arts and Science College, Coimbatore

Abstract- *The growth of social media leads to enormous interest among the users who use internet today. The data that we obtain from the social networking site are used for many purpose like prediction, Sentiment Analysis etc. In this paper we take the networking site twitter. There are huge number of tweets received every year and to analyze these tweets we use sentimental analysis. So, handling Big data and for analyzing we use Hadoop*

Keywords- Hadoop, Sentiment analysis, Big data, social networking site.

I. INTRODUCTION

Twitter is an social network which is extensively used for posting short messages to explicit people‘s views and opinions. Sentimental Analysis is a form of data mining that measures the opinion of people by use of natural language processing (NLP), Analyzing the text, Computational linguistics to identify and observe the subjective information in the social media. The total number of tweets (messages) updating are increasing for every year. It is difficult to handle this vast data.

To study this big data Hadoop technology is used. Hadoop is an ascendable open source framework. Hadoop technology is used to execute actions on stored data in a useful way. MapReduce is a processing in Hadoop where it allows to parallel processing of the data stored in HDFS. HDFS is the storage of Hadoop technology that allows to dump any kind of data across the cluster. In this paper, people‘s opinion shared in social media in that we are selecting a particular person and analysing their opinions by positive, negative and neutral comments.

II. PROPOSED ARCHITECTURE

Steps involved in architecture:

a. Data source

In our country, there are around million of twitter users. The main source of data is that the tweets posted about the service provider.

b. Hadoop

Apache Hadoop is the open source framework that performs big data in the efficient way. It contains HDFS file system to store data and MapReduce engine to process data.

c. Data collection

The tweets posted on one year are used to study the sentiments. The coding is performed/designed in python or Java.

d. Naïve Bayes Classification

The Naïve Bayes is the efficient method to implement the study of classification. It helps to know the frequency of the words.

e. Training with Mahout

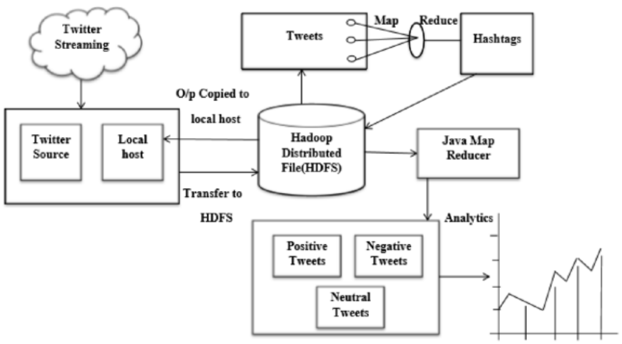
Mahout is the source that helps to train the data set by converting it into file format of hadoop sequence.

f. Data Cleaning & Preprocessing

Data Preprocessing plays a main role in sentimental analysis. That it maintains data cleaning to avoid repetition of words.

g. Data Analysis

The classification of tweets are positive, negative and neutral based on keywords. This analysis is used to present the solution of sentimental Analysis.



III. MODULES OF SYSTEM

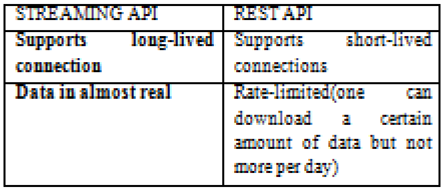
The system has the modules as follows:

Data Streaming

By twitter streaming API we obtain real time tweets. We use twitter data to classify and train the classifier. There are two API‘s that twitter handles:

- Stream API
- Rest API

The difference between Steaming API and Rest API are,



STREAMING API	REST API
Supports long-lived connection	Supports short-lived connections
Data in almost real	Rate-limited(one can download a certain amount of data but not more per day)

Preprocessing

In this paper, the tweets are observed as text data. At the beginning we need to delete or ignore the retweets that will cause a distort in the classification process. Then the symbols and punctuations are removed to avoid the inconsistency and the unnecessary symbols can effect the accuracy of the analysis process.

Sentiment Polarity Analysis

MapReduce is the model with parallel programming, thus the Sentimental Analysis algorithm based on Naïve Bayes is adapt to apply into MapReduce Model. We prefer to apply Naïve Bayes classifier with an English lexical dictionary sentiword net.

Visualization

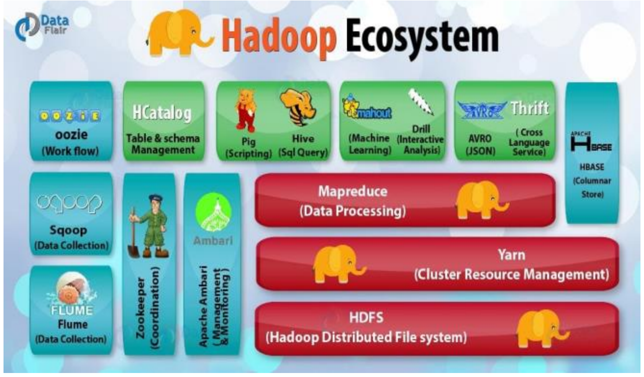
Different Visualization techniques are used to display the tweets. Now, we are preferring Business Intelligence (BI) tool for visualization.

Evaluation Metrics

In this we are explaining the efficiency of information from the observation.

IV. METHODOLOGY ADOPTED

There are three main components of the ecosystem of Hadoop:


Data Flair
Hadoop Ecosystem
oozie (Work flow)
HCatalog Table & schema Management
Pig (Scripting)
Hive (Sql Query)
mahout (Machine Learning)
Drill (Interactive Analysis)
AVRO (JSON)
Thrift (Cross Language Service)
HBASE (Columnar Store)
Sqoop (Data Collection)
Zookeeper (Coordination)
Ambari
Apache Ambari (Management & Monitoring)
Mapreduce (Data Processing)
Yarn (Cluster Resource Management)
FLUME Flume (Data Collection)
HDFS (Hadoop Distributed File system)

Hadoop Common

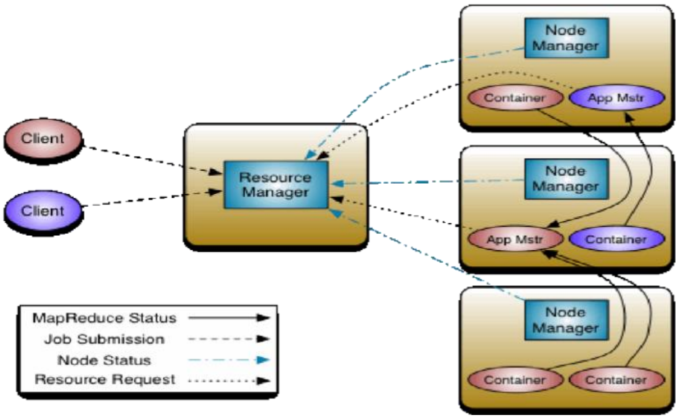
Within the Hadoop ecosystem, Apache foundation has the well-defined set of libraries that can be used by other sectors. For example, if HBase and Hive need to process HDFS they want to make JAR files(Java archives) that are stored in Hadoop common.

Hadoop Distributed File System (HDFS)

HDFS component build various copies of data section to the distributed across different clusters for secure and fast data process. HDFS includes of three major nodes:

- ➢ Name Node
- ➢ Data Node
- ➢ Secondary name Node.

HDFS works on a Master Slave Architecture Model where the Name Node perform as the Master node for maintaining a record of the strong cluster and the Data node perform as a Slave node counting up to the numerous systems inside the Hadoop cluster.



Mapreduce

The important basis of action behind MapReduce is that the _Map' task sends a query for performing to several nodes in the Hadoop cluster and the _Reduce' task collects all the results of the output into a single value. Map is the job that in Hadoop ecosystem takes input data and splits into independent chunks and output of the Map job will become the input of the Reduce job. Since the Reduce job combines Mapped data tuples into smaller set of tuples. Thus, the both input and output of the job are collected in file system. MapReduce performs with allotting jobs, supervising jobs and re-performs the failed task. MapReduce is a framework that gives the compute node and also the HDFS file system gives the data node.

V. IMPLEMENTATION

- Gather unstructured data from Social network.
- Real-Time processing with a sentimental analysis engine based on keyword search.
- Accumulate prepared data (with sentiment) in NoSQL database.
- Analyze sentiments from NoSQL to visualization layer.
- Represent/visualize information with a tool.

VI. RESULT DISCUSSION

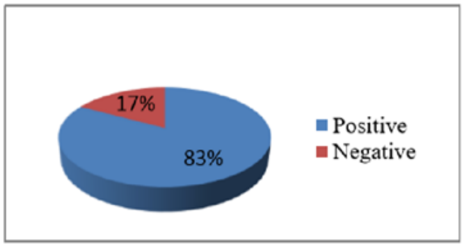
Twitter is a generally useful social network for sharing comments on different subjects in the form of short messages. In this paper, tweets have been gathered and transformed into a training set by using a python script. The tweets are gathered by using hashtag, which are meant to refer the comfort and discomfort level of consumers with respect to the service provider. After the training set has been processed, data is interpret by uploading it on HDFS and Naïve Bayes classification is accomplished.

Thus the training set is transformed into vectors using tfidf weight (term frequency x document frequency). This is made to allow weight to each one phrase in the word list generated from the set could be done by considering the phrase frequency, which finds the frequency of each phrase in the paper. The second condition is Inverse Document Frequency which process that the lesser the existence of the phrase in all documents, the more is the phrase value in this element.

VII. ANALYSIS

The study on Twitter exhibit that around 17% of the point of view were negative, 83% were positive point of view about a well known person Narendra Modi is taken dataset.

Figure represent a pie chart of the overall point of view of the people.



VIII. CONCLUSION

This paper terminates that different classes have been evaluated on tweets of well-known person. It is also effective in gathering the point of view of everyone when it comes to deserve topics associated to any fields. In our case study, we can more to compare the work of different providers and judge which is the best.

As the future service, we can also compare the analysis of different peoples and conclude that who is the best. By Hadoop MapReduce and Naïve Bayes algorithm, we can maintain a ordinary computerized method to access what people's information from social networks and analysing it using Big data techniques.

REFERENCES

[1] G. Vinodhini and R. Chandrasekaran, “Sentiment analysis and opinion mining: A survey,” International Journal, vol. 2, no. 6, 2012.

[2] H. Cui, V. Mittal, and M. Datar, “Comparative Experiments on Sentiment Classification for Online

Product Reviews." In Proceedings of AAAI-06, 2006, pp.1265-1270.

[3] Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010, 2010.

[4] Neethu M S and Rajasree R, "Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India IEEE – 31661.

[5] Y. Mejova, "Sentiment analysis: An overview," Comprehensive exam paper, http://www.csuioedu/˜ymejova/publications/CompsYelena Mejova. Pdf [2010-03], 2009.

[6] https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/

[7] https://blog.algorithmia.com/sentiment-analysis-with-twitter/

[8] https://www.digitalvidya.com/blog/twitter-sentiment-analysis-introduction-and-techniques/

[9] https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed