Sentiment Analysis on Product Reviews using Hadoop

Jalpa Mehta
M.Tech (Computer Science),
Assistant Professor
Shah & Anchor Kutchhi Engineering College
Mumbai, India

Jayesh Patil
B.E. Student,
Dept. of Information Technology
Shah & Anchor Kutchhi Engineering College
Mumbai, India

Rutesh Patil
B.E. Student,
Dept. of Information Technology
Shah & Anchor Kutchhi
Engineering College Mumbai, India

Mansi Somani
B.E. Student,
Dept. of Information Technology
Shah & Anchor Kutchhi
Engineering College Mumbai, India

Sheel Varma
B.E. Student,
Dept. of Information Technology
Shah & Anchor Kutchhi
Engineering College
Mumbai, India

ABSTRACT
Most of the e-commerce sites ask their customers to provide relevant reviews on their products which could help other customers to decide their choice. A slew of reviews is being generated on a daily basis due to an increase in the usage of e-commerce sites. A potential customer may need to go through thousands of reviews before arriving at a firm decision, which is time-consuming. The project elaborated below aims at reducing this time constraint, by providing an effective summarization of reviews in a manner suitable for users. Usage of MapReduce technique provided by Apache Hadoop is highly emphasized for processing reviews. The summarization of reviews is limited to attributes that the potential customers might be interested while looking for the particular product. In this paper, the technique used for the same is described which substantially reduces time complexity when implemented.

General Terms
Algorithms

Keywords
Sentiment Analysis, Opinion Mining, Product Reviews, Hadoop, MapReduce, OpenNLP, SentiWordNet.

1. INTRODUCTION
The enhancement in the field of e-commerce has led to a revolutionary change in the trading process. People's viewpoint have shifted from traditional commerce to e-commerce in the past years. In order to generate more traffic and increase in sales, merchants have enabled customers to share their opinion of the product. Consequently, the reviews are generated at an enormous rate. Merchants provide "votes" mechanism wherein the potential customer vote reviews which he/she considered helpful. Such highly voted reviews are then surfaced at the top of review list so that the potential customer gets the gist of the products while perusing fewer reviews. Instead of utilizing manual efforts to vote helpful reviews, sentiment analysis can be aimed to automate the process of rating on feature - based opinion summarization of reviews.

1.1 Sentiment Analysis
Sentiment analysis is a subfield of Artificial intelligence focused on parsing the given text and proposed its opinion in terms of positive, negative or neutral text. Feature - based opinion summarization identifies the features in the given review and expresses the sentiment relevant to that feature. A simple example to illustrate features in sentence would be as follows:

"*The display quality of the phone is fantastic.*"

"*The battery life though is draining fast.*"

Here, "display" and "battery life" should be considered as features in the above sentences respectively. By using such summarization, a potential customer might be able to narrow down his choices of the product if he's interested in specific features and also ease him in comparing the products.

1.2 Apache Hadoop
The Apache Hadoop project is an open-source software build for scalable and distributed computing. It provides processing techniques that allows for large scale processing of data on clusters of computers [5]. Hadoop works when the input/output in the format of Key-Value pairs. Let us consider an example to illustrate this:

Key	Value
1010	*Hello world!*

Here, '1010' is referenced as a key and 'Hello world!' as the value. Keys are not necessary to be integers only, Strings are also allowed. As for the gamut of this paper, we are interested in the Map Reduce technique. The name is derived from the steps it performs, 'MAP' step – implemented in mapper class and 'REDUCE' step – implemented in reducer class. In the 'MAP' step, the input is divided into smaller sub-problem

creating a tree structure. The output of this step produces multiple different keys and values. The 'REDUCE' step, however, is responsible for combining the output. The output produced has only distinct keys and all its values combined with a delimiter as a separator. The usage of Hadoop for sentiment analysis has proven to be highly effective.

2. RECENT WORKS

A comparative study between different methodologies has been reviewed and analyzed including subjectivity detection, feature selection for opinion mining, and different machine learning approaches[1]. Various mechanism has been implemented until now, which includes bags of words, training corpus, document level, sentence level and feature-level opinion mining[3]. Different polarity measures exist according to the external system wherein sentimental analysis is utilized. The linguistics feature and domain relevant features are essential for providing the better classification of text[2]. Hence, in this system, consideration the gamut of keywords associated with the feature is essential for successful classification. The algorithm explained revolves around the expansion and better understanding of the model proposed by Dave, Lawrence and Pennock[8].

3. PROPOSED TECHNIQUES

An overview of transfer of reviews file is depicted in figure 1. The e-commerce website generates the review of various products (for defining scope, limited products, i.e. mobile phones are used). The review file is then stored in an SQL database through a JDBC connectivity. The processing mechanism on this file takes place in Hadoop system (single cluster) and the output generated is displayed back in the e-commerce website in the form of the progress bar.

Figure 1. Architectural Overview

In order to process data, review data should be present. Web scraping which is defined as the process for gleaning information for the Internet is performed on various e-commerce platform for extracting reviews to act as training data. Once training data is acquired, next algorithmic steps could be performed on the reviews text file. The review text file is in the following format, which eases processing in Hadoop Architecture.

"*PRODUCT_NAME<tab separator>REVIEW1*"

"*MOTO_G<tab separator>The display quality of the "phone" is fantastic.*"

The proposed mechanism for feature – based opinion summarization takes place in the Hadoop system and its architectural overview is illustrated in figure 2.

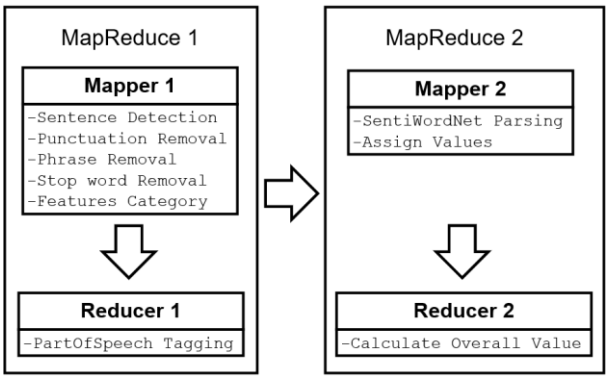

Figure 2. Procedural Overview

3.1 First Map-Reduce

3.1.1 Sentence Detection

A review is not necessarily always written in a single line. Most of the time, it is in a form of paragraphs. Sentence Detection allows detection and segregation of sentences from the paragraphs which can then be processed.

3.1.2 Punctuation Removal

Punctuations and special characters are to be removed from the sentences such that only alphabets and number are left in the sentences. The sentences are also entirely converted into lower cases.

3.1.3 Phrase Removal

Phrases such as 'could have been', 'hope it will be' are removed and replaced by a negation word.

3.1.4 Stop Words Removal

Stop words are considered as meaningless words which are filtered out to reduce the processing time. This list consists of the preposition, conjunctions, articles, etc.

3.1.5 Feature Category

It represents searching for features related words in the sentence and then classifying in the same feature cluster. For example, the review data set is parsed for keywords/ feature (such as display: display, screen, gorilla glass, resolution, color, pixels) which are generated by finding frequent item set through Apriori Algorithm[7]. After finding such word in the sentence, the sentence is classified in the cluster of only those features (such that in display cluster, only sentence relevant to display will be stored).

3.1.6 Parts-Of-Speech(POS) Tagging

The POS tagging model is applied to the sentences thereby providing part of speech of each word in the sentences.

Apache's OpenNLP has been used to perform Sentence Detection and POS tagging. POS tagging model uses Maximum Entropy Model for analyzing information gain on training data and provides parts of speech tags to new sentences. The reason for removing 'opinion changing phrases' before stopwords can be understood by the given example.

"The memory of the phone could have been better."

In this case, if the stop words are removed directly then the opinion of sentence changes. In the given example, the opinion is negative as they expect memory to be better, but if stopwords are removed then the remaining words are: 'memory phone better'. This, when processed for sentiment gives an influence on positive sentiment which is

contradictory. Hence, using phrase removal before stopword removal acts as a solution, so that a negative word can be substituted in such cases and the remaining words left after phrase removal and stopword removal are 'memory phone not better', which gives a sense of negative influence in the sentence. So the first Map-Reduce provides an output of POS tagged sentences placed in the feature cluster to the second Map-Reduce.

3.2 Second Map-Reduce

Sentimental words can be defined as describing words. In English Language, such describing word can be clubbed under two categories: Adjectives (which describes noun, pronoun such as, good phone), and Adverbs (which describes verbs, such as, a fast processor). Keeping track of such describing word which are present in the processed sentences received from first Map-Reduce affects later stages. After performing POS tagging, following steps are performed:

3.2.1 Words Classifier

According to Penn Treebank POS tags, all variations of 'JJ' represent adjectives and 'RB' represents adverbs. Such POS tags are searched in the processed sentences are collected to provide values of opinion generated.

3.2.2 SentiWordNet Values

With the help of SentiWordNet[4], an open source lexical resource, the Objective, Negative, and Positive scores of the words under consideration are procured. The maximum score amongst these scores are utilized.

3.2.3 Calculate Overall Value

The score obtained are then averaged to get an accurate estimation of the opinion expressed relevant to the feature.

The problem arises while dealing with adjectives/adverbs preceded by a negative word. The following sentence is an example of such case.

"The phone camera is not good"

To solve this, the value procured from the adjectives, in this case the value associated with the word 'good' = 0.75 is multiplied by -1 to indicate the negative influence on positive word. The values thus calculated are then displayed on the website. Table 1 provides a detailed illustration of these steps.

Table 1: Illustration of Map-Reduce processes

Steps	Output in Key-Value pair
Input	*MOTO_G<tab separator>The display quality of the "phone" is fantastic and awesome. The music ratio of the phone….could have been better. While on the 'other hand' the battery is not so good.*
Sentence Detection	*MOTO_G<tab separator>The display quality of the "phone" is fantastic and awesome.* *MOTO_G<tab separator>The music ratio of the phone….could have been better.* *MOTO_G<tab separator>While on the 'other hand' the battery is not so good.*
Punctuation Removal	*MOTO_G<tab separator>the display quality of the phone is fantastic and awesome* *MOTO_G<tab separator>the music ratio of the phone could have been better* *MOTO_G<tab separator>while on the other hand the battery is not so good*
Phrase Removal	*MOTO_G<tab separator>the display quality of the phone is fantastic and awesome* *MOTO_G<tab separator>the music ratio of the phone not better* *MOTO_G<tab separator>while on the other hand the battery is not so good*
Stop Words Removal	*MOTO_G<tab separator>display quality phone fantastic awesome* *MOTO_G<tab separator>music ratio phone not better* *MOTO_G<tab separator>hand battery not good*
Feature Category	*MOTO_G,DISPLAY<tab separator>display quality phone fantastic awesome* *MOTO_G,SOUND<tab separator>music ratio phone not better* *MOTO_G,BATTERY<tab separator>hand battery not good*
Parts-Of-Speech Tagging (JJ-Adjective & RB Verb)	*MOTO_G,DISPLAY<tab separator>NN,display NN,quality NN,phone JJ,fantastic JJ,awesome* *MOTO_G,SOUND<tab separator>NN,music NN,ratio NN,phone RB,not RB,better* *MOTO_G,BATTERY<tab separator>NN,hand NN,battery RB,not JJ,good*
Words Classifier	*MOTO_G,DISPLAY,ADJECTIVE<tab separator>fantastic, awesome* *MOTO_G,SOUND,NEGATIVEADJECTIVE<tab separator>better* *MOTO_G,BATTERY,NEGATIVEADJECTIVE<tab separator>good*
SentiWordNet Values	*MOTO_G,DISPLAY,ADJECTIVE<tab separator>* 0.241, 0.25 *MOTO_G,SOUND,NEGATIVEADJECTIVE<tab separator>* -0.416 *MOTO_G,BATTERY,NEGATIVEADJECTIVE<tab separator>* -0.25
Calculate Overall Value	*MOTO_G,DISPLAY<tab separator>* 0.2455 *MOTO_G,SOUND<tab separator>* -0.416 *MOTO_G,BATTERY<tab separator>* -0.25

4. CONCLUSION

The system developed aims to achieve an efficient mechanism for summarizing the reviews posted by customer to help other potential customer. It enable many ecommerce websites in the need of time to substitute there 'upvote' system for surfacing helpful reviews with the proposed system which doesn't involve manual intervention in rating process. The system provide information in a graphical progress bar adjacent to each features where each score is displayed by the length of the bar as shown in fig 3 and fig 4. We assert that usage of techniques and mechanism provided by Hadoop System such as Key – Value pair and MapReduce significantly reduces the time complexity of system with such intensive processing.

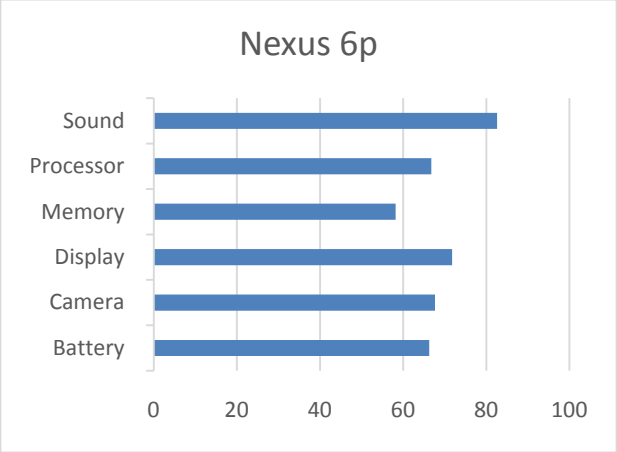

Figure 3: Graphical Representation of Output

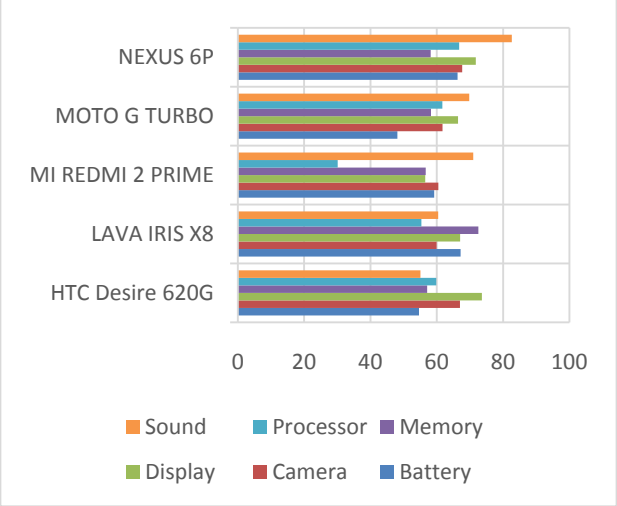

Figure 4: Collaborative Representation for Comparison between Products

5. FUTURE WORK

In future work, these techniques and rating process can be improved by taking into consideration the usage of slangs term used by people. Features can also be clubbed together according to the score as good, neutral, and bad. Spam reviews can be detected and removed from the list to increase the overall efficiency (Algorithm can be developed to check whether features are present in the reviews posted or not).

6. REFERENCES

[1] S. Chandrakala and C. Sindhu. "Opinion Mining and Sentiment Classification: A Survey" ICTACT Journal on Soft Computing, October 2012, Volume: 03, issue: 01, ISSN: 2229-6956.

[2] Yun Niu, MSc, Xiaodan Zhu, MSc, Jianhua Li, MSc and Graeme Hirst, PHD. "*Analysis of Polarity Information in Medical Text*" AMIA 2005 Symposium.

[3] Nidhi Mishra, C. K. Jha, PHD. "Classification of Opinion Mining Techniques". International Journal of Computer Applications (0975 – 8887), Volume 56– no.13, October 2012.

[4] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining" LREC 2010.

[5] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N, Prasad. M. R., *"Analysis of Big Data using Apache Hadoop and Map Reduce"* Volume 4, Issue 5, May 2014, India.

[6] Minqing Hu and Bing Liu, "*Mining Opinion Features in Customer Reviews*", American Association for Artificial Intelligence, 2004.

[7] Othman Yahya, Osman Hegazy, Ehab Ezat, "*An Efficient Implementation Of Apriori Algorithm Based On Hadoop - MapReduce Model*", International Journal of Reviews in Computing, ISSN: 2076-3328.

[8] Kushal Dave, Steve Lawrence, David M. Pennock, "*Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*", ACM 2003.

[9] Kavita Ganesan, ChengXiang Zhai & Evelyne Viegas. "Micropinion Generation: An Unsupervised Approach to Generating Ultra-Concise Summaries of Opinions" (2012) Lyon, France, April 16–20.

[10] Tingting Wei, Yonghe Lu c, Huiyou Chang, Qiang Zhou, Xianyu Bao "*A semantic approach for text clustering using WordNet and lexical chains*" (2014) China, 18 October.

[11] Sinno Jialin Pany, Xiaochuan Niz, Jian-Tao Sunz, Qiang Yangy and Zheng Chen (2010) North Carolina, USA, April 26–30. "*Cross-Domain Sentiment Classification via Spectral Feature Alignment*".

[12] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N, Prasad .M .R *"Analysis of Bidgata using Apache Hadoop and Map Reduce"* Volume 4, Issue 5, May 2014, India.