# Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm

Huma Parveen
Dept. of Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, India 490024
humaparveen812@gmail.com

Prof. Shikha Pandey
Dept. of Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, India 490024
shikhamtech2008@gmail.com

***Abstract*—In the last few years, use of social networking sites has been increased tremendously. Nowadays, social networking sites generate a large amount of data. Millions of people conveniently express their views and opinions on a wide array of topics via microblogging websites. In this paper, we will discuss the extraction of sentiment from a famous microblogging website, Twitter where the user posts their views and opinion. We have done sentiment analysis on tweets which help to provide some prediction on business intelligence. We use Hadoop Framework for processing movie data set that is available on the twitter website in the form of reviews, feedback, and comments. Results of sentiment analysis on twitter data will be displayed as different sections presenting positive, negative and neutral sentiments.**

***Keywords*—Hadoop, MapReduce, Twitter, Sentiment Analysis, Naïve Bayes.**

## I. Introduction

Comments, reviews and opinion of the people play an important role to determine whether a given population is satisfied with the product, services. It helps in predicting the sentiment of a wide variety of people on a particular event of interest like the review of a movie, their opinion on various topic roaming around the world. These data are essential for sentiment analysis [2]. In order to discover the overall sentiment of population, retrieval of data from sources like Twitter, Facebook, Blogs are essential.

For the sentiment [5] analysis, we focus our attention towards the Twitter, a micro-blogging social networking website. Twitter generates huge data that cannot be handled manually to extract some useful information and therefore, the ingredients of automatic classification are required to handle those data. Tweets are unambiguous short texts messages that are up to a maximum of 140 characters. By the use of Twitter, millions of people around the world to be connected with their family, friends and colleagues through their computers or mobile phones. The Twitter interface allows the user to post short messages and that can be read by any other Twitter user. Twitter contains a variety of text posts and grows every day. We choose Twitter as the source for opinion mining simply because of its popularity and data mining.

The Existing Database is not able to process the big amount of data within specified amount of time. Also, this type of database is limited for processing of structured data and has a limitation when dealing with a large amount of data. So, the traditional solution cannot help an organization to manage and process unstructured data. With the use of Big Data technologies like Hadoop [4] is the best way to solve Big Data challenges.

## II. Limitations of Available Systems and Tools for Analytics

The limitations of available systems are not sufficient to deal with the complex structure of the big data. In this section, we present some of the limitations that are present in the existing system.

1) The available systems like Twitter-Monitor and Real Time Twitter Trend Mining System require extensive data cleaning, data scraping and integration strategies that will ultimately increase the overhead [9].

2) For real time analytics, the available system is inefficient.

3) It is very time consuming process to analyze the huge amount of data in a short period of time.

The proposed method helps to eliminate all the drawbacks mentioned above.

## III. HADOOP AND MAPREDUCE ARCHITECTURE

In this section we first present architecture of HDFS and then we explain the working of MapReduce.

### *A. HDFS Architecture*

It Hadoop enables the application to work in a distributed environment. There may be thousands of distributed component working together to accomplish a single task. Generally, the huge log files are distributed over various clusters known as HDFS [6] cluster (Hadoop distributed file system). HDFS is able to store huge amount of data. Hadoop helps to create the cluster of machines and perform parallel work among them. Hadoop operates cluster without losing data or interruption of any work. HDFS helps to manage cluster by breaking incoming files into small chunk called block.

978-1-5090-2399-8/16/$31.00 ©2016 IEEE

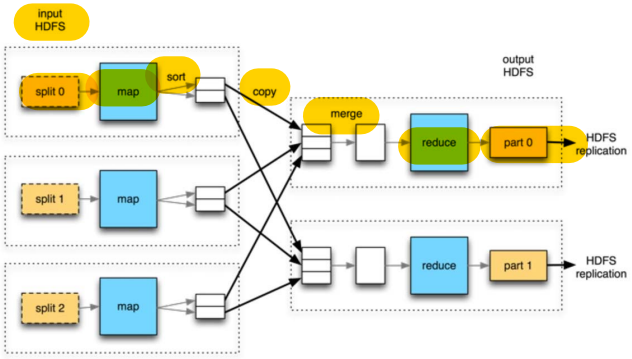


Fig. 1: Architecture of MapReduce process

1) Name node and data node: Name node stores the information about metadata which maps to the data-node for actual data. Data node simply contains the actual data.

2) Data Replication: HDFS stores each file as a sequence of blocks. These blocks are replicated to various racks on HDFS for fault tolerance. The block size and replication factor can be configured from the configuration file of Hadoop.

3) Racks: Racks are the collection of data-node. The data nodes which belong to the same network can be treated as one rack. If one of the data nodes crashes, the replica of that data-node which is present on another node starts moving to the failed data node.

### *B. MapReduce Architecture*

The Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage [8].

Fig.1 shows the architecture of MapReduce. MapReduce is a programming model for the processing of huge data. It is divided into two phases, the map and reduce phase. It allows the specific application to run in parallel so that the task is accomplished in less period of time. MapReduce jobs are controlled by the JobTracker [3]. JobTracker simply schedules the jobs submitted by the user and provide the mechanism to monitor the jobs.

## IV. METHODOLOGY

Social networking sites acquired immense popularity and interest with the people around the world. Twitter is one of the effective tools for any business intelligence to get information about what people are talking and reacting about the topics that are roaming around the world. A twitter helps to engage the users and directly communicates with them and in response, users to provide word-of-mouth marketing for companies by discussing the product quality. With the limited resources and knowing about no one can target directly to the destination consumers, the business intelligence can be more efficient in their policy of marketing by being very selective about consumers choice they should reach out to. Fig.2. shows the steps involved in processing of twitter data.

| Emoticons | Examples | | | |
|---|---|---|---|---|
| Smile | :d | :-d | :) | :-) |
| | | | | |
| Wink | ;) | ;-) | | |
| Laugh | :D | :-D | =D | =-D |
| Surprised | :o | :-o | :O | :-O |
| Playful | :p | :-p | :P | :-P |
| Sad | :( | :-( | :[ | :-[ |
| Confused | : | :- | :/ | :-/ |
| Embarrassed | :*> | | | |
| Cool | B\| | B-\| | 8\| | 8-\| |
| Angry | x( | x-( | X( | X-( |
| Love | :x | :-x | :X | :-X |
| Sleep | x-) | X-) | x-\| | X-\| |
| Cry | :,( | :”( | :’( | |

Table. I: List of emoticons

### *A. Fetching Twitter Data using Twitter API*

Develop a twitter API [10] for downloading the tweets. The Twitter API directly communicates with the Source and Sink. The Authentication keys and tokens are established that helps in communication over Twitter Server. The source is twitter account and the sink is HDFS (Hadoop Distributed File System) where all the tweets are stored.

### *B. Pre-processing of tweets*

The data coming out from twitter contains various non-sentiment contents such as website link, emoticons, white spaces, hashtag etc. which should be removed before processing it so that the sentiment generated are accurate. Preprocessing includes:

1) *Removal of URL's*: Twitter data consists of different type of information. If any user posted any link which is none of the use for sentiment analysis. Therefore, URL should be removed from the tweet.

2) *Removal of special symbol*: There are various types of symbols used by the user such as punctuation mark (!), full stop (.) etc. which does not contain sentiment. Therefore, special symbols should be removed from the tweet.

3) *Converting emoticons*: Table. I. shows the various emoticons used for conversion. Nowadays emoticons become a way for the user to express their views, feeling, and emotion. Emotions play a big role in the sentiment analysis. Therefore, convert the whole emoticons into its equivalent word by which we can do the analysis efficiently.

4) *Removal of Username*: Every Twitter user has a unique username, therefore, anything is written by a user can be indicated be writing their username proceeding by @. This type is denoted as proper nouns. For example, @username. This also has to be removed for effective analysis.

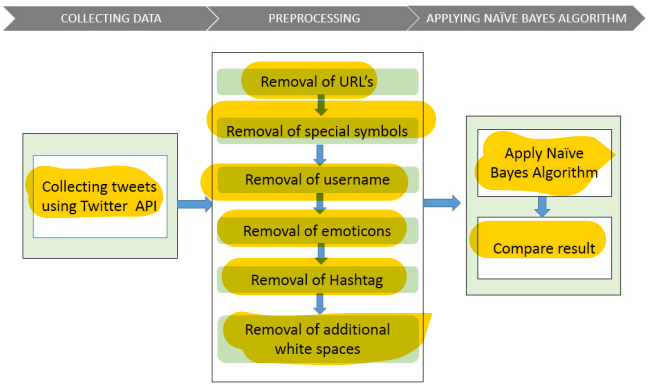


Fig. 2: Proposed System Architecture

5) *Removal of Hashtag*: A hashtag is a prefixed with the hash symbol (#). Hashtag are used for naming subjects or phrases that are currently in trend. For example, #google, #twitter.

6) *Removal of additional white spaces*: There may be consists of extra white space in the data and it needs to be removed. By removing white spaces the analysis to be done more efficiently.

### C. Applying Naive Bayes Algorithm

The Naive Bayesian Classification [7] represents a supervised learning method as well as a statistical method for classification. It is probabilistic model and it permit us to capture uncertainty about the model in a principled way by determining probabilities. It helps to solve diagnostic and predictive problems. This Classification is named as Naive Bayes after Thomas Bayes, who proposed the Bayes Theorem of determining probability. Bayesian classification provides useful learning algorithms and past knowledge and observed data can be combined. It helps to provide a useful perspective for understanding and also evaluating many learning algorithms. This helps to determine exact probabilities for hypothesis and also it is robust to noise in input data.

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \quad (1)$$

P(C | X) is posterior probability,
P(X | C) is likelihood,
P(C) is class prior probability,
P(X) is predictor prior probability.

## V. Framework Implementation of Naïve Bayes Algorithm

In this section, we present the implementation of our Hadoop framework for efficiently executing Naive Bayes algorithm. Our proposed mechanism extends Hadoop to implement map and reduce phase.

To implement Naïve Bayes algorithm we need a trained SentiWordNet [1] dictionary which is available online. It consists of collection of different word with its synonym and its polarity. The synonym represents the similar word meaning which will be having same polarity. The polarity represents the positivity of the word in the context of the sentence.

We have to input 2 files to the mapper

- Twitter Dataset which contains the comments and review of the user.
- SentiWordNet dictionary which contains the polarity of the different words.

The proposed methodology for applying Naïve Bayes algorithm splits into 2 phase, Map and Reduce Phase.

### A. Map Phase

The working of Map phase consists of two major tasks. First, creating a hash map for retrieval of polarity of each words. Secondly, processing the overall polarity of the tweets by applying Naïve Bayes algorithm. The map() method in MapReduce phase reads the content of the SentiwordNet dictionary from a file and transform into the Hash map for key-value based polarity retrieval of words. From here, the polarity of each word is stored in the hash map for faster processing.

Now, the map() method read tweets line by line from the file. Map method parses each and every word and generate tokens. Each tokens has polarity available in the hash map. The polarity are fetched for each word and calculate the overall polarity of a single tweets using probabilistic model.

### B. Reduce Phase

The reduce() method collects the overall polarity of each tweets and transform into 5 different categories as extreme positive, positive, extreme negative, negative and neural. The reduce() method iteratively work to collect various sentiments and based on polarities it classify and write the output on HDFS.

| Sentiments | Count without emotions | Count with emoticons |
|---|---|---|
| Extreme Positive | 130 | 177 |
| Positive | 59 | 90 |
| Extreme Negative | 45 | 42 |
| Negative | 30 | 26 |
| Neutral | 136 | 65 |

Table. II: Sentiments of tweets with and without considering emoticons

## VI. RESULT

Firstly the data is downloaded from the twitter. They are stored into the HDFS for analysis. Before evaluation the tweets, we need to be pre-processed in order to remove the noise from the data. We evaluate performance of our algorithm by comparing the result with and without considering emoticons. We observed in our result that when we perform pre-processing without considering emoticons the tweets

contains sentiment in the form of emoticons are simply ignored by Naive Bayes algorithm and hence when we perform pre-processing by considering emoticons, the results are much more accurate than the previous one.

From table II, we observed that the sentiments which is neutral are in great number while emoticons are not considered. But when pre-processing is used with emoticons the neutral tweets are considerably decreases due to conversion of emoticons. Hence performance of Naive Bayes algorithm increases by converting the emoticons by assigning its equivalent word.

## VII. CONCLUSION AND FUTURE WORK

Twitter Data in the form of opinion, feedback, reviews, remarks and complaint are treated as big data and it cannot be used directly. These data first convert as per requirement. In this paper, we discussed pre-processing of data to remove noise from the data. We have implemented sentiment analysis for movie data set, on Hadoop framework and analyzed with large number of tweets. This type analysis will definitely help any organization to improve their business productivity. The analysis of twitter data are done on various perspective like Positive, Negative and Neutral sentiments on tweets. It also provide the fast downloading approach for efficient Twitter Trend Analysis. Tweets can also be useful in prediction of product sales, quality of services offered by company, feedback of users etc. Hence, the future scope in the sentiment analysis for the other social networking websites like Facebook, Google Plus etc.

## REFERENCES


[1] "SentiWordNet Dictnary", http://sentiwordnet.isti.cnr.it/ [accessed 05 May 2016]

[2] F.Neri, C.Aliprandi, F.Capeci, M.Cuadros, T.By, "Sentiment Analysis on Social Media", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2012, pp. 919 - 926

[3] "Job Tracker", https://wiki.apache.org/hadoop/JobTracker

[4] M. Bhandarkar, "MapReduce programming with apache Hadoop", IEEE International Symposium on Parallel & Distributed Processing (IPDPS), 2010, pp. 1

[5] Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas, "Sentiment Analysis on Social Media", 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

[6] J.Shafer, S.Rixner, A.L.Cox, "The Hadoop distributed filesystem: Balancing portability and performance", IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS), 2010, pp.122-133

[7] J.Ren, S.D.Lee, X.Chen, B.Kao, R.Cheng, D.Cheung, "Naive Bayes Classification of Uncertain Data", Ninth IEEE International Conference on Data Mining, 2009. ICDM '09, pp. 944 - 949

[8] J. Dean, S. Ghemawat, "MapReduce: simplified data processing on large clusters", The 6th Symposium on Operating Systems Design and Implementation, OSDI-04, USENIX Association, 2004, pp. 107 – 113

[9] Gaurav D Rajurkar, Rajeshwari M Goudar, "A speedy data uploading approach for Twitter Trend And Sentiment Analysis using HADOOP", HADOOP, 2015 International Conference on Computing Communication Control and Automation

[10] "Twitter API", https://dev.twitter.com/rest/public/ [accessed 05 May 2016]