IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

ANALYSIS OF SOCIAL DATA USING BIGDATA ANALYTICAL TOOLS

Dr. Vineet Richhariya[1], Mr. Jay Prakash Maurya[2], Sakshi Agrawal[3]

[1]Prof. & Head, Dept. of CSE, LNCT, Bhopal(M.P), [2]Prof. Dept. of CSE, LNCT, Bhopal(M.P)
[3]M.Tech(CSE),LNCT, Bhopal(M.P)

ABSTRACT

Social Networking Service (SNS), is a platform to provide social relations among individuals who share common interest. Twitter has become very popular. Millions of users post their comments on twitter; they specify their view on current affairs. Daily large amount of row data is available and which can be helpful for industrial or business purpose. Hence the twitter data can be analyzed and used for different businesses which will helpful for decision making. Twitter sites generates petabytes of data per day so to store and process such huge amount of data using traditional tools and technique is not suitable. In this paper gives a way of analysis of twitter data. To store, categories & process large sentiments we are using Hadoop an open source framework and to analyze the twitter data we uses bigdata analytical tools.

Keywords-- *Hadoop, sentiment analysis,social data, opinion mining, bigdata analytical tools.*

I. INTRODUCTION

We live in a society where the textual data on the Internet is growing at a rapid pace and many companies are trying to use this deluge of data to extract people’s views towards their products. Online social network platforms, with their large-scale repositories of user-generated content, can provide unique opportunities to gain insights into the emotional “pulse of the nation”, and indeed the global community. A great source of unstructured text information is included in social networks, where it is unfeasible to manually analyze such amounts of data. There is a large number of social networks websites that enable users to contribute, modify and grade the content, as well as to express their personal opinions about specific topics.[1] Some examples include blogs, forums, product reviews sites, and social networks, like Twitter (http://twitter.com/). Twitter (San Francisco, CA, USA) is a micro blogging site that offers the opportunity for the analysis of expressed mood, and previous studies have shown that geographical, diurnal, weekly, and seasonal patterns of positive and negative affect can be observed.

Micro blogging and more particularly Twitter is used for the following reasons:

- Micro blogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people’s opinions.
- Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.

- Twitter’s audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.
- Twitter’s audience is represented by users from many countries .

1.1 HADOOP

The Apache Hadoop[9] project develops open-source software for scalable, reliable, distributed computing. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using a thousands of computational independent computers and large amount (terabytes, petabytes) of data. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. Apache Hadoop is good choice for twitter analysis as it works for distributed huge data. Apache Hadoop is an open source framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different clusters nodes. In short, Hadoop framework is able enough to develop applications able of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data. Hadoop MapReduce is a software framework [8] for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

II. LITERATURE REVIEW

In [1] This paper author investigates the problem of predicting Twitter hashtags popularity level. A data set of more than 18 million tweets containing 748 thousand hashtags has been prepared by using Twitter’s rest API. Early adoption properties including profile of tweet authors and adoption time series are used to predict a tag’s later popularity level. The followers count and tweets count are two such characteristics related to adopters' profile. On the other hand, two types of frequency domain analyses are used to augment the simple mean and standard deviation characteristics of the adoption time series. Fourier transform (FT) spectrum and wavelet transform (WT) spectrum are considered in this study. Experimental results show that WT spectrum improves the prediction result of viral hashtags while FT spectrum does not.

Online social networks provide communication channels to spread an idea, behavior, style or usage throughout the world village. Twitter is a special online service that provides both social network and microblog functions. Posting tweets through devices from desktop to mobile is the main activity of the microblog function, while following and retweeting offer the social network function. Users post tweets by encoding topics in the form of hashtags, which are summarized by Twitter to make a list of current trending tags.

In [2], the author describes that Big data analytics has attracted intense interest from all academia and industry recently for its attempt to extract knowledge, information and wisdom form big data. Big data and cloud computing, two of the most important trends that are defining the new emerging analytical tools. Big data analytical capabilities using cloud delivery models could ease adoption for many industry, and most important thinking to cost saving, it could simplify useful insights that could providing them with different kinds of competitive advantage. Many companies to provide online Big Data analytical tools some of the top most companies like Amazon Big data Analytics Platform ,HIVE web based Interface, SAP Big data Analytics, IBM InfoSphere BigInsights, TERADATA Big Data Analytics, 1010data Big Data Platform, Cloudera Big Data

Solution etc. Those companies analyze huge amount of data with help of different type of tools and also provide easy or simple user interface for analyzing data.

Judith Sherin Tilsha S, Shobha M.S [4] (2015) A Survey on Twitter Data Analysis Techniques to Extract Public Opinion. Using machine learning algorithm ,a feature vector is constructed with the emotion describing words from tweets and are fed to the classifier that classifies the sentiment or opinion. It said that various twitter data analysis techniques that are based on dictionary and that are using the machine learning approaches.

Ramesh R, Divya G, Divya D, Merin K Kurian [5] (2015), Big Data Sentiment Analysis using Hadoop. The main focus of the research was to find such a technique that can efficiently perform Sentiment Analysis on Big Data sets. In this paper Sentiment Analysis was performed on a large data set of tweets using Hadoop and the performance of the technique was measured in form of speed and accuracy. The experimental result shows that the technique exhibits very good efficiency in handling big sentiment data sets.

G.Vinodhini , RM.Chandrasekaran [7] (2012), Sentiment Analysis and Opinion Mining: A Survey. An accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict online customer’s preferences, which could prove valuable for economic or marketing research. Till now, there are few different problems predominating in this research community, namely, sentiment classification, feature based classification and handling negations. This paper presents a survey covering the techniques and methods in sentiment analysis and challenges appear in the field.

III. OBSERVATION

Hadoop and bigdata analytical tools, for getting raw data from the Social Network, we may use Hadoop online streaming tool such as Apache Flume, apache kafka. By utilizing this tool only, we are going to configure everything, which we wanted to get (data) from the Social Network. Mainly we want to set the configuration model and also want to define what information that we want to collect form Social Network. All these will be stored into our HDFS (Hadoop Distributed File System) in our own prescribed format. From this unrefined data we are going to refined the data using analytical tools and than start analysing these social data to predict or to help in decision making.

IV. PROBLEM DEFINITION

The project focuses on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis. The tweets are important for analysis because data arrive at a high frequency and algorithms that process them must do so under very strict constraints of storage and time. Because these tweets generates the huge information related to different area like government, election, etc. millions of tweets is generated every day and which is very useful in decision making because every one is share their view and opinions on issues or variety of topics. The analysis of twitter data gives real view or different user opinions regarding what they think and to analysis these data provide a better way for making any decision.

V. PROPOSED WORK

For analysing these large and complex data required a power tool, we are using hadoop[6] which is a open source implementation of mapreduce, a powerful tool designed for deep analysis and transformation of very large data.

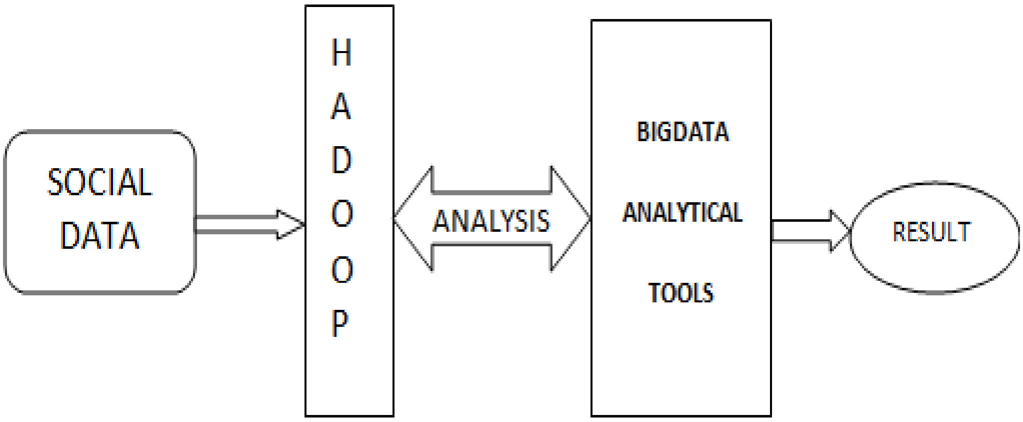

Figure1. Workflow Diagram

This paper we design algorithm for handling the problems raised by the larger data volume and the dynamic data characteristics for finding and performing operation on social media data sets. For analysing first we used standard platform as hadoop on single node ubuntu machine to solve the challenges of big data through MapReduce framework [12] where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling ,After that we can use bigdata analytical tools to refine such unstructured data and analyse the social data using bigdata analytical tools.

VI. PROPOSED METHODOLOGY

Our Steps or Algorithm Steps will follow*:*

First we can create a social account and than we can use social API's [3] for fetching real time social data and store it into HDFS.

1. For fetching social data we can use bigdata tools such as apache flume and kafka through which we can authenticate our keys and start fetching fetching data from social sites.
2. After fetching data, the data is store into HDFS (Hadoop Distributed File System), which is very reliable for storing such huge amount of data.
3. After storing data into HDFS, we can pre-process the data because from the social sites an unstructured raw data is coming, which is very difficult to analyze such kind of unstructured data, so we can first pre-process the data and convert it into some structure form.
4. After pre-processing we can start analyzing such huge amount of social data.

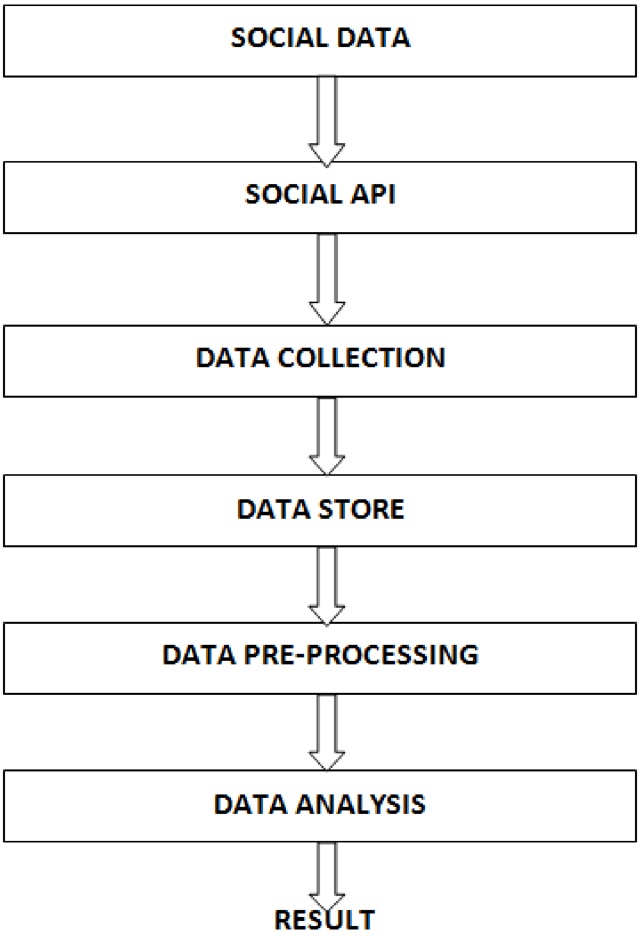

Figure 2. Analysis Steps

VIII. CONCLUSION

In this paper, we introduced bigdata analytical tools through which we can do analysis of social data. With the help of bigdata analytical tools on top of the hadoop we can easily handle the large amount of complex data.

REFERENCES

[1] Shing H. Doong, “Predicting Twitter Hashtags Popularity Level”, in 2016 49th Hawaii International Conference on System Sciences, IEEE, DOI 10.1109/HICSS.2016.247.

[2] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, “Big Data and Advanced Analytics Tools”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[3] "Twitter's API --- HowStuffWorks." *HowStuffWorks*. N.p., n.d. Web. 24 Oct. 2014.

[4] Judith Sherin Tilsha S , Shobha M S, “A Survey on Twitter Data Analysis Techniques to Extract Public Opinion”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 11, November 2015, pp 536-540.

[5] Ramesh R, Divya G, Divya D, Merin K Kurian , “Big Data Sentiment Analysis using Hadoop “, (IJIRST)International Journal for Innovative Research in Science & Technology,Volume 1 , Issue 11 , April 2015 ISSN : 2349-6010

[6] Praveen Kumar, Dr Vijay Singh Rathore,” Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014, pp 7123-7126.

[7] G.Vinodhini , RM.Chandrasekaran, “Sentiment Analysis and Opinion Mining: A Survey” , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012 ISSN: 2277 128X.

[8] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.

[9] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/