IJEE 0973-7383 Volume-9 • Number-1 Jan -June 2017 pp. 201-207 available online at www.csjournals.com

Sentiment Analysis of Twitter and Facebook Data Using Map-Reduce

Shruti Gupta, Ashutosh Pandey, Prof. K.K.Paliwal
Deptt. of Computer Science, P.I.E.T, Panipat,India

Abstract: Big data is an assortment of large data sets where data is present either in structured or unstructured form. With the advent of social media websites, online user opinions are acquiring more heed of researchers because beneficial information about different subjects is available on social media. Twitter and Facebook has become very popular among billions of people who share their views and opinions about different topics. Thus, these websites are amusing source of data for opinion mining or sentiment analysis. Social media websites produces a huge volume of data every day, this data can be used to find the sentiments of people on a given topic or product. In this paper we proposed a system which involves collecting data from social network using the Twitter and Facebook API's. Then, the challenges of big data are solved using Hadoop through map reduce framework where the complete data is mapped and reduced to smaller sizable data to ease of handling and finally includes analysing the collected data and represent the results through graphs.
Keywords: sentiment analysis, Hadoop, Map-Reduce.

I. INTRODUCTION

Big data is an assortment of large data sets where data is present either in structured or unstructured form. Some well-known social media websites such as Twitter, Facebook, Yahoo!, LinkedIn, etc. generates an enormous quantity of unstructured and structured data daily. This rapid growth of data leads to some challenges like extraction of useful information from big data and processing of large dataset.

Vast amount of data generated by social networks is difficult to analyze [1]. This data will go waste if left unused while this data can be used for the purpose of understanding customer behavior, marketing, etc. Different methodologies and algorithms are used for the prediction of sentiments which will extract opinions from internet and help in decision making of customers.

Sentiment analysis [2]is a kind of natural language processing for tracing the mood of people about a particular topic, service or product. It includes construction of a system to accumulate and examine opinions about product made in blog, post, tweets, reviews, orcomments.

1.1 Sentiment Analysis methods

There are main following methods to perform Sentiment analysis and opinion mining:

i. Machine learning
ii. Lexicon based method.

1.1.1 Machine learning approach

Different machine learning algorithms are used in machine learning techniques such as Support Vector machine, Naïve bayes, maximum entropy, neural networks and many more; for classifying text. Machine learning approach is further classified as supervised and unsupervised learning techniques. In supervised learning a lots of labelled training data sets are used and when the training data sets are harder to find then unsupervised method is adapted [3].

1.1.2 Lexicon based approach

In lexicon based method, text classification is done by using sentiment lexicon, which is a collection of pre-recognised and pre-compiled sentiment word. Lexicon based method is divided into dictionary based and corpus based methods. In dictionary based approach, sentiment polarity can be determined by comparing given word with a list of words present in the dictionary which consist of synonyms and antonyms. Corpus based approach depends on statistical and semantic methods for classifying the polarity of sentiment words in a large corpus [3].

1.2 Sentiment Analysis Levels

Usually, there are 3 levels of sentiment analysis that has been studied; document, sentence, and entity or aspect level [4].

Document level: Overall sentiment of a complete document is decided in this level. For instance, if review of product is given, the task is to decide whether it convey an overall negative or positive opinion regarding the product [4]. The job is to verify whether the whole document is negative, positive or neutral.

Sentence level: The job at this phase is limited to sentences and test if each sentence conveyed a negative, positive or neutral opinion [4]. Firstly, sentence is classified as objective or subjective and then sentences which are subjective are categorised as positive, negative or neutral.

Aspect and entity level: This level [4] is more challenging than the other two. Aspect level analyses the opinions instead of analysing paragraphs, documents, phrases or sentences. It provides finer-grained analysis for each aspect. Opinion is a phrase which consists of a target/subject and sentiment on the target. This idea is used in entity level analysis. This helps in understanding sentiment analysis problem better.

1.3 Hadoop and Map-Reduce Algorithm

Big data is a pool of large data sets where data is present either in structured or unstructured form. A large amount of data is generated by social media [1]; analysis of such volume of data is not an easy task. The data generated can be used in marketing and understanding customer behaviour. In last few years, different methods and techniques have been used for handling and analysing big data. Among so many processing tools that are used to handle and analyse big data Hadoop is one of them.

Open- source software is developed by apache Hadoop for reliable, distributed, scalable computing. The library itself is designed to detect and handle failures at application layer instead of relying upon hardware. Thus, delivering a highly available service on top of a cluster of computers, each of which may be prone to failures [5].

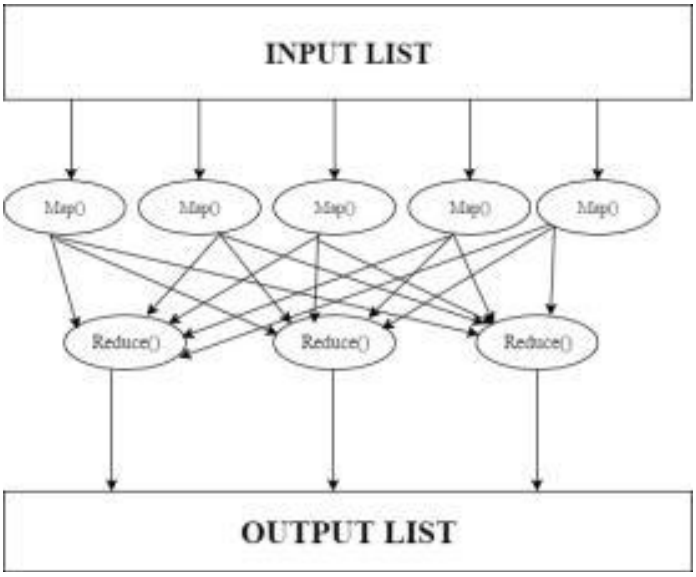

Fig.1Map-Reduce produce an aggregate value from given input list. [5]

Hadoop is a programming framework used to support the processing of large data sets in distributed computing environment [6]. The processing pillar in hadoop system is the Map-Reduce framework. A Map-Reduce algorithm usually divides the input dataset into two separate classes: Mapper class & Reducer class, which are handled separately. The Mapper takes a set of data and transforms it into an alternative set of data, where the whole data set is converted into key/value pairs (tuples). The output from a mapper class is given as input to reducer class and combines those data tuples into a minor set of tuples. As the term Map-Reduce suggests, the mapper task is always performed before the reduce job.

This paper is organized as follows: section 2 describes the literature review that has been done in past for sentiment analysis then in section 3 motivation for the research followed by the proposed approach for this research in section 4 then implementation and results in section 5 finally the conclusions and future scope are explained in section 6.

II. LITERATURE REVIEW

This section shows the literature review done on different techniques to perform sentiment analysis. Xiaoqian Zhang et.al [7] exemplified a creation of polarity-shifted product review corpus in the sentence-level is created. All the sentences with polarity shiftare analysed and categorised into five types: hypothesis, Opinion, time, target and holder. Experimental studies demonstrated that in text reviewing the polarity shifting phenomenon appears very commonly.

Pang B. and Lee L. [8] projected a peculiar machine-learning technique to calculate the sentiment polarity of subjective part of the document using text categorization methods. They examined the relation between polarity classification and subjectivity detection. Their experiments revealed the graph-cut formulation to be a better methodology for both Naïve Bayes and SVM. They achieved extremely statistically significant enhancement from 82.8% to 86.4%.

Modha J. S. et al. [9] proposed an approach to handle objective as well as subjective sentences and find opinion from them. In their proposed system they followed following steps:

(i) Firstly they classified sentences as opinionated and non-opinionated. (ii) Then, they classified opinionated sentences into subjective or objective. (iii) Third step is to classify subjective sentences into negative, positive or neutral category. (iv) Then, classify objective sentences into positive, negative or neutral.

They provided context or sentiment orientation as and when needed.

Borade A. J. et al. [10] used review mining techniques for presenting product ranking system. On the basis of collected user reviews, ranking of different types of products is provided. Three issues are considered while determining product scores: Product review, product popularity and product release month.

Khuc et al. [11] outlined a large-scale distributed system for real-time sentiment analysis on Hadoop. There are two components in their system: a sentiment classifier and lexicon builder. These components are implemented using a map-reduce framework and distributed database. A method is introduced to combine sentiment lexicon with machine learning algorithm and improvement in accuracy is observed.

Manoj Kumar Danthala [12] suggested an idea for analysing tweets. Big data that is, tweets, are handled using Apache Hadoop. This analysis can be used in predictive analytics, text analytics and sentiment analysis.

MahimaGoyal [13] used Sentiwordnet dictionary for calculation of the sentiments. They used unsupervised dictionary based approach in their work. Their field of research is limited to tourism domain only.

III. MOTIVATION

The tremendous expansion of documentary information on the network in past few years leads fundamental revolution in individual's life. People use different microblogging sites to express their views and opinions about different subjects/ topics. This generates a vast assortment of opinions in the form of texts. These assortments need to be analysed to know the efficiency of the subject or service. Therefore, there is a need to develop a system which can examine the opinions of people. This system will help in decision making of customer and will also provide a feedback to seller so that they will know the area to work on.

IV. PROPOSED SYSTEM

The proposed system/tool allow users to use a simple search bar to search for any services, products or any current topic and the engine of that application is to crawl over the internet collecting all comments, reviews, tweets related to the user's search keyword.

First, the system involves collecting data from social network using the Twitter and Facebook API's. Then, the system uses hadoop platform to solve the challenges of big data through Map-Reduce framework where the complete data is mapped and reduced to smaller sizable data to ease of handling and finally includes analysing the collected data and represent the results through graphs.

Map-Reduce code snippet: Following is a code snippet used in the system showing use of map reduce algorithm. Two classes are formed one is mapper class and other is reducer class. Mapper class takes input data and passes it to reducer class and then result is shown by reducer class.

```
@Override
    public void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {

        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {

            String token = tokenizer.nextToken();

            context.write(new LongWritable(Long.parseLong(tokenizer.nextToken().toString())), new Text(token));

        }
    }


public void map(LongWritable key, Text value, Context context) throws IOException,InterruptedException
        {
            pwords= readLines("C:\\eclipse\\workspace\\TweeterWebApp\\pos.txt");
            nwords= readLines("C:\\eclipse\\workspace\\TweeterWebApp\\neg.txt");
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            while (tokenizer.hasMoreTokens()) {
                    String token = tokenizer.nextToken();
                    for(int i=0;i<pwords.length;i++)
                    {
                        if(token.startsWith(pwords[i].toLowerCase())){
                            posmpos+1;
                            word.set(token.toLowerCase());
                            context.write(word, one);
                        }
                    }

                    for(int i=0;i<nwords.length;i++)
                    {
                        if(token.startsWith(nwords[i].toLowerCase())){
                            negneg+1;
                            word.set(token.toLowerCase());
                            context.write(word, one);
                        }
                    }

                    for(int i=0;i<neuwords.length;i++)
                    {
                        if(token.startsWith(neuwords[i].toLowerCase())){
                            neutral= neutral+1;
                            word.set(token.toLowerCase());
                            context.write(word, one);
                        }
                    }
```

Fig 2: Mapper code snippet

```
@Override
protected void reduce(LongWritable key, Iterable<Text> trends, Context context)
            throws IOException, InterruptedException {

        for (Text val : trends) {

            context.write(new Text(val.toString()),new Text(key.toString()));
        }
}


@Override
protected void reduce(Text key,Iterable<IntWritable> values,Context context)
        throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values) {
        sum += val.get();

    }
    context.write(key, new IntWritable(sum));
}
```

Fig 3: Reducer code snippet

3.2 FLOW CHART:

A flowchart [14] is a pictorial exemplification of the sequence of steps and decisions needed to perform a process. There is a connection between every consecutive step with the help of directional arrows. This helps anyone to outlook the flowchart and logically tracks the process from beginning till the halt state.

Following is the flowchart for the system. This gives an outlook of the process followed in the system. Starting from fetching comments, tweets and posts by the users till the time you can save the results.

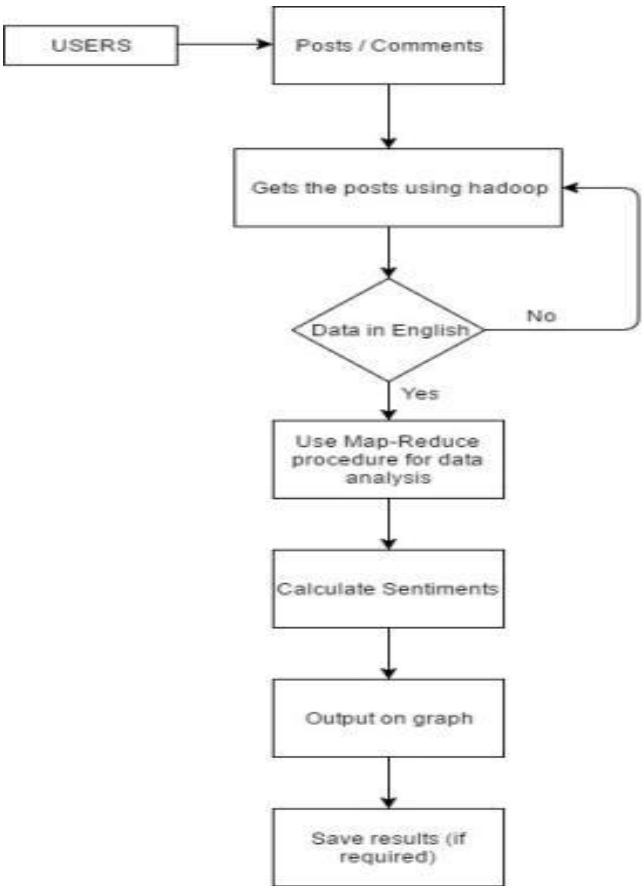

Fig 4: Flow chart showing flow of the system.

V. IMPLEMENTATION

The core objective of the project is:-

1) Data Collection: Facebook and Twitter streaming APIs are used for collecting large amount of content.

2) Storage: The data is stored in some format for further processing. This data is stored in a certain format so as to form key value pair which is needed to feed to mapper class in map-reduce programming approach.

3) Data Processing and analysis: Data collected over a period of time is processed by using java and map reduce programming model by Apache Hadoop. Mapper class takes data and form key value pairs and reducer class refine these key value pairs. Further the output of reducer phase is analysed.

4) Data Representation: Data is represented in the form of pie chart and bar graph showing positive, negative or neutral views.

SCREENSHOTS OF IMPLEMENTATION

Fig.5: Log-in page

Fig. 6 Keyword to be analysed is entered in this page.

VI. RESULTS

In this paper, we have taken into consideration Facebook as well as twitter data. In earlier work only twitter data was use for calculating the sentiments. With the help of this system, sentiment polarity (positive, negative or neutral) of posts and tweets on different topics can be analysed. This will help in increasing the efficiency a more than one microblogging site is used in this system.

Given below is a graph showing the output obtained when different keywords are used for analysing the sentiments of users.

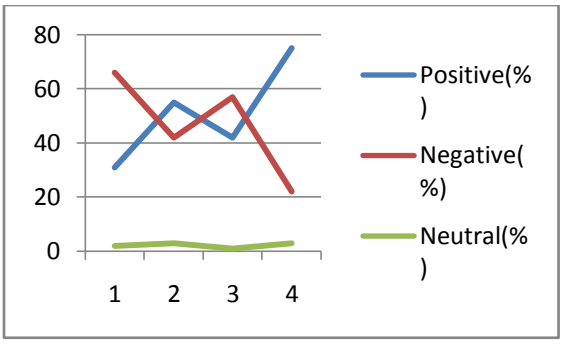

Fig.7: Output obtained when varied input were given.

VII. CONCLUSION & FUTURE SCOPE

Using sentiment analysis, we can differentiate poor quality content from superior quality content. Nowadays people express their views and opinions on any product or service over social media. So the data over social media should be utilised in a proper way. Our system will use this big data and analyse the views and opinions of people into positive, negative or neutral. Only English language is considered in this system. In future use of multiple languages can be done moreover combination of native language with English is very common over internet this can also be considered in future. Sarcastic comments are very common over social media and they can change the context of the sentences. In this paper, sarcastic comments are not considered so there is a scope of working on this type of comments.

REFERENCES

[1]Mahalakshmi R and Suseela S. Big-SoSA: "Social Sentiment Analysis and Data Visualization on Big Data", International Journal of Advanced Research in Computer and Communication Engineering. Vol. 4, Issue 4, April 2015.

[2] G. Vinodhini and RM.Chandrasekaran."Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 6, June 2012.

[3] "On Social Sentiment and Sentiment analysis", (Last visited in Dec 2016). [online]. Available: http://www.mssqltips.com/sqlservertip/3180/big-data-basics--part-4--introduction-to-hdfs/

[4] Bing Liu. "Sentiment Analysis and Opinion Mining.Synthesis Lectures on Human Language Technologies", Morgan and Claypool Publishers, 2012.

[5] "Hadoop Tutorial", (Last visited in Nov. 2016). [online]. Available:" http:// www.tutorialpoint.com/hadoop/hadoop_mapreduce.htm"

[6] Harshawardhan S. Bhosale and Prof.Devendra P. Gadekar."Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Vol. 4, Issue 10, October 2014.

[7] Xiaoqian Zhang; Shoushan Li; Guodong Zhou; Hongxia Zhao, "Polarity Shifting: Corpus Construction and Analysis", Asian Language Processing (IALP), pp.272, 275, 15-17 Nov. 2011

[8] Pang, B. & Lee, L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (pp. 271). Association for Computational Linguistics. July 2004
[9] Jalaj S. Modha, gayatri S. Pandi and Sandip j. Modha, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering, pp. 91-97, vol.3, Issue 12, December – 2013.
[10] Abhijit J. Borade, Imran R. Momin, Dattatray B. Ghogare and S. Pratap Singh, "Sentimental Analysis to Rank Web Products", IJSRD - International Journal for Scientific Research & Development, vol. 3, Issue 02, 2015 | ISSN (online): 2321-0613
[11]V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan, "Towards building large-scale distributed systems for Twitter sentiment analysis", In Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12), pp. 459–464, March 2012.
[12] Manoj Kumar Danthala. "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal of Core Engineering & Management. Vol. 1, Issue 11, Feb. 2015.
[13] MahimaGoyaland Vishal Bhatnagar. "A Classification Framework on Opinion Mining for Effective Recom*mendation Systems", Collaborative Filtering Using Data Mining and Analysis*,2016.
[14] "What is flow chart?" (Last visited in Dec. 2016). [online]. Available: "https://www.smartdraw.com/flowchart/"