IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Sentiment Analysis of Android Application Feedback Data using Hadoop

Prakash R. Andhale[1], Prof. S.M. Rokade[2]

M.E. Student, Department of Computer Engineering, SVIT, Nasik, India[1]

Assistant Professor, Department of Computer Engineering, SVIT, Nasik, India[2]

Abstract: In today's highly developed word every person in society are express themselves via various number of medium on the internet every minute. At the every minute the large number of heterogeneous data is generated. This data may be in text format which are retrieved from forums and the social media websites. This data is also called as unstructured data. The thought of users related to the particular application or the topics like politics, economics, social affairs and products. These are extract from various type of technologies for the highly importance to make the forecast for one-to-one consumer marketing. In this paper we propose the analysis of the "Is selected android application genuine or fraud by android application users through their in order to mine what they think. Hence we are using hadoop framework for sentimental analysis which will process the large amount of data on a hadoop cluster faster.

Keywords: Sentiment analysis, Information extraction, Big Data, Hadoop, Stemming Algorithm, Tokenization, NLP.

INTRODUCTION

Sentiment analysis is nothing but the extract the of the users or the opinion mining. The process of computationally identifying and categorizing inhibits in a token of text, especially in order to determine the writer's attitude towards a particular topic or product. Sentiment Analysis is the process of detecting the contextual polarity of text. In other words, it determines whether a piece of writing is positive, negative or neutral.

Android Application and Ratings Data:
Android application feedback data is nothing but the of particular application data received from users. The million number of application are available on Google play store and the each application hits the huge number of and ratings in the range of Zettabyte per years. This huge amount of raw data can be used for industrial or business purpose by organizing according our requirements and process.

About this Paper:
In this paper, we are going to implement a system in Hadoop which analyses Android application feedback data (i.e and ratings data) where cluster of nodes will be formed. Android application and ratings data is in the form of comments and number which are nothing but sentiments that is opinions, feelings of people. This data will be collected by using Google API.

By analyzing this data, our system will give output in the form of positive, negative and neutral counts of and finally predict whether given application is genuine or fraud. In this case, it makes the use of data dictionary for classifying the data. This data can be used further according to particular application. And this analyzed data can be represented in the form of line and dotted graphs.

Motivation:
Today we are living in the world which is surrounded by 99% of data. There are different microblogging sites where users express their views about different products these views are nothing but opinions of people and it will go waste if it is not used in proper way so there is a need to use opinions of people in improving productivity, usefulness, functionality of particular product or application or technique or any entertainment resource.

Hence, there is a need to develop a product which cans analyses opinions of people. This product will be useful in increasing market value of industries as well as satisfy needs of customers.

BRIEF DESCRIPTION

Need of Sentiment Analysis:
Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Social media monitoring tools like Brand-watch Analytics make that process quicker and easier than ever before. The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election. The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

increase in negative feedback to the music used in one of their television adverts.

OBJECTIVES

- **Data Retrieval:** The huge amount of data is collected from Google API.
- **Storage:** This data is stored in Hadoop distributed file system for further processing like mapper in Map-Reduce programming approach. The data stored in HDFS system.
- **Data Processing:** Data collected over a period of time is processed by using java and distributed processing software framework developed by Apache Hadoop and using map reduce programming model and Apache hive frame work.
- **Data Analysis:** The output obtained from reducer phase is analyzed.
- **Data Representation:** Representation of classified data in the form of graphs.
- At the end we will get the outcome in the form of classified that is Positive, Negative and Neutral and on the basis of this finally predict the result whether the selected android application is genuine or fraud.

This project will mainly analyse the predefined stored Android application feedback data and classify it based on polarity.
Analysis of data consists following steps:

1. Tokenization:
All the words in a comment are broken down into tokens. This is the tokenization process. For example, '@Sameer That is an awesome application!' is broken down into individual tokens such as '@Jack', 'That', ‗is‘, 'an', 'awesome', 'car'. Emoticons, abbreviations, hashtags and URLs are recognized as individual tokens. Each word in a comment is separated by a space. Therefore, on encountering a space, a token is identified.

2. Normalization:
The normalization process verifies each token and performs some computing based on what kind of token it is.

- If the token is an emoticon, its corresponding polarity is taken into account by searching the emoticon dictionary.
- If the token is an acronym, it is checked in the acronym dictionary and the full form is stored as individual tokens.
- Intensifiers such as 'AWESOME' are converted into lowercase and the token is stored as 'awesome'.
- Spelling of character repetitions such as 'veryyyyy' are first corrected into 'very' and then stored as 'very'.
- The normalization process also discards all those tokens which, in no way, contribute to the sentiment of such tokens are called stop word. It also discards URL‘s.

For analysing the reviews, we have to take polarity into consideration using various types of dictionaries.

1) Lexical Dictionary: It mainly consists of most of the English words which will help us to analyse the reviews by matching the word in the reviews with the words in the lexical dictionary. It also consists of idioms, phrases, headwords and multiword.
2) Acronym Dictionary: It is used to expand all the abbreviations and acronyms which will further generate words which can be analysed using lexical dictionary.
3) Emoticon Dictionary: Reviews containing emoticons can be analysed by using this dictionary. Emoticons are basically the textual portrayal of the reviews mode which conveys some meaning.
4) Stop Words Dictionary: These are the words in the reviews which do not have any polarity and they need not be analysed. So they are eliminated and tagged as stop words. We maintain a dictionary with the list of all stop words for example able, are, both, etc.

Sentiment Classifier:
The reviews are broken down into tokens where each token is assigned polarity which is a floating point number ranging

A. Positive Reviews: Positive reviews are the reviews which show a good or positive response towards something. For example reviews such as-
It was an inspiring movie!!!‖ or ―Best movie ever‖.

B. Negative Reviews: Negative reviews can be classified as the reviews which show a negative response or oppose towards something.
For Example reviews such as ―Waste of time‖ or ―Worst movie ever‖.

C. Neutral Reviews: Neutral reviews can be classified as the reviews which neither show a support or appreciate anything nor oppose or depreciate it. It also includes reviews which are facts or theories.
For example reviews such as ―Earth are round‖.

Classification:
At the end system will classify the android application feedback data into Positive, Negative, Neutral reviews with the help of data dictionary

RELATED WORK

Over the last decade, there has been an explosion of work exploring various aspects of sentiment analysis: detecting subjective and objective sentences; classifying sentences as positive, negative, or neutral; detecting the person expressing the sentiment and the target of the sentiment; detecting emotions such as joy, fear, and anger; visualizing sentiment in text; and applying sentiment analysis in health, commerce, and disaster management. Surveys by Pang and Lee (2008) and Liu and Zhang (2012) give a summary of many of these approaches.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Semi-supervised and automatic methods have also been proposed to detect the polarity of words. Hatzivassiloglou and McKeown (1997) proposed an algorithm to determine the polarity of adjectives. SentiWordNet (SWN) was created using supervised classifiers as well as manual annotation (Esuli & Sebastiani, 2006). Turney and Littman (2003) proposed a minimally supervised algorithm to calculate the polarity of a word by determining if its tendency to co-occur with a small set of positive seed words is greater than its tendency to co-occur with a small set of negative seed words. Mohammad, Dunne, and Dorr (2009) automatically generated a sentiment lexicon of more than 60,000 words from a thesaurus. We use several of these lexicons in our system. In addition, we create two new sentiment lexicons from using hashtags and emoticons.

Since manual annotation of data is costly, distant supervision techniques have been actively applied in the domain of short informal texts. User-provided indications of emotional content, such as emoticons, emoji, and hashtags, have been used as noisy sentiment labels. For example, some user feedback their with emoticons as labeled data for supervised training. Emoticons such as :) are considered positive labels of the comment and emoticons such as :(are used as negative labels or use certain seed hashtag words such as #cute and #sucks as labels of positive and negative sentiment.

PROPOSED SYSTEM

Architecture:

Following gives the description of blocks used in system architecture.

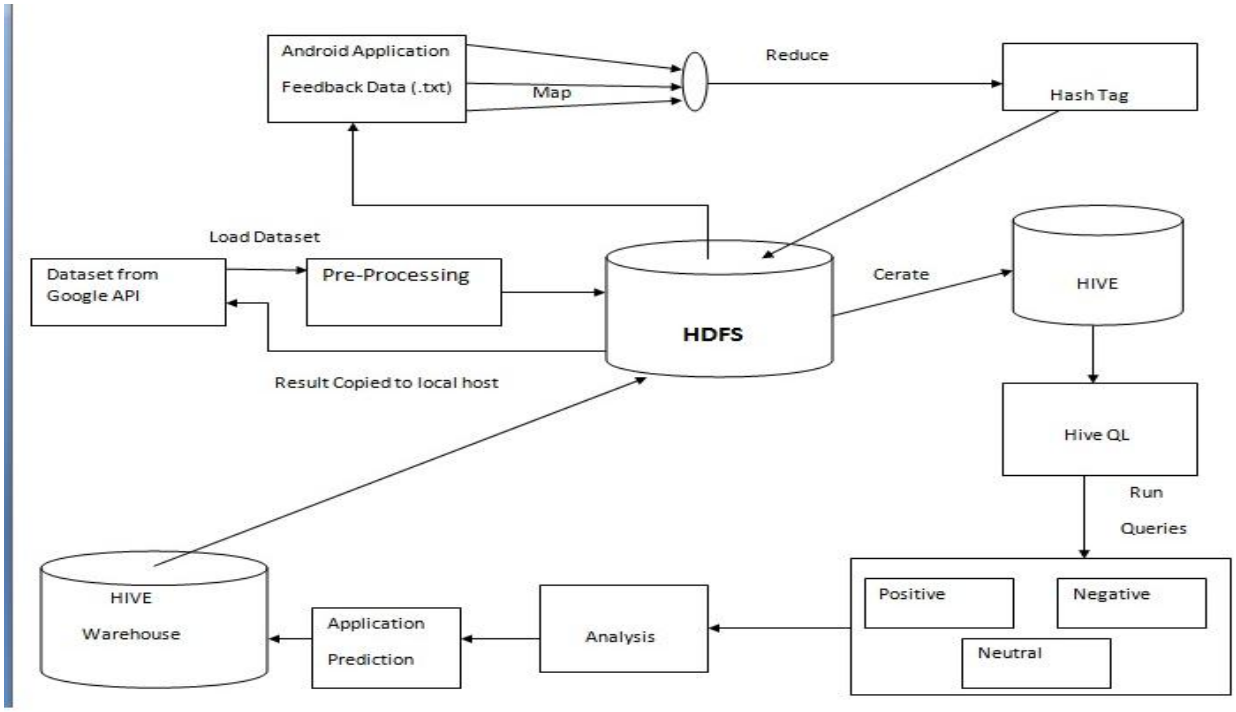


Fig.1 Overall Architecture of Proposed System

- Data Pre-Processing: The dataset which is used in our proposed system initially is in unstructured form. By using string tokenization process and Porter Stemming algorithm convert this unstructured data into structured form.
- HDFS: Stored structured data in hadoop distributed file system for further processing like Map-Reduce for sentiment analysis. The positive, negative and neutral word base analyses are done in map-reduce process.
- HIVE ETL Tool: The Hive is a ETL tool use for internal query processing.
- Analysis: In analysis module there are various type of analysis will be done on the basis of previous word base sentiment analysis and users given rating.
  - Application Leading Event: It shows rank of the selected application.
  - Ranking Based Analysis: It shows how many days the selected application is in top 300 application.
  - Session Based Analysis: It shows how many sessions are done for particular application.
  - Rating based Analysis: It shows the rating which is given by users to particular application.
  - Cosine Similarity: It compares the result to the expected result.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

- HIVE warehouse: This is a HIVE warehouse, where store the final prediction for each application.

EXPERIMENTAL SETUP

For the experiment, we take the android application feedback data as input. Initially it is in unstructured format (i.e text form). We are considering this text file as dataset. This dataset is freely available on Google API. All experiment is performed using i3 processor 4GB RAM. The Operating system is Ubuntu 12.4 LTS. We are using Java for programming. This will give the result of reviews in form of positive, negative and neutral comments as follows.

App Name	Total Review Extracted	Positive Sentiment	Negative Sentiment	Precision Positive Sentiment	Precision Negative Sentiment
Instagram	340	289	51	0.85	0.15
Twitter	478	348	130	0.72	0.27
Snapchat	148	72	76	0.48	0.51
Facebook Lite	389	247	142	0.63	0.36
MX Player Pro	1047	748	299	0.71	0.28
VLC for Android	145	41	104	0.28	0.71
poker	79	19	60	0.24	0.75

Fig.2 Dataset after Processing

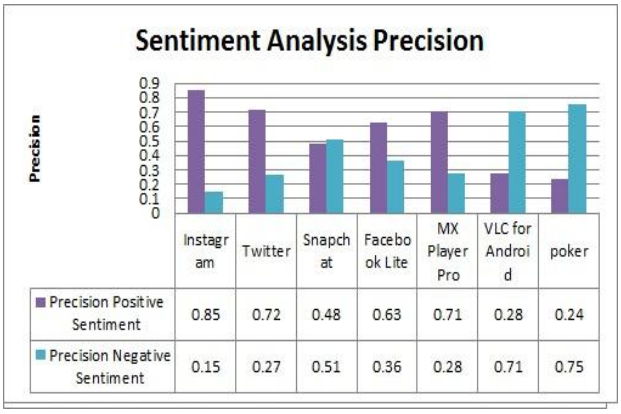


Fig.3 (a) Sentiment Analysis precision

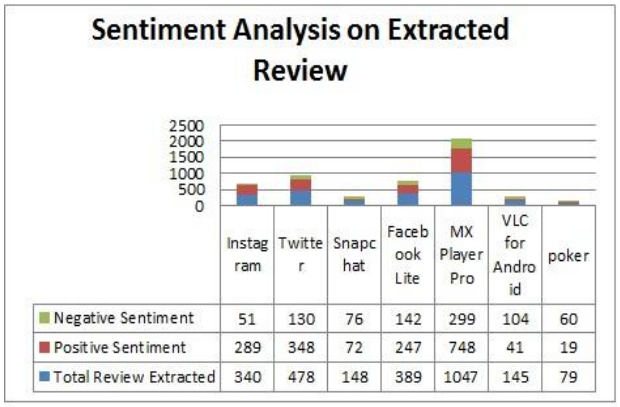
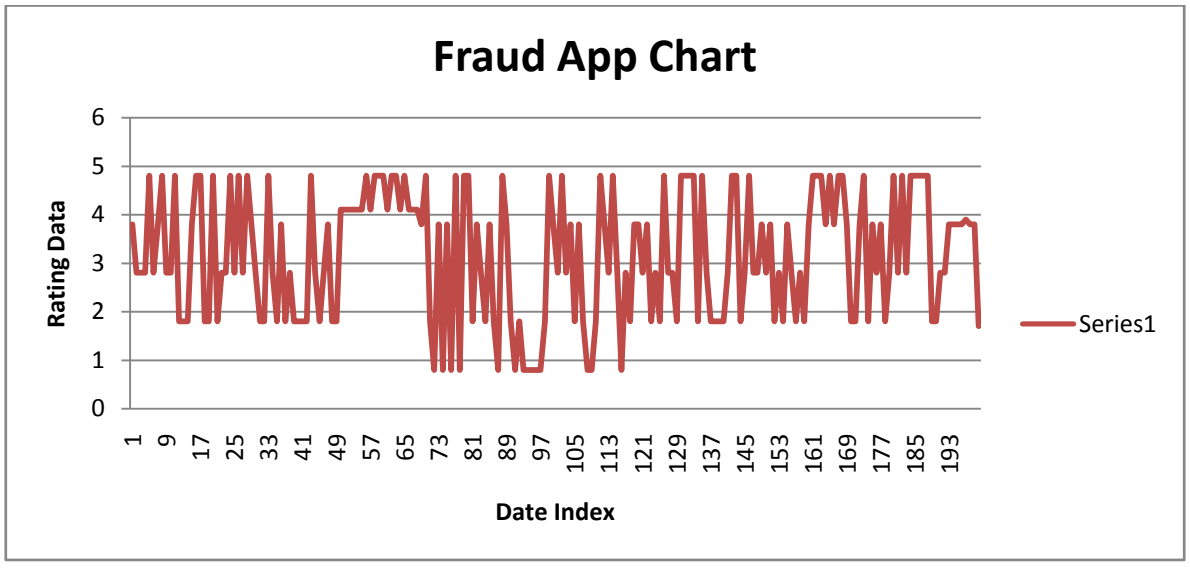


Fig (b) Sentiment analysis on extracted Review

Application Leading Event: It shows rank of the selected application.



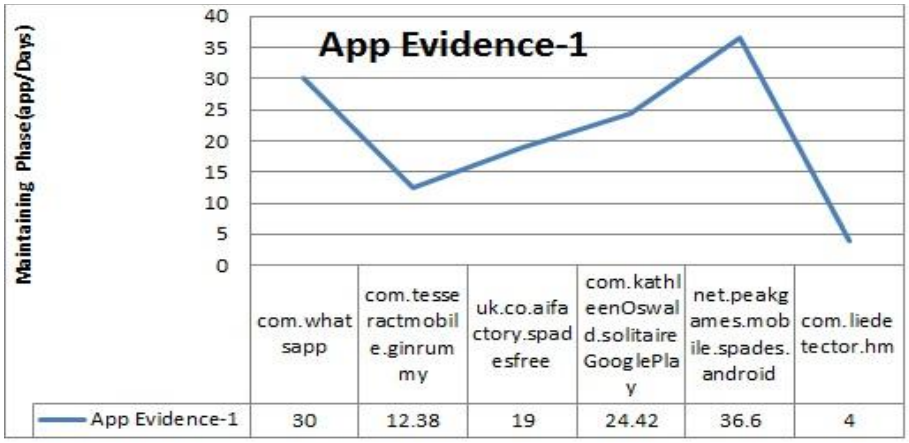
DOI 10.17148/IJARCCE.2016.51206

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

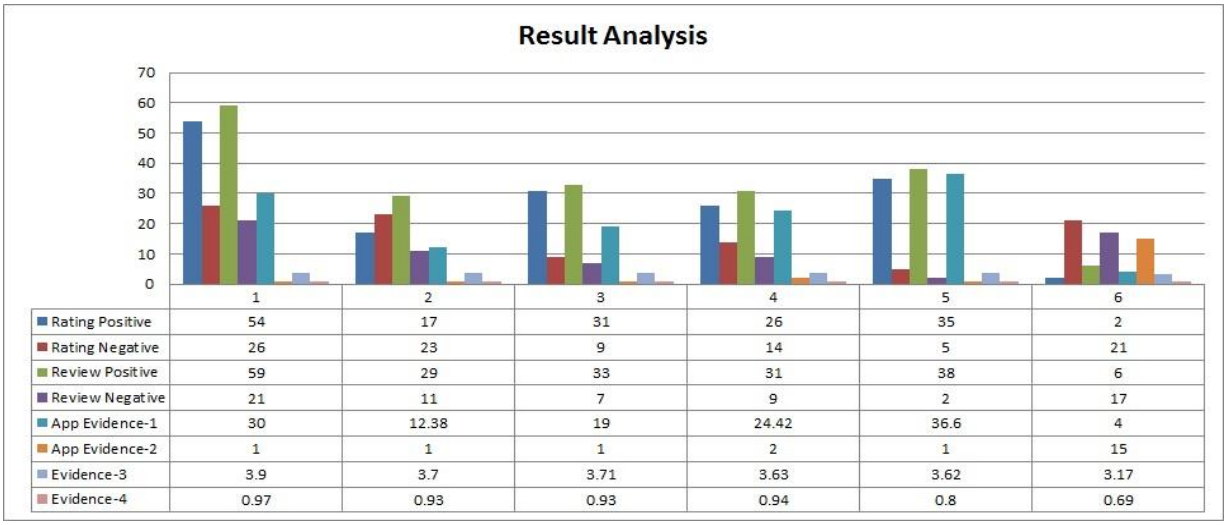
IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Session Based Analysis: It shows how many sessions are done for particular application.



Result Analysis: Combine analysis of the android application prediction as follows.



	1	2	3	4	5	6
Rating Positive	54	17	31	26	35	2
Rating Negative	26	23	9	14	5	21
Review Positive	59	29	33	31	38	6
Review Negative	21	11	7	9	2	17
App Evidence-1	30	12.38	19	24.42	36.6	4
App Evidence-2	1	1	1	2	1	15
Evidence-3	3.9	3.7	3.71	3.63	3.62	3.17
Evidence-4	0.97	0.93	0.93	0.94	0.8	0.69

Simulation Results:

1. Pre-Processing: In this firstly we convert the unstructured file into structured form as shown in fig.4.

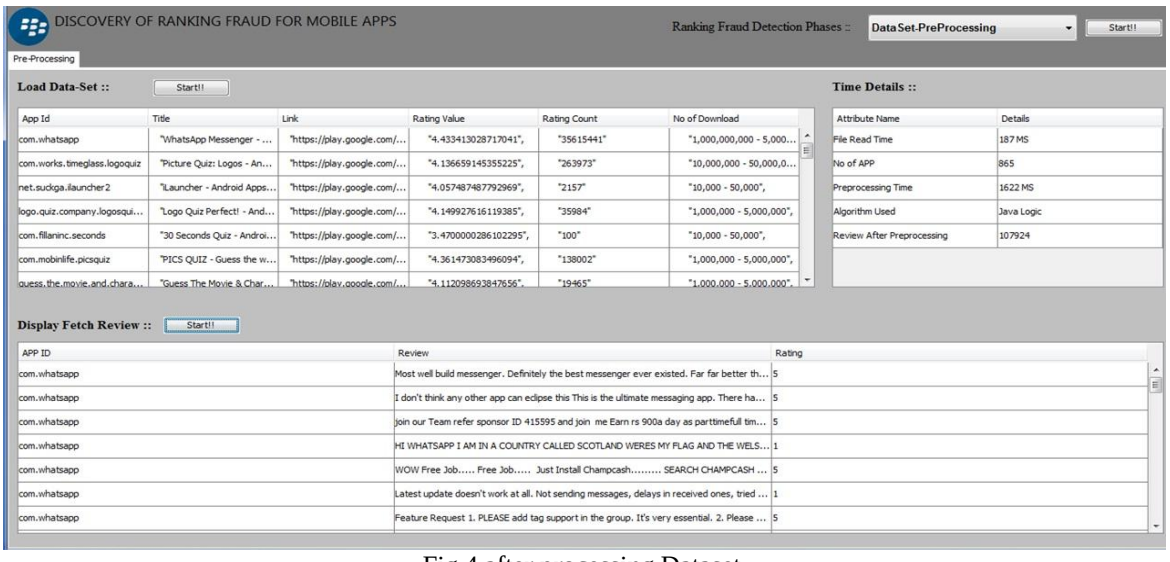


Fig.4 after processing Dataset

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

2. Map-Reduce: In this module word based sentiment analysis will be done.

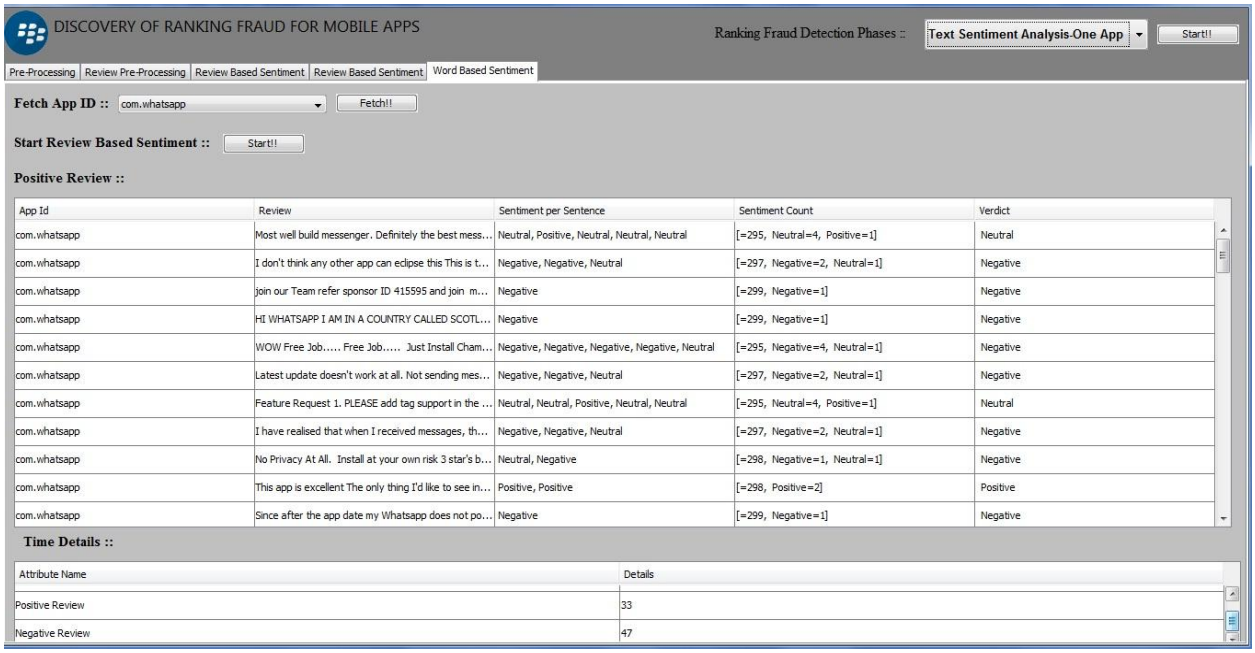
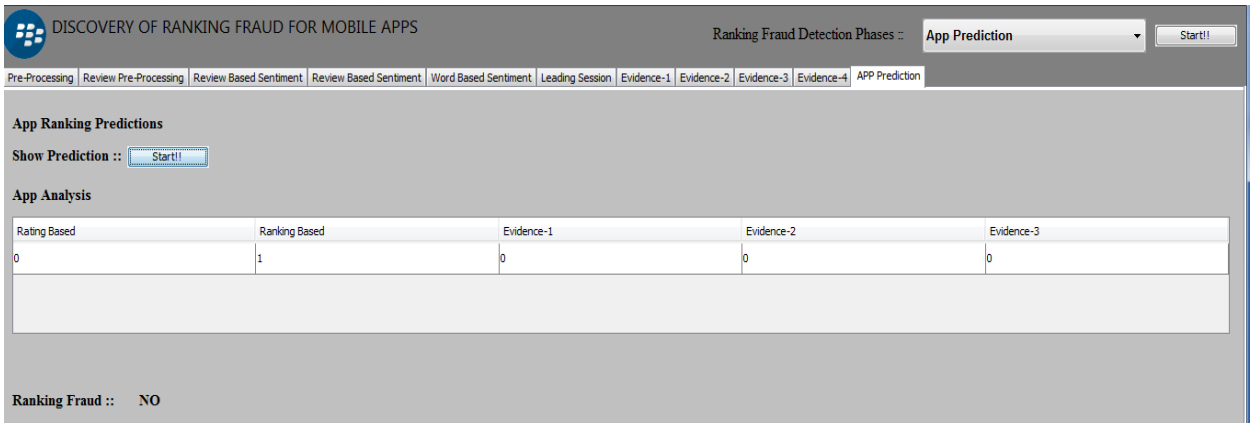


Fig.5 Map-Reduce

3. Result:



CONCLUSION AND FUTURE WORK

This project will give us hands on experience of handling and parallel processing of huge amount of data. Data collection process will introduce us to Java Google API. We will get exposure to work with prominent parallel data processing tool: Hadoop Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyses growth rapidly. This project will help us not only to gain knowledge about installation and configuration of hadoop distributed file system but also map reduce programming model. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any – the ability to analyses sentiment, or sentiment analysis. The future of this data analysis field is vast. This project not only analyses the sentiments of the user but also computes other results like the user with maximum friends/followers, top application etc. hence hadoop can also be effectively used to compute such results in order to determine the current trends with respect to particular topics. This can be very useful in the marketing sector.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE,Gong-Qing Wu, and Wei Ding, Senior Member, IEEE "Data Mining with Big Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1 97-107, Jan-2014.

[2] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang, ―SentiView: Sentiment Analysis and Visualization for Internet Popular Topics‖, IEEE Transactions On Human-Machine Systems, Vol. 43, No. 6, November 2013 .

[3] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, ―Sentiment Analysis of Twitter Data‖, Department of Computer Science, Columbia University .
[4] Jianshu Weng, Ee-Peng Lim, Jing Jiang, Qi He, ―TwitterRank: Finding Topic-sensitive Influential Twitterers‖, WSDM'10, February 4–6, 2010, New York City, New York, USA Copyright 2010 ACM.
[5] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, ―Twitter Sentiment Analysis: The Good the Bad and the OMG!‖, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Rushabh Mehta, Dhaval Mehta, Disha Chheda, Charmi Shah and Pramila M. Chawan, ―Sentiment Analysis and Influence Tracking using Twitter‖ in International Journal of Advanced Research in Computer Science and Electronics Engineering, Vol 1, Isuue 2, May 2012.
[6] Bo Pang and Lillian Lee, ―Opinion Mining and Sentiment Analysis‖, Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) Aditya Pal & Scott Counts, ―Identifying Topical Authorities in Microblogs‖, WSDM'11, February 9–12, 2011, Hong Kong, China, Copyright 2011 ACM.
[7] Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution of conserved relational states in dynamic networks, Knowledge and Information Systems, December 2012, Volume 33, Issue 3, pp 603-630.
[8] J. Mervis, "U.S. Science Policy: Agencies Rally to Tackle Big Data," Science, vol. 336, no. 6077, p. 22, 2012.
[9] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

BIOGRAPHIES

Mr. Prakash Rangnath Andhale has received B.E degree in Computer Engineering, Nasik. He is currently pursuing Master Degree in Computer Engineering at SVIT, Chincholi, Nasik, India .His research interest includes Data Mining, DBMS.

Prof. S.M. Rokade presently working as assistant professor in the Department of Computer Engineering at SVIT, Chincholi, Nasik.