International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 4 Issue 5, May 2015

Market Sentiment Analysis for Popularity of Flipkart

Mr. Sagar Nadagoud
Department of Computer Science and Engineering,
Channabasaveshwara Institute of Technology
Gubbi, Karnataka, India.

Mr. Kotresh Naik.D
Department of Information Science and Engineering,
Channabasaveshwara Institute of Technology
Gubbi, Karnataka, India.

Abstract —**as of now we know present industries and some survey companies are mainly taking decisions by data obtained from web. As we see WWW is a rich collection of data that is mainly in the form of unstructured data from which we can do analysis on those data which is collected on some situation or on a particular thing. In this paper, we are going to talk how effectively sentiment analysis is done on the Flipkart data which is collected from the Twitter using Flume.**

Twitter is an online web application which contains rich amount of data that can be a structured, semi-structured and un-structured data. We can collect the data from the twitter by using BIGDATA eco-system using online streaming tool Flume.

And doing analysis on Twitter is also difficult due to language that is used for comments. And, coming to analysis there are different types of analysis that can be done on the collected data. So here we are taking sentiment analysis, for this we are using Hive and its queries to give the sentiment data based up on the groups that we have defined in the HQL (Hive Query Language). Here we have categorized this sentiment analysis into 3 groups like tweets that are having positive, neutral and negative comments

Keywords - **Analysis, BIGDATA, Flipkart, Comment, ,Flume, Hive, HQL, Sentiment Analysis, Structured, Standard Core NLP, Data Dictionary, Semi-Structured, Twitter, Tweets, Un-Structured, WWW (Word Wide Web).**

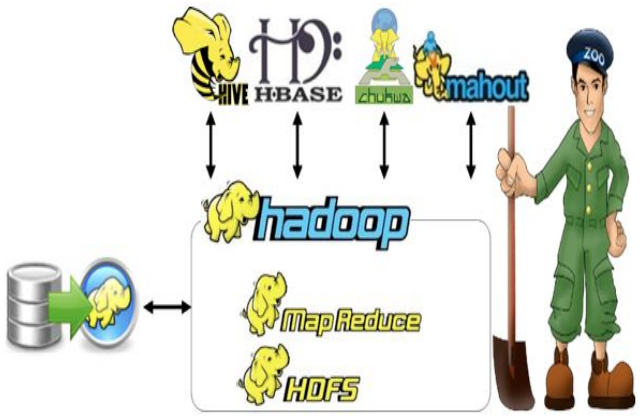
I. INTRODUCTION

From 20th century onwards this WWW has completely changed the way of expressing their views. Present situation is completely they are expressing their thoughts through online blogs, discussion forms and also some online applications like Facebook, Twitter, etc.

If we take Twitter as our example nearly 1TB of text data is generating within a week in the form of tweets. So, by this it is understand clearly how this Internet is changing the way of living and style of people. Among these tweets can be categorized by the hash value tags for which they are commenting and posting their tweets.

So, now many companies and also the survey companies are using this for doing some analytics such that they can predict the success rate of their product or also they can But, to calculate their views is very difficult in a normal way by taking these heavy data that are going to generate day by day.

Figure 1: Describes clearly Apache Hadoop Ecosystem.



The above figure shows clearly the different types of ecosystems that are available on Hadoop so, this problem is taking now and can be solved by using BIGDATA [13] Problem as a solution.

- And if we consider getting the data from Twitter [1] one should use any one programming language to crawl the data from their database or from their web pages.
- Coming to this problem here we are collecting this Flipkart data by using BIGDATA online streaming Eco System Tool known as Flume
- And also the shuffling of data and generating them into structured data in the form of tables can be done by using Apache Hive [7].

II. RELATED WORK

A. What is Sentiment Analysis

Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to obtain writers feelings expressed in positive or negative comments, questions and re-quests, by analysing a large numbers of documents.

For example: I am so happy today, good morning to everyone", is a general positive text. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall functionality of a document. Sentiment analysis is also known as opinion mining.

Basically, Sentiment Analysis is the task of identifying whether the opinion expressed in a text is Positive or Negative. Natural language processing (NLP) is a held of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.

B. Need of Sentiment Analysis

According to a recent statistics by the Social Media tracking company Technorati, four out of every users of Internet use social media in some form. This includes friendship networks, blogging and micro-blogging sites, content and video sharing sites etc. It is worth observing that the World Wide Web has now completely transformed into a more participative and co-creative Web.

It allows a large number of users to contribute in a variety of forms. The fact is that even those who are virtually novice to the technicalities of the Web publishing are creating content on the Web. In fact the value of a Website is now determined largely by its user base, which in turn decides the amount of data available on it. It may perhaps be true to say that Data is the new Intel inside.[1]

For example, a user looking for a hotel in a particular tourist city may prefer to go through the reviews of available hotels in the city before making a decision to book in one of them. Or a user willing to buy a particular model of digital camera may be look at reviews posted by many other users about that camera before making a buying decision.

Sometimes users prefer to write their experiences about a product or service as form of a blog post rather than an explicit review. However, in both case the data is basically textual. Popular sites like carwale.com, imdb.com are now full of user reviews, in this case reviews of cars and movies respectively. [3].

Fortunately we have a solution to this information overload problem which can present a comprehensive summary result out of a large number of reviews. The new Information Retrieval formulations, popularly called sentiment classifiers, now not only allow to automatically label a review as positive or negative, but to extract and highlight positive and negative aspects of a product/ service.

Sentiment analysis is now an important part of Information Retrieval based formulations in a variety of domains. It is traditionally used for automatic extraction of opinions types about a product and for highlighting positive or negative aspects/ features of a product.

It is widely believed that Sentiment analysis is needed and useful. It is also widely accepted that extracting sentiment from text is a hard semantic problem even for human beings. So in general, Sentiment Analysis will be useful for extracting sentiments available on Blogging sites, Social Network, Discussion Forum in order to been t both company and customer/user.

C. Existing System

As we have already discussed about the older way of getting data and also performing the sentiment analysis on those data. Here they are going to use some coding techniques for crawling the data from the twitter where they can extract the data from the Twitter web pages by using some code that may be written either in JAVA, Python etc. For those they are going to download the libraries that are provided by the twitter guys by using this they are crawling the data that we want particularly. [1]

After getting raw data they will filter by using some old techniques and also they will find out the positive, negative and moderate words from the list of collected words in a text file. All these words should be collected by us to filter out or do some sentiment analysis on the filtered data.[2],[5].

These words can be called as a dictionary set by which they will perform sentiment analysis. Also, after performing all these things and they want to store these in a database and coming to here they can use RDBMS [12] where they are having limitations in creating tables and also accessing the tables effectively.

III. IMPLEMENTATION

Here we are implementing sentimental analysis for flipkart data which is fetched from twitter blog, the problem is sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level about Flipkart. Whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral of Flipkart. To implement an algorithm for automatic classification of text into positive, negative or neutral. Sentiment Analysis to determine the attitude of the mass is positive, negative or neutral towards subject of interest. Graphical representation of the sentiment in form Pie-Chart.

A. Proposed System

As it can have seen existing system drawbacks, here we are going to overcome them by solving this issue using Big Data problem statement. So here we are going to use Hadoop and its Ecosystems, for getting raw data from the Twitter we are using Hadoop online streaming tool using Apache Flume [11]. In this tool only we are going to configure everything that we want to get data from the Twitter.[3] For this we want to set the configuration and also want to define what information that we want to get form Twitter. All these will be saved into our HDFS (Hadoop Distributed File System)[10] in our prescribed format. From this raw data we are going to create the table and filter the information that is needed for us and sort them into the Hive Table.

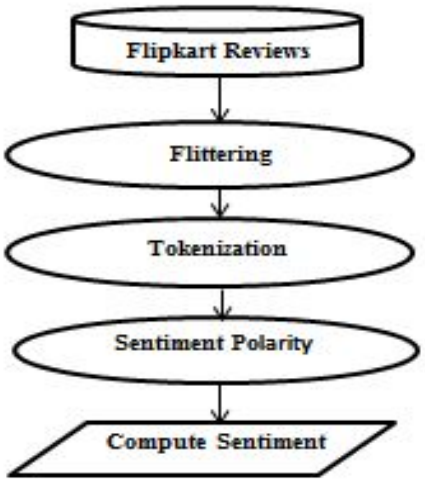
And from that we are going to perform the Sentiment Analysis by using some UDF's (User Defined Functions) by which we can perform sentiment analysis by taking Stanford Core NLP[9] as the data dictionary so that by using that we can decide the list of words that coming under positive, moderate and negative.

a. Natural language Processing Approach:

Natural Language processing approach uses Data Dictionary. This consists of positive, negative words for each of the term occurring in document. The implementation done by extracting the adjectives out of the sentence and then searching it in the dictionary to find out its positive, negative score. In this way the total net score of the sentence is calculated and whichever is greater (either positive or negative) becomes the review for the sentence.

Following figure shows the basic implementation architecture of Sentiment Analysis using Natural Language Processing Approach.

Figure 1: Sentiment Analysis using NLP Approach



b. User defined Functions (UDF's):

Clean up tweets (Filtering):

```
CREATE VIEW tweets_simple AS
SELECT
 id, cast (from_unixtime( unix_timestamp(concat( '2015 ',
substring(created_at,5,15)), 'yyyy MMM dd hh:mm:ss')) as
timestamp) ts,
text,user.time_zone
FROM tweets_raw;
```

Separate tweets into different tokens:

```
Create view l1 as select id, words from tweets lateral view
explode (sentences (lower (text))) dummy as words;
```

Separate tweets into individual words that comes for same tweet id:

```
Create view l2 as select id, word from l1 lateral view
explode (words) dummy as word;
```

For assigning polarity:

```
Create view l3 as select
id,
l2.word,
case d.polarity
when 'negative' then -1
when 'positive' then 1
else 0 end as polarity
from l2 left outer join dictionary d on l2.word = d.word;
```

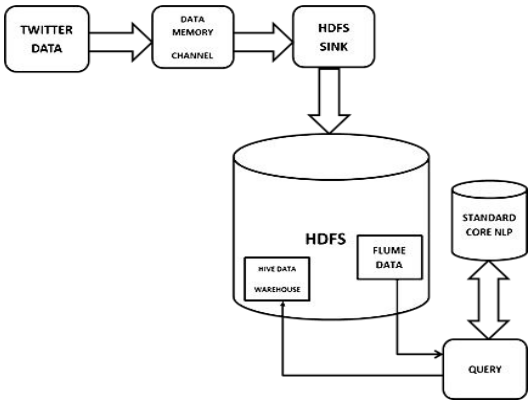
For computing sentiment:

```
Create VIEW tweets_sentiment as select
id,
case
when sum( polarity ) > 0 then 'positive'
when sum( polarity ) < 0 then 'negative'
else 'neutral' end as sentiment
from l3 group by id;
```

B. Methodology

The following figure shows clearly the architecture view for the proposed system by this we can understand how our project is effective using the Hadoop ecosystems and how the data is going to store in the form of Flume, also how it is going to create tables using Hive also how the sentiment analysis is going to perform [6].

Figure 2: Architecture diagram for proposed system.



As we have seen the procedure how to overcome the problem that we are facing in the existing problem that is shown clearly in the proposed system. So, to achieve this we are going to follow the following methods:

- Creating Twitter Application.
- Getting data using Flume.
- Querying using Hive Query Language (HQL).

B. Phase 1: Creating Twitter Application

First of all if we want to do sentiment analysis on Twitter data we want to get Twitter data first so to get it we want to create an account in Twitter developer and create an application by clicking on the new application button provided by them.[3] After creating a new application just create the access tokens so that we no need to provide our authentication details there and also after creating application it will be having one consumer keys to access that application for getting Twitter data. The following is the figure that show clearly how the application data looks after creating the application and here it's self we can see the consumer details and also the access token details. We want to take this keys and token details and want to set in the Flume configuration file such that we can get the required data from the Twitter in the form of twits.

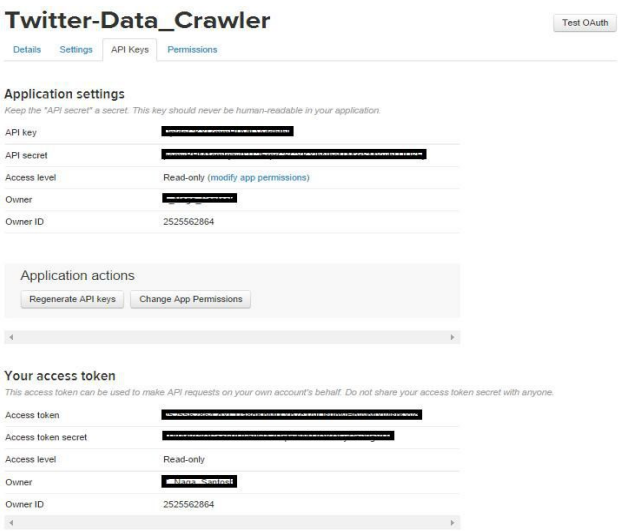


Figure 3: Creating Twitter application from Twitter Developer.

The figure show clearly the application keys that are generated after creating application and in this keys we can see the top two keys are the API key and API secret. And coming to the reaming two keys it is nothing but know as the access tokens that we want to generate it by ourselves by clicking the generate access token. After clicking that we can get the two keys that are our account access token and coming to that one is Access token and the other one is the Access token secret.

C. Phase 2: Getting data using Flume

After creating an application in the Twitter developer site we want to use the consumer key and secret along with the access token and secret values. By which we can access the Twitter and we can get the information that what we want exactly here we will get everything in JSON format and this is stored in the HDFS that we have given the location where to save all the data that comes from the Twitter. The following is the configuration file that we want to use to get the Twitter data from the Twitter.

Figure 4: Flume configuration files for Twitter data.

```
TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type =
org.apache.hadoop.sentiment.analysis.TwitterSourceComments
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = [redacted]
TwitterAgent.sources.Twitter.consumerSecret =
[redacted]
TwitterAgent.sources.Twitter.accessToken = [redacted]
[redacted]
TwitterAgent.sources.Twitter.accessTokenSecret =
[redacted]
TwitterAgent.sources.Twitter.keywords = hadoop, big data,
analytics, bigdata, cloudera, data science, data scientiest,
business intelligence, mapreduce, data warehouse, data
warehousing, mahout, hbase, nosql, newsql, businessintelligence,
cloudcomputing
TwitterAgent.sources.Twitter.filter = false

TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://localhost:8020/user/flume/twittertweets/%Y/%m/%d/%H
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

D. Phase 3: Querying using Hive Query Language (HQL)

After running the Flume by setting the above configuration then the Twitter data will automatically will save into HDFS [4] where we have the set the path storage to save the Twitter data that was taken by using Flume. The following is the figure that shows clearly how the data is stored in the HDFS in a documented format and the raw data those we got form the Twitter is also in the JSON format that is shown clearly below:

Figure 5: Twitter data in HDFS (Hadoop Distributed File System).

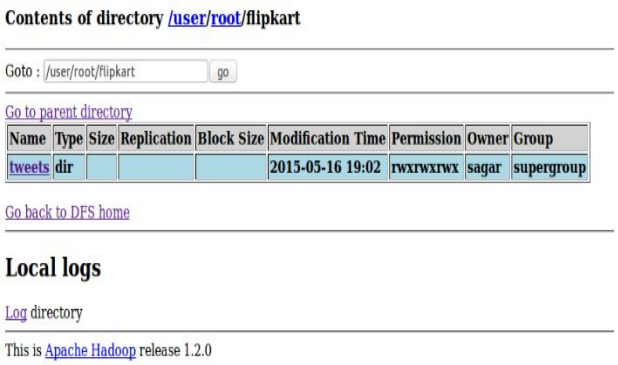
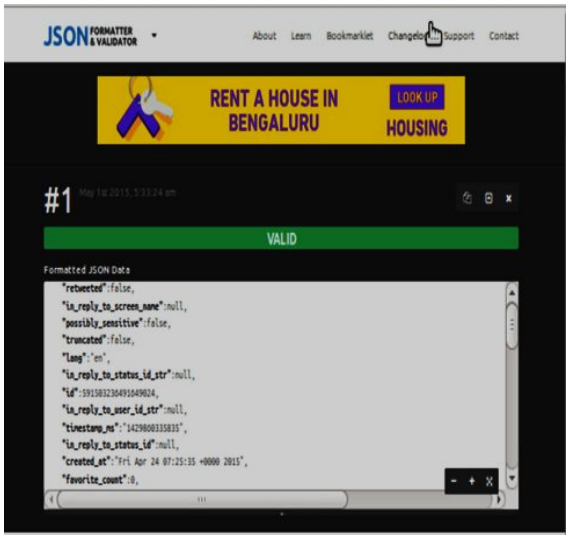


Figure 6: Twitter data in JSON format.

Figure 7: Validating JSON data for HQL.



From these data first we want to create a table where the filtered data want to set into a formatted structured such that by which we can say clearly that we have converted the unstructured data into structured format. For this we want to use some custom serde concepts. These concepts are nothing but how we are going to read the data that is in the form of JSON format for that we are using the custom serde for JSON so that our hive can read the JSONdata [8] and can create a table in our prescribed format.

Figure 8: HQL Query for creating Tweets table

```
CREATE EXTERNAL TABLE tweets (
Id BIGINT,
created at STRING,
source STRING,
favorite BOOLEAN,
retweet_count INT,
retweeted_status STRUCT<
text: STRING,
user:STRUCT<screen_name:STRING,name:STRING>>,
entities STRUCT<
urls:ARRAY<STRUCT<expanded_url:STRING>>,
user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
hashtags:ARRAY<STRUCT<text:STRING>>>,
text STRING,
user STRUCT<
screen_name:STRING,
name:STRING,
friends_count:INT,
followers_count:INT,
statuses_count:INT,
verified:BOOLEAN,
utc_offset:STRING,
time_zone:STRING>,
in_reply_to_screen_name STRING,
year int,
month int,
day int,
hour int
)
ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
LOCATION '/user/root/today/tweets';
```

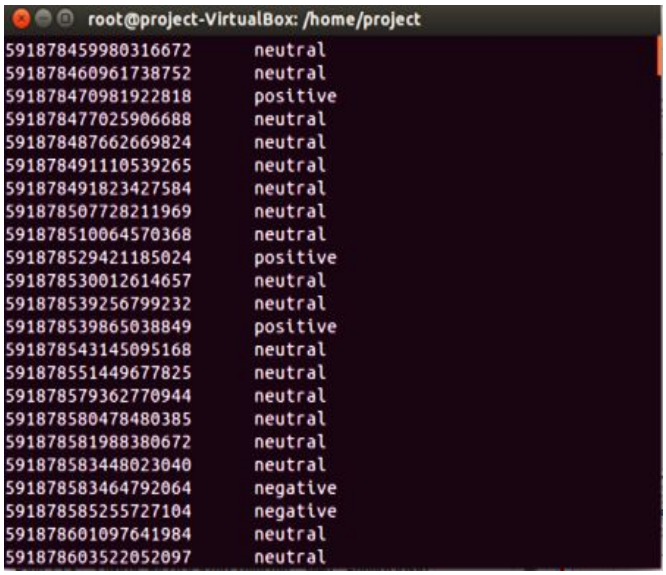
Figure 9: Creating sentiment table and performing sentiment analysis.

```
Create VIEW tweets sentiment as select
Id,
Case
When sum (polarity) > 0 then 'positive'
When sum (polarity) < 0 then 'negative'
Else 'neutral' end as sentiment
From 13 group by id;
```

Also we are using another UDF's (User Defined Functions) for performing the sentiment analysis on the tables that are created by using Hive.[7]From that we can perform the sentiment analysis. And acquire the results where a new table is created by partition concept such that all the comments that are having positive will go into the positive partition and all the comments that are having moderate will go into moderate partition and finally all the comments that are having negative will go into negative partition.

The following figure shows clearly how the rating is done and how the data is partitioned into 3 types.

Figure 10: Result after performing the sentiment analysis.



IV. ACCURACY

The overall accuracy of project is determined by time required to access from various modules i.e. accessing from nlp, word net and sentiwordnet. As all components are in series i.e. used one after the overall, theoretically the overall accuracy of the program is the product of accuracy of all its modules .We tested our implementation on the standard dataset of flip kart provided by twitter.

The accuracy of our project after running on this data set is as following:

Sentiment	Count	Correct	%	Tolerance
Positive	151	102	74.01	-0.01
Negative	131	95	68.94	+0.09
Neutral	253	235	78.08	-+0.05

So, overall accuracy is the mean i.e. 73.67.

V. RESULT ANALYSIS

In this we are plotting graph for twitter data of flip kart. As shown below graph this represents polarity score as positive, negative and neutral (1,-1, 0) and we are plotted here country names in this view for the country information.

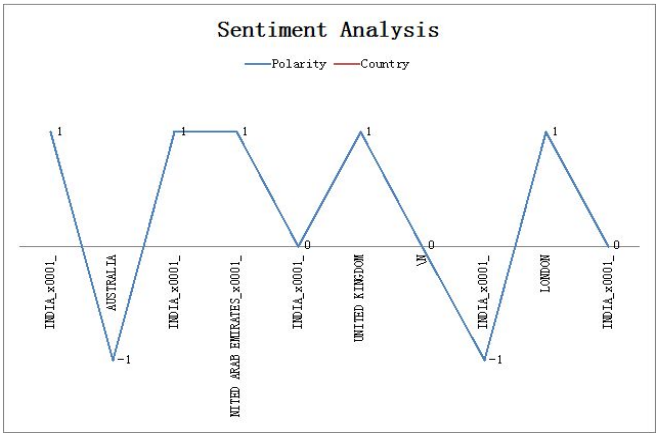


Figure 11: Graph for sentiment analysis of Flipkart tweets.

VI. CONCLUSION AND FUTURE WORK

There are various ways to get Twitter data or any other online streaming data where they want to code lines of coding to achieve this. And, also they want to perform the sentiment analysis on the stored data where it makes some complex to perform those operations. Coming to this paper we have achieved by this problem statement and solving it in BIGDATA by using Hadoop and its Eco Systems. And finally we have done sentiment analysis on the Twitter data that is stored in HDFS [4]. So, here the processing time taken is also very less compared to the previous methods because HadoopMapReduce and Hive are the best methods to process large amount of data in a small time.

In this paper it has shown the way for doing sentiment analysis for Twitter data. Also, we can do this by using Oozie by creating a work flow so that we can give a time slang such that it will work based upon that time we allocated for performing a particular work. Also at last we can also visualize the word map i.e., the most frequent words that are used in positive, moderate and negative fields by using R language to visualize.

VII. ACKNOWLEDGMENTS

We are grateful to express sincere thanks to our faculties who gave support and special thanks to our department for providing facilities that were offered to us for carrying out this project.

REFERENCES

[1] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.

[2] Tang, H., Tan, S., Cheng, X., A survey on sentiment detection of reviews, Expert Systems with Applications: An International Journal, v.36 n.7, p.10760-10773, September, 2009.

[3] A. Pak and P. Parouek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010.

[4] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.

[5] Bahrainian, S.A., Dengel, A., Sentiment Analysis using Sentiment Features, In the proceedings of WPRSM Workshop and the Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Atlanta, USA, 2013.

[6]"Sentimental Analysis", Inc. [Online]. Available: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis [Accessed 23 March 2013].

[7] (Online Resource) Hive (Available on:http://hive.apache.org/).

[8] (Online Resource)http://jsonlint.com/

[9](OnlineResource)http://nlp.stanford.edu/software/corenlp.shtml.

[10]T. White, "The Hadoop Distributed File system," Hadoop: The Definitive Guide, pp. 41-73, GravensteinHighwaNorth, Sebastopol: O'Reilly Media, Inc., 2010.

[11] (Online Resource) http://flume.apache.org/

[12]S. W. Ambler. Relational databases 101: Looking at the whole picture.www.AgileData.org, 2009.

[13] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big Data: The Next Frontier For Innovation, Competition, And Productivity", May 2011.

First Author: Sagar Nadagoud was born in Karnataka, India. He received the B.E Degree in Computer Science and Engineering from Visvesvaraya Technological University Belgaum Dist., India in 2013 and M.Tech Degree in also same branch and University. His research interests are in the area of Semantic Web and Big Data Analytics.

Second Author: Kotresh Naik.DM-tech was born in Karnataka, India. He is working as Asst.Prof. In Information Science Engineering Dept., CIT, Gubbi, Tumkur (district), Karnataka, INDIA

