Sentimental Analysis of Social Media Using R language and Hadoop

Sayali Shinde
IT-SAKEC
Mumbai, India.
sayali.shinde@sakec.ac.in

Shrutika Kaware
IT-SAKEC
Mumbai, India.
shrutika.kaware@sakec.ac.in

Chintal Gala A
Assistant Professor
IT-SAKEC
Mumbai, India.
chintal.gala@sakec.ac.in

Ronak Thakkar
IT-SAKEC
Mumbai, India.
ronak.thakkar@sakec.ac.in

Sonali Rai
IT-SAKEC
Mumbai, India.
sonali.rai@sakec.ac.in

Abstract— The way of expressing people's views, opinions and Sentiments about others has been changed due to the growth of Technology of World Wide Web. Mostly people use blogs, Social sites, online discussions etc. for expressing their views. Massive amount of data is generated because of this. Companies face many problems in storing this massive data. This paper helps in analysis of sentiments of the data fetched from twitter using R language which either collects the information in the form of positive, negative or neutral score. After this the fetched data is pre-processed i.e. all the slang words, misspelled words etc. are removed. Using R language and Hadoop Connector we perform the analysis of twitter data which has size of TB's also known as big data. R language and Hadoop tool are the two different platforms on which the performance estimation will be based on.

Keywords: *Twitter API, Rhadoop, HDFS, rhdfs, Sentimental Analysis, RClient, twitteR.*

*****

I. Introduction

Sentimental analysis means to check the taste, views and interest of people regarding celebrity, politicians or some other topic. Basic thing in sentimental analysis is to classify their mood in different category like positive, negative and neutral. Twitter is one such well known micro-blogging site getting around 500 million tweets per day [1]. Each user has a daily limit of 2,400 tweets and 140 characters per tweet [2]. Twitter also serves as a huge platform for users to know more and get direct comments about a product or a service in which they are interested [3]. It is difficult to understand and extract information from this data. So, we need such tools and Technologies that can efficiently store and process unstructured and big data. There are different techniques and tools are available that can handle this type of data and produce meaningful information but in this paper we study R language and Rhadoop Tool. Social media is a web-based and mobile-based internet application that will allow the creation, access and exchange of user-generated content that is ubiquitously accessible. Besides social networking media like twitter and facebook, the term social media to encompass really simple syndication (RSS) feeds, blogs, wikis and news, all typically yielding unstructured text and accessible through the web. Sentimental analysis of twitter is quite difficult[4] as compared to other general sentimental analysis due to the presence of slang words, misspelling in tweets, short length, and some graphics type words etc.

Big data analyst, Scientists, Engineers use R language for Statistical computing, Graphics and analysis purpose. R is a most popular open source platform with different version on Windows, Linux and mac OS.R is a comprehensive statistical

platform provides approximate 5000 packages and offers data analytics techniques. It is a powerful platform for data analysis and exploration. In some cases when size of data is large and it exceeds from its physical memory, then it performs very slow and gives poor results. Therefore,[5]Rhadoop is introduced which stores whole data into Hadoop (Hdfs) and R is used for fetching data from (HDFS file system) and performs analysis on that data.

Social analytics collects and analyses consumer opinions and convert them into insights and help businesses in identifying areas of customer satisfaction or any customer grievance for the product. It also provides a quick feedback to marketing campaigns, so as to analyse campaign that will be received well by the consumers. Social analytics acts as a new channel between consumers and industries. So the proposed model fulfils the needs of companies by analysing the data efficiently and delivering results. There are many Analytical Tools and Dashboards present in the market such as SocialBro, Tweet

Stats, Twenty feet, Twtrl and, Topsy etc. but are very costly and inefficient.

A. AIM OF THE PROJECT

"SENTIMENTAL ANALYSIS OF SOCIAL MEDIA USING R LANGUAGE AND HADOOP"classifies the polarity of fetched tweets in the form of blog, sentence, or feature/aspect level—whether the expressed opinion in a blog, a sentence or an entity feature/aspect is positive, negative, or neutral.

SCOPE OF THE PROJECT

This technique will improve the efficiency and will be able to address all the needs of the different users. It will help the users to fetch the data from twitter and to know about the sentiments regarding the same. For example: If a user wants to decide whether to watch a particular movie or not then he/she can fetch the data related to that movie and can make decision on the basis of the sentiments.

C. PROBLEM DEFINITION

The available system 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. The available systems are inefficient for Real Time Analytics. The available methods and systems undergo time consuming process and the proposed work eliminates all those drawbacks mentioned above.

II. Steps For Sentimental Analysis

A. Twitter Authentication

The[6] user connects to the Twitter Streaming API using the developer's username and password of twitter account. As soon as the username and password is authenticated and handshake is done with twitter API, it provides some methods through which we crawl the data from twitter server. The resultant stream of tweets is stored into some files that can be used for analysis purpose.

B. Pre-processing

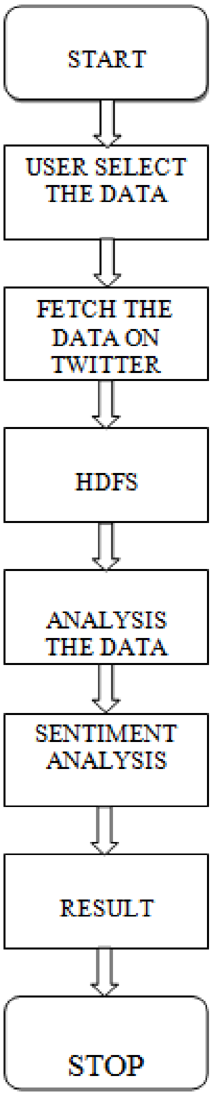
Thetweets data contain a lot of misspelling in words, graphics tags, the presence of slang words etc. So, it is very necessary to do pre-processing of data and removable of all repeated words and hash tags. These words have been corrected to the best possible effort by applying a probabilistic model based on Bayes' theorem which shows roughly 86.67% accuracy [7].

C. Sentimental Analysis

The final step is to check the sentiments of different tweets of every user on the basis of positive score, negative score or in between them. The main objective is to use the stored documents that contain positive and negative words list. The function uses above List and check the similar words that user uses in their sentences. Thus, it is helpful for determining the opinion of the sentiment for the users.

III. Design & implementation Flowchart

A **flowchart** is a type of diagram that represents an algorithm, workflow or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows. This diagrammatic representation illustrates a solution model to a given problem. Flowcharts are used in analysing, designing, documenting or managing a process or program in various fields.



:

Using R Language and Hadoop: Rhadoop[6]

Step1. Installing Rhadoop

As Rhadoop is a connector of R and Hadoop, we need Hadoop and R installed on our machine in the following sequence:

1. Installing Hadoop.
2. Installing R.
3. Setting up environment variables.
4. Installing rJava.
5. Installing Rhadoop

Step2. Load all the Rhadoop and R packages in the R console to setup the system for performing analysis on the data. Start the Hadoop from the Linux terminal and off the safe mode of Hadoop.

Step3. Crawl the tweets using search twitter() method and save them into a file that have an extension of .csv.

Step4. Put the file into the Hadoop HDFS (Hadoop Distributed File System) from where we can access it using some commands that are provided by rhdfs packages.

Step5. Apply the Sentimental function of R language.

Step6. Fetch the required file from the HDFS whose data size is largeusing the command mentioned below**.**

dd< - hdfs.read.text.file("/user/hduser/twitterlist.csv") – (Read file from HDFS)

Step7. Final Step is to calculate the score and compared it with the R language.

IV. Result Obtained

The general sentiment derived from the dataset regarding the three trending topics IPL, Politics and movies were, as follows: a total of 153 tweets were regarded as positive, 92 as negative and 298 as neutral for Instagram. 214 tweets were classified as positive, 147 as negative and 424 as neutral for facebook. As previously mentioned, a magnitude of 0 was considered as neutral valence, greater than 0 was considered as positive valence while less than zero was considered as negative valence. Table I displays a few of the statistics obtained.

College	Ratio of positive to negative tweets
Instagram	1.66
Facebook	1.45

TABLE I. STATISTICS ON THE SENTIMENTS EXTRACTED FROM TWEETS

V. Conclusion

Sentiment analysis is an effective way of classifying the opinions formulated by people regarding any topic, service or product. Automation of this task makes it easier to deal with the massive amount of data being produced by social websites like Twitter on a real-time basis.

Naïve Bayes outperforms Support Vector Machine for the purpose of textual polarity classification which is interesting because the model used by Naïve Bayes is simple (use of independent probabilities) and the probability estimates produced by such a model are of low quality. Yet, the classification decisions made by the Naïve Bayes model portray a good accuracy because each time a decision with the higher probability is being made [8].

VI. References

[1] Twitter Usage/Company Facts, https://about.twitter.com/company

[2] Posting a tweet, https://support.twitter.com/articles/15367-posting-atweet

[3] King R. A., Racherla P. and Bush V. D., What We Know and Don't Know about Online Word-of-Mouth: A Review and Synthesis of the Literature, Journal of Interactive Marketing, vol. 28, issue 3, pp. 167-183, August 2014

[4] BalakrishnanGokulakrishnan, PavalanathanPriyanthan, ThiruchittampalamRagavan,NadarajahPrasath and AShehanPerera," Opinion Mining and Sentiment Analysis on a Twitter Data Stream," The International Conference on Advances in ICT for Emerging Regions, May, 2012

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

[6] Sunny Kumar, Paramjeet Singh and Saveta Rani, "Sentiment Analysis on Social Media,"Computer Science Engineering and TechnologyGianiZail Singh Campus College of Engineering and Technology, GZSCCET Bhatinda, India

[7] Segaran T. and Hammerbacher J., Beautiful Data: The Stories behind Elegant Data Solutions, Beijing: O'Reilly, 2009

[8] Manning C. and Raghavan P., Introduction to information retrieval, New York: Cambridge University Press, 2008.