International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 10 (2015) pp. 24905-24916
© Research India Publications
http://www.ripublication.com

Market Based Trends Analysis Using 'Sentiments Analysis'

Rohit Gore[a*] and Swati Ahirrao[b*]

[a]M.Tech student, Symbiosis International University, Pune, 412115, India
[b]Assistant Professor, Symbiosis International University, Pune, 412115, India
Email: rohit.gore@sitpune.edu.in
Email: swatia@sitpune.edu.in

Abstract:

Organization whether mobile based or any other competitive industry is always eager to know the feedback of their customers on their latest launch. Social media is one of the means for organizations to collect its customer's opinion about their product and its features. It is a social space where everyone is given freedom of expression in various forms like posts, tweets, blogs, micro blogs etc. Hence social media is one of the most popular means which targets on customer's views and all latest market trends. This paper emphasizes on collection of this social data only in form of reviews, feedbacks and comments of people regarding various products and tries to figure out products report and provides updates on choice of customers. The technology and platform used for collection of social media data is Flume which stores the data into HDFS (Hadoop Distributed File System) in unstructured format. Map Reduce for data mining technique is used for conversion of unstructured data into structured. The term sentiment analysis takes into account extraction of sentiment related information from various social media, which deals with expression of various kinds of sentiment on various commercial products; thereby social media becomes one of the mostly likely platforms for promoting a product.

Keywords: Big Data, Flume, Map Reduce, Sentiment Lexicon, lexicon-based approach

1. Introduction

In this competitive market every organization is struggling for its survival. 'Survival of the fittest' is possible only for those organizations who try to understand and keep

themselves updated with their customer's feedback on various products. Social media is a space where a person can express their views in the text format on various products and ongoing issues. So organizations target social media to gather views of customers for their products. This social media is accumulating lots and lots of big data every day, especially increase in size of textual data. This paper focuses its attention on extraction of twitter based data using flume and store it into HDFS and performing the sentiment analysis on it to determine the trend of product.

There are different frameworks for analyzing sentiment based information in web (social media) each serving a different purpose to society. Posts, tweets, comments, blogs in social media are various valuable assets for organization to identify their prospective customer needs. For example a tweet may be used for purposes like examining the feedback of customer on its product and their inbuilt features. The areas of concern for sentiment analysis are subject, strength of sentiment detection and polarity, where subject deals with the topic of given text, sentiment strength hunts for positive and negative words in sentiment, polarity deals with whether the text is positive, negative or neutral. A method is also used that includes the feature based lexicons to improve the accuracy of sentiment analysis which shows the sentiment based results on product along with review about their features.

The Big Data which is collected across different social media is in unstructured form and it needs to be converted to proper structured format, this can be done using data mining technique of Map Reduce. Now the structured data available is basically in form of tweets and is of different mobile handsets like Samsung, Lenovo, Apple etc. To retrieve data particularly of Samsung based handsets, Map Reduce automatically employs parallel K Means clustering Algorithm to generate Samsung handset based clusters; this is the output at end of first run of parallel K Means Algorithm. After second run of parallel K Means Algorithm featured data clusters are generated. Advantage of this clustering technique compared to other clustering tools like Weka, Excel Miner & Rapid Miner etc. is that processing of data is faster in a distributed system; else considerably more time would have been deployed. Also clustering tools like Weka, Excel Miner & Rapid Miner etc. can be deployed only for limited data and not for such Big Data. A feature based lexicon dictionary is also proposed to deal with some of new terms and concepts like Blurry Image etc.

2. Related work

There are two approaches to improve lexical sentiment analysis for the social web by allowing a general algorithm to be modified for a specific topic. The first method considers the mood of the posts within the topic and the second method recognizes and adds topic-specific terms to the general sentiment lexicon. Although sentiment analysis often focuses on reviews of movies or consumer products they most likely form a small fraction of part for social web.

The lexicon-based approach [5] depends on finding the opinion lexicon which is used to analyse the text. There are two methods in this approach. The dictionary-based approach which depends on finding opinion seed words, and then searches the

dictionary of their synonyms and antonyms. The corpus-based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods. There is a brief explanation of both approaches' algorithms and related articles in the next subsections.

AFINN is lexicon based on the Affective Norms of English words proposed by Bradley & Lang [1]. AFINN lexicon focused on the languages used in micro blogging platform. Positive words are scored from 1 to 5 and negative words are scored from -1 to -5. The lexicon includes total 2,411 English words.

Bing Liu's Opinion Lexicon is polarity oriented and is formed by 2006 positive words and 4683 negative words.

Sentiwordnet lexicon is a lexical resource for sentiment classification introduced by Baccianella et al [2] that is improved by Esuli and Sebastiani [3] extension of wordnet the well-known English lexical database where words are clustered into groups of synonyms known as synsets.

Sentiment 140 lexicon [6] is a corpus of 1.6 million tweets with positive and negative emoticons was used to calculate the sentiment words. The tweet collection is the same as one used to train sentiment 140 method.

3. Tools and technology

I. Apache Flume: Flume is an Apache project, mainly used to move large amount of streaming data into HDFS. Apache Flume is distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amount of data into the HDFS.

Component of Apache Flume:

- **Source** – the entity through which data enters into Flume. Sources either actively poll for data or passively wait for data to be delivered to them. A variety of sources allow data to be collected, such as log4j logs and syslogs.
- **Sink** – the entity that delivers the data to the destination. A variety of sinks allow data to be streamed to a range of destinations. One example is the HDFS sink that writes events to HDFS.
- **Channel** – the conduit between the Source and the Sink. Sources ingest events into the channel and the sinks drain the channel.

In order to extract the twitter data the flume agent named as Twitter. Agent is to generated

Source is the name of the project created on the dev.twitter.com. Twitter. channel is an integer value initialised by name Memchannel. Twitter. sink is initialised a HDFS which contains the path HDFS

II. Map Reduce: MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks. The nature of this programming model and how it can be used to write

programs which run in the Hadoop environment is explained by this model. Hadoop [4] is an open source implementation for this environment. Map and Reduce are two functions. The main job of these two functions are sorting and filtering input data. During Map phase data is distributed to map per's machines and by parallel processing the subset it produces <key,value>pairs for each record. Next shuffle phase is used for repartitioning and sorting that pair within each partition. So the value corresponding to same key is grouped into {v1, v2…} values. Last during Reduce phase reducer machine processes subset <key, {v1, v2,}>pairs parallel in the final result which is written to distributed file system. The input reader divides the input into blocks (in practice typically 64 MB to 128 MB) and the framework assigns one block to each Map function. The input reader reads data from stable storage and generates key/value pairs. The Map function takes a series of key/value pairs, processes each, and generates zero or more output key/value pairs. The input and output types of the map can be different from each other. The map function breaks the line into words and output a key/value pair for each word. The reducer function can iterate through values that are associated with that key and produce an output. Output writer writes the output of the reducer to the stable storage.

4. Proposed Work

I. Parallel K Means: K-means clustering is commonly used for a number of classification applications. Because k-means is run on such large data sets, and because of certain characteristics of the algorithm, it is a good candidate for parallelization. The goal of this project was to implement a framework in java for performing k-means clustering using Hadoop MapReduce. In this problem, we have considered inputs of twitter data sets which contain tweets regarding the Samsung handsets and desired clusters of size 3. Once the k initial centers are chosen, the distance is calculated (Euclidean distance) from every point in the set to each of the 3 centers & point with the corresponding center is emitted by the mapper. Reducer collects all of the points of a particular centroid and calculates a new centroid and emit.

Termination Condition:
When difference between old and new centroid is less than or equal to 0.1
Algorithm:
Step1: Initially randomly centroid is selected based on data. In our implementation we used 3 centroids.
Step2: The Input file contains initial centroid and data.
Step3: In Mapper class "configure" function is used to first open the file and read the centroids and store in the data structure(we used ArrayList)
Step4: Mapper read the data file and emits the nearest centroid with the point to the reducer.

Step5: Reducer collects all this data and calculates the new corresponding centroids and emit.
Step6: In the job configuration, we are reading both files and checking
if difference between old and new centroid is less than 0.1 then
Convergence is reached
else
Repeat step 2 with new centroids.

II. **Stop Words Removal:** Many times, it makes sense to not index "stop words" during the indexing process. Stop words are words which have very little informational content. These are words such as: and, the, of, it, as, may, that, a, an, of, off, etc. Studies have shown that by removing stop words from the index, you may benefit with reduced index size without significantly affecting the accuracy of a user's query. Care must be taken however to take into account the user's needs. For example, the phrase "to be or not to be" from Hamlet is composed entirely of stop words. Most of the internet's search engines eliminate all the stop words from their indexes. By eliminating stop words from the index, the index size is typically reduced by about 33% for a word level index. For a record level index or IDF level index, then eliminating stop words is not typically done as they will not add significantly to the index size.

III. **Stemming:** It is the process of reducing derived words to their stem, or root form. Stemming programs are commonly referred to as stemmers or stemming algorithms. A simple stemmer looks up the inflected form in a lookup table; this kind of approach is simple and fast. The disadvantage is that all inflected forms must be explicitly listed in table.eg. "developed", "development", "developing" are reduced to the stem "develop".

IV. **Feature Based Specific Lexicon:** Previous studies show that the sentiment of several studies are feature dependent. For a mobile the feature like 'long life battery' is positive and 'low life battery' related to negative feedback. The context being discussed is significant when one want to determine related sentiment based on the study. Feature based lexicon is being made which indicate both specific domain as well as sentiment in the particular domain. For feature 'Picture quality of Samsung S4' clearly express positive feedback and blurry express negative feedback/sentiment for the camera quality.

V. **Data Filtration:** HDFS contains all tweets comments related with to features as well as related sentiment based information This approach contains the two steps:

- Extract all tweets that occurs in training sentences about particular mobile features and that may be noun, adjective, verbs, adverbs as well as negated forms. This builds initial list of lexicons for mobile features.

Second step for text in the list for each camera features, if the tweet does not contains words from the list, remove the list.

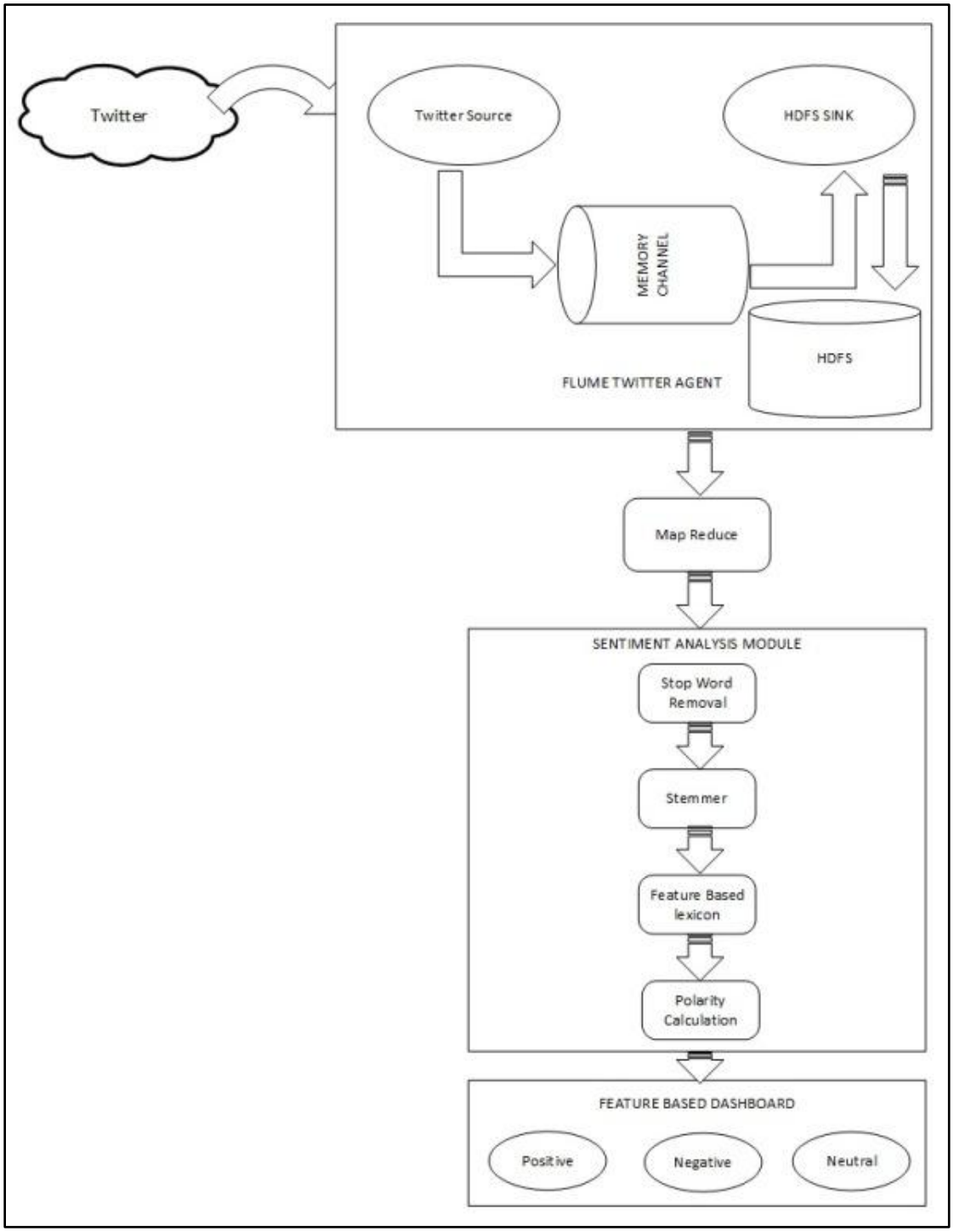

Fig.1. System Flow

VI. Improved dictionary: The strength of the sentiment analysis is based on the strength of the dictionary. Feature based lexicons needs finding sentences that

are conceptually correlated that are not necessarily synonymous or antonyms. Example: sharp and clear are conceptually correlated to camera picture qualities but they are not synonyms from linguistics point of view, so solution is to use dictionaries which contains lexicons for mobile feature such as picture quality positive lexicon[clear, sharp, bright, sunny] while picture quality negative lexicon [dark, dim, grey, blurry].

Next step is classifiers are trained to analyse the feature based tweets from the database and second steps results are aggregated together to generate the final prediction.

VII. Algorithm: Sentiment Calculation

Data: Pre-processed twitter data.
Input: Feature Based lexicons (*SentiList*), Stop word removal.
Output: Positive, Negative and Neutral.
*If***stop words** is present
Find all the **stop words** in tweets
Using *dictionary* and add them to *List*.

end
*If***stemming words** is present
Find all the **stemming words** in tweets
Using *dictionary* and add them to *List*.
SentiSum=0;
end
foreach word in the *SentiList* do
SentiSum=*SentiSum*+sentiment of
word;
end
SentiType= "Neutral";
IfSentiSum>0 then
SentiType="Positive";
end;
IfSentiSum<0 then
SentiType="Negative";
end;
return*SentiType*;

Sentiment calculation is the aggregation of the sum of the sentiment bearing entities of the tweet. Entities can be text, emoticons, hash tags and raw data. The sentiment calculation algorithm is shown in Algorithm. The sentiment calculation is based on a set of heuristics built on the sentiment orientation of the words. Negation words are extracted from the sentence. The presence of the negation words indicates negative sentiment. If the sentence contains a negation word, then other steps can be

skipped and sentiment is blindly assigned as negative. Next, sentiment words are extracted. The sentiment polarity of the word can be changed due to negation words that occur in proximity (2 word distance). If a sentiment word is not present, then the sentiment negation word becomes additive to the negative sentiment list. The sentence "I cannot deal it with Samsung galaxy" has the negation word "not" and it does not contain a sentiment word. So the negation word just gets added to the negative sentiment word.

5. Results:

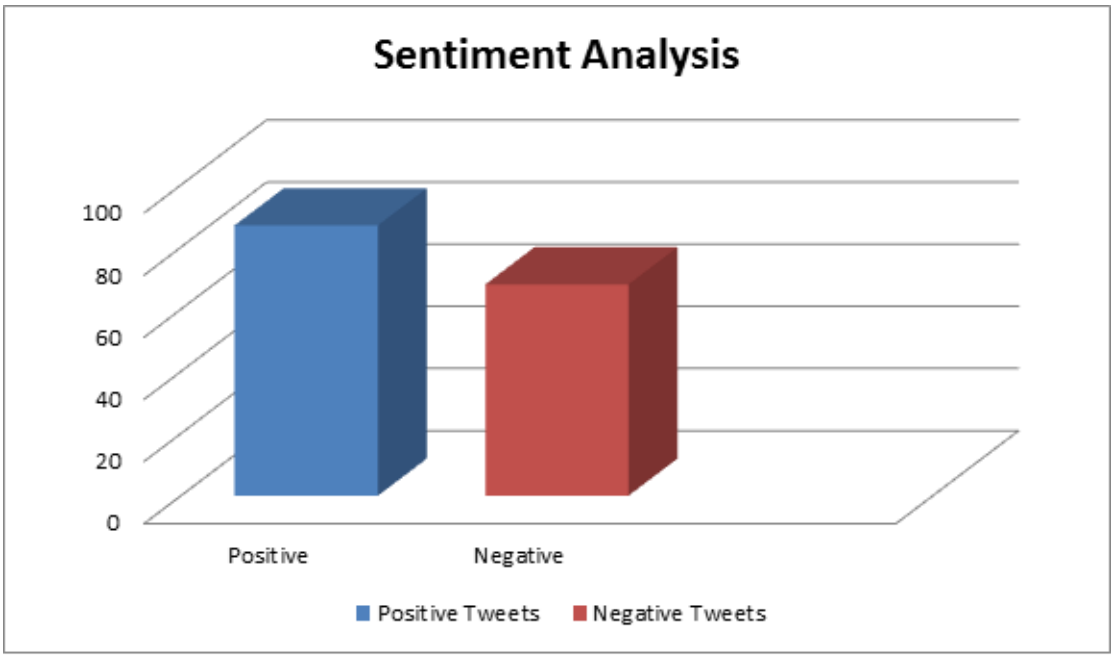

Fig.2. Sentiment Score without Feature Based Lexicons.

There are the two experiments that are performed on the Samsung handsets tweets which are extracted using the flume. In first experiment traditional dictionary approach is used on tweets of mobile and their features. In second experiment information determined in the feature based lexicons are introduced into the sentiment dictionary and applied the sentiment on tweets of mobile and their feature. Fig 2 shows the results without the feature based lexicons.

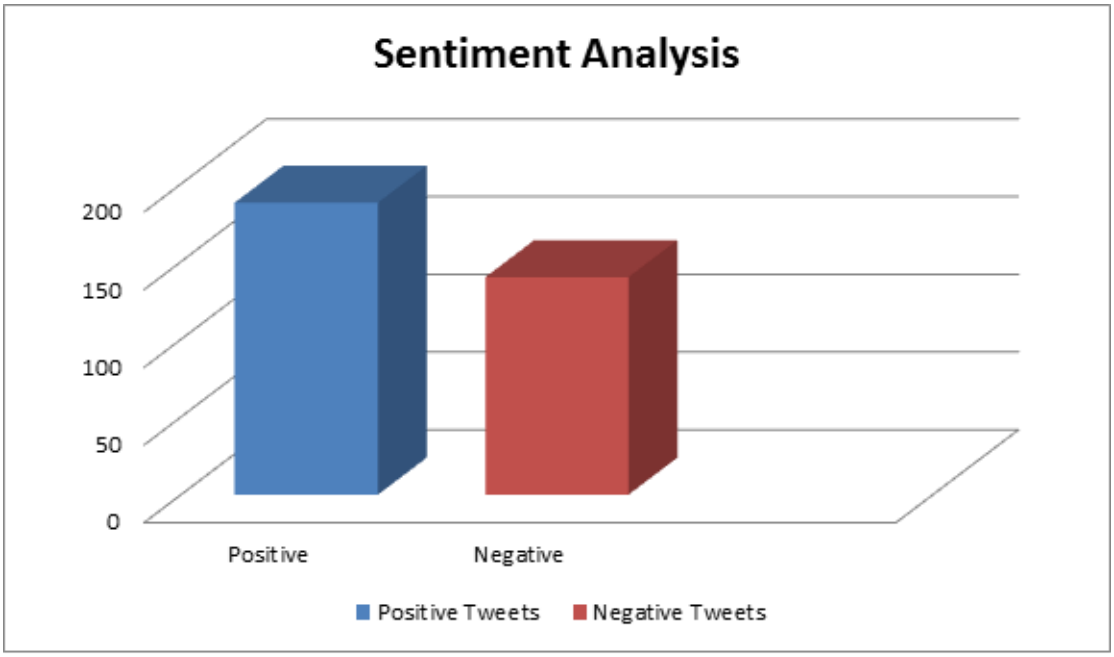

Fig.3. Sentiment Score with Feature Based Lexicons.

Fig 3 shows the results with the feature based lexicons. As the feature based lexicons contains the corpus for the each and every feature of the mobile so it is gives a good results.

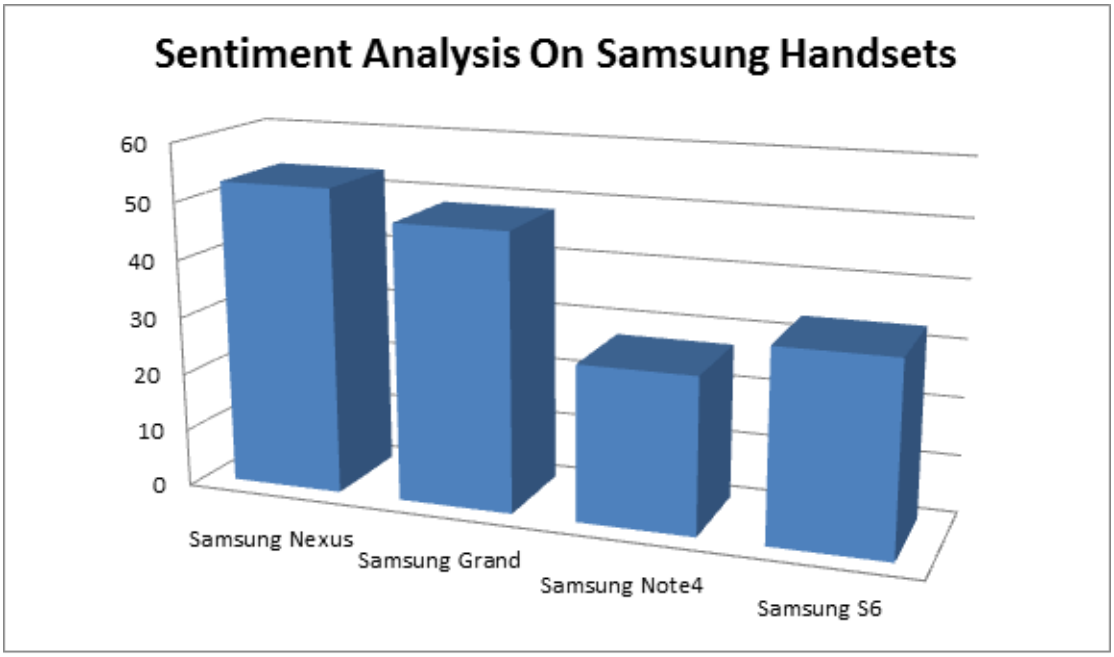

Fig.4. Trend Analysis

Fig.4 was a survey conducted which shows Trend Analysis about positive sentiment analysis for Samsung handsets, the results reveals as can be seen in above fig.4 as well that Samsung Note 4 has lesser number of positive views in comparison to other handsets. When a feature based analysis was done on Samsung Note 4 handset to reveal the reason behind such trend, it was observed that some of the features of handset like 'battery life' is comparatively less everlasting also 'Front Camera Quality' of handset is blurry in comparison to other handsets of Samsung. The results of feature based analysis can be seen in fig.5.

```
mysql> select * from camera;
+------+-----------------------------------------------------------------------
---------------+----------+
| SNO  | tweet
               | location |
+------+-----------------------------------------------------------------------
---------------+----------+
|    1 | It is surprising how good looking Galaxy Note 4 is. front camera image
quality is dark |          |
|    1 | Samsung Galaxy Note 4 with blurred front camera
               |          |
+------+-----------------------------------------------------------------------
---------------+----------+
2 rows in set (0.00 sec)

mysql> select * from battery;
+------+-----------------------------------------------+----------+
| SNO  | tweet                                         | location |
+------+-----------------------------------------------+----------+
|    1 | Samsung Galaxy Note 4 with low battery life |          |
+------+-----------------------------------------------+----------+
1 row in set (0.00 sec)

mysql>
```

Fig.5. Feature based Analysis of Samsung Note 4

6. Conclusion

The data mining technique - MapReduce employs a data mining clustering technique to generate clustered data based on Samsung handsets tweets. But this clustered data is still voluminous, having considerably large size and need to be compacted in order to extract meaningful information for sentiment analysis of customers about a particular product. So in order to deal with this issue of size of data, a feature based dictionary was proposed as a solution, which contains compact, precise and significant data which can be used for analysis purpose. The information in sentiment lexicons is mainly domain specific and can be improved through accuracy of proposed sentiment analysis through data filtration and improved dictionary. Finally a novel method has been proposed to incorporate the lexicon information to improve the sentiment based analysis.

7. Future Work

To extend the scope of this project, a Multi Lingual Dictionary can be proposed in the proposed system. This dictionary will not only act as a language translator for different languages, but will also increase the efficiency of organizations in terms of time and profit. An organization will come across the different cultures and mindsets which will increase their vision about different behavioral pattern of customers. This will help in further improvement of product in terms of quality and as per customer's choice, which will increase market value of organizations and their long term relations with their customers. Also it is not restricted to any geographic boundaries.

8. References

[1] M.M. Bradley, P.J. Lang, Affective Norms for English Words (ANEW) Instruction Manual and Affective Ratings, Technical Report C-1, The Center for Research in Psychophysiology University of Florida, 2009.

[2] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valletta, Malta, 2010.

[3] A. Esuli, F. Sebastiani, Sentiwordnet: a publicly available lexical resource for opinion mining, in: Proceedings of the 5th Conference on Language Resources and Evaluation, 2006.

[4] Benjamin C. M. Fung, Ke Wang, Philip S. Yu, "Top-Down Specialization for Information and Privacy Preservation", Proceedings of the 21st International Conference on Data Engineering (ICDE 2005), 1084-4627/05 $20.00 © 2005 IEEE.

[5] WalaaMedhat, Ahmed Hassan, HodaKorashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal Volume 5, Issue 4, December 2014, Pages 1093–1113.

[6] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, Technical report Stanford University, 2010.

[7] Aaron K. Baughman, Wesley Chuang, Kevin R. Dixon, Zachary Benz, and Justin Basilico," DeepQA *Jeopardy!* Gamification: AMachine Learning Perspective", IEEETRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES, VOL. 6, NO. 1, MARCH 2014.

[8] Alvaro Ortigosa, José M. Martín, Rosa M. Carro," Sentiment analysis in Facebook and its application to e-learning", Elsevier :Computers in Human Behavior 31 527–541,2014.

[9] Ana Carolina E.S. Lima, Leandro Nunes de Castro," A multi-label, semi-supervised classification approach applied to personality prediction in social media", Elsevier: Neural Networks 58 122–130,2014.

[10] Dayong Wang, Steven C.H. Hoi, Member, IEEE, Ying He, and Jianke Zhu," Mining Weakly Labeled Web Facial Images for Search-Based Face

Annotation", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.