Product Review Analysis using Big Data Analytics

[1]Jovitha Christina .R.H [2]Mary Caroline Nikitha [3]Kavitha .V
[1,2]Research Scholar [3]Assistant Professor
[1,2,3]Department of Information Technology Engineering
[1,2,3]Loyola-ICAM College of Engineering and Technology
Chennai, India

Abstract

Sentiment Analysis is the process of using text analytics to mine various data sources for opinions. Often, sentiment analysis is done on the data that is got from the Internet and from various social media platforms. Because the content collected from the internet is unstructured, we need tools that can process and analyze this disparate data. Hence we make use of Big Data to handle the different sources and formats of the structured and unstructured data. In particular consumer reviews of a product are given in textual format, we first parse the reviews and classify them into positive and negative and then send these datasets to the Hadoop File System (HDFS) to analyze them. This helps the purchaser to have some knowledge about the product's pros and cons and decide which product to buy.

Keyword- Big Data, HDFS, HIVE, Key Generation, Sentiment Analysis, Sqoop

I. INTRODUCTION

User reviews are something that the majority of customers will want to see before deciding, to make a purchase. They are proven sales-drivers. They help to eliminate any doubts potential customers may have about a product, or can help in selecting a product .However Fake reviews and false ratings are the most annoying of them. They compromise the customer's trust in the product and lower the quality of the shopping experience .Thus in our project, we generate a random key for each purchaser who buys the product so that, only the people with deeper perspective and experience with the product can enter that key to give their reviews .This helps in avoiding fake reviews from being published. Since the amount of reviews got may be large in number, we make use of big data to analyse them. Big Data, in general refers to the large amount of data which are structured, semi-structured or unstructured that has the potential to be mined or analysed for information. We make use of Hadoop tools like Sqoop and HIVE in our project. Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases like MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. HIVE is a tool that resides on top of Hadoop to summarize Big Data, and makes querying and analysing easy.

II. LITERATURE SURVEY

In this paper, the authors Yubo Chen and Jinhong Xie argues that online consumer review, a type of product information created by users based on personal usage experience, can serve as a new element in the marketing communications mix and work as free "sales assistants" to help consumers identify the products that best match their idiosyncratic usage conditions

M. Gamon demonstrates that it is possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. He shows that by using large feature vectors in combination with feature reduction, hw can train linear support vector machines that achieve high classification accuracy on data that present classification challenges even for a human annotator. He also shows that, surprisingly, the addition of deep linguistic analysis features to a set of surface level word n-gram features contributes consistently to classification accuracy in this domain.

D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura empirically show that their method of sentiment classification of sentences using word-level polarity, almost always improves the performance of sentiment classification of sentences especially when they have only small amount of training data.

Bo Pang and Lillian Lee focused on methods that seek to address the new challenges raised by sentiment aware applications, as compared to those that are already present in more traditional fact-based analysis. They include material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rise to

S. Li, R. Xia, C. Zong and C.Huang performed experiments on data sets from both topic-based and sentiment classification tasks which shows that their new method is robust across different tasks and numbers of selected features.

III. OBJECTIVE

Our project's main objective or focus is to ensure that the customers benefit from a fulfilling shopping experience on its platform by getting reviews only from people who have used it. The random key that is generated, prevents others from writing fake reviews about the products online. The goal of the system is to provide detailed analysis of the product.

IV. PROPOSED SYSTEM

In this paper, we propose a simple yet efficient algorithm called Clustering algorithm to analyse the dual bags (positive and negative reviews).Then we make use of Sqoop to transfer the data between the MySQL database to the Hadoop File System (HDFS) and that datasets are analysed using HIVE.

V. SYSTEM REQUIREMENTS

A. Hardware Requirements

An I3 1.3 Intel quad-core processor is used which is a chip with four separate cores that read and execute CPU instructions such as add, move data and branch. Within the chip, each core operates in conjunction with other circuits such as cache, memory management, and input/output ports. A 1.3 GHZ quad-core CPU will have each CPU running at 1.3GHZ.

A 4GB RAM is also used which is a form of computer data storage which stores the instructions of the frequently used programs to increase the speed of a system and allows data items to be read or written in almost the same amount of time irrespective of the physical location of data inside the memory.

B. Software Requirements

1) ECLIPSE

Eclipse is an integrated development environment (IDE) used in programming, and is the most widely used Java IDE .It contains a base workspace and an extensible plug-in system for customizing the environment. Eclipse is written mostly in Java and its primary use is for developing Java applications, but it may also be used to develop applications in other programming languages through the use of plugins.

2) MySQL Database

MySQL is an open source relational database management system (RDBMS) which is mainly used for developing web-based software applications.

3) Windows 10 OS

Windows 10 operating system harmonizes the user experience and functionality between different classes of devices. Windows 10 incorporates multi-factor authentication technology based upon standards developed by the FIDO Alliance. It includes improved support for biometric authentication through the windows hello and passport platforms. The passport platform allows networks, software and websites to authenticate users using either a PIN or biometric login to verify their identity, without sending a password.

4) Ubuntu (OS)

Ubuntu is the most popular operating system running in hosted environments, so–called "clouds", as it is the most popular server Linux distribution .The Ubuntu project is publicly committed to the principles of open-source software development where people are encouraged to use free software, study how it works, improve upon it, and distribute it.

5) Hadoop Framework

Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users. Hadoop Distributed File System (HDFS) is a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster.

VI. SYSTEM DESIGN

System Design involves identification of classes their relationship as well as their collaboration. In objector, classes are divided into entity classes and control classes. The Computer Aided Software Engineering (CASE) tools take advantage of Meta modelling that is helpful only after the construction of the class diagram. The analyst creates the user case diagram. The designer creates the class diagram.

But the designer can do this only after the analyst creates the use case diagram. Once the design is over, it is essential to decide which software is suitable for the application.

VII. ALGORITHM

A. User Module

1) Input
User Login name and Password

2) Output
(If valid cloud user) Open the user window otherwise error page.

B. Admin module

1) Input
Admin Login name and Password

2) Output
Opens the admin window

C. Key Generation

1) Input
User purchases the products.

2) Output
Random Key is generated

D. Review Sharing

1) Input
User enters the random key

2) Output
User can share the reviews online

E. Data integration with sqoop

1) Input
Given dataset from MySQL

2) Output
Data transferred to HDFS

F. Data Analytics with Hive

1) Input
Data stored in HDFS

2) Output
Required Analysis Report

VIII. SYSTEM ARCHITECTURE

A system architecture is a conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system.
System Architecture for product review analysis using Big Data analytics is given below

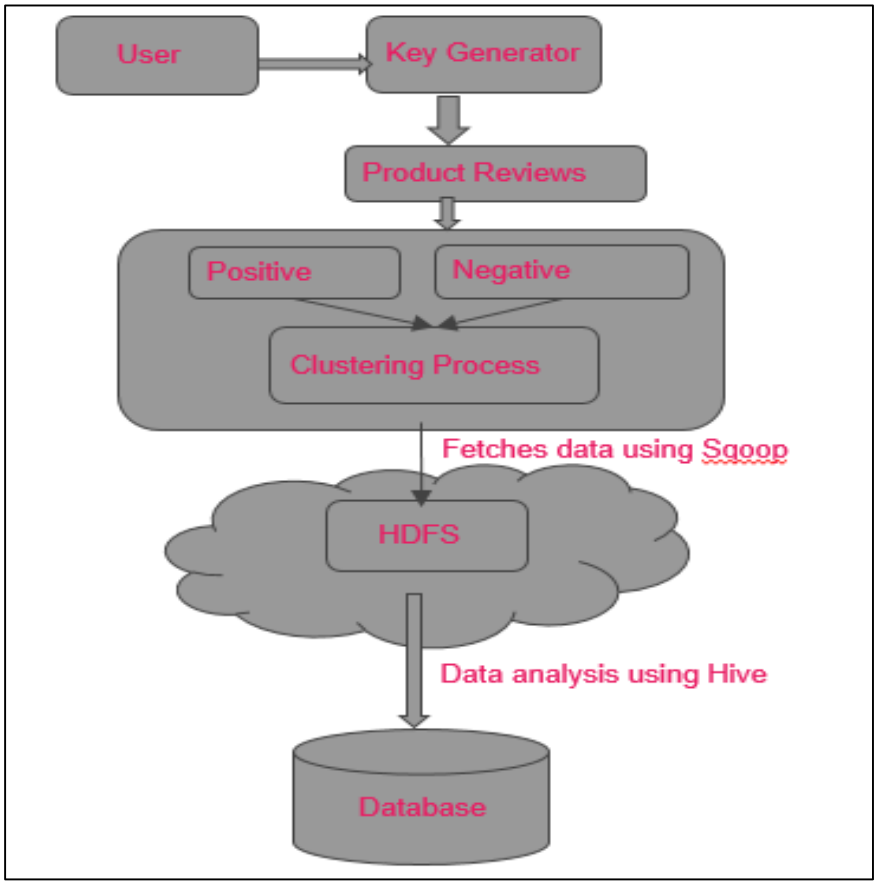


Fig. 1: Product Review Analysis using Big Data analytics

IX. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system.

X. MODULES

A. Modules List

1) Admin module
2) User module
3) Key generation & Review Sharing
4) Review Analysis
5) Data integration using Sqoop
6) Data analysis using Hive

B. Modules Description

1) Admin Module

This module stores and updates the information of the products. The admin logins into the site using his id and password and prevents unauthorized users from gaining access to it. He updates about the product if there are any changes and will also be able to view the reviews of the customers.

2) User Module

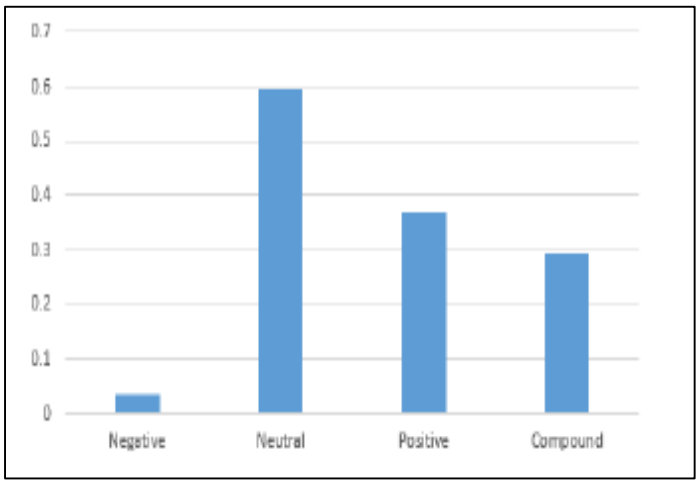
The user registers into the system to search for the desired product. If he enters any invalid username or password, he cannot enter into the login window as it will show some error message. This is to prevent unauthorized users from entering into the system. The user's login details will be stored in the database.

3) Key Generation & Review Sharing

This module is used to help the purchaser to share their opinions or reviews about the purchased product. A random key is generated for each buyer so that they can enter this key while sharing their reviews online. This prevents fake reviews from people who have not used this product.

4) Review Analysis

This module generates a graph on the comparison of the products based on the analysis of the reviews got from the users who have used the products.



5) Data Integration Using Sqoop

The main aim of this module is to transfer the datasets from the MySQL database to the Hadoop file system (HDFS) using Sqoop. Sqoop is a command-line interface application for transferring data between relational databases and Hadoop.

6) Data Analysis Using Hive

Hive is a data warehouse system for Hadoop. It runs SQL like queries called HQL (Hive query language) which gets internally converted to map reduce jobs. Hive was developed by Facebook. Hive supports Data definition Language (DDL), Data Manipulation Language (DML) and user defined functions. In this module, the dataset is analysed using HIVE tool and display the results.

C. Modules Implementation

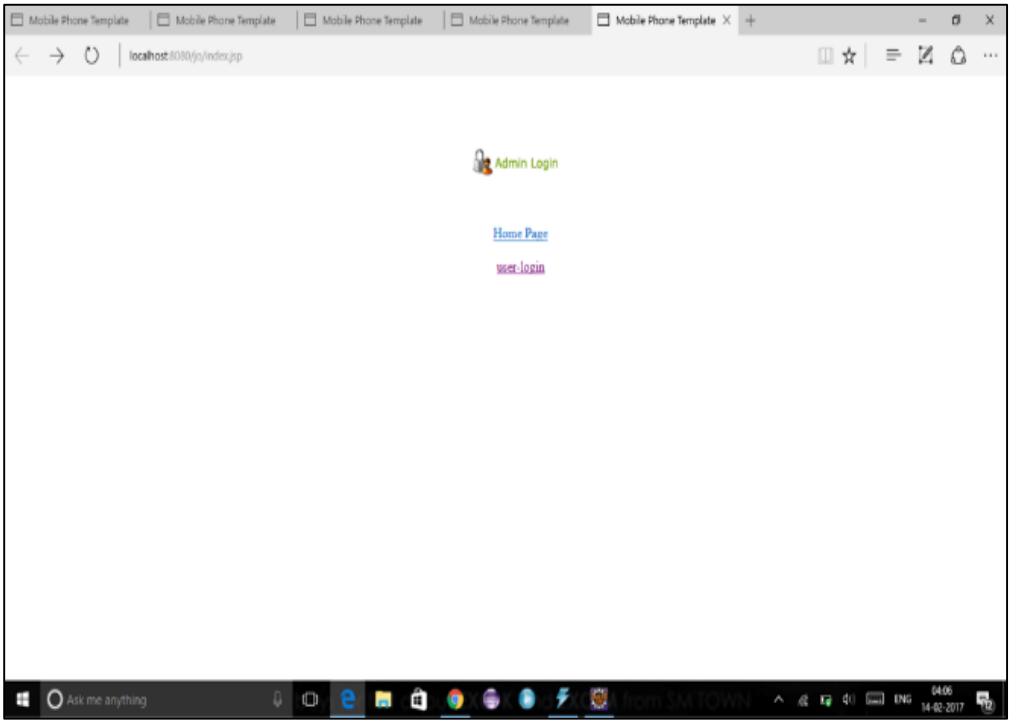


Fig. 2: Home Page

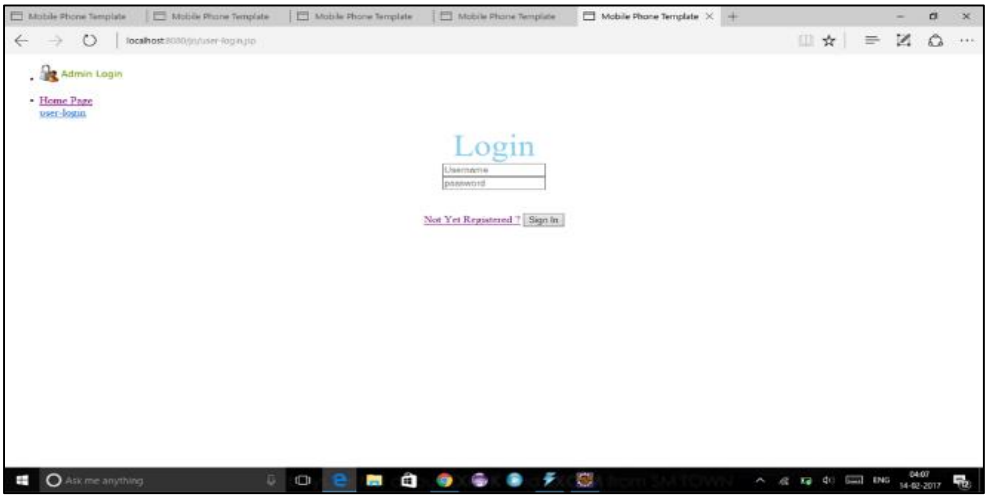


Fig. 2.2: User Login Page

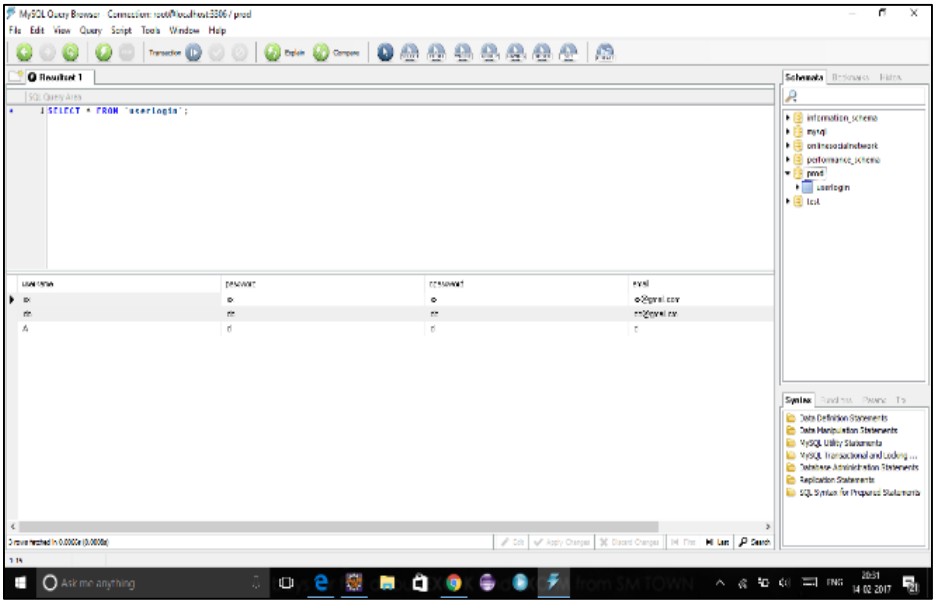


Fig. 2.3: MySQL Database Connection- User login

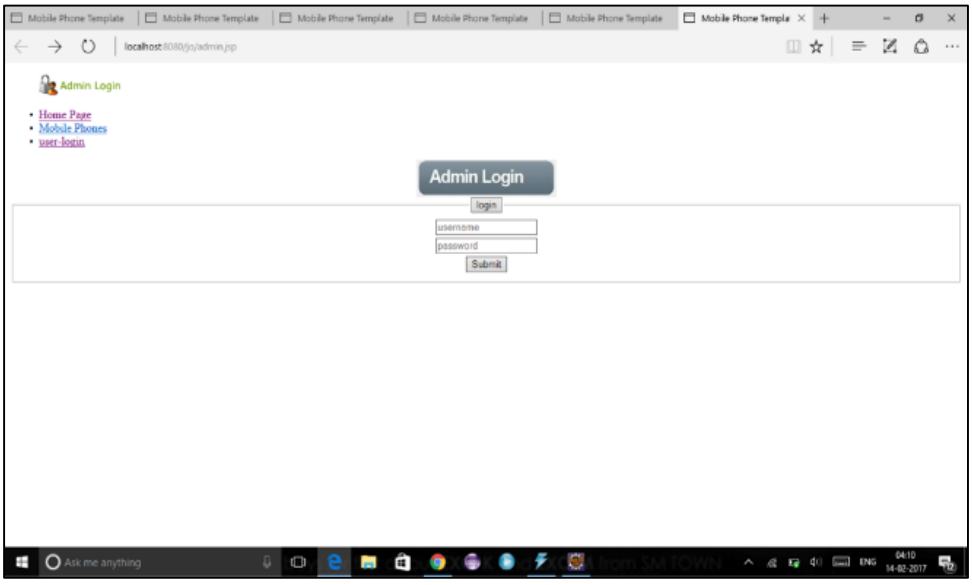


Fig. 2.4: Admin Login Page

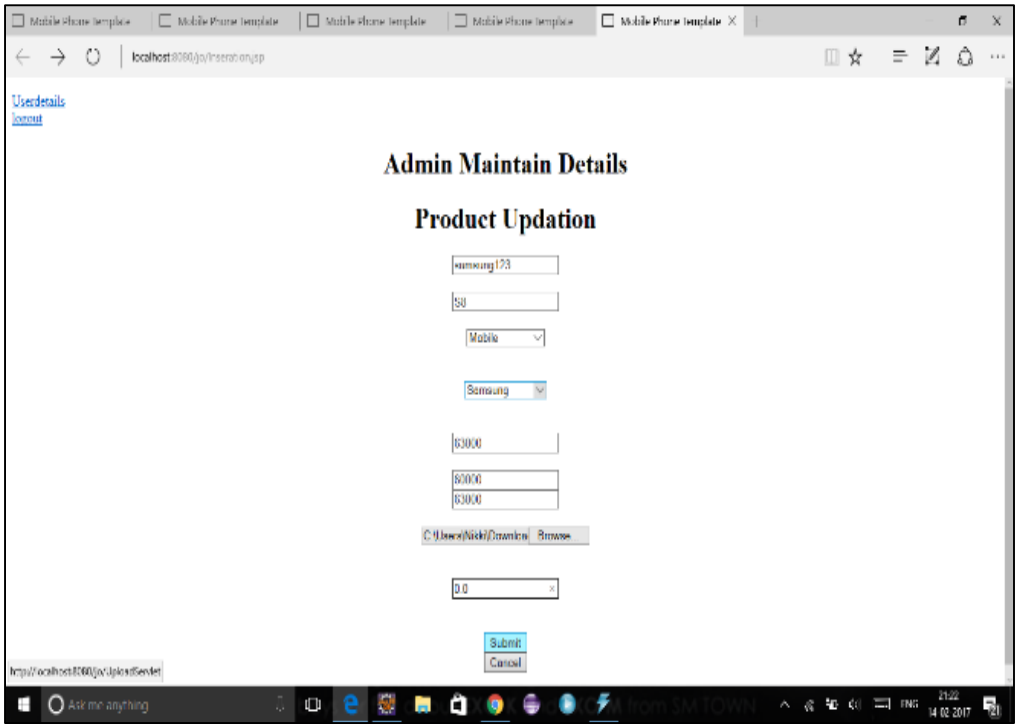


Figure 2.5: Product Updation

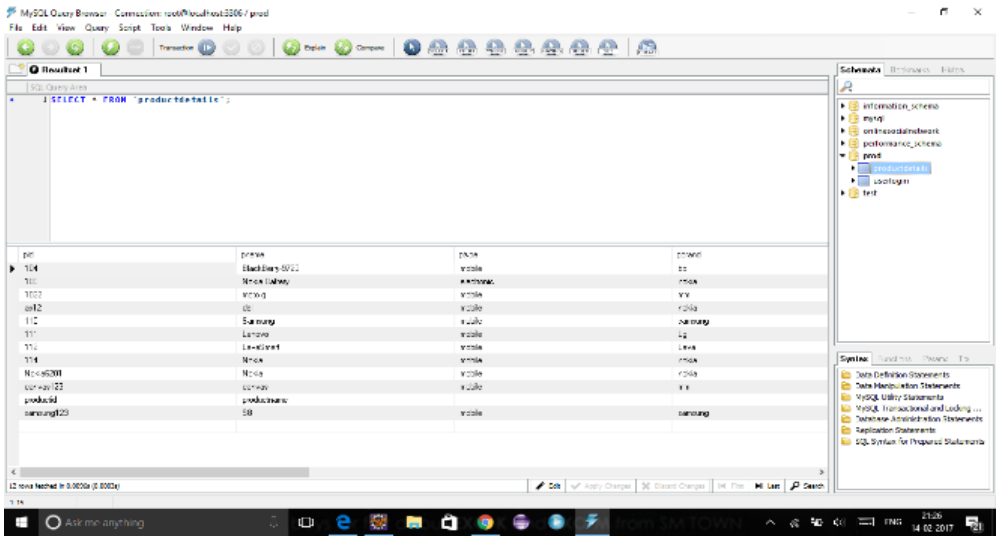


Fig. 2.6: MySQL Database Connection-Admin

XI. CONCLUSION

All reviews are valuable, and a mix of both positive and negative reviews helps to improve consumer trust in the opinions they read. Our project's main objective is to ensure that the customers benefit from a fulfilling shopping experience on its platform by getting reviews only from people who have used it. We make sure that the detailed analysis of the product is sufficient enough for the purchaser to quickly make a decision to buy the product without second thoughts.

XII. FUTURE ENHANCEMENTS

In the future, we can generalize the Clustering algorithms to a wider range of sentiment analysis tasks. We also plan to consider more complex polarity shift patterns such as transitional, subjunctive and sentiment-inconsistent sentences. This might be of greater advantage in the future.

REFERENCES

[1] Bing Liu,"Exploring User Opinions in Recommender Systems", Proceeding of the second KDD workshop on Large Scale Recommender Systems and the Netflix Prize Competition", Aug 24, 2008, Las Vegas, Nevada, USA.
[2] Lina Zhou, Pimwadee Chaovalit, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii International Conference on system sciences, 2005.
[3] B. B. Khairullah Khan, Aurangzeb Khan, "Sentence based sentiment classification from online customer reviews," ACM, 2010.
[4] Tomasz Jacha, Ewa Magieraa, Wojciech Froelicha "Application of HADOOP to Store and Process Big Data Gathered from an Urban Water Distribution System " Procedia Engineering 119 on 2015,pp.1375 – 1380
[5] Y. Chen, S. Alspaugh, D. Borthakur, R. Katz, "Energy efficiency for large-scale mapreduce workloads with significant interactive analysis", in: Proceedings of the 7th ACM European conference on Computer Systems, EuroSys '12, 2012,pp. 43–56.
[6] F. L. Gang Li, "A clustering-based approach on sentiment analysis," IEEE, 2010.
[7] M. X. Y. S. Haiping Zhang, Zhengang Yu, "Feature- level sentiment analysis for chinese product reviews," IEEE, 2011.
[8] Suresh Ramanujam,R. Nancyamala, J.Nivedha, J.Kokila "Sentiment analysis using big data"
[9] Andry Alamsyah, Marisa Paryasto, Feriza J. Putra, Rizal Himmawan, "Network text analysis to summarize online conversations for marketing intelligence efforts in telecommunication industry", Information and Communication Technology (ICoICT) 2016 4th International Conference on, pp. 1-5, 2016.

