Enhanced Classification Analysis for Product Based Customer Reviews using Big Data

Shwetha RP
Department of Information Technology
Jerusalem College of Engineering, Pallikaranai, Chennai

Shelma S
Department of Information Technology
Jerusalem College of Engineering, Pallikaranai, Chennai

Lilly Sheeba S
Department of Information Technology
Jerusalem College of Engineering, Pallikaranai, Chennai

Abstract

In recent days many web applications tend to collect product reviews from customers to ascertain the satisfaction level on specific products. Big Data analysis can go hand in hand with product reviews in order to bring about useful information that can assist executives and managers in making high end decisions. Big Data analysis not only enables executives to get relevant data in less time but also enables them to carry forward fraudulent analysis, customer segmentation and product recommendations. The proposed system is a web application that takes product reviews from customers as input and performs Enhanced Classification Analysis (ECA) on the reviews using Hadoop to categorize reviews as positive and negative feedbacks. The categorized and segregated positive reviews of products with better comments will be displayed to incoming customers as reports while marketing products. This analysis enables both manufacturers to predict public opinion of their product and also customers to make better decisions and incorporate improved services.

Keywords: Product Reviews, Big Data, Fraudulent Analysis, Customer Segmentation and Product Recommendations

I. INTRODUCTION

Initially data analytics was employed to analyze large volumes of stored data generated by an organization to extract useful information which helps managers to make better decisions and thereby enhance the profit incurred from products they manufactured. Now with exhaustive growth among competent organizations, new technologies and tools should be at hand to analyze huge volumes of information stored in various heterogeneous sources to predict future business trends.

Big data is widely applied in the fields of marketing, governance, healthcare, defense and social media analytics. Big data analytics is extremely useful in the manufacturing, since it can forecast demand changes, and match the supply as required. It also provides organizations the ability to segment customers based on many socioeconomic characteristics.

The characteristics of big data include volume, variety, velocity, and veracity. Velocity is defined as rate at which the data is created and volume refers to the data size. Variety refers to different formats and data types. Veracity defines the quality of data whether the data is good, bad or undefined.

Reviews have become an integral part of online shopping nowadays. Consumers are ready to provide reviews so that others will not repeat their mistakes. Large scale analysis will reveal information that will be extremely helpful to the customer as well as the seller. The output of this analysis involves calculating the number of positive and negative reviews and recommending products to similar customers. This helps customers to make quicker purchasing decisions and allows manufacturers to understand public opinion of their products. It also saves time since they need not read all the reviews.

II. RELATED WORK

Topics usually were mined from documents by using the classic topic model LDA. In recent years knowledge-based topic models were used which had the requirement of users providing domain knowledge prior to mining. Research was done to find a solution for lifelong or incremental learning. One of the solutions was to mine two types of knowledge mustlinks and cannotlinks. The results were found to be dramatically better than state-of-the-art mining algorithms.

One of the methods used for calculating the relative importance of textual units is the Extractive TS. It depends on sentence salience and identifies the most important sentences from a document or more. Salience is measured using the presence of selective important words. An advanced method for computing sentence importance is by using eigenvector centrality when sentences are graphically represented.

Text analysis also includes sentiment analysis. It is the process of determining opinion in a text. Various classifiers are available for sentiment analysis of user opinion. It is easy to classify text into classes of interest using such classifiers.

A particular application of product recommendation was shown in the travel industry of Taiwan. The idea was that lead users of a product would be able to influence customers more to buy that particular product. For the travel industry experiment, the

lead users chosen were travel bloggers to endorse trip itineraries designed by travel agencies. Hierarchical regression analysis was used to find the relationship between the two. The results showed that this relation is subject to culture and quality of communication available.

Former research in the area of product recommendation includes trying to derive demographic information from mentions of product adopters from online reviews. Product adopters are those who have bought the product for someone else. These adopters are then categorized into user groups based on their physical locations. The results are then incorporated by recommending products to those in similar user groups using regularized matrix factorization.

There are over 190 million tweets received from twitter and companies receive more product surveys, and reviews about the product from the customers. These tweets and reviews are used as a source to identify customer

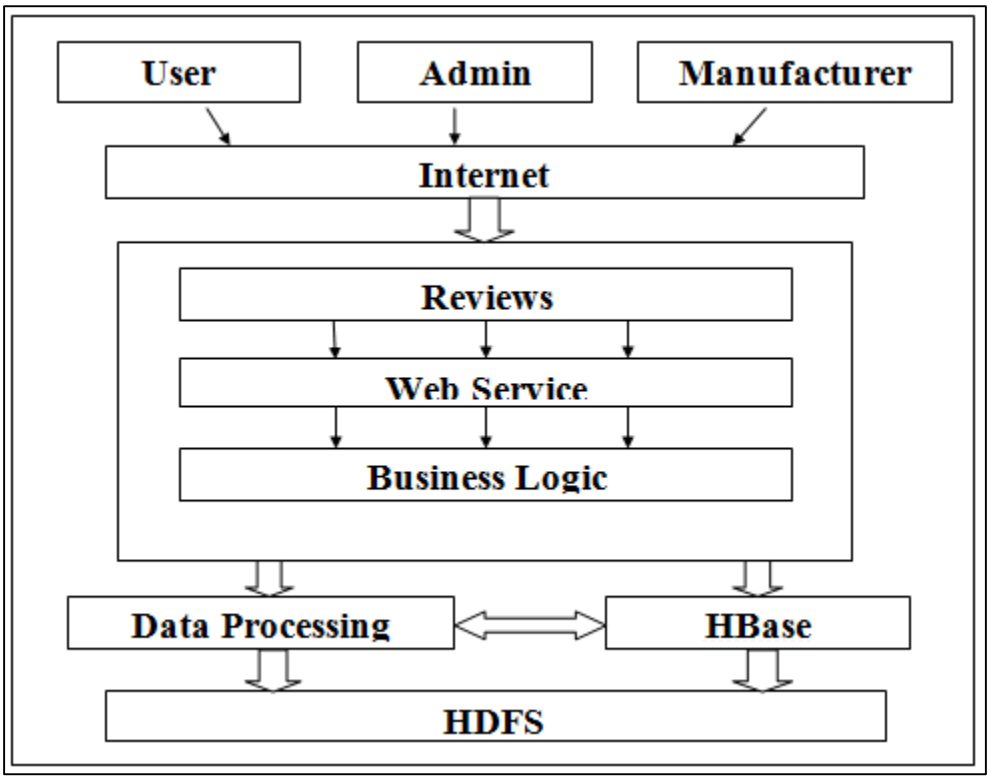
III. PROPOSED SYSTEM

The proposed system allows user groups comprising of administrator, common user and manufacturers including top management. The web service here acts as an interface between the user and the back end where large volumes of customer reviews are stored. The web service performs analysis on the customer reviews to classify them as positive reviews and negative reviews. While positive reviews expose the satisfaction level of the customer, negative reviews assist the manufacturers in enhancing the quality of the product.

In the proposed system the user are allowed to purchase products and give reviews about the products they purchased. The admin are allowed to update the product details and delete the product, the manufacturer is allowed to view the customer reviews of their products and this review helps the manufacturer to improvise their product and services. This is clearly depiced in Fig.1. The goal is to develop a classifier which performs sentiment analysis, by categorizing the users comment as positive or negative. The positive comments are used to recommend to new customers to assist them in shopping online.

The Apache Hadoop software is a framework that processes the large set of data in a cluster of computers. It is a way to store and process data which are distributed in various systems. In Hadoop, processing is done to the huge data in a cost effective manner. The apache HBase is a non relational database (distributed database) and a table is provided to store and process data. HBase runs on the top of HDFS as in Fig

Hbase is a distributed light weight database resides at the top of HDFS. HDFS is Distributed File System used to store unstructured data and HDFS has one Name Node and many Data nodes, Name Nodes are the Master and the Data Nodes are the slaves.



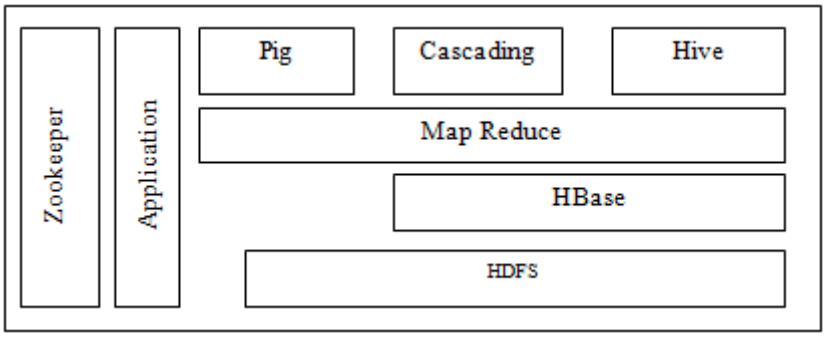


Fig. 2: Hadoop Structure

IV. CONCLUSION

In this paper we have proposed a system that makes the process of analyzing online reviews easier and more accurate. A novel model is proposed that not only segregates comments but also recommends it to future customers.

REFERENCES

[1] Rutuja Tikait, Ranjana Badre, Mayura Kinikar ,N.2014 .Product Aspect Ranking Techniques: A Survey Information Review An ISO 3297: 2007 Certified Organization, Vol. 2, Issue 11.
[2] Ganu, G., Kakodkar, Y., and Marian, A. 2013. Improving the quality of predictions using textual information in online user reviews. Information Systems 38(1):1–15.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. the Journal of Machine Learning research, 3:993,1022, 2003.
[4] Ganu, G.; Elhadad, N., and Marian, A. 2009. Beyond the stars: Improving rating predictions using review text content.In Proceedings of the 12th International Workshop on the Web and Databases (WebDB).
[5] Giering, M. 2008. Retail sales prediction and item recommendations using customer demographics at store level. ACM SIGKDD Explorations Newsletter 10(2).
[6] Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 168–177.
[7] Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), 263–272.
[8] Jamali, M., and Ester, M. 2009. Trustwalker: a random walk model for combining trust-based and item-based recommendation. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), 397–406
[9] http://www.snia.org/sites/default/education/tutorials/2013/fall/BigData/SergeBazhievsky_Introduction_to_Hadoop_MapReduce_v2.pdf
[10] http://web.cs.wpi.edu/~cs561/s12/Lectures/6/Hadoop.pdf