SENTIMENT ANALYSIS ON PRODUCT REVIEWS USING BIG DATA

[1]Mrs. J. Samatha, [2]Vaibhav Lahoti, [3]Anusha.V, [4]Deekshitha. A
[1]Asst.Professor,[2]BE Scholar,[3] BE Scholar,[4] BE Scholar
[1,2,3,4]Computer Science and Engineering, Matrusri Engineering College, Hyderabad, India

Abstract : Most of the e-commerce sites raise their customers to supply relevant reviews on their merchandise that might facilitate alternative customers to choose their alternative. A slew of reviews is being generated on a each day attributable to a rise within the usage of ecommerce sites. a possible client may have to travel through thousands of reviews before inward at a firm call, that is long. Many users give reviews for the single product. Such thousands of review can be analyzed using big data effectively. The results can be presented in a convenient visual form for the non-technical user. Thus, the primary goal of research work is the classification of customer reviews given for the product in the map-reduce framework

Index Terms: **Sentiment Analysis, Opinion Mining, Product Reviews, Hadoop, MapReduce, Sentiwordnet.**

I. INTRODUCTION

The growth in the field of e-commerce has semiconductor diode to a revolutionary modification within the commerce method. People‟s viewpoint[2] have shifted from ancient commerce to ecommerce within the past years. so as to get a lot of traffic and increase in sales, merchants have enabled customers to share their opinion of the merchandise. Consequently, the reviews area unit generated at a high rate. Merchants give “votes” mechanism[1] whereby the potential client vote reviews that he/she thought-about useful. Such extremely voted reviews area are then surfaced at the highest of review list so the potential client gets the gist of the product whereas perusal fewer reviews. Rather than utilizing manual efforts to vote useful reviews, sentiment analysis is aimed to modify the method of rating on feature - based mostly opinion summarization[3] of reviews.

1.1 Big Data

Big data is data set whose size is beyond the ability of typical database software tools[6] to capture, store, manage, and analyze. Such data come from everywhere: pictures and videos, online records, and comments from social media etc. Big data are not just about sheer volume in terabytes though. Other important aspects have been emphasized in addition to volume, including variety, velocity.

1.2 Sentiment Analysis

Sentiment analysis is a subfield of Artificial intelligence focused on parsing the given text and planned its opinion in terms of positive, negative or neutral text. Feature – based opinion summarization identify the features in the given review and expresses the sentiment appropriate 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 likely customer might be able to slight down his choices of the product if he is interested in specific features and also ease him in comparing the products.

1.3 Apache Hadoop

The Apache Hadoop project is an open-source software build for scalable [4] and distributed computing. It provides giving out techniques [8] that allows for large scale giving out 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 essential to be integers only, Strings are also allowed. As for the range of this paper, we are interested in the Map Reduce technique. The name is derived from the steps it performs, “MAP‟ step – implement in mapper class and “REDUCE‟ step – implement 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[9], however, is liable 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 valuable.

II. METHODOLOGY

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

Figure 1. Architectural Overview

Sentiment classification is to be done in Hadoop environment. SentiWordNet[10] is used to assign sentiment values using mapper job. SentiWordNet is a dictionary which has a positive, negative and objective score for each sentimental word. It has around 19000 adjectives and noun which expresses emotion. In order to practice data, review data should be present. Web scraping which is defined as the process for gleaning information for the Internet is performed on a variety of ecommerce platform for extract reviews to act as training data. Once training data is acquired, next algorithmic steps could be processed on the reviews text file. The review text file is in the following format, which eases processing in Hadoop Architecture[8].

"*PRODUCT_NAME<tab separator>REVIEW1*"
"*MOTO_G<tab separator>The display quality of the*
"phone" is fantastic."

The methods introduced are built on opinion words that are commonly used in expressing positive or negative sentiment. This type of approaches uses SentiWordNet[10] , tagged corpora with a positive score and a negative score along with part of speech tags. [12] proposed a novel lexicon based approach for opinion mining and also their own published lexicon dataset[11].

The proposed mechanism for feature – based opinion summarization takes place in the Hadoop[6] system and its architectural overview is illustrated in figure 2.

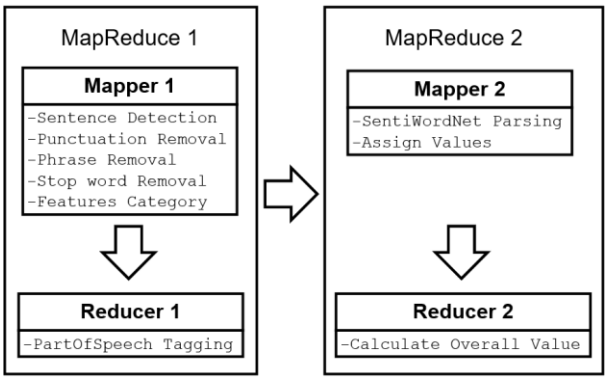

Figure 2. Procedural Overview

2.1 First Map-Reduce

2.1.1 Sentence Detection

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

2.1.2 Punctuation Removal

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

2.1.3 Phrase Removal

Phrases such as „could have been", "hope it will be" are detached and replace by a negation word.

2.1.4 Stop Words Removal

Stop words are considered as worthless words which are clean out to reduce the processing time. This listing consists of the preposition, conjunctions, articles, etc.

2.1.5 Feature Category

It represents search for features associated words in the sentence and then classify in the same feature group. For example, the evaluation 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 classify in the cluster of only those features (such that in display cluster, only sentence related to display will be stored).

2.1.6 Parts-Of-Speech(POS) Tagging

The POS tagging model is applied to the sentences thereby provided that part of speech of each word in the sentences. Apache"s OpenNLP has been used to do Sentence Detection and POS tagging. POS tagging model use Maximum Entropy Model to analyze information gain on exercise data and provides parts of speech tags to new sentences. The cause for removing „opinion altering phrases" before stop words can be known by the given example.

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

In this case, if the stop words are detached directly then the opinion of sentence change. In the given example, the review is negative as they guess memory to be better, but if stopwords are detached then the remaining words are:

„memory phone better". This, when process for sentiment gives an influence on positive sentiment which is contradictory. Hence, using phrase elimination before stopword removal acts as a clarification, 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 provide an output of POS tagged sentences located in the feature cluster to the second Map-Reduce.

2.2 Second Map-Reduce

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

2.2.1 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 generate.

2.2.2 SentiWordNet Values

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

2.2.3 Calculate Overall Value

The score taken are then averaged to get an precise estimation of the opinion described relevant to the feature. The problem arise while dealing with adjectives/adverbs preceded by a negative word. The subsequent sentence is an example of such case.

"The phone camera is not good"

To solve this, the value procure from the adjectives, in this case the value related with the word „good" = 0.75 is multiplied by -1 to designate the negative influence on positive word. The values thus calculated are then displayed on the website. Table 1 provides a detailed picture of these steps.

Table 1: Illustration of Map-Reduce processes

Steps	Output in Key-Value pair
Input	*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	*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.*
Punctuation Removal	*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*
Phrase Removal	*the display quality of the phone is fantastic and awesome* *the music ratio of the phone not better* *while on the other hand the battery is not so good*
Stop Words Removal	*display quality phone fantastic awesome* *music ratio phone not better* *hand battery not good*
Feature Category	*display quality phone fantastic awesome* *music ratio phone not better* *hand battery not good*
Parts-Of-Speech Tagging (JJ-Adjective & RB Verb)	*NN,display NN,quality NN,phone JJ,fantastic JJ,awesome* *NN,music NN,ratio NN,phone RB,not RB,better* *NN,hand NN,battery RB,not JJ,good*
Words Classifier	*fantastic, awesome* *better* *good*
SentiWordNet Values	0.241, 0.25 -0.416 -0.25
Calculate Overall Value	0.2455 -0.416 -0.25

III. CONCLUSION

The proposed method in this paper aims how to improve the quality of sentiment analysis on textual product reviews using Hadoop framework. Also, the methodology is based on training and testing will improve the accuracy of results of analysis. The focus is on the use of open source technologies mainly. However, proposed system has remarkable practical applications for both individual customer and service provider. The individual customer takes its advantage for decision making and service provider can take benefit to improve the quality of service as well as for new product design.

IV. FUTURE WORK

In future work, these technique and rating process can be enhanced by taking into consideration the usage of review 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. For future we will be planning to analyse other formats of data posted on product reviews like any images(JPEG,GIF).Analyse data from videos(MP4) posted as reviews.

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] Nidhi Mishra, C. K. Jha, PHD. "Classification of Opinion Mining Techniques". International Journal of Computer Applications (0975 – 8887), Volume 56– no.13, October 2012.
[3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining" LREC 2010.
[4] 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.
[5] Kushal Dave, Steve Lawrence, David M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", ACM 2003.
[6] Kavita Ganesan, ChengXiang Zhai & Evelyne Viegas. "Micropinion Generation: An Unsupervised Approach to Generating Ultra-Concise Summaries of Opinions" (2012) Lyon, France, April 16–20.
[7] 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".
[8] 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.
[9] Ding, X., Liu, B. and Yu, P. A Holistic Lexicon-Based Approach to Opinion Mining. Proceedings of the first ACM International Conference on Web search and Data Mining(WSDM'08),2008
[10] SentiWordNet, http://sentiwordnet.isti.cnr.it/
[11] Enock Kanyesigye, Sumitra Menerea, "Sentiment Analysis Of Reviews Using Hadoop",IJARIIE Vol-2,Issue 2,2016
[12] MugdhaJinturkar Pradnya Gotmare,"Sentiment Analysis of Customer Review Data Using BigData:A Survey",International Journal of Computer Applications, Emerging Trends in Computing 2016.