Study of Sentiment Analysis In Twitter Data Using Hadoop

Mr. Shailesh P. Thakare, **Mr. Sanjay V. Dhopte**

***Abstract*- Social media websites provides a platform to users to communicate with friends, family, and colleagues; also it gives them an easy and effective way to talk about their favorites. This unstructured data and communication is important for businesses; how customer understand their brand, and allow them to actively make business decisions to maintain their image. Twitter is a social networking platform which contains large amount of data that can be a structured, semi-structured and un-structured. This huge amount of raw data is used by businesses for deciding strategy according to consumers need and demand. In this paper presents the overview of the sentiment analysis and also discusses the steps of sentiment analysis. It also discusses the HADOOP for the analyzing the twitter data which is also known as a big data.**

***Keywords*- Social media, structured, semi-structured, un-structured data, HADOOP, big data, sentiment analysis**

I. INTRODUCTION

Due to heavy use and reliance on the internet of the businesses, peoples and many automation systems generates huge amount of data every day. In addition to that; teenagers, celebrities, and business uses social media to share their views, ideas and opinions. So it may be very easy for organizations to know about customers demand, opinion about something, and peoples view towards their product by extracting huge data on social media. Now a day's social media become a main source of data due to its popularity among large number of users. Social media contains huge amount of data about many things but it is very difficult to extract and analyze it. This is where the need of automatic categorization becomes apparent. Subjective data is analyzed generally in this case. There are a large number of social media websites that enable users to contribute, modify and grade the content. Users have an opportunity to express their personal opinions about specific topics. The example of such websites include blogs, forums, product reviews sites, and social networks. In this case, twitter data is used. Media like twitter contain prevalently short comments, like status messages on social networks like twitter. Additionally many web sites allow rating the popularity of the messages which can be related to the opinion expressed. 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 areas and groups.
- Twitter's audience is represented by users from many countries.

As the audience of micro blogging platforms and services grows every day, data from these sources can be used in opinion mining and sentiment analysis tasks.

II. SENTIMENT ANALYSIS

Sentiment analysis; also known as opinion mining; is process of computationally recognize and classifying the opinions which indicate the some portion of text, especially in order to decide the writer's or users attitude or opinion towards a particular product or thing. Sentiment Analysis is the process of discovering of the text, it determines whether a chunk of data is positive, negative, neutral or something else [2] [4].

Steps of Sentiments Analysis:

1) Collection of data

2) Preparation of data

3) Detection of sentiments

4) Classification of sentiment

5) Output

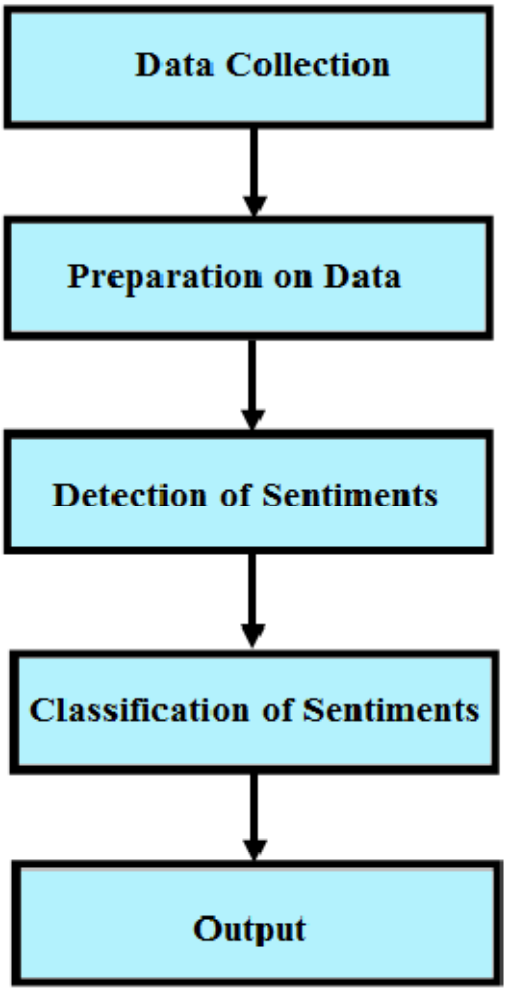

Figure 1: Steps of Sentiment analysis

III. TWITTER DATA SET

Twitter is a popular micro blogging site where users create status called tweets. Tweets are having a max limit of 150 characters in length. People post short messages, use various short forms, use emotions and other characters and symbols that express special meanings of the sentences. These tweets show someone's opinions about particular topic, thing or product. Emoticons are the facial expressions represented pictorially [1].The use hash tags to mark a particular topics is common. This increases the visibility of the tweet. Twitter messages prove be a big data source for classifying sentiment. The Twitter data is available for public access through streaming API. Streaming or media streaming is a method for transferring data so as to process it as a static and continuous stream. If a user has a Twitter account, then he or she can create an App on Twitter to collect the related tweets [3], [5].

IV. METHODS OF COLLECTING AND ANALYZING DATA

HADOOP is an open source framework which is freely available for everyone; which usually deals with processing of large data across clusters of commodity hardware. The Hadoop framework does the distribution of the data and various tasks of computation. These tasks are independent; hence the entire nodes may restart. It ideally provides linear scaling and is used for the design of cheap, commodity hardware. Its programming model is where in the end-user only has to write map-reduce tasks. Most of the data available from twitter for analysis is unstructured and the rest part is structured [6] [7].

Some of the major components in HADOOP are:

4.1 Hadoop Distributed File System

HDFS is block like file system where file is broken into fixed sized blocks and stored across a single node cluster of HDFS. It works very well with Hadoop Map-Reduce which allows data to be read and computed upon locally whenever possible. HDFS provides streaming read of the twitter data. As a result data has to be written to the HDFS once; it can then be read several numbers of times. It can be deployed on cluster with cheap commodity hardware [8][9].

Mapping: The Map function runs first and is mostly used for filtering, transforming, or parsing the data. The output of the Map is given as an input to the Reduce function.

Reduce: The Reduce function is mostly used for summarizing the data from the Map function.

Hive: The Apache Hive works along with HDFS; it includes the data in the database. The Hive query language is used in working with the data.

Dashboard: The data collected must be presented in graphical format. Hence the visualization tools like Tableau or excel sheets are used to make the data presentable.

V. CONCLUSION

Since use of social media is very common and popular among people now a days it is very easy for organizations to understand their demands and needs, by exploring social media data where peoples share their opinions and views about something. Sentiment analysis may help organizations to serve customers better if they make use of it. Twitter is very popular social media site where one can found huge amount of text data about people's opinions and views which can be used to develop applications related with sentiment analysis. Since HADDOP is open source framework and easily available, processing of huge amount of data collected from twitter make it easy and simple. If researchers and developers make use of HADOOP and twitter data then it may be possible to develop accurate and effective sentiment analysis applications.

REFERENCES

1. Dmitry Davidov, Oren Tsur& Ari appoport "Enhanced SentimentLearning Using Twitter Hashtags and Smileys," Coling 2010: Poster VolAugust 2010, pp 241–249, Beijing.

2. Luciano Barbosa, "Robust Sentiment Detection on Twitter from Biased and Noisy Data" Coling 2010: Poster Vol, August 2010pp 36–44,Beijing.

3. Anna Jurek, Yaxin Bi, Maurice Mulvenna" Twitter Sentiment Analysis for Security-Related Information Gathering", 2014 IEEE Joint Intelligence and Security Informatics Conference pp 48-55

4. Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, and Xiaofei He" Interpreting the Public Sentiment Variations on Twitter" IEEE Transactions On Knowledge And Data Engineering, VOL. 26, NO. 5, pp 1158-1169, MAY 2014.

5. Nishad patil, Sayalitingre, Kalyani thorat, Swapnil shivshetty "A Survey Paper On Twitter Sentiment Analysis Using portat Stemming Algorithm" international journal of science and research VOL 4 Issue 10, pp 1707-1708 OCTOBER

6. Apporv Agarwal, Jasneet Singh Sabarwal, "End to End Sentiment Analysis of Twitter Data".

7. Theresa Wilson, Joanna Moore, Efthymios Kouloumpis,"Twitter Sentiment Analysis-The Good, the Bad and the OMG".

8. Apoorv Agarwal, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data".

9. Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis".

10. Andrea Esuli, Fabrizio Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining" .