SENTIMENT ANALYSIS USING BIG DATA

R. Suresh Ramanujam Ph.D
Assistant Professor
Department of Information Technology
Sri Manakula Vinayagar Engineering College
Puducherry, India
sureshramanujam78@gmail.com

R. Nancyamala
Department of Information Technology
Sri Manakula Vinayagar Engineering College
Puducherry, India
nancyamala11@gmail.com

J. Nivedha
Department of Information Technology
Sri Manakula Vinayagar Engineering College
Puducherry, India
nivedha25594@gmail.com

J. Kokila
Department of Information Technology
Sri Manakula Vinayagar Engineering College
Puducherry, India
kokila.koki19@gmail.com

1. ABSTRACT

The Web has become an excellent source for assembling consumer opinions. There are now several Web sites containing such opinions, e.g., customer reviews of products, forums, discussion groups, and blogs. This paper focuses on online customer reviews of products. It makes two contributions. First, it proposes a framework for analyzing and comparing consumer opinions of competing products in map and reduce environment for better analysis. Second, a new technique based on lexicon based technique is proposed to extract neutral reviews and restrict them from being categorized under positive or negative. Experimental results show that the technique is highly effective and smash existing methods significantly.

2. INTRODUCTION

E-Commerce has been expanded widely due to the usage of Internet. It not only offer service through internet to the consumers but also allows them to provide their view or opinion about that product and its services. Such numerous consumer reviews contain rich and valuable knowledge and have become important resource for both consumers and firms [1]. Consumers normally view the online reviews and overall ranking about the product before purchasing them. On the other side the organization would get feedback from the reviews which would help them to improve their development, marketing and consumer relationship management. Normally the review may be in structured and unstructured form. The analysis of this data for extracting sentiment or user’s opinion is a great deal. There are many challenges in the field of sentiment analysis. The most common challenges are Firstly, Word Sense Disambiguation (WSD), a classical NLP problem is often encountered. Secondly, addressing the problem of sudden deviation from positive to negative polarity. Thirdly, negations, unless handled properly can completely mislead. Fourthly, keeping the target in focus can be a challenge. Fifthly, handling the misspelled word by finding with correct spelling by calculating the nearest distance method.

Organizations use sentiment analysis to understand how the public feels about

978-1-4673-6524-6/15/$ 31.00 © 2015 IEEE

something at a particular moment in time, and also to track how those opinions change over time.

An enterprise may analyze sentiment about:

- A product – For example, does the target segment understand and appreciate messaging around a product launch? What products do visitors tend to buy together, and what are they most likely to buy in the future?
- A service – For example, a hotel or restaurant can look into its locations with particularly strong or poor service.
- Competitors – In what areas do people see our company as better than as (or weaker than) our competition?
- Reputation – What does the public really think about our company? Is our reputation positive or negative?

In Sentiment analysis, the neutrality is handled in various ways, depending on the technique that is being used. In lexicon-based techniques the neutrality score of the words is taken into account in order to either detect neutral opinions (Ding and Liu, 2008) or filter them out and enable algorithms to focus on words with positive and negative sentiment (Taboada et al, 2010). On the other hand when statistical techniques are used, the way that neutrals are handled differs significantly. Some researchers consider that the objective (neutral) sentences of the text are less informative and thus they filter them out and focus only on the subjective statements in order to improve the binary classification (Bo Pang and Lillian Lee, 2002).In other cases they use hierarchical classification where the neutrality is determined first and sentiment polarity is determined second (Wilson et al, 2005). Finally in most academic papers of sentiment analysis that use statistical approaches, researchers tend to ignore the neutral category under the assumption that neutral texts lie near the boundary of the binary classifier. Moreover it is assumed that there is less to learn from neutral texts comparing to the ones with clear positive or negative sentiment.

Koppel and Schler (2006) showed in their research both of the above assumptions are false. They suggested that as in every polarity problem three categories must be identified (positive, negative and neutral) and that the introduction of the neutral category can even improve the overall accuracy. Their work was primarily focused on SVM and they used geometric properties in order to improve the accuracy of their three binary classifiers.

3. ANALYZING FRAMEWORK

The analyzing framework consists of data collection using Apache flume (A tool to obtain data from various web sites and store in HDFS), data analyzing using Hadoop (A framework where one can store and analyze huge volume of data) and data visualization using Rstudio (An IDE for viewing statistical data graphically).

3.1. DATA COLLECTION

Flume is used to obtain data from any social web site and store in HDFS. Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store. The summary of concepts of Flume is

Event: A byte payload with optional string headers that represent the unit of data that Flume can transport from its point of origination to its final destination.

Flow: Movement of events from the point of origin(i.e.)from twitter source to their final destination(i.e.) HDFS is considered a data flow, or simply flow.

Client: An interface implementation that operates at the point of origin of events(i.e.)flume is configured and delivers them to a Flume agent. Clients typically operate in the process space of the application they are consuming data from.

Agent: An independent process that hosts flume components such as sources, channels and sinks, and thus has the ability to receive, store and forward events to their next-hop destination. Eg.: We have configured our agent as twitter agent.

Source: An interface implementation that can consume events delivered to it via a specific mechanism. For example, The data that is collected from twitter source is stored in HDFS . When a source receives an event, it hands it over to one or more channels.

Channel: A transient store for events, where events are delivered to the channel via sources operating within the agent. An event put in a channel stays in that channel until a sink removes it for further transport. For example The flume that configured uses the memory channel to get data from twitter source and store it in a sink(i.e.) HDFS . Channels play an important role in ensuring durability of the flows. The size of the channel here is fixed as 1000MB and the channel waits for getting the data and stores it in a buffer space. When it exceeds 1000MB then the data is transferred into sink.

Sink: An interface implementation that can remove events from a channel and transmit them to the next agent in the flow, or to the event's final destination. Sinks that transmit the event to its final destination are also known as terminal sinks. The Flume HDFS sink is an example of a terminal sink.

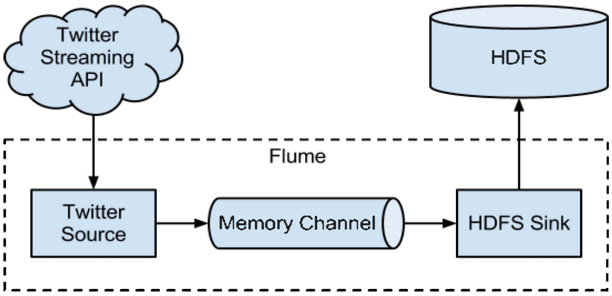


Fig 1: Conceptual data flow flume

3.2. DATA ANALYZING

Hadoop is apache open source software (java framework) which runs on a cluster of commodity machines. Hadoop provides both distributed storage and distributed processing of very large data sets. Hadoop is capable of processing **big data** of sizes ranging from Gigabytes to Petabytes. **Hadoop architecture** is similar to master/slave architecture. The master being the namenode and slaves are datanodes. The namenode controls the access to the data by clients. The datanodes manage the storage of data on the nodes that are running on. Hadoop splits the file into one or more blocks and these blocks are stored in the datanodes. Each data block is replicated to 3 different datanodes to provide high availability of the hadoop system. The block replication factor is configurable.

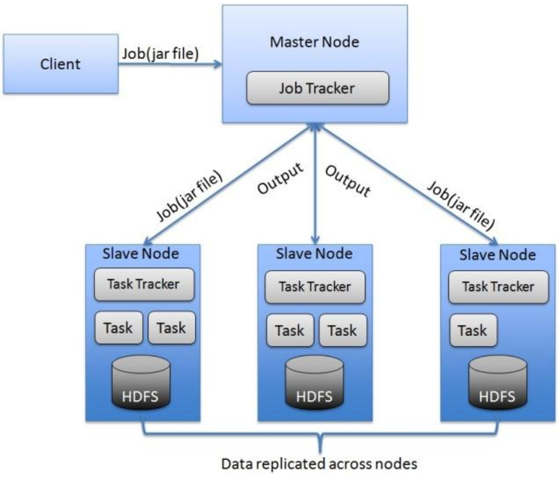


Fig 2: Data processing in hadoop framework

978-1-4673-6524-6/15/$ 31.00 © 2015 IEEE

As said earlier, Hadoop is used for both storing and analyzing. The data obtained from flume is stored in HDFS which servers as storage purpose for hadoop. The collected reviews are analyzed using Naïve Bayes algorithm in mapreduce environment which servers as analyzing purpose for hadoop.

In mapreduce technique the algorithm should be divided between the mapper and the reducer as follows:

Mapper

```
positiveWords = load positive words
negativeWords = load negative words
for each tweet:
    parse the tweet
    date  = date of the tweet down to the minute
tweetWords = all the words in the tweet text
positiveCount = 0
negativeCount = 0
    for candidate in "Google glass":
        if candidate is in the text:
            if a positive word is in the text:
positiveCount = positiveCount + 1
            if a negative word is in the text:
negativeCount = negativeCount + 1
positiveRatio = positiveCount / count of all words
negativeRatio = negativeCount / count of all words
emit date, candidate, positiveRatio - negativeRatio
Intermediate – sort keys
```

Reducer

```
for each key:
    sum = sum of the values associated with this key
    n   = number of values
 emit key, sum/n
```

3.3. DATA VISUALIZATION

R is the name of the programming language and RStudio is a convenient interface that provides the statistical computing and graphics for representing the data.

RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server.

4. EVALUATION

4.1. EVALUATION OF SENTIMENT IDENTIFICATION & CLASSIFICATION

The collected reviews about a particular product are read by the mapper. The reviews are analyzed only if it is in English else they are ignored. This is because the trained data set which we use contains only English sentiment words. The analyzed English reviews are tokenized which means splitting the sentence into words by using the blank space left between each words. These tokenized words are parsed separately. Each word is mapped with a trained dataset for identifying the sentiment word present in the review. If present then the next step is to classify that sentiment word under positive, negative or neutral. This is done by assigning values for each sentiment word based on their level of sentiment. Using hashMap technique the value of the particular sentiment word is returned.

4.2. EVALUATION OF SENTIMENT ANALYSIS

The sentiment scale ranges from 10 to -10. Based on the level of the sentiment word the metrics is assigned. The analysis is made by summing up the sentiment values. In our experiment we have summed the values based on each hour. Eg.: For an hour if 10 peoples have given their feedback, and in that if 7 are positive, 2 negative and 1 neutral then the sentiment value is:

$$S = 7+ (-2) - 1 = 4$$

Thus in general, the sentiment value is calculated as:

$$S_v = P_v + N_v - Ne_v$$

Where:

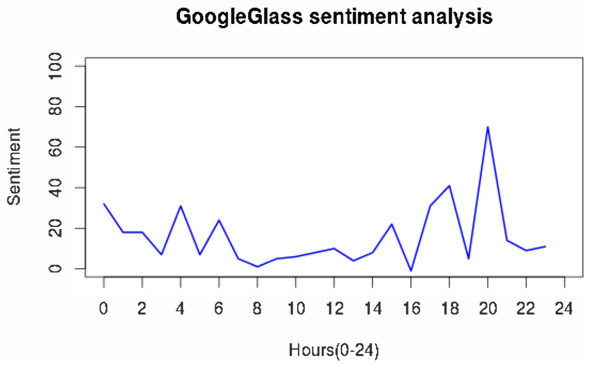
S_v – Sentiment value

P_v – Positive value

N_v – Negative value

Ne_v – Neutral value

The analyzed output for the product "Google glass" is given. It is an hourly based analyses which can be extended weekly and further monthly also. From the output one can evaluate the downfall of the product at which period. Also this analyses helps to promote the product for the next cycle.



5. CONCLUSION

In this article, we have proposed a framework for analyzing and comparing consumer opinions in mapRreduce environment for better analysis and a new technique to extract neutral reviews and restrict them from being categorized under positive or negative. The framework contains three main components, i.e., sentiment identification, sentiment classification and data visualizing. The algorithm simultaneously identifies and classifies the sentiment values. The entire framework is implemented in hadoop environment which results in better output of analyzing semi-structured data and huge data.

6. REFERENCE

[1] A. Ghose and P.G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining Text and reviewer characteristics," IEEE Trans. Knowl. Data Eng., vol.23, no. 10, pp. 1498-1512. Sept.2010.

[2] Erik Cambria, Bjorn Schuller, Yunqing Xia, Catherine Havasi, "Knowledge based approaches to concept level – Sentiment Analysis: New Avenues in Opinion Mining and Sentiment Analysis," Published by the IEEE Computer Society, March/April 2013.

[3] Paul Ekman, "Emotion in the Human Face", Cambridge University Press, second edition, 1982.

[4] Lisa Hankin, "The effects of user reviews on online purchasing behavior across multiple product categories", Master's final project report, UC Berkeley School of Information, 2007.

[5] Michael Wiegand and Alexandra Balahur, "A Survey on the Role of Negation in Sentiment Analysis", Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, 2010.

[6] S. chandrakala and C. sindhu, "opinion mining and sentiment classification: a survey," ICTACT journal on soft computing, volume: 03, issue: 01, ISSN: 2229- 6956(online), Oct 2012.

[7] M. Koppel and J. Schler (2005) "The Importance of Neutral Examples for Learning Sentiment". In *IJCAI*