IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A NOVEL MACHINE LEARNING ALGORITHM TO ANALYZE SENTIMENT DATA ON SOCIAL MEDIA NETWORKS TOWARDS PREDICTION OF ATTITUDE OF THE USER

PRATHEEPKUMAR.P
Assistant Professor,
Department of Computer Science
Nift-Tea College of Knitwear Fashion, Tirupur

PRIYADHARSINI.M
Assistant Professor,
Department of Computer Science
Nift-Tea College of Knitwear Fashion, Tirupur-

Abstract

The Sentiment Analysis is the process use to determine the semantic orientation of the reviews. There are many algorithms are exists for the sentiment classification. Support vector machines are a specific type of machine learning algorithm used for many statistical learning problems, such as text classification, spam filtering, face and object recognition, handwriting analysis and countless others. Ideas and opinions of peoples are influenced by the opinions of other peoples. Lot of research is going on analysis of reviews given by peoples. 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 researches because beneficial information about different subjects is available on social media. In this paper we proposed a system which involves collecting data from social network using the Google, Flipkart, Twitter and Facebook API's. Then, the challenges of big data a resolved using Hadoop through MapReduce framework where the complete data is mapped and reduced to smaller sizable data to ease of handling and finally includes analyzing the collected data and represent the results through graphs.

Keywords : ***HDFS, Sentiment analysis, Pig, MapReduce, Hive***

1. INTRODUCTION

With the wide growth in the practice of social media, the amount of information is available as users' preference regarding any product, services provided by various organizations or with respect to any political issues. Micro blogs, forums are also available wherein online users, can express their opinions. As mobile devices can access network easily from anywhere, social media is becoming more and more popular. The number of people using the social media is increasing day by day as they can share their own feeling every day and reviews are created in large - scale. Every minute opinion, reviews are being expressed online and a potential user rely on these reviews, opinions, feedback given by various other users to make decisions with respect to purchasing an item or developing a software when it comes to an organization that provides services. Analyzing these reviews, opinions or feedback in this scenario is of utmost importance. It seems evaluating these reviews, opinions are not as easy as it appears to be, and it requires performing sentiment analysis[7]. Sentiment analysis greatly helps us in knowing the customer behavior. The biggest challenge is to process the social data which are in unstructured or semi-structured form. The former technologies fail to process the data in this form in an effective way. So, there is a need for highly optimized, scalable and efficient technology to process the abundant data that are being produced at a high rate. The social media

data produced will be either unstructured or semi-structured. Hadoop and MapReduce framework effectively analyzes the unstructured and semi-structured form data.

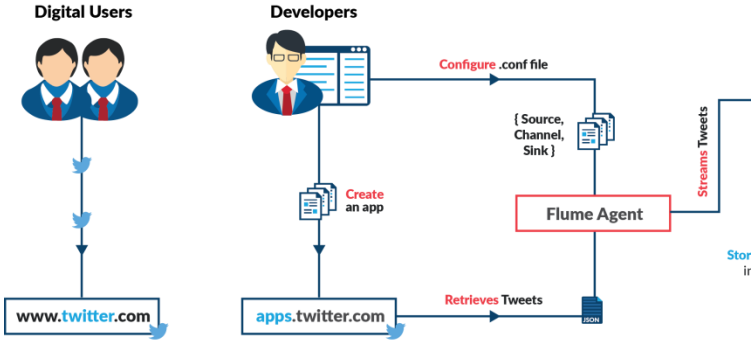

Figure 1: Sentiment analysis using Hadoop and MapReduce

2. PROBLEM DEFINITION

This paper focuses on using online social media like Facebook(2017), Twitter(2017), the most popular micro blogging platform, for the task of sentiment analysis. The tweets are main for analysis because data arrive at a high rate and algorithms that process them must do so under very strict constraints of storage and time. It will be shown how to automatically collect a corpus for sentiment analysis and opinion mining purposes. All public posts posted on social media are freely available through a set of APIs provided by blogs. Using the corpus, a sentiment classifier, is constructed that is able to determine positive, negative and neutral sentiments.

3. RELATED WORKS

Sentiment analysis is exceptionally prevalent these days. Heaps of research is going around the natural language processing information extraction and data mining. Many researchers are exploiting different methods and tools to obtain the sentiment analysis. Any system must have good speed and efficiency too. Without the speed and efficiency, a system might not be suitable for performing the sentiment analysis on the vast amount of data that is used for the opinion mining.

Sunil B. Mane and his colleagues provided[1] efficient way of doing sentiment analysis using Hadoop. Their approach was focused on the speed. Technique provided by them processed vast amount of data on a Hadoop cluster faster in real time.

Jayshree Khairnar [2] discussed about using Support Vector Machin and LSA to perform sentiment analysis. They found that there is still need of improvements in terms of efficiency and accuracy. Hence the reason they suggested to use MapReduce. Their method gave better efficiency for producing the result.

Jeffrey Shafer [3] discussed the causes of performance bottlenecks in Hadoop. He identified three different problems like software architecture bottlenecks, portability limitation and portability assumption. Jeffery found that the problem of bottlenecks is more related to HDFS than to MapReduce. Sentiment analysis result can be shown using different types of graphs and
charts.

Changbo Wang presented another [4] representation framework for doing analysis and visualizing and verifying the sentiments from web source. In SentiView, different methods of visualization to show the output of sentiment analysis have been added.

Ruchika Sharma proposed a system [5] in which her emphasis was on improving the accuracy of sentiment analysis. Her method of using Multiple kernel gave better accuracy of 90% and 92%. Though her conclusion was of using multiple kernel with some different Machine leaning algorithms does not improve on accuracy.

Dhiraj Gurkhe [6] tried different datasets for his proposed methods. Datasets were mainly from the social networking sites. Dhiraj concluded that his system gives best results with Unigram detection.

4. SENTIMENT ANALYSIS

Sentiment is defined as an expression or opinion about any object or any aspect. Analyzing, investigating, extracting users' opinion, sentiment and preferences from the subjective text is known as sentiment analysis. The main focus of sentiment analysis is parsing the text. In simple
terms, sentiment analysis can be defined as detecting the polarity of the text. Polarity can be positive, negative or neutral. It is also referred to as opinion mining as it derives opinion of the user. Opinions vary from user to user and sentiment analysis greatly helps in understanding users' perspective. Sentiment can be,

Direct opinion as the name suggests the opinion about an object is given directly and the opinion may be either positive or negative. For example, "The video clarity of the mobile phone is poor" expresses a direct opinion.

Comparison opinion, it is a comparative statement which consists of comparison between two identical objects. The statement, "The picture quality of camera-x is better than that of camera-y" is one possible example for expressing a comparative opinion. Sentiment analysis is performed at three different levels:

- Sentiment analysis at sentence level identifies whether the given sentence is subjective or
- objective. Analysis at sentence level assumes that the sentence contains only one opinion.
- Sentiment analysis at document level classifies the opinion about the particular entity. Entire document contains opinion about the single object and from the single opinion holder.
- Sentiment analysis at feature level extracts the feature of a particular object from the reviews and determines whether the stated opinion is positive or negative. The extracted features are then grouped and their summarized report is produced.

4.1 MapReduce

The MapReduce paradigm enables the writing of applications in an effective manner and also processing of huge sets of data is efficient with Hadoop MapReduce [8]. The MapReduce paradigm has two different tasks:

The Map Task: The Map task captures the input and this input data are divided into pair of data. This data is further divided into tuples to form a key/value pair.

The Reduce Task: The input to the Reduce task is the output from the Map task. All the divided tuples in the Map task is combined to form smaller set of tuples. Map Task is followed by Reduce Task.

The MapReduce component of the Hadoop framework schedules monitors the tasks and also re-executes the failed task. The MapReduce paradigm has a single Job Tracker and one Task Tracker that acts as master and slave respectively. The master Job Tracker directs the slave Task Tracker to execute the task and also it manages the resource, tracks the resource distribution, consumption and availability. On the other hand the Task Tracker provides the status information to the Job Tracker.

5. PROPOSED ALGORITHM

I. Design Considerations:

- 1.Collected unstructured, structured, semi-structured data from online social media sources.
- 2.Store collected tweets and comments into HDFS.
- 3.Normalize it and Store processed data in HDFS.
- 4.implement MapReduce technique and Sentiment Analysis is done
- 5.Visualize with a tool of choice.

II. Sketch of the Proposed Algorithm:

The proposed system has the following modules:

a) Extracting The collected Tweets and comments Preprocessing
b) Sentiment Divergence Analysis
c) Visualization

Sample Twitter Dataset

Number of users	17,069,982
Number of tweets	476,553,560
Number of URLs	181,611,080
Number of Hashtags	49,293,684
Number of re-tweets	71,835,017

PSEUDO CODE

Step 1: Start.
Step 2: Extracting the tweets and comments using Facebook and twitter API by giving a keyword (#tag).
Step 3: Storing it to the Hadoop Distributive File System.
Step 4: Normalizing the collected tweets and storing those tweets in HDFS.
Step 5:implement MapReduce technique .
Step 6: Finding the divergence of the words and classifying it according to the polarity.
Step 7: Displaying the results.
Step 8: Stop.

5.1 Collection of data

A wide range of data is available which can be helpful to classify emotions in tweet as positive and negative. We have collected millions of tweets and comments from various sources and description of those dataset is given below.

• **Sanders Twitter and Facebook Dataset:**

The Sanders dataset consists of about 9986. Each of these tweet was manually labeled as either positive or negative according to the emotion contained in it.

5.2 Evaluation Metrics

To evaluate our experiment results by using following Information Retrieval matrices.

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- F - measure = 2 * Precision * recall / (Precision + recall)
- Accuracy = TP + TN /(TP + TN + FP + FN)

Where

- TP – True Positive
- FP – False Positive
- TN – True Negative
- FN – False Negative

Table 1: Tweets and comments analysis of online social media

Parameter	Total	Positive	Neutral	Negative
Words	9986	4895	1495	3596
Percentage	100	49.01	14.97	36.01

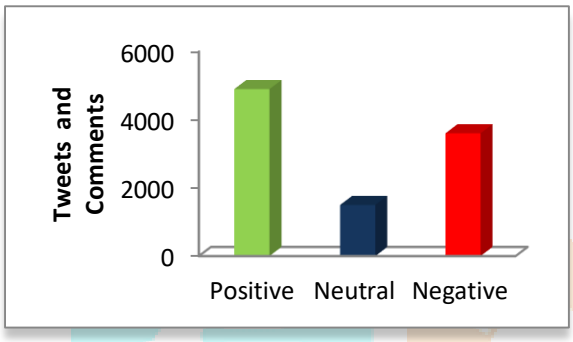

Figure 2: Tweets and comments analysis of online social media

6. CONCLUSION

There is lot of possible research to be done in the sentiment analysis. We tried to cover most of the important aspects of the sentiment analysis. Getting the tweets closely related to the aspects can give improved data sets. Negation handling is very important in sentiment analysis and should have better output. Since the negation handling can really impact the result of sentiment analysis. Future work can be done on implementing different foreign language dictionaries as current approach supports only English Language. Finding diverse emotions like angry, tedious, thrilling from sentiment analysis can also be considered. Special ways to produce the output visually can be thinking of too.

REFERENCES

1.Sunil B. Mane et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014.

2.JayashriKhairnar, MayuraKinikar "Sentiment Analysis Based Mining and Summarizing Using MapReduce" 27th IEEE Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2015, 4081-4085

3.Shafer, Jeffrey, Scott Rixner, and Alan L. Cox. "The hadoop distributed file system: Balancing portability and performance." Performance Analysis of Systems & Software (ISPASS), 2010 IEEE International Symposium on. IEEE, 2010.

4.Wang, Changbo, et al. "SentiView: Sentiment analysis and visualization for internet popular topics." Human - Machine Systems, IEEE Transactions on 43.6 (2013): 620-630.

5.Ruchika Sharma and Amit Arora. Article: Improve Sentiment Analysis Accuracy using Multiple Kernel Approach. International Journal of Computer Applications 71(20):12-15, June 2013. Full text available

6.DhirajGurkhe, Niraj Pal and Rishit Bhatia. Article: Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification. International Journal of Computer Applications 99(13):1-4, August 2014.

7.Jai Prakash Verma,Bankim Patel, Atul Patel "Big Data Analysis: Recommendation System with Hadoop Framework."Computational Intelligence and Communication Technology(CICT)(2015):92-97.

8.Aditya Bhardwaj, Vanraj, Ankit Kumar, Yogendra Narayan, Pawan Kumar, "Big Data Emerging Technologies: A CaseStudy with Analyzing Twitter Data using Apache Hive."Recent Advances in Engineering and Computational Sciences(RAECS)(2015):1-6.

9.Han Hu, Yonggang Wen, Yue Gao, Tat-Seng Chua, and Xuelong Li "Toward an SDN-Enabled Big Data Platform for Social TV Analytics," *IEEE Network.*,vol.29,pp. 4349,Sept-Oct.2015.