Sentiment Analysis in Twitter Data using Hadoop: Review

Ms. Rupali A.Raut[1], Mr. Chetan A. Mugal[2], Ms. Pallavi H.Dhole[3]

[1,3]*Department of Computer Science & Technology,* [2]*Department of Computer Engineering,*
Amravati University,

Abstract - Social media gives users a platform to communicate effectively with friends, family, and colleagues, and also gives them a platform to talk about their favourites. 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 data. This huge number of raw data is used for business purpose by making the business strategy according to consumer need and processing. In this paper gives the overview of the sentiment analysis and also discuss the steps of sentiment analysis. This paper also discuss the HADOOP for the analysing 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

Day by day, the growth of textual data on the internet tremendously. Different industries are trying to use this huge textual data for extracting the people's views towards their products. Today social network media is a main source of data or information. It is very difficult to analyse the large amount of data. This is where the need of automatic categorization becomes apparent. Subjective data is analysed 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 social and interests 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. The Sentiment analysis process of computationally recognize and classifying the opinions indicate the some portion of text, especially in order to decide the writer's attitude towards a particular product. Sentiment Analysis is the process of discovering of the text, it determines whether a piece of writing is positive, negative or neutral [2][4].

Steps of Sentiments Analysis or Opinion Mining

1) Data Collection
2) Preparation on Data
3) Detection of sentiments
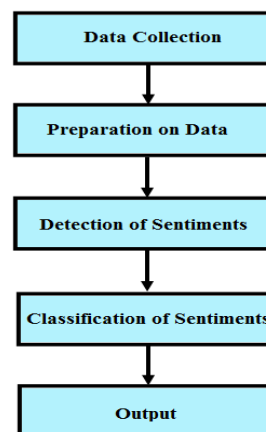4) Classification of sentiment
5) Output

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 emoticons and other characters that express special meanings of the sentences. These tweets present opinions about different topics. 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. HADOOP

HADOOP is an open source framework which is freely available for everyone. Hadoop is a framework 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. The 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].

Following are the some major components in HADOOP:

A. Hadoop Distributed File System - HDFS is block like file system. Here file is broken into fixed sized blocks and stored across a single node cluster of HDFS. HDFS 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].

B. 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.

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

D. Hive - The Apache Hive works along with HDFS. The Hive includes the data in the database. The Hive query language is used in working with the data.

E. 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

Sentiment analysis is a very wide branch for research. In this paper covered some of the important aspects. In this paper gives the overview of the sentiment analysis and also discuss the steps of sentiment analysis. This paper also discuss the HADOOP for the analysing the twitter data which is also known as a big data.

VI. REFERENCES

[1]. Dmitry Davidov, Oren Tsur& Ari appoport "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," Coling 2010: Poster Vol August   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 2015

[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"