IMPLEMENTATION OF A NOVEL SENTIMENT ANALYSIS TECHNIQUE FOR TWITTER DATA USING RHADOOP

Thasnim K M[1], M Sudheep Elayidom[2]

[1]*M.Tech Scholar, Division of Computer Engineering, School of Engineering, Cusat, Kerala, India*
[2]*Associate Professor, Division of Computer Engineering, School of Engineering, Cusat, Kerala, India*

Abstract

Sentiment analysis is a method of mining knowledge from sentiments or opinions of others about some specific domain or a problem. It is widely applied in product reviews, movie reviews, Twitter, Micro blog etc. Sentiment analysis or in other words, opinion mining of the tweets available on a micro blogging site such as Twitter offers insights to identify products and services that are relevant to users. Before choosing a product, consumers' usually go for others opinion about a particular product or a brand. Our specific objective is to identify the popularity of different smart phone brands using RHadoop . This is done by identifying the polarity of tweets derived and assign sentiment scores accordingly to each brand. The comprehensive evaluation of users response is found more challenging for some brands compared to others since the availability of tweets varies significantly for various brands over time. This paper uses a much advanced technology that integrates the Big Data framework-Hadoop, and a very popular visualization tool called R using the RHadoop package. From the literature it is being observed that very few works are done using this integrated framework. The problem with existing technology is that it uses only Hadoop but my work overcomes the problem of scalability and speed along with providing visualization techniques using RHadoop.

Keywords: *Product Review, Sentiment Analysis, Opinion Mining, Smart Phone Brands, Polarity of Tweets, RHadoop*

----------***----------

1. INTRODUCTION

Todays world is full of competitions; competition among the companies for marketing same product under different brand names; competition among the consumers for making a better choice while selecting a product or a service and so on. This has lead companies, marketing agencies and even consumers or the users to go for systematic and untraditional ways to link with the market and update with day-to-day pulse. The social media now has emerged into a giant that conquers and rules our day -to day lives . It can be described as a platform that the consumers' utilize to share information about their likes and dislikes, their comments and views on products, brands, services, and any political events or issues. Among different social media sites available like Twitter, Facebook, Google+ etc. Twitter is a popular micro blogging site because we have got access to comments and reviews of persons you don't necessarily know (or will ever know) in person.

Twitter sentiment analysis mainly concentrates on accessing twitter data via twitter streaming API 's and identify the sentiments or opinions by investigating tweets related to products that appear in Twitter. The objective is realized by identifying products and the keywords are passed along with hash tags so that Twitter is searched for these products and product reviews are returned. Following that sentiment analysis is done by extracting opinion expressions. Predefined sentiment words :positive and negative words along with their synonyms are identified and tweets are searched for identifying the sentiments hidden under the thoughts.

Electronic products are a common domain where a number of customer reviews appears. Among the electronic products, mobile phones are a major domain in which customers have a keen interest. Before making a choice on buying a product online, it has become a habit that the users go through reviews about the product. But, going through each and every review is a time consuming and a tedious task. There comes the solution of providing sentiment analysis of reviews so that a user can get the overall analysis or rating of the product by opinion mining each and every tweet related to the product found in a popular micro blogging site such as Twitter. Using this overall analysis the customer can decide whether to buy the product or not. The use of applying sentiment analysis on Twitter product review can also help manufacturers and marketers so that they can review the needs of users and customize or rectify if any defects immediately being notified by the consumers.

2. DATA EXTRACTION

Twitter Data specifically tweets related to mobile phones are extracted via Twitter 'API' (Application Programming Interface). An API is mainly used for a program an application to perform a task, usually by retrieving or modifying data. Usually programs communicate to the Twitter API over HTTP, the same protocol that any browser uses to visit and interact with web pages.

Twitter APIs can be classified as Streaming API, Search API and REST API. Each of these provide tweets according to user requirement. If we want to stream real time tweets as they happen , then they go for **Streaming API.** The **Search API** allows us search old tweets with severe limitations and

the **REST API** allows you to collect user profiles, friends, and followers. Inorder to get access to Twitter REST and Streaming API, first we should register in the Twitter developer tab and need to understand the working of OAuth . Through this OAuth authentication our program gets the permission to make API calls.

2.1 Twitter Search Keywords

In addition to being able to pull in tweets from a @username and Twitter list, search can be done directly via twitter API's using a variety of keywords including search terms and hash tags just as you can with Twitter Advanced search [7]. Simply enter your search query in the Twitter field.

Here are a few examples:

- Twitter search -- containing both "Twitter" and "search"
- "@SAMSUNG" -- containing mentions of @ SAMSUNG
- flight :(with #NYC filter -- containing "flight" and with a negative attitude, tagged #NYC
- sunny -today -- containing "sunny" but not "today"
- "mobiles" -- containing mentions of "mobiles"
- hilarious filter:links -- containing "hilarious" and linking to URLs
- dogs OR cats -- containing either "dogs" or "cats" or both
- #HTC -- containing HTC as hash-tag
- "Happy hour" -- containing the exact phrase "happy hour"

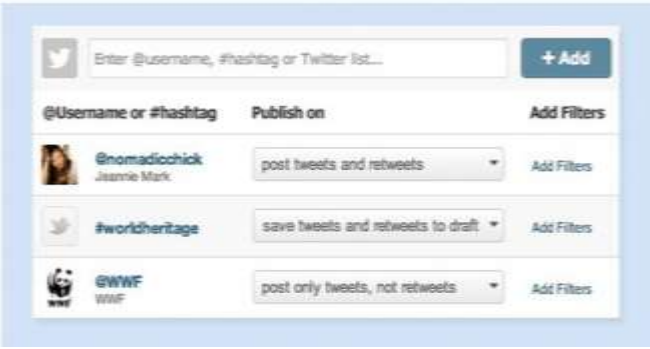

Fig 2.1 Search Tags

3. MAP REDUCE FRAMEWORK

Hadoop is an open source framework that provides parallel processing capabilities along with providing speed and scalability. Here, we concentrate on the integration of a well known statistical analysis and visualization tool named R with Hadoop. R, a very popular statistical analysis tool in data mining, provides a number of in-built functions and packages for various data operations. But the thing is, R will not help loading all the 'Big Data' into main memory. Hence as a solution, Hadoop may be find optimal to load the data as 'Big Data'. For the analysis of dataset, first and foremost thing is that R loads it into the memory, and if it is large enough, it will fail loading dataset with the following exceptions- "cannot allocate vector of size x" [5,13]. Hence, as a solution we can integrate R with Hadoop thereby magnifying the processing power of R on combining it with the power of a Hadoop cluster. By doing so, we can utilize R algorithms along with data processing over map reduce framework.

The advantages of R and Hadoop integration can be summarized as follows: Since statisticians and data analysts frequently use the R tool for data exploration as well as data analytics, Hadoop integration is a big boon for processing large-size data.

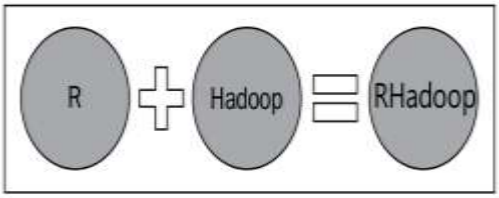

Fig 3.1 RHadoop

Similarly, data engineers who use Hadoop tools, such as system, to organize the data warehouse can perform such logical analytical operations to get informative insights that are actionable by integrating with R tool[5,63]. There are three ways to link R and Hadoop : RHIPE, RHadoop, Hadoop streaming. In this work we are applying RHadoop concept for integrating R and Hadoop.

Two main Hadoop concepts are: **HDFS** and **Map Reduce**. HDFS stands for Hadoop Distributed File System. The MapReduce framework consists of a single master -Job Tracker and one slave- TaskTracker per cluster-node[6]. An entire job is split into tasks and the master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master. When it comes to the concept of Hadoop these core concepts can be transformed as: R Map-Reduce(**rmr)** and R -HDFS (**rhdfs).**

RHadoop

RHadoop is a popular integration tool that allows running a Map Reduce jobs within R. RHadoop is a collection of five R packages that allow users to manage and analyze data with Hadoop. The packages include rhdfs, rmr2, rhbase, plyrmr and ravro. Of these rmr2 and rhdfs are basically needed to set up RHadoop.

- **rmr2** –a package that allows R developer to perform statistical analysis in R via Hadoop Map-Reduce functionality on a Hadoop cluster. We need to install this package on every node in the cluster.
- **rhdfs** – an interface between R and HDFS. This package provides basic connectivity to the Hadoop Distributed File System. R programmers can browse, read, write, and modify files stored in HDFS from within R

Using these packages Map and Reduce tasks can be easily accomplished in R without making considerable changes in R code.

4. LEXICON-BASED RHADOOP SENTI_SCORE (LRS) ALGORITHM

Twitter the second biggest social network generates 347,222 Tweets each minute or 21 million Tweets per hour. Hence data processing becomes an extreme overhead. An excellent and sensible solution to the problem is implementation in Map-Reduce Framework.

Our aim is to design an and Implement an effective methodology for customer satisfaction in mobile phone purchase by sentiment analysis of tweets. For that we are introducing an implementation technique which involves RHadoop Known as **LRS** Algorithm.

Lexicon-based Rhadoop Senti_Score (LRS)Algorithm

Input : Tweets extracted from Twitter related to mobile phone reviews

Output : Calculation of Sentiment score for various mobile products

Steps:

1. For each *phones* in given list
2. Load n no. of ***Tweets*** from Twitter API
3. Convert *Tweets* from Json format to text format
4. Set up a Opinion-Lexicon-English dictionary; a dictionary consisting of positive and negative sentiment words
5. Clean the extracted *Tweets* for each phone
6. Initialize a map-reduce environment at the back-end
7. Copy cleaned *Tweets* to hdfs
8. Input key-value pair is mapped using mapper function sentiment score is calculated using Opinion-Lexicon-English
9. Scores are reduced using Reducer function
10. Reduced Data is passed from dfs to R
11. Final data frame is displayed

The entire sentiment analysis process can be split into meaningful subtasks . Firstly, Extraction of tweets from twitter API . Secondly, Data Preparation which involves data preprocesing to extract sentiment words.Thirdly identifying a suitable opinion lexicon dictionary which identifies clearly positive and negative sentiment words. Finallly R and Hadoop Integration which does the overall task of sentiment analysis starting from moving the data from local directory to distributed file system(**dfs**), spliting the tasks , replicating the data nodes , assign each task to task tracker which is monitored by job tracker and finally calculation of sentiment score by dictionary lookup and applying the normalization process. Introducing map-reduce frame work along with a popular computational platform and a visualization tool such as R (RHadoop) helps us to process a huge amount of Tweets from Twitter in a few seconds.

This methodology of using RHadoop for twitter sentiment analysis for evaluating the popularity of different mobile phone brands is a novel approach which decreases the processing speed considerably and is scalable and provides high performance computing. In the RHadoop context R will R will take care of data analysis operations along with data visualization capabilities [5,14]and Hadoop will take care of Big Data storage as well as increased computation power against distributed data.

5. RESULTS AND ANALYSIS

Given below is a snap shot of partial results obtained while testing the Sentiment analysis App using the **LRS** algorithm. The system is tested in Ubuntu 14.04 OS. The App is developed using RHadoop which is an integration of R and Hadoop. R version 3.2.3 and Hadoop 2.7.2 is used for our App.

The scores assigned range from -5 to +5 based on the intensity of sentiment portrayed in the tweet analyzed using Opinion-Lexicon-English dictionary. Scores ranging from -5 to +5 shows human response towards various mobile phone products via Twitter. A screen shot of partial output obtained is shown to analyze the response towards phones such as Note5, Nexus5, iPhone6, One-plus etc. Score -5 shows worst comment(most negative) and Score +5 shows the best comment (most positive). Given below is a screen-shot of table showing frequency distribution for each score. Score '0' shows that the sentiment or opinion is neutral.

Table 6.1: Table showing phones,scores and their frequencies

	phone	score	count
1	note5	0	393
2	note5	1	469
3	note5	2	25
4	note5	3	5
5	note5	4	2
6	nexus5	0	900
7	nexus5	1	49
8	nexus5	2	13
9	nexus5	3	8
10	note5	-1	54
11	note5	-2	17
12	note5	-3	2
13	note5	-5	1
14	iphone6	0	695
15	iphone6	1	139
16	iphone6	2	27
17	iphone6	3	2
18	nexus5	-1	22
19	nexus5	-2	6
20	nexus5	-3	2

6. CONCLUSION

Sentiment Analysis is now a days a vast field of measuring human interaction and response towards various fields and commodities. Among them analyzing human response towards various products available in the market is a very popular means of measuring market response for a new product so that manufacturers can customize or rectify if any defects immediately being notified by the users. The

algorithm **LRS** discussed is an efficient method for Sentiment Analysis of mobile phone products. From the experiments which I have conducted it is being observed that very few works are done using this integrated framework.

For the work, we have chosen twitter as the medium for data collection and extracted tweets posted in Twitter related to product reviews. For the time being, from the entire product domain we have taken only mobile phone reviews for our case study. This information gained can be used for analyzing immediate market reactions and it is an appropriate tool both for the consumers as well as manufacturers. The problem with existing technology is that Twitter sentiment analysis is done only using R or only using Hadoop and lacks the integration of these two. But my work overcomes the problem of scalability and speed along with providing visualization techniques using the integrated tool RHadoop.

REFERENCES

[1] http://www.ideahatching.com/2011/06/whats-your-twitter-handle-2/

[2] Deepali Arora, Kin Fun Li and Stephen W.Neville, "Consumers' Sentiment Analysis of popular phone brands and operating system preference using Twitter data: A feasibility study", *IEEE 29th International Conference on Advanced Information Networking and Applications*, 2015.

[3] K. Ann Jurek, Yaxin Bi and Maurice Mulvenna, "Twitter Sentiment Analysis for Security-Related Information Gathering",*IEEE Joint Intelligence and Security Informatics Conference*, 2014.

[4] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede "Lexicon-Based Methods for Sentiment Analysis", *Association for Computational Linguistics 2011*, Volume 37, Number 2 .

[5] Vignesh Prajapati, "Big Data Analytics with R and Hadoop," *Copyright © 2013 Packt Publishing*, ISBN 978-1-78216-328-2.

[6] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

[7] R Nithish, S Sabarish, M Navaneeth Kishen, A.M. Abirami and Dr. A. Askarunisa ," An Ontology based Sentiment Analysis for mobile products using tweets", F*ifth International Conference on Advanced Computing(ICoAC)*, 2013.

[8] Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde, "Real Time Sentiment Analysis of Twitter Data Using Hadoop",*(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 5 (3), 3098 3100, 2014.

[9] http://help.rebelmouse.com/hc/en-us/articles/201810397-Twitter-Search-and-Hashtags.

[10] Carlo Lipizzi , Luca Iandoli , Jos Emmanuel , Ramirez Marquez ,"Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers reactions to the launch of new products using Twitter streams ", *International Journal of Information Management*, 35 (2015) 490503, Elsevier, 2015.