Use of Hadoop Framework for Web Based Sentiment Analysis

Akshay R. Kalambate[1] Mayur R. Mane[2] Zilu Rane[3] Prof. Pralhad S. Gamare[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Mumbai University, RMCET Ambav, Maharashtra 415804, India

Abstract—Current era is of social networking sites, petabytes of data is generated daily on web. Millions of people are posting their likes, dislikes, comments daily on social networking sites. In this system, we are proposing a model that will extract the sentiment from a famous micro blogging site, Twitter, where users post their opinions for everything. Twitter is an online web application which contains lots of data that can be a structured or semi-structured or un-structured data format. We can collect the data from the twitter by using Apache Hadoop(BIG DATA) eco-system using online streaming tool Flume. There are different types of analysis that can be done on the collected data. So here we are taking sentiment analysis, for this we are choose to use Hive and its queries to give the sentiment data based up on the groups that we have defined in the HQL (Hive Query Language).Proposed model uses modified version of Naïve Bayes machine learning algorithm. Our modifications introduce neutral class by eliminating class conditional independence assumption of Naïve Bayes classifier by considering probability intersection between positive and negative classes. Algorithm results are improved by reducing words in tweet to their root form through mechanism of pre-processing before passing them to sentiment analyser. Hence, proposed system classifies tweets as positive, negative or neutral with respect to a query term. This system may prove useful for the enterprises who want to know the feedback about their product brands or the customers who want to improve their productivity and this system may also can be beneficial for election exit polls.

Key words: Hadoop Framework, Web Based Sentiment Analysis

I. Introduction

The current trend of social networking in the World Wide Web is observed greatly; especially of micro-blogging. Since there is no limit to the range of information carried by tweets and texts, often these short messages are used to share opinions that people use, about what is going on in the world around them.

In social media monitoring, Sentiment analysis is extremely useful, as it allows us to gain an overview of the wider public opinion. The different applications of sentiment analysis are wider and powerful. The ability to extract valuable information from social data is a practice that is being widely adopted by organisations, across the world. The Obama administration used sentiment analysis, to study public opinion to policy announcements and campaign messages ahead of 2012 presidential election.

In short, at Present situation peoples are expressing their thoughts through some online applications like Facebook, Twitter, WhatsApp, etc. As we concerned with Twitter, in Twitter more than 1TB of text data is generating every week in the form of tweets posted by all world peoples. But, analysing all this Tweets is very difficult as these huge data that are going to be generated day by day. This problem is taking now and can be solved by using concept of BIG DATA[1], that is Apache Hadoop ecosystem.

II. Hadoop Architecture

Hadoop make use of HDFS for data storage purpose. Each cluster of Hadoop consists of different nodes. Hence, HDFS architecture is broadly divided into following three nodes,

- Name Node.
- Data Node.
- HDFS Clients(Edge Node).

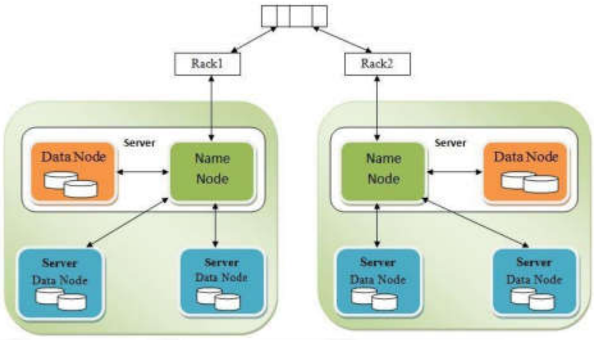

Fig. 1: Architecture of Hadoop

A. Name Node:

Name Node is also known as master node, which contains the information about or we can say that meta data about the all data node and there address(use to talk) and many other configuration like replication of data.

B. Data Node:

In simple words, Data Node is one type of slave node in the Hadoop, which is used to save the data. There is a task tracker in data node. It is used to track the ongoing jobs on the data node and the jobs which coming from name node.

C. HDFS Clients:

Hadoop Architecture is based on HDFS- Hadoop Distributed File System. The data is equally (ideally) distributed on each node in the Hadoop system. When client want to fetch or add or modify or delete some data from Hadoop, then Hadoop system collect the data from each node and do the meaningful actions as per requirement.

1) Advantages:

- Distributed data and computations, and HDFS store large amount of information.
- Simple programming model. So, sentiment analyser can be implemented easily.
- Quick recovery from system failures.
- Once data written in HDFS can be read several times, increases redundant tweets extraction.

III. Literature Review

In recent years a lot of work has been done in the field of ‘Sentiment Analysis’ by number of researchers. Work in this field started since the beginning of current century. In its early stage, it was intended for binary classification, which assigns opinions or reviews to bipolar classes such as positive or negative.

Go et al[2] started one of the early researches in this area, where the authors try a novel approach to automatically classify sentiments in tweets. They have used distant learning methods. They had classified the tweets ending with positive emoticons, one such like :-), as positive; and tweets ending with negative emoticons such as :-(, as negative. But the system uses unigram approach which fails during determination of neutral sentiments.

There are various approaches to retrieve data from Twitter dataset. Traditional approach includes writing a program in suitable languages, to get data from Twitter database. The obtained unstructured data is to be filtered to get desired structured data, again by using programs. This is tedious way and mostly time and resource inefficient. The data to be processed is then stored in RDBMS, which has limitations for creating tables and accessing it.

Paper[3] mentions the future work of using Oozie to perform sentiment analysis task with certain time span and result visualisation task.

IV. PROPOSED SYSTEM

The proposed system models the machine learning approach for sentimental analysis of Twitter feeds using Hadoop software framework. It improves accuracy of Naïve Bayes by adding Neutral class and by probability intersection of positive and negative classes.

A. System Architecture

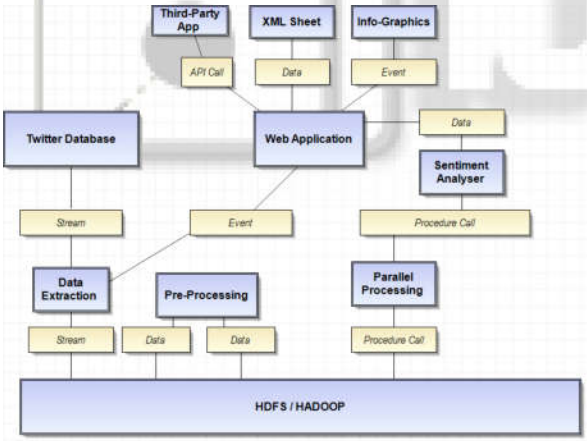

Fig. 2: Architecture of proposed system

B. Outline of System

1) Validate the Keyword

System uses Web interface to get keyword from user and validate the keyword so, only english keyword can be passed to tweet extraction module.

2) Tweets Extraction

Pass given keyword to Twitter API that find tweets related to keyword and return back to extraction module in JSON format. All above task will be done through flume that will store this Tweet Set into HDFS.

3) Pre-processing

The quality of the data affects the results and therefore in order to improve the quality, the raw data is pre-processed. It deals with the preparation that removes the repeated words and punctuations and improves the efficiency of analysis algorithms.

- Cleaning: By using a list of cut off patterns, we omit contact addresses and formatting in order to extract only the textual components, smileys.
- Tokenisation: Each tweet is split into sentences and single words named tokens.
- Stop word removal: Words without a deeper meaning, such as the, is, of, are named stop words and can thus be removed. We use a list of stop words.
- Part-of-Speech Tagging: It involves identification of verb, adverb, nouns from the sentence.
- Stemming: In computational linguistics, stemming refers to the process that reduces inflected words to their stem.

4) Parallel Processing

Structured tweets are now passed to sentiment analyser module. Oozie will take care that sufficient tweets are ready to continue task with sentiment analyser.

5) Sentiment Analyser

Now, sentiment analyser module can analyse that tweets by using modified Naïve-Bayes classifier. We will use special scoring models for smileys and normal text. Hadoop supports HiveQL to query preprocessed tweets.

C. Naïve Bayes Classifier

The Naive Bayes classifier is based on the Bayes' theorem. It is a probabilistic model. It calculates the probability of a tweet belonging to a specific class such as positive or negative. This assumes that all the features are conditionally independent. Probabilities were calculated using formula as,

$$P_{NB}(c|d) = \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)}$$

Here, class c is assigned to tweet *d*, where, *f* represents a feature and n_i(d) represents the count of feature f_i found in tweet *d*. There are a total of m features. Parameters P(c) and P (f|c) are obtained through maximum likelihood estimates which are incremented by one for smoothing. Pre-processed data along with extracted feature is provided as input for training the classifier using naïve Bayes. Once the training is complete, during classification it provides the polarity of the sentiments.

1) Output

Analysed output will be in the form of XML sheet. Proposed system also makes use of infographics to visualise the output.

V. CONCLUSION

Data and networks are becoming trends. This project shows how we can use Petabytes of data that generated daily to improve our life standard. The proposed system designs an API that can be used in Prediction of Market Trends, Infotainment, Stock markets and lots more.

REFERENCES

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, “Big Data: The Next Frontier For Innovation, Competition, And Productivity”, May 2011.

[2] A. Go, R. Bhayani, and L. Huang, -Twitter Sentiment Classification using Distant Supervision, Stanford, Technical, 2009.

[3] "Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive"

[4] International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 8, October 2014.By, Penchalaiah.C, Murali.G,