Sentiment Analysis of Twitter Data through Big Data

Anusha.N
Computer Science and Engineering
Sai Vidya Institute of Technology
Bangalore, India

Divya.G
Computer Science and Engineering
Sai Vidya Institute of Technology
Bangalore, India

Ramya.B
Computer Science and Engineering
Sai Vidya Institute of Technology
Bangalore, India

***Abstract*— The rise of social media leads to tremendous interest among the internet users nowadays. Data from these social networking site is used for many puposes like prediction, marketing, sentiment analysis etc. In this paper we are considering the social media site-Twitter for analyzing the sentiments because huge number of tweets received every year could subjected to sentiment analysis. So, to handle these Big Data and for analysis we are using Hadoop .**

Keywords—Twitter; Hadoop, Sentiment analysis; Big Data; Social networking site;

I. INTRODUCTION

Twitter is a widely used platform for posting comments and people can express their views and opinions. Sentiment analysis refers to use of natural language processing, text analysis to computational linguistics to identify and extract subjective information in source material. Number of tweets received every year is increased. It is hard to process this huge data.

To analyze this big data we are using the Hadoop technology in this paper .Hadoop is a scalable open source framework where Hadoop technology helps us to perform operations on distributed data in an efficient manner. Hadoop contains a programming model called Map Reduce where it provides an associated implementation for processing and generating big data sets with parallel, distributed algorithm on a cluster. In this paper, we are taking a opinions of the people on a well known person. People expressed their views about the person which helps us to analyze the positive, negative and neutral comments .

II. PROPOSED ARCHITECTURE

Our proposed architecture includes different methods/steps like:

1.Data Source: There are around million of twitter users in India as per statistics. Thus posts tweeted about the service provider are the main source of data.

2.Hadoop/Mahout: Apache Hadoop is an efficient and scalable open source framework that processes big data in a distributed manner. It consists of HDFS file system and Mapreduce engine.

3.Data Collection: One year posted comments or data are taken to analyze the sentiments. A program is designed in Python/Java.

4.Naïve Bayes Classification: Naïve Bayes gave an effective method to carry out the study of classification. It is used to know the word frequency.

5.Training with Mahout: Here we train the data set using Mahout by converting it into hadoop sequence file format.

6.Data Cleaning and Pre-Processing: Text pre-processing is an important phase for sentiment analysis. It contains Data Cleaning No Repetition Text Correction.

7.Data Analysis : The positive, negative or neutral tweets are analyzed based on key words. The classification is analyzed to find the results of sentiment analysis.

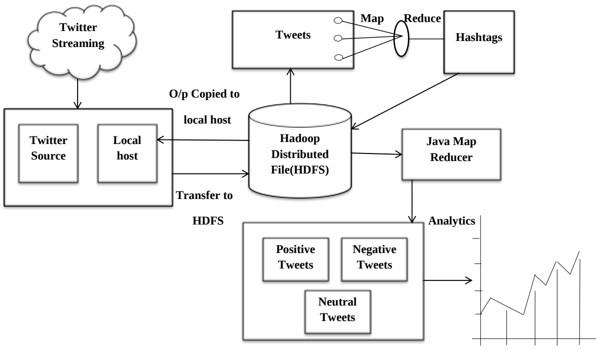

Fig 1: Architecture of proposed system

III. MODULES OF PROPOSED SYSTEM

The proposed system has the following modules;

1.Data Streaming : Extracting real time tweets using Twitter Streaming API .For classification and training the classifier we need Twitter data. For this purpose we make use of API's twitter provides. Twitter provides two API's; Stream API1 and REST API2.

The difference between Streaming API and REST APIs are: Streaming API supports long-lived connection and provides data in almost real -time. The REST APIs support short-lived connections and are rate-limited (one can download a certain amount of data but not more per day).

2.Preprocessing : In this phase, the tweets are available as text data and each line contains a tweet. Initially we clean up or remove retweets as that will induce a bias in the classification process. We need to remove the punctuations and other symbols that doesn't make any sense as it may result in inefficiencies and may affect the accuracy of the overall process.

3. Sentiment Polarity Analysis: MapReduce is a new parallel programming model, hence the classical Naive Bayes based sentiment analysis algorithm is adjusted to fit into Map Reduce model. we choose to employ a Naive Bayes classifier and empower it with an English lexical dictionary SentiWordNet

4.Visualization : Tweets are presented using several different visualization techniques. Here we are using BI (Business Intelligence) tool for visualization.

5.Evaluation Metrics : Here we are evaluating the effectiveness of information.

1V. PROPOSED METHODOLOGY

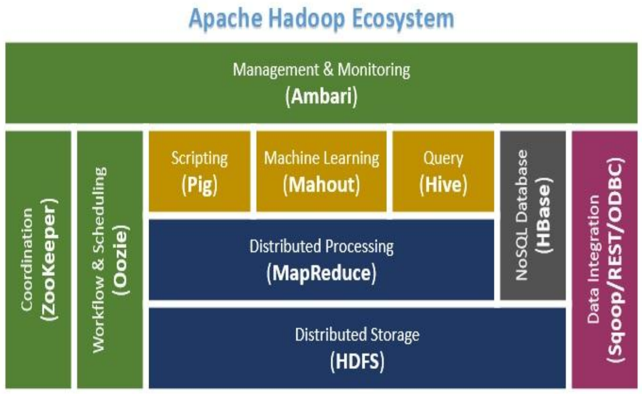

Fig 2: Apache Hadoop Ecosystem

The Hadoop Ecosystem comprises of 4 core components –

1. Hadoop Common-Apache Foundation has pre-defined set of utilities and libraries that can be used by other modules within the Hadoop ecosystem. For example, if HBase and Hive want to access HDFS they need to make of Java archives (JAR files) that are stored in Hadoop Common **.**

2) Hadoop Distributed File System (HDFS) - HDFS component creates several replicas of the data block to be distributed across different clusters for reliable and quick data access. HDFS comprises of 3 important components-NameNode, DataNode and Secondary NameNode. HDFS operates on a Master-Slave architecture model where the NameNode acts as the master node for keeping a track of the storage cluster and the DataNode acts as a slave node summing up to the various systems within a Hadoop cluster.

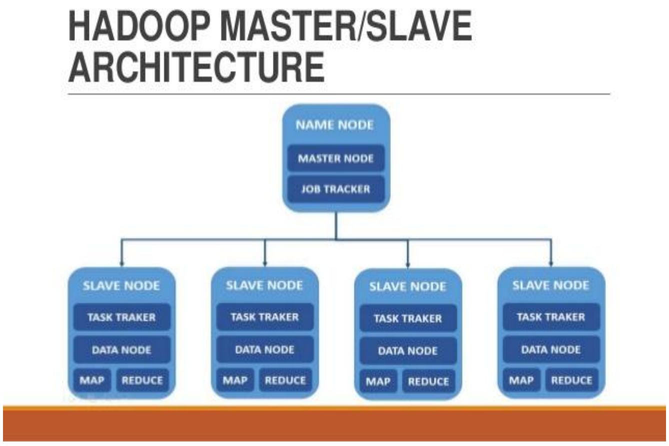

Fig 3: Hadoop Master/Slave Architecture

3) MapReduce-The basic principle of operation behind MapReduce is that the "Map" job sends a query for processing to various nodes in a Hadoop cluster and the "Reduce" job collects all the results to output into a single value. Map Task in the Hadoop ecosystem takes input data and splits into independent chunks and output of this task will be the input for Reduce Task. In The same Hadoop ecosystem Reduce task combines Mapped data tuples into smaller set of tuples. Meanwhile, both input and output of tasks are stored in a file system. MapReduce takes care of scheduling jobs, monitoring jobs and re-executes the failed task.MapReduce framework forms the compute node while the HDFS file system forms the data node.

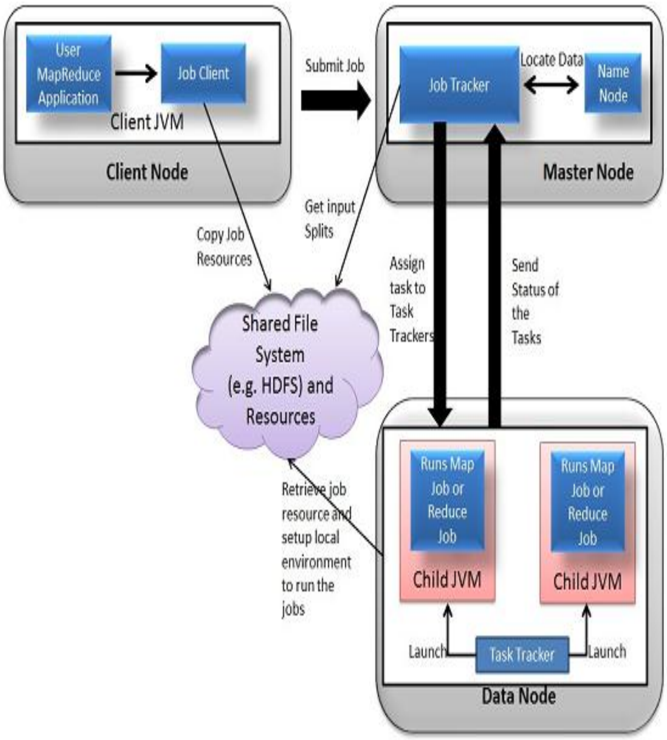
User MapReduce Application
Job Client
Client JVM
Client Node
Submit Job
Job Tracker
Locate Data
Name Node
Master Node
Copy Job Resources
Get input Splits
Shared File System (e.g. HDFS) and Resources
Assign task to Task Trackers
Send Status of the Tasks
Retrieve job resource and setup local environment to run the jobs
Runs Map Job or Reduce Job
Child JVM
Runs Map Job or Reduce Job
Child JVM
Launch
Task Tracker
Launch
Data Node

Fig 4: Apache Hadoop Task Tracker

V. IMPLEMENTATION

1. Collect unstructured data from Social Media sources.
2. Real-Time Processing with a sentiment analysis engine based on keyword search.
3. Store processed data (with sentiment) in NoSQL database.
4. Extract sentiments from NoSQL to visualization layer.
5. Visualize with a tool of choice.

VI. RESULT AND ANALYSIS

A *Results Discussion*

Twitter is a widely used social platform used to post comments on various topics in the form of short status messages. In this thesis, tweets have been collected andconverted into a training set by using a python script. The tweets are collected by using hashtag, which are meant to judge the satisfaction and dissatisfaction level of customers with respect to the service provider. After the training set has been prepared, data is analyzed by uploading it on HDFS and Naïve Bayes classification is carried out.

The training set is thus converted into vectors using *tfidf weight (*term frequency x document frequency). This is done to assign weight to every term in the word listcreated from the setcould be done by calculating the Term frequency, which finds the frequency of each term in the document. The second aspect is Inverse Document Frequency which means that the lesser the presence of a term in all documents, the more is the term value or weight in this matter .

B *System Requirements*

AMD quad core 1.5GHZ processor
4GB RAM
500 GB hard disk
1000 Mbps Ethernet connection switch
Operating system-Ubuntu 14.01
Java version-1.7.0
Hadoop version-1.0.3
Mahoutversion-0.8

C *Analysis*

The analysis showed that around 17% of the opinions were negative, 83% were positive opinions about a well known person Narendra Modi in taken dataset. Fig 5 shows a pie graph of the overall opinions of the people.

VII.CONCLUSION AND FUTURE ENHANCEMENT

This project concludes that various classes have been analysed on tweets of well known person.It is also useful in guaging the opinions of people when it comes to diserve topics related to any fields. In our case study, we can further compare the services of various providers and judge which one is the best.

As future work, we can further compare the reviews of various persons and judge who is the best. Hadoop map-reduce and naïve algorithm, we can provide a simple automated method to evaluate what people information from social networks and analyzing it using Big Data techniques has left behind the traditional think.

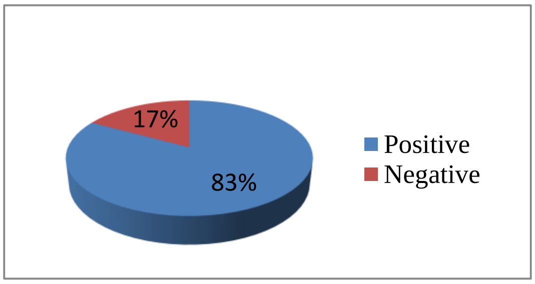

Fig 5: Overall Opinions for a data

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up Sentiment Classification using Machine Learning Techniques". In Proceedings of the Empirical Methods on Natural Language Processing, Pennsylvania, 2002, pp. 79-86.

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

[3] Y. Mejova, “Sentiment analysis: An overview,” Comprehensive exampaper,http://www.csuioedu/˜ymejova/publications/CompsYelena Mejova. Pdf [2010-03], 2009.

[4] Rui Xia, Chengqing Zong, Shoushan Li, “Ensemble of feature sets and classification algorithms for sentiment classification”, Information Sciences 181 (2011) 1138–1152.

[5] 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.

[6] P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss,” Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1997

[7] Z. Niu, Z. Yin, and X. Kong, “Sentiment classification for microblog by machine learning,” in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, 2012.

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

[9] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[10] 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.