Real Time Big Data Analytics Using Sentimental Analysis

[1]Prof.Akshay Agarwal, [2]Mr.Akshay Karangutkar [3]Mr.Vallabh Mahajan, [4]Mr.Hitesh Pargi

[1] Asst.Professor,[2,3,4]UG Student,[1,2,3,4]Computer Engineering Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

[1]akshay1661@gmail.com,[2]akshaykarangutkar4@gmail.com,[3]vallabhmahajan1@gmail.com,

[4]hiteshpargi23@gmail.com

Abstract- People go for online social media as it is easy to convey their opinions, to have an up-to-date knowledge about the ongoing trends on a daily basis. Twitter is the biggest and renowned online social media gets a large number of tweets each day on various topics. This huge volume of raw information can be used for Social, Industrial, Economic, or Government approaches by arranging as per our need and processing. Hadoop is the best tool for analyzing the twitter data as it processes the huge sets of data in parallel [1]. Since twitter contains a variety of opinions on various topics it is necessary to analyze these opinions to know the customer behavior [3]. The pig scripts are written to extract tweet from the raw nested twitter data.

Keywords-Hadoop, Big Data, HDFS.

I. INTRODUCTION

Social media has achieved a lot of light in the few years. Twitter is an important platform which people are taking up to prompt their views and opinions about any topic. Twitter is the social media sites to give an opportunity to peoples to express the ideas and opinions about the particular topic for modification of a models of any particular topic and has results large data sets.

If the data is of small size then it is easy to extract the useful information, but if the size of data is huge then it is quite difficult to analyze what that data actually intends. [3] All the traditional big data technique proposed earlier have gradually failed performing effectively with increase in dataset size. Recently the powerful tool that has proved to be efficient in processing the large sets of data is Hadoop, which is considered to be efficient for distributed processing as well as distributed storage of large sets of data.

II. AIMS AND OBJECTIVE

The major function of the proposed system is to get the opinion of people on any product, current trending topic and political views etc. from social media. User will categorize tweets in 3 types:

1. Positive Tweets.
2. Negative Tweets.
3. Neutral Tweets.

Basically, User will fetch tweets and for analysing our tweets it will use Hadoop ecosystem. This system will be analysing the trending hashtags and will generate a report based on analysis. Based on the categorization of positive, neutral and negative tweets of one Twitter account/hashtag, User can compare it with other more than one Twitter account/hashtag. Thus, User can compare more than one Twitter account/hashtag and analyse its sentiments.

III. LITERATURE SURVEY

Sentimental analysis is implemented by many researchers to support various types of platforms. Most of them are customized for particular platform [2].

3.1 Large-Scale Machine Learning at Twitter.

In this paper there is a study of Twitter's amalgamation of tools into its Apache Hadoop-based and Apache Pig-centric analytics platform. It provides a base-line for classification accuracy from content, given only data with large amounts.

3.2 Towards Large-Scale Twitter Mining for Drug-Related Adverse Events

This paper defines a method to search drug users and possible adverse events by evaluating twitter tweets utilizing NLP and to create SVM classifiers.

3.3 Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier

Technology of machine learning are mostly used in sentiment organization since their capability to "learn" from the dataset to forecast or support choice making with comparatively high accuracy.

However, once the dataset is big, some algorithms might not scale up well.

3.4 A Framework for Massive Twitter Data Extraction and Analysis

In this paper the abilities of this platform are illustrated with two study cases in Spanish, one related to a high impact event (the Boston Terror Attack), and another one related to regular political activity on Twitter. The first case study involves the activity on Twitter around a high impact event, the Boston Terror Attacks.

In this case, they tracked a hash tag. The second case study was focused on regular Twitter usage, tracking the activity around well-known Spanish political actors, i.e politicians, political parties, journalists and activist organizations as well.

3.5 Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction

In this paper there is a prospect of making calculation of stock market basing on sorting of data approaching from micro blogging Twitter platform. Calculations were made for Apple Inc. in order to confirm that sufficiently large datasets would be retrieved Only tweets in English are used in this research work.

IV. EXISTING SYSTEM

There is a real demand for instant techniques for fetching and processing actual time tweets and also for performing sentiment analysis in a fault-tolerant manner. The proposed architecture uses the Hadoop framework to handle the streaming real-time tweets.

The Flume component communicates with the streaming API of twitter and fetches the tweets that match the keywords. The fetched tweets are highly nested and highly unstructured that is stored in HDFS in JSON format. These stored tweets in HDFS are then loaded Apache PIG, where further processing and summarization of sentiment takes place.

In order to determine the opinion of the given tweet, the tweets are loaded into one of Hadoop supported component known as Apache PIG, where in these tweets are passed through a series of PIG scripts for sentiment classification. The tweets are either positive or negative. Following are the approaches to detecting sentiment in tweets:

1) Loading the twitter data

2) Extracting the Tweets

3) Tokenizing the tweets

4) Loading the Dictionary

5) Rating the tweets

6) Classification of tweets

V. COMPARTIVE STUDY

Table no 5.1: Comparative analysis

Sr no.	TITLE	METHODOLOGY	REMARKS
1	Large scale machine learning at twitter	• Simple logistics regression classifier • Hashed byte 4-grams as features	Polarity sorting tests displayed accuracy in the range 77% to 82% with varying data size set.
2	Towards large-scale twitter mining for drug connected adverse events.	• Define a method to search drug users and possible adverse actions by analysing the content of twitter message.	The predication exactness on average over 100 repetitions was gauged to 0.74 and the mean AUC value is 0.82
3	Scalable Sentiment Classification for Big Data Analysis using Naïve Bayes Classifier	• Implemented NBC to accomplish fine grain control of the analysis technique for Hadoop implementation.	Resulted in an 80.85% avg accuracy
4	"A Framework for Massive Twitter Data Extaction and Analysis	• Tracking activity around well- known Spanish political actors. • The framework is executed in python, but classifier and Tester run on NodeJs	The deduction is that the best trainers has 1 g included and at least score between 2 and 4
5	Sentiment analysis of Twitter data within big data distributed environment for stock prediction	• Discusses Stock market prediction. • Naïve Bayes technique was chosen employing SentiWordNet.	Consider large value volumes of data caused in decision to put on a map reduce version of Naïve Bayes algorithm

VI. PROBLEM STATEMENT

The main drawback of the technique described is that, this method cannot be used for real-time analytics. Hadoop was initially established for batch processing. The actual thought of MapReduce is engaged towards batch and not period of time. But to be honest, this was solely the case at Hadoop's starting, and currently you've have got many opportunities to use Apache Hadoop in a very additional period of time method.

VII. PROPOSED SYSTEM

In Our experimental study, user retrieves the Twitter data (as twitter is a very important source of data) for sentimental analysis with Apache Hadoop Ecosystem. The generalized method of performing sentimental analysis is explained in following flowchart. This system uses Apache Spark for real time analytics. Apache Spark also helps to achieve faster processing as it uses in-memory computation. The workflow for our system will be as follows:

Import the tweets to Hadoop HDFS (Hadoop Distributed File System) using Twitter API via Apache Flume.

After that, the system will load the tweets in Apache Hive/Pig along with a sentiment analysis dictionary to give rating to each word in the tweets. Once ratings are given to each word in the tweets, system will apply following machine learning algorithms to classify the tweets in positive, negative and neutral form.

K- means clustering.

Support vector machine.

Naïve Bayes algorithm.

As all the 3 algorithms can't give the same results of classification, the system will select the optimum algorithm for performing sentiment analysis.

VIII. ALGORITHM

1. K means clustering::

Define K (number of clusters).

Define the Distance Measure.

Select Initial Seeds.

Execute Algorithm.

Check the Output for

Cluster Contribution

R-squared (Overall and for all the variables)

RMSSTD for each of the Cluster

Repeat the procedure for different K if criterion not satisfied

2. Support vector machine::

Prepare a dataset of tokenised tweets.

Normalize the dataset.

Select activating function.

Apply SVM algorithm on dataset.

Perform cross-validations.

Train the SVM model.

Test the SVM model.

Evaluate model performance.

IX. MATHEMATICAL MODEL

The Naive Bayes Classification denotes a supervised learning technique and statistical method for classification. It is model and it give license us to capture ambiguity about the model in a way by defining probabilities. It helps to resolve diagnostic and analytical problems. It helps to provide a beneficial perspective for empathetic and also evaluating many learning algorithms.

$$P(C|X) = \frac{P(X|C).p(C)}{P(X)}$$

P (C | X) is posterior probability,

P (X | C) is likelihood,

P(C) is class prior probability,

P(X) is predictor prior probability.

X. SYSTEM ARCHITECTURE

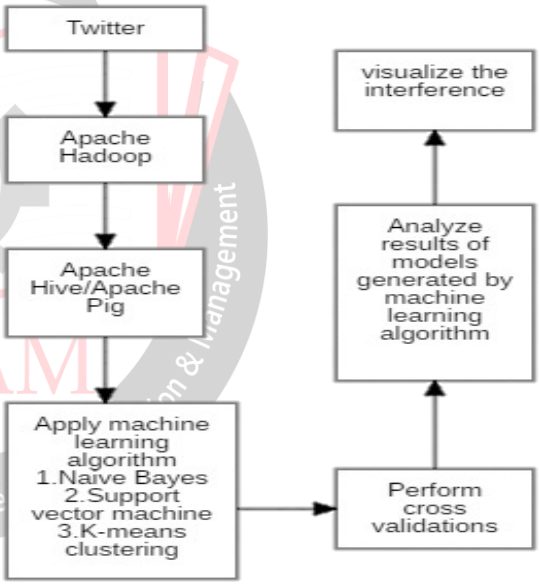

Fig.1: System Architecture

1. Data Collection:

Big data is nothing but a lot of data. Alone, small amount of data can't give you much perception. But petabytes of data, joint together with complex mathematical models and great computing power, can create inference that humans aren't capable of producing. Twitter is a source of data. Unlike additional social platforms, nearly each user's tweets are totally public and pull able. As user require real time analytics of tweets, there should be constant flow of tweets in our HDFS.

2. Perform Sentimental Analysis:

Apache Flume is used to import tweets from twitter API to HDFS. The output from Apache Flume is in JSON format. But Apache Hive does not understand JSON format. A custom SerDe can be written by user which recites the JSON data in and decodes the objects for Hive. After converting JSON data into readable format, user will make use of spell check algorithm to update spellings of each and every word according to grammar. Using Apache Hive, assign polarities to tweets and categorize them into positive, neutral and negative. The tokenization for the tweets will be done in Apache Hive. Apache Hive uses sentiment analysis dictionary like sentiword. As soon as the tokenization is done, we use machine learning algorithms like Naïve Bayes, K-means clustering, SVM to give polarities to the tweets.

3. Data Visualization:

The c.s.v/t.s.v file containing the classification of positive, neutral and negative tweets will be used by data visualization team to get some meaningful information from the data. Graphical representation of information and data is called Data visualization. Data visualization is one more form of visual art that grasps our attention and keeps our eyes on the message. When user see a chart, user quickly see trends and outliers. If user can see something, user internalize it quickly. It's storytelling with a purpose. If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be. As the “age of Big Data” kicks into high-gear, visualization is an increasingly key tool to make sense of the trillions of data generated every day. The plainest graph could be too boring to catch any notice or it make tell a powerful point; the most stunning visualization could utterly fail at conveying the right message or it could speak volumes. The data and the visuals need to work together, and there's an art to combining great analysis with great storytelling.

XI. ADVANTAGES

1) Adjust marketing strategy.
2) Measure ROI of your marketing campaign.
3) Develop product quality
4) Improve customer service
5) Crisis management
6) Lead generation

XI. DESIGN DETAILS

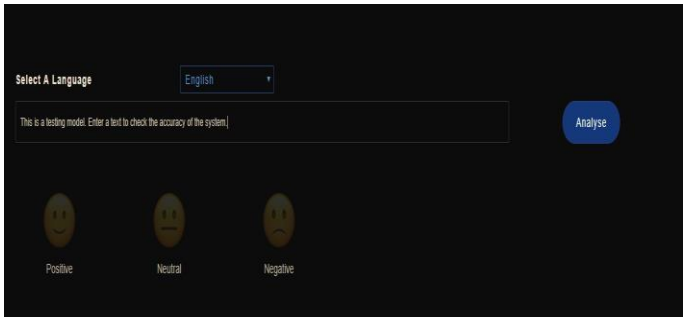

Fig.2: GUI

XII. CONCLUSION

Thus, we have tried to implement paper Rashid Kamal, Munam Ali Shah, Asad Manif, J Ahmad “Real-time Opinion Mining of Twitter Data using Spring XD and Hadoop,2017.” Thus, an application to make sentiment analysis is implemented. It can classify tweets and positive, neutral and negative sentiments.

REFERENCES

[1] Anisha P Rodrigues, Archana Rao, Niranjan N Chiplunkar “Sentiment Analysis of Real Time Twitter data using Big data Approach”.
[2] Monu Kumar, Dr. Anju Bala “Analyzing Twitter Sentiments Through Big Data”.
[3] Sunny Kumar, Paramjeet Singh, Shaveta Rani “Sentimental Analysis of Social Media Using RLanguage and Hadoop: Rhadoop”.
[4] Changhua Yang, Kevin Hsin-Yih Lin and Hsin-Hsi Chen 2007” Emotion classification using web blog corpora”.
[5] Panda, Ganapati, and Babita Majhi. "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements."
[6] Khushboo R. Shrote, Prof. A.V.Deorankar, “Review Based Service Recommendation for Big Data”.Advances in Electrical, Electronics,Information, Communication and BioInformatics (2016):470-474.
[7] Kusum Yadav, Manjusha Pandey, Siddharth Swarup Rautaray, “Feedback Analysis Using Big Data Tools”, Business Industry and Government (ICTBIG) (2016):1-5.
[8] Tare, Mohit, et al. "Multi-class tweet categorization using map reduce paradigm." International Journal of Computer Trends and Technology (IJCTT) 9.2 (2014): 78-81.
[9] https://www.tutorialspoint.com/apache_flume/fetching_twitter_ data.html