ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Sentimental Analysis of Twitter User using big data and Hadoop

[1] Hemant J. Kamble, [2] Jyoti M. Ingale, [3] Bhagyashri R. Posture, [4] Ganesh S. Ghuge
Department of Computer Engg

Abstract: **- With rapid growth in user of social media data for sentiment analysis of people. Handling such a huge amount of unstructured data is a tedious task so we are using Hadoop Technology. Sentiment Analysis is nothing but Opinion Mining which is used to specify the Nature of particular person by Analyzing its Timeline on Twitter. The result is categorized in the form of Positive, Negative and Neutral by the Decision Dictionary. We came on the result by the comment and retweets of the user. We are using Naïve Based Classification Algorithm for classification of data. In this paper, we propose a method that does sentiment analysis on twitter dada using Hadoop and big data.**

***Keywords*: Big data; Hadoop; Naïve Based Classification Algorithm; Sentiment analysis; Twitter.**

I. INTRODUCTION

Now a day, there are millions of users on Social Media who are active daily and generate huge amount of data daily (In TB's and PB's). Therefore, it is difficult to store and manage that unstructured data hence now it has become a tedious task to do. We are going to use Hadoop technology to maintain this huge amount of data. This data will be stored in HDFS (Hadoop Distributed File System) format. Hadoop is a platform, which is use to store distributed and computational data. We are using Hadoop technology for Sentiments Analysis of particular user using his/ her data from social media account. Here we are using Twitter's data for analysis because Twitter has become an important platform, which people are taking up to express their views and opinions about any topic [1]. There are 328 Million active users on twitter in 2017 till now with generated 500 million tweets per day, which is very huge amount of data. Now we will do the Sentiments Analysis of Twitter User with the help of his/her Twitter data using Hadoop Technology.

Twitter data is in the form of huge data i.e. unstructured data. This data is nothing but big data. High volume, high variety and high velocity these terms are come in big data. Data can be in the form of Structured, Semi structured or Unstructured. In this project we are working on unstructured data, which is generated by particular user on their own timeline on twitter. There are different sources of Big data such as Social Media, Log files, Sensor data, Docs, Business Apps, etc.

Social Media: In this high volume and high velocity, data are generated at every day. This data is in the form of comments reactions, emotions, etc. This data can be use to detect trends, analyze the sentiment of particular brand or person.

Docs: It uses only archived data. Docs don't use APIs.

Media: In this data is in the structured format and use APIs.

Sensor data: High volume, variety and velocity of data can be generated. This data is nothing but Temperature, Geolocation, Noise, Engagement, Biometrics and many more. This data is generated by different organization.

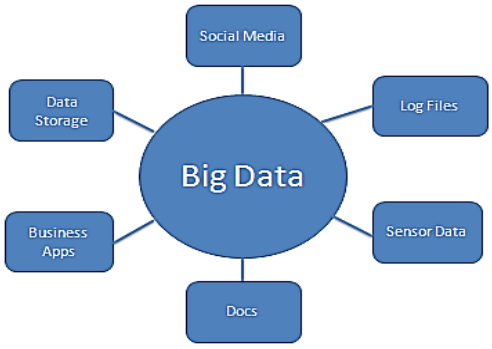


Figure: Sources of Big Data

So, There are many sources of big data as shown in figure, we have chosen the Social Media for Sentiment Analysis of user. Currently Facebook, Twitter, LinkedIn, etc. provides platform to post their view publicly. We choose Twitter because it has only 280 characters for tweet. Among all the social media twitter is widely use for expressing view on certain topic or particular post.

There are different classification algorithm, which is use for classification purpose such as Naïve Bayes Classifier

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Algorithm, K means clustering algorithm, SVM (Support Vector Machine Algorithm), Decision Trees, etc.

We are using Naïve Bayes Classifier Algorithm for classification of particular user by their post, comment or reaction. This algorithm is fast to predict the resultant data as compare to other classification algorithm. This algorithm mostly use where large dataset is present. If the input variables are in well structured then this algorithm is perform well. This algorithm is use for Sentiment Analysis, Document Categorization and Email Spam Filtering.

The paper is organized as follows. Section II is Related Work. Section III is Methodology, which gives brief information of our proposed architecture and algorithm using Hadoop. Section IV is designing section, which shows overall design of our project. In the last section, we are concluding the project.

II. RELATED WORK

A lot work has been done in the field of sentiment analysis using Hadoop. Here are various ways to gather the user's data and do the sentiment analysis on it.

Monu et al [1] in this paper they have presented an approach in which users preferences based on various products where analyzed using the Hadoop technology. In this paper, they have cleared that data from social networking sites can be used for number of purposes so they have used it for sentiment analysis. The millions of tweets receive every year put the subjected to sentiment analysis using Hadoop on cloud environment. In this paper they have propose a method that does sentiment analysis on tweets in cloud environment.

Anurag et al [2] they have proposed a paper in which they have said that there is a rapid growth in social media in recent years, the researcher get attracted towards the use of social media data for sentiment analysis of people or particular product or person or event. For sentiment analyses, they have used Data Mining classifiers and they have used k-nearest neighbor classifier, which gives very high predictive accuracy. So there is no need to use ensemble of classifiers for sentiment predictions of tweets.

Neethu et al [3] they have propose a approach that Twitter is generating a vast amount of data daily. Here they have done the analysis of electronics products posts using the Machine Learning approach. They have present the new feature vector for classifying the tweets in terms of Positive or Negative to extract the opinion about product.

III. METHODOLOGY

This paper presents a mechanism to predict overall sentiments of particular user by analyzing their timeline. Figure 1 shows architectural diagram of our system. Raw training tweets are collecting by Twitter API. After that, various methods are used to clean that raw data. For collecting raw tweets and data, same methods are used and generate training datasets. After that, various classification algorithms are used.

A. Architecture

The data and tweets collected from Twitter API. This data is stored in HDFS. Some training dataset were generated by applying Naïve Bayes Classifier Algorithm and Decision Tree.

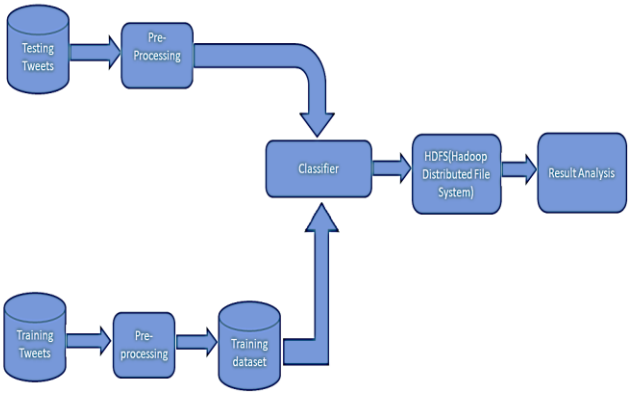


Figure 1: Hadoop Architecture

There are some tweets, which are not in proper format for this; we used various methods for cleaning the tweets. All special characters as well as common verbs removed from the tweets. Duplicate tweets also get removed. In training dataset, there is no any retweet present.

B. Classifier Algorithm

There are various algorithms which can be use for sentiment analysis. Here are two algorithms.

I.Decision Tree Algorithm:-

The Decision Tree Algorithms are represented in the leaves. It is one of the predictive modelling approaches. It is used of data mining and machine learning. There models where the target variable. They can take a discrete set of value are called as classification. In this tree structure are leaves represent class labels and tree branches represented conjunction of features that those class labels. There is two-type method in classification and regression tree. In regression, tress is similar to the classification tree. It is analysis of when the predicted outcome can be consider real number and classification tree analysis is when the predicted outcome is the class to which the data belong. In some cases tree do not work very well smooth boundaries they are not best. It is unstable and high variance.

II. K Nearest Algorithm:-

It is the very simple algorithm. In this algorithm they use a database that stores all available cases and classifies new

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

cases based on a similarity measure. This algorithm referred as a non-parametric machine learning algorithm. It makes number of assumption about the functional problem which are get solved. The non- parametric method is used for classification and regression. In K Nearest Neighbour algorithm classification of the output is generates a class membership and In this object its classified by majority vote of neighbours object, which will be assigned to class of most common K Nearest Neighbour and k is positive integer. It is small and the output of K Nearest Neighbour regression is the property value for the object. This value is average value of K Nearest Neighbour. In this algorithm need to determine value of parameter of K (It is Nearest Neighbour) and distance based learning is not clear, reduce the computational cost.

III. Naïve Bayes Algorithm:-

This is a learning and classification method of data. The name of algorithm came from Thomas Bayes, Who proposed Bayes theorem. The Naive based studied frequently since 1950. It is also known as text retrieval community. It is popular method for text classification, categorization, learning combine data.it is solved by judging the document to belonging one category or the other (like spam or legimate,sports or politics) with word frequency as a feature. Naive based suited for data when dimension of input is high. It uses the method of maximum like hood. In spite over simplified assumptions. It often performs better in many complex real world situations only it required small amount of training data to estimate parameter. Bayes theorem is not a single algorithm but it is collection of algorithm that follows a same rule to classification on every feature, which should classified independently on that basis of any other feature commonly useful in spam detection and document classification. It is simple to understand and build.

It is easily trained even with a small dataset. It is fast! It is not sensitive to irrelevant features. It assumes every feature is independent, so does not mean that each feature is new case.

Consider a class variable y and a dependant feature x1 through an

So the Bayes theorem state that,

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Bye using naive assumption we can say that

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

For I the relation in simplified form

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

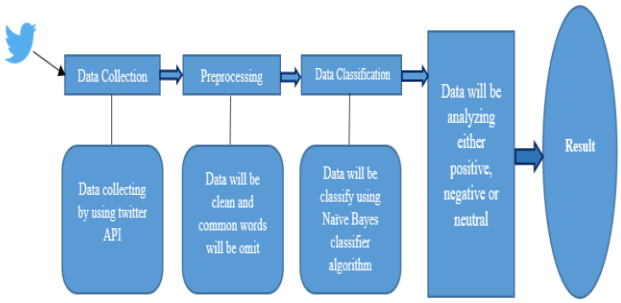
$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Hence, we can use maximum possibilities to estimate p(y) and p (x1\y)

So the class y is training set.

We have use Naïve Bayes algorithm among of these three algorithms. We chosen the Naïve Bayes Algorithm because, It is faster than other algorithms. It is specially use for the big amount of data. It has some probability method for classifying the data. It gets the output more accurately as well as more faster. It classifies the data or the input on the basis of tested data. It's specially used for input data which have high dimension.

IV. DESIGN



Data Collection: We collects the data or the input of twitter.
Preprocessing: Omits the common words or verbs like is, are, now, its etc. and remains data proceed.
Data Classification: It classifies the data with Naïve Bayes Algorithm and analyze it in Positive, Negative or Neutral.
Then it stores the data and shows the results directly.

V. CONCLUSION

We have proposed a approach, that it's easy and better to do the sentiments analysis of twitter data using Hadoop and big data with Naïve Bayes Algorithm. We can now do the sentiment analysis of a particular person based on their twitter data, in terms of Positive, Negative or Neutral. With help of Hadoop we can store huge amount of data and sorted it easily. We can use this technique (Sentiment Analysis) for future use or analysis of other things like

electronic products, or any trending posts etc. and we can do it on any other data.

REFERENCES

1. Monu Kumar And Dr. Anju Bala,"Analyzing Twitter Sentiments Through Big Data," IEEE Trans. on 2016 International Conference on Computing for Sustainable Global Development (INDIACom).

2. A. P. Jain and V. D. Katkar, "Sentiments analysis of Twitter data using data mining," 2015 International Conference on Information Processing (ICIP), Pune, 2015, pp. 807-810

3. M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-5.

4. Beiming Sun, Vincent TY Ng, "Analyzing Sentimental influence of Posts on Social Networks", Proceeding of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design.

5. Ye Wu, Fuji Ren, "Learning Sentimental Influence In Twitter", International Conference on Future Computer Sciences and Application, 2011.