www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20320-20323
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.30

Spam Reviews Detection Using Hadoop

Akshay Chavan[1] , Omkar Darekar[2], Omkar Kulkarni[3], Yash Jain[4]
[1,2,3,4]B.E Student,
Dept. of Information Technology
Shah & Anchor Kutchhi Engineering College
Mumbai, India
akshayc278@gmail.com, omkardarekar85@gmail.com, omkark1495@gmail.com, yashjain199522@gmail.com

***ABSTRACT*-Product reviews are now widely used by individuals and organizations for their decision making. However, due to the reason of profit or fame, people try to manipulate the system by opinion spamming (e.g., writing spam reviews) to promote or demote some target products. For reviews to reflect genuine user experiences and opinions, such spam reviews should be detected. The project elaborated below aims at filtering the spam reviews, by providing an effective method. Usage of MapReduce technique provided by Apache Hadoop is highly emphasized for processing reviews. In this paper, the technique used for the same is described which substantially reduces time complexity when implemented.**

Keywords-***Outlier Detection, POS Tagging, Sentimental Analysis, Spam Reviews.***

1. INTRODUCTION

The enhancement in the field of e-commerce has led to a revolutionary change in the trading process. People's view point 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. But, since there are no constraints on posting review on an e-commerce website, some people write spam reviews. These people have ulterior motives and hence their reviews are not genuine. Since, it is not possible to detect and filter spam reviews manually. There needs to be a system which can detect such spam reviews in order to protect the customer's interest and also the maintain the true value of the product. The project makes first attempt to investigate opinion spam in reviews and proposes a technique to identify spam review using sentimental analysis and statistics. The system is to be develop in such a way that depending on outlier aspects the reviews will get separated

1.1 Parts-Of-Speech Tagging

Using OpenNLP package provided by Stanford University, each review in the word is assigned a part of speech such as noun, verb, adjective, etc.

1.2 Sentimental Analysis

Sentiment analysis [2] 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.

The fake negative reviewers are seen to over-produce negative emotion terms relative to the truthful reviews in the same way that fake positive reviewers over-produced positive emotion terms. Therefore, fake reviewers exaggerate the sentiment. We can use the above observation to a lot a sentiment score to each review using lexical resource and then try to find out which review sentiment deviates from normal sentiment.

1.3 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 .Hadoop works when the input/output in the format of Key-Value pairs. Let us consider an example to illustrate this:

Key	Value
1200	Hello world!

Here, "1200" 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 number of studies in the past have focused on traditional spam detection in e-mail and on the web. However, only recently there are studies examined the opinion spam. Jindal and Liu performed some of the first studies of this nature. (Jindal, 2007)[1] They focused on three types of disruptive opinion spam, including spam containing advertisements and other non-related text. While these types of spam may be distracting where they are easily detectable by human readers.

On the other hand, the focus of our paper is detecting opinion spam, written with the specific intent of misleading customers which may be difficult for humans to detect.
Spam detection provides an unusual scenario in the assessment of human-created data, since machine-based methods have been shown to outperform human judges.
In this paper, we are interested in hotel reviews. Even if we can borrow some ideas from previous studies, their clues are not sufficient enough to define hotel review spammers. Hence, there is a need to look for a more sophisticated and complementary framework based on sentimental analysis and outlier detection.

Hu and Liu[12] summarized a list of positive words and a list of negative words, respectively, based on customer reviews. The

positive list contains 2006 words and the negative list has 4783 words. Both lists also include some misspelled words that are frequently present in social media content. Sentiment categorization is essentially a classification problem, where features that contain opinions or sentiment information should be identified before the classification. For feature selection, Pang and Lee suggested to remove objective sentences by extracting subjective ones. They proposed a text-categorization technique that is able to identify subjective content using minimum cut. Gann et al. selected 6,799 tokens based on Twitter data, where each token is assigned a sentiment score, namely TSI(Total Sentiment Index), featuring itself as a positive token or a negative token.

3. PROPOSED SYSTEM

The e-commerce website contain lots of reviews about products or services .The data is stored in data repository .

Figure 1. Architectural Overview

The primary source of such data is online data repositories for education purpose. So the data can downloaded from repository. The processing mechanism on this file takes place in Hadoop system (single cluster).

3.1 Data Gathering

The dataset which we are using for our project is been downloaded from UCI Machine Learning Repository. From this Repository we are going to use Hotel Reviews to find whether they are spam or not. This dataset contains sentences extracted from user reviews on a given topic. Example topics are performance of Toyota Camry and sound quality of ipod nano , etc. In total there are 51 such topics with each topic having approximately 100 sentences (on the average). The reviews were obtained from various sources - Tripadvisor (hotels), Edmunds.com (cars) and Amazon.com (various electronics).

3.2 Hadoop

The purpose of using Hadoop for this project is that we a larger vision of expanding this system which can handle any amount of data and can process it at faster rate. The user reviews are generally in large numbers and a normal processor maybe unable to process it.

Hadoop is the core platform for structuring Big Data [7], and solves the problem of formatting it for subsequent analytics purposes.

First Map-Reduce

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.

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.*

3.2.1.1 Phrase Removal

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

3.2.1.2 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.

3.2.1.3 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

3.2.1.4 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 stop words can be understood by the given example. “The infrastructure of hotel 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 stop words are removed then the remaining words are: “infrastructure better”. This, when processed for sentiment gives an influence on positive sentiment which is contradictory. Hence, using phrase removal before stop word 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 stop word removal are “infrastructure 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.

3.2.2 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:

3.2.2.1 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.

3.2.2.2 SentiWordNet Values

With the help of SentiWordNet [6], 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.

3.2.2.3 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 hotel infrastructure 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 saved in file.

3.2.3 Third Map-Reduce

3.2.3.1 Sort the data

In order to process the obtained data we first have to sort the available values in an ascending order . After which IQR outlier detection method is applied on sorted data.

3.2.3.2 Outlier Detection

DOI: 10.18535/ijecs/v6i2.30

To identify the reviews that are deviating from the normal sentiment, we use the concept of inter-quartile range to detect the outliers.

The IQR [11] is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

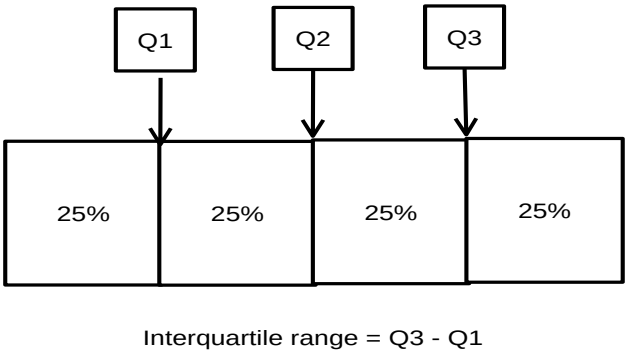

Figure 2. IQROverview

Q1 – First quartile
Q2- Second quartile
Q3-Third quartile

Since, reviews will be converted to numeric values after sentimental analysis, if the review score is greater the Q3+1.5IQR or less than Q1-1.5IQR. Then, it will be marked as spam review.

Interquartile Range =Q3 – Q1

Spam Review value > Q3+1.5*IQR and

Spam Review value < Q3-1.5*IQR are separated as spam .

3.2.4 Fouth Step :Map Reviews

3.2.4.1 Map the unique ID

The Output from Third Map-Reduce Then maps to previous file containing the unique ID and respective sentimental value by setting reducer job as "0"(No Reducer Job),then it check for Spam review condition that is

Spam Review value > Q3+1.5*IQR and

Spam Review value < Q3-1.5*IQR

And forward the unique ID from one Map function to further Map function given below.

3.2.4.2 Map the review

Now here the unique ID is act as key to Map the Review from input file (setting Reducer Job as "0"No Reducer Job)and then reviews get extracted from that file to new file .

3.2.5 Display reviews

The output file is then displayed which contain spam reviews.

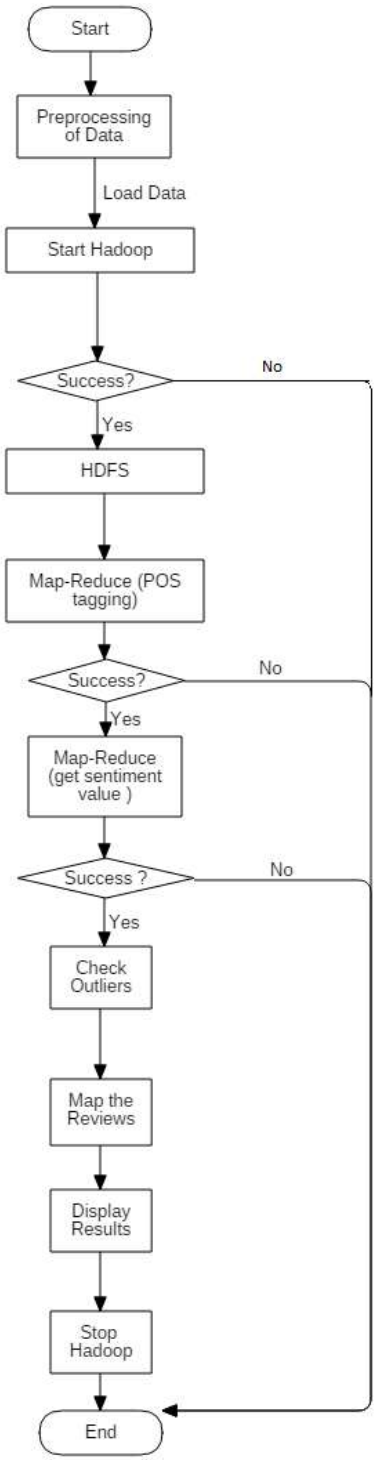

Figure 3 : The above figure depicts the flow of the system using HDFS and Map-reduce

4. CONCLUSION

In the proposed system the business analyst will be able to differentiate whether the review is spam or genuine. So, the system that we are developing will make an attempt to maintain the integrity of the reviews that are posted online.It

will avoid false promotion or deliberate de-motion of a particular product by identifying the reviews which deviate from the general sentiments of the people about that product. This system will equally benefit the businesses as well as the consumers. 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.

5. FUTURE WORK

In future work, this can use to optimize the data for proper operations ,so results will be exact .This will leads to increment in efficiency and accuracy of task which is using such processed data .

6. REFERENCES

[1] N. Jindal and B. Liu. Analyzing and Detecting Review Spam. ICDM2007.

[2] B. Pang, L. Lee & S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. EMNLP'2002.

[3] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," in ICWSM, 2013.

[4] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?" in ICWSM, 2013.

[5] J. Mehta , R. Patil , J. Patil , M.Somani , S.Varma "Sentimental analysis on product Reviews using Hadoop" in IJCA,2016 .

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

[7] 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.

[8] Jindal, N. & Liu, B. Review Analysis. Tech. Report, 2007.

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

[10] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In CIKM, 2010.

[11] Hodge, V.J. and Austin, J. orcid" A survey of outlier detection methodologies"2004

[12] M. Hu & B. Liu. Mining and summarizing customer reviews. KDD'2004.