Improvement in Sentiment Analysis of Twitter Data Using Hadoop.

Aishwarya Kotwal¹, Priyanka Fulari², Dipali Jadhav³ & Ratika Kad⁴
¹,²,³ Student, Sinhgad Academy of Engineering, Pune, India
⁴ Asst. Professor, Sinhgad Academy of engineering, Pune, India

Abstract – Twitter is a blogging site where people post various comments and views related to various products. Thus the opinions of the audiences can be captured from here and can be used to extract the sentiments of the people. Thus the sentiments can be analyzed and used to improve the product. This will be beneficial to the companies and other organizations to improvise or make the changes needed in their products and thus ultimately increase the profit. This is a part of business intelligence.

Goals:
1. Stream tweets from twitter.
2. Process the tweets
3. Analyse the tweets

Keywords - Big Data, Hadoop, HDFS, Map Reduce, Sentiment data, Twitter.

NOMENCLATURE
HDFS - Hadoop Distributed File System
NLP- Natural Language Processing
SVM - Support Vector Machines
JSON - Java Scrip Object Notation
API- Application Programming Interface

I. INTRODUCTION

Twitter helps people to tweet their view. This Twitter data is one of the best places to have a deep study and analyze data. Twitter data always contains important information for a wide variety of applications. Thus scientists interpret the human feelings. The businessmen can improvise the sale of a product if they are familiar with the statistics of how customers responses to it.

Sentiment Analysis is a vital Natural Language Processing field that comprises of obtaining the writer’s sentiments regarding range of tweets on products which are written on the twitter via various posts or comments. Thus Sentiment Analysis helps to determine the reactions of a user or a group of users on a topic and categorizing their opinion as positive, negative or neutral.

It can help producers in deciding the strategies for the launch of new products based on the responses of the previous versions of that product in various localities. It can also be used to detect impassioned or intense comments and use of slang language or spam detection. Sentimental Analysis can be phrase based where the sentiment of a single phrase is taken in to consideration, sentence based where the sentiment of the whole sentence is calculated.

Sentiment Analysis is generally carried out in following steps.

1. Initially analyze the subject of the sentiment.
2. Working on polarity of the sentiment
3. Eventually assign the degree of the polarity is assigned along with the help of a sentiment score

Twitter produces extremely large amount of data every day in the form of tweets. This data is mainly unstructured or structured and is termed as big data.

Thus sentiment data hence is now referred to as big data and needs advanced technology that have the capability to deal with such large amounts of data efficiently. There come various challenges while dealing with this high volume of data like processing of large data sets, extraction of useful information from online generated data sets etc. The term Big Data is globally used for collection of Datasets that are huge and complex, these huge Datasets makes it difficult to process by adopting traditional way of data processing techniques. The challenges related to Big Data provide a chance to understand the data patterns and helps in prediction of events and results. Hadoop is of the processing tools that can provide best solution to analyze and process large data sets.

II. LITERATURE SURVEY

A. EXISTING SYSTEM

Barbosa and Feng put forth that use n-grams on tweet data
may stop the classification perform as the large number of infrequent words in Twitter. Agarwal et al. Proposed the POS features, and the lexicon features along with micro blogging features. Not just combining various features, but also designed some tree representation of tweets to inculcate many categories of features in just one succinct representation.

B. WORKING

A. The Sentiment analysis process

i) Collect twitter data
ii) Prepare text
iii) Find the sentiments
iv) Analyze the sentiment
v) Output

i) Collect twitter data:
The first step in sentiment analysis involves collection of data from user. These data are disorganized, expressed in different ways by using different vocabularies, slangs, context of writing etc. Manual analysis is almost impossible. Therefore, text analytics and natural language processing are used to extract and classify [11].

ii) Prepare text:
This step involves cleaning of the extracted data before analyzing it. Here non-textual and irrelevant content for the analysis are identified and discarded

iii) Find the sentiments:
All the extracted sentences of the views and opinions are studied. From this sentences with subjective expressions which involves opinions, beliefs and view are retained back whereas sentences with objective communication i.e facts, factual information are discarded.

iv) Analyze the sentiment:
Here, subjective sentences are classified as positive, negative, or good, bad or like, dislike [1]

v) Output: The main objective of sentiment analysis is to convert unstructured text into meaningful data. When the analysis is finished, the text results are displayed on graphs in the form of pie chart, bar chart and line graphs. Also time can be analyzed and can be graphically displayed constructing a sentiment time line with the chosen value (frequency, percentages, and averages) over time automatically as possible.

III. PROPOSED SYSTEM

PROCCES FOLLOWED

Creating twitter API:
We have created a twitter API. Twitter provides access to tweets using this API.

Collecting the twitter data:
The data related to particular keywords is fetched using flume. It is a channel through which data will flow an collected into HDFS.

Refining the tweets:
We have taken only the fields which we require and not all fields

Analysing using hive:
Using hive we have analysed the data

Presenting:
The analysed data is presented in a understandable format using dashboard.

IV. METHODOLOGY

Flume:
Apache Flume is used in the streaming of large amounts of twitter data. This streamed real-time data from Twitter and is then stored in the Hadoop Distributed File System (HDFS). The Flume sends tweets from the Twitter data stream to the database.

HDFS (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[15] 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

**Map:**
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. The Hive 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 [19] or excel sheets are used to make the data presentable.

### V. FUTURE SCOPE

There are various ways to obtain Twitter data other online streaming data where they want to code lines of coding to achieve this. And, also they want to perform the sentiment analysis on the stored data where it makes some complex to perform those operations. Coming to this paper we have achieved by this problem statement and solving it in BIGDATA by using Hadoop and its Eco Systems. And finally we have done sentiment analysis on the Twitter data that is stored in HDFS. So, here the processing time taken is also very less compared to the previous methods because Hadoop Map Reduce and Hive are the best methods to process large amount of data in a small time.

### VI. CONCLUSION

The Analysis of Twitter Sentiment reduces the latency and price by using the Big Data technology called as ‘HADOOP’.This system is very cost efficient and secure. So Apache, the Open Source Product provides the platform that can have Twitter Analysis by using HADOOP. The speed of the system will be increased.

### VII. REFERENCES


[5] “Twitter as a Corpus for Sentiment Analysis and Opinion Mining