Sentiment Analysis on Amazon Alexa using Big Data Analytics

[1] Mrs. S. Geetha, [2] P. Jothi Lakshmi
[1]Associate Professor, [2]Research Scholar
[1,2]Department of Computer Science
[1,2]Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India-620002

Abstract: Traditional data analytics systems collect huge static volumes of data and periodically process those data. On the other hand, streaming analytics systems avoid putting data at rest, and process it as it is coming to the system or as it becomes available, cutting the data analytics time significantly and minimizing the time a single data spends in a processing pipeline. In recent years, batch processing was thriving with offline data processing. Both Hadoop and Batch processing have progressed with time to become excellent offline data processing platforms for Big Data. Recently, things have changed drastically, as many use cases across various domains started calling for near-real-time/ real-time response on Big Data for faster decision making in their respective field. Hadoop was not suitable enough for those use cases. The rate at which organizations were analyzing their data was slow, and the data was losing its value exponentially over time. In this project work, the revenue of the company has developed with finding the reviews patterns of the Amazon Alexa among the users by using the following steps: Hadoop installation, Dataset collection, Pre-processing using Chi-Square analysis and how Flume can take log files as source and it can store it directly to file system like HDFS and the Amazon Alexa data analyzation with the implementation of Kafka stream engine in this project work in this research work. This work aims to relate user's ratings and reviews to discover how beneficial and good a product. User ratings are collected and are analyzed based on different categories (datasets) which gives an insight as to which product performs good and what are the problems associated to a certain non-performing product**.**

IndexTerms **- Data Mining, Big Data Analytics, Pre-processing, Hadoop Distributed File System, Amazon Alexa, Review Mining, Chi-Square analysis.**

I. INTRODUCTION

Extremely huge datasets that cannot be analyzed by a normal database software to find useful patterns, meaningful insights, behavior or decisions usually related to human activities is termed as Big Data[1].Data is proliferating day by day in terms of its size, complexity, variety and veracity. A set of new technologies and techniques is needed to uncover and integrate meaningful insights from a large, complex and diversified dataset. The data gathered from web, industries, research works, medicines or any other source participates in the creation of a large structured or unstructured dataset. Big Data is a mixture of both unstructured data (e-mails, manuals, documents, records etc.) and structured data (CSV's and XLS's files).

The Apache Hadoop is a framework that can scale up thousands of machines to process and store large Datasets [4] [9]. Apache Hadoop does not depend upon hardware to deliver high computation and performance; it's built in library can handle the problems [5]. HDFS, Hadoop Distributed File System is a java-based architecture that is responsible for storing large volumes of data [8] [13]. HDFS can store upto 198PB of data and each cluster capable of processing around 4000 servers that can further support millions of files [9].

MapReduce is a fast processing fault-tolerant paradigm that can process huge amounts of data. MapReduce splits the data into two procedures, Map() and Reduce() [8]. The Map() procedure is responsible for taking the initial values and and producing intermediate key values while Reduce() procedure is responsible for processing the intermediate key values to the final results [10]. MapReduce works on the basis of Master-Slave architecture. Hbase is database management systems that runs above HDFS and provides read and write access [1]. It is similar to Google's big table that is used to provide random access to huge structured data. Hbase does not support structured query language like SQL. Hbase is a column-oriented database i.e. it stores data as columns of data rather than rows of data, column oriented database is used for online Analytical Processing (OLAP) [7]. Apache Hive, initially developed by Facebook, used by many other companies like Netflix and Financial Industry [11]. It helps in data summarization, query and analysis of structured data by working on top of Apache Hadoop. Apache Hive allows the SQL developers to Hive Query Language that are similar to SQL commands [14]. Apache Hive further breaks these HQL statements into Hive Jobs for further execution through the Hadoop We could either run them in Hive shell (command line interface) and Java Database Connectivity. Hive is not appropriate for the applications that require very fast write operations [14]. It is a read based tool that cannot be used for transaction and other purposes that require very high rate of write operations [11].

Apache Pig is a platform developed at Yahoo Research for analyzing large data sets [11]. The SQL-like scripting language used by Apache Pig is called Pig Latin. Pig's jobs can be executed in MapReduce or Apache Spark. It allows users to perform desired data manipulations and build complex applications in Hadoop with the help of User Defined Functions (UDF) [1] [5]. Sequence of steps can be assigned to define collective result. SQL can be used for analysis of small data sets but the manipulation of large data sets usually requires the involvement of experienced programmers with Pig Latin Approach [2]. Pig Latin ease programmers from writing MapReduce functions for low level operations. The attributes are in the form of int, double, float, long and char array. Pig programs can be executed in Script, Grunt and Embedded which works on both MapReduce and Local mode [8]. Pig Latin helps us to accomplish highly complex processing tasks due to its Extensive nature [10]. Depending upon the requirements our execution can be automatically optimized by the system. Due to its parallel processing it can simultaneously process multiple computations [4].

Amazon, an American e-commerce and cloud computing company was founded in July, 1994. It is known for Earth's biggest store of CDs, books, a utomobile spare parts, Kids toys, electronics, Hardware etc. It is also known for manufacturing consumer electronics – Amazon Kindle, Amazon Alexa, Echo and many more. Amazon also let authors and publishers to publish and make their books available at the Kindle Store, with "Amazon Publishing" publishing arm on it [3].

II. RELATED WORKS

Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh [15] Sentiment Analysis and Opinion Mining is a most popular field to analyze and find out insights from text data from various sources like Facebook, Twitter, and Amazon, etc. It plays a vital role in enabling the businesses to work actively on improving the business strategy and gain an in-depth insight of the buyer's feedback about their product. It involves computational study of behavior of an individual in terms of his buying interest and then mining his opinions about a company's business entity. This entity can be visualized as an event, individual, blog post or product experience. In this paper, Dataset has taken from Amazon which contains reviews of Camera, Laptops, Mobile phones, tablets, TVs, video surveillance. After preprocessing we applied machine learning algorithms to classify reviews that are positive or negative. This paper concludes that, Machine Learning Techniques gives best results to classify the Products Reviews. Naïve Bayes got accuracy 98.17% and Support Vector machine got accuracy 93.54% for Camera Reviews.

Bose, Rajesh, et al [16] Today, people are exchanging their thoughts through online Web forums, blogs, and different social media platforms. Sometimes, they are giving reviews and opinions on different products, brand, and their services. Their reviews toward a product not only improve the product quality but also influence purchase decisions of the consumers. Thus, product review analysis is a widely accepted platform where consumer can easily aware of their requirements. In this experiment, we track 568,454 fine food reviews of 74,258 products and 256,059 users on Amazon over a period of ten years. To analyze the result, we select six most popular products and users based on the plain text review, and NRC emotion lexicon is used which can be categorized eight basic emotions and two sentiments. Word cloud also help our research to make comparisons between the eight emotion categories. Our results show that how sentiment analysis will help to identify the consumers' behaviors and overcome those risks to meet the consumers' satisfaction.

Khan, Aurangzeb, et al [17] The user reviews about the different products posted on social media sites, provide an opportunity to opinion mining researchers to develop applications capable of performing comparative opinion mining on different products. Therefore, it is an important task of investigating the applicability of different supervised machine learning algorithms with respect to classification of comparative reviews. In this work different machine learning algorithms are applied for performing multi-class classification of comparative user reviews into different classes. The results show that Random Forest outperforms amongst all other classifiers used in the research.

Pandey, Avinash Chandra, Saksham Raj Seth, and Mahima Varshney [18] In this paper, NLTK has been used which is a Python toolkit to harness the power of generating information from the huge text datasets available. Sampled data from Amazon Alexa has been collected which is further processed using SentiWordNet 3.0 and TextBlob to remove noise and irrelevant data. Thereafter, Gaussian naive Bayes algorithm along with TextBlob has been used to detect sarcasm in dataset. The performance of the proposed method is compared with naïve Bayes, decision tree, and support vector machine. From the experimental results, effectiveness of the proposed method is observed.

Devi, DV Nagarjuna, et al [19] In this study, we proposed an algorithm named ASSAY (which means Analysis), to find the polarity at the document level. In our algorithm, initially we classify the reviews of each domain using naive Bayes and Support Vector Machine (SVM) algorithms which are in machine learning approach and then find the polarity at document level using HARN's algorithm which comes under lexicon-based approach. In this algorithm, we use TextBlob for Parts of Speech (POS) tagging, where NV-Dictionary, ordinary dictionary, and SentiWordNet are used for extracting the polarities of features. Here, we combine both machine learning and lexicon-based approaches to calculate the result at document level accurately. In this way, we get the result about 80–85% more accurately than HARN's algorithm which is proposed in lexicon-based approach.

III. PROBLEM STATEMENT

In the previous study similarities measurements like Text similarity, Semantic similarity, Sosine similarity, Jaccard similarity, Euclidean distance, Minkowski distance, Pearson correlation, Spearman correlation, Tanimoto coefficient are considered for the sentiment measurement in the text documents. The lexical or lexicon approach is a method for teaching dictionary based described by Michael Lewis in the early 1990s. The basic concept of this approach respites an idea that the significant part of education involves understanding and produce lexical phrases as chunks. In this pattern of language like grammar as well as consume meaningful set of words at their dumping.

- Lack in accuracy
- Increased error rates
- Less precision

IV. PROPOSED METHODOLOGY

In the proposed system, Data Mining techniques like Feature Selection, Association Rule Mining are used to improve the Decision Making. Instead of utilizing RDBMS, the analytical process has been done on the Alexa Review dataset for the business intelligence purpose. In this proposed system, Numeric to Nominal filtering method is used to convert the numeric values into nominal values. Then the feature selection technique like Chi-Square attribute selection is used to reduce the size of the feature space. Support Vector Machine is used for the classification process. This is used to analyse the behaviour of the customer with their customer review to make sure the loyalty of the customer in the business market by using HDFS with Hive functionality and analyzation with Kafka Stream Engine.

4.1Pre-Processing using Chi-Square analysis

Feature selection is an example of the common prominent and frequent techniques in data pre-processing and has converted a crucial part of the machine learning method it has also distinguished as attribute selection, variable subset or variable selection in statistics and machine learning. It is the process of identifying relevant and eliminating irrelevant features, redundant or noisy data. This method rushes up data mining algorithms, improves predictive accuracy and understandability. Irrelevant features are those that provide no valuable knowledge, and irrelevant features present no further information than the presently selected features. Chi-Square analysis has used in this project work to carry out the pre-processing step [20][21][22].

- In this module, a rank-based feature selection technique called Chi-Square analysis has done.

- Chi-Square analysis is the hypothetical testing method, which is used to choose the best features among the dataset using the measurement of null hypothesis and alternative hypothesis.
- With the degree of freedom, observed frequency and expected frequency, the best feature subset has obtained using chi-square method.
- Feature selection technique has carried out to improve the classification accuracy of the alexa review dataset.

3.2 J48 Classification Technique

A decision tree partitions the input space of a data set into commonly select regions, each of which is apportioned a label, a value or an action to portray its statistics points. The decision tree mechanism is apparent and we can monitor a tree structure easy to understand how the decision is made. A decision tree is a tree structure comprising of internal and external nodes associated by branches [14]. Moreover, an internal node is a decision-making unit that appraises a decision function to regulate which child node to follow subsequently. On the other hand, the external node has no child nodes and it is related to a brand or value that symbolizes the certain data that indicates to its being visited. However, many decision tree construction algorithms involve a two-step process. First, a very huge decision tree is developed. At that moment, to lessen the large size and overfitting the data, in the second step, the given tree is pruned. The trimmed decision tree that is accustomed to categorization drives is known as a classification tree. To form a decision tree, we must compute entropy and information gain [23][24].

3.3 ID3 Classification Technique

ID3 is a modest decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm [15] is to create the decision tree by retaining a top-down, greedy exploration over the given sets to test each attribute at every tree node. To Facilitate and select the character that is most suitable for categorizing a given set, and introduce a metric---information gain. To identify a positive method to categorize a learning set, what is needed to do to reduce the questions asked (i.e. minimizing the depth of the tree). In consequence, some function which can calculate the questions deliver the most balanced splitting. Therefore, the information gain metric is such a task [25][26][27].

3.4 CART Classification Technique

Decision Trees are frequently used in data mining with the aim of producing a model that forecasts the value of a target (or dependent variable) based on the values of several inputs (or independent variables) [16]. At present, it depicts CART decision tree methodology. The CART or Classification & Regression Trees methodology was presented in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone as a canopy term to refer to the subsequent kinds of decision trees:

- **Classification Trees**: where the objective variable is definite and the tree is practiced to detect the "class" within which an objective variable would probably fall into.
- Regression Trees: where the objective variable is constant and the tree is deployed to forecast it's value.

The CART algorithm is designed as a structure of questions, the responses to which control the next question if any should be. The outcome of these queries is a tree like an organization where the finishes are fatal nodes at which point there are no more queries.

The main elements of CART (and any decision tree algorithm) are:

1. Guidelines for excruciating data at a node based on the value of one variable;
2. Preventing rules for determining when a branch is terminal and can be fragmented no more; and

Finally, a forecast for the objective variable in each terminal node

3.5 HDFS with Hive

Hadoop Hive is an open source SQL-based distributed warehouse system which is proposed to solve the problems mentioned above by providing an SQL-like abstraction on top of Hadoop framework. Hive is an SQL-to-MapReduce translator with an SQL dialect, HiveQL, for querying data stored in a cluster. When users want to benefit from both MapReduce and SQL, mapping SQL statements to MapReduce tasks can become a very difficult job. Hive does this work by translating queries to MapReduce jobs, thereby exploiting the scalability of Hadoop while presenting a familiar SQL abstraction.

- In this module, the installation of Hadoop takes place. The command JPS (Java Process Status) has used to check the running application in the Hadoop environment [22].
- The starting of Distributed File System (DFS) and YARN (Yet Another Resource Navigator) has done in this module.
- Hive is used in this module to run the MapReduce jobs.
- MapReduce is used in this module to filter the size of the reviews with appropriate rating and stars

3.6 Analyzation with Kafka Stream Engine

Apache Kafka quickly routes real-time information to consumers, and is a message broker that provides seamless integration. The two-collateral objective of it is to not block producers and to not let the producers know who the final consumers. he few examples of Apache Kafka use cases are, it commits logs, does log analysis, stream processing, record user activity, etc. Kafka streams, compared with other streaming engines, are relatively new. It is fast, scalable, fault tolerant, and offers high throughput. Mainly, it is used for log processing. It provides an API similar to messaging system, and allows applications to consume log events in real time.

- In this module, Apache Kafka is used to analyse the streaming reviews of amazon alexa from twitter.
- By sharing the twitter privacy keys, the tweets about the Amazon Alexa has extracted and analysed by using Kafka.
- Kafka in this module has used to record the activity of the tweets

V. RESULT AND DISCUSSION

Table 1 contributes to the performance analysis of the J48 classification technique for the novel dataset and pre-processed dataset (Chi-Squared processed dataset). From table 1, the pre-processed dataset stretches high accuracy of 95.34%, augmented Kappa

Statistic, Rate, Precision, True Positive F-Measure, and ROC area. Reduced error rates like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Absolute Error (RAE), Root Relative Squared Error (RRSE). Table 2 portrays the performance examination of the ID3 classification method for the novel and pre-processed dataset. From table 2, it is vibrant that the pre-processed dataset in the ID3 classification method achieves highly in features. Table 3 describes the performance analysis of the CART method for the novel dataset and pre-processed dataset.

Table 1: Performance analysis of the J48 classification techniques for the original dataset and pre-processed dataset

Performance Metrics	Type of Dataset	
	Original Dataset	**Pre-Processed Dataset**
Accuracy	77.3367 %	95.34 %
Kappa Statistic	0.5672	0.9279
MAE	0.2397	0.2267
RMSE	0.3248	0.2802
RAE	124.2287 %	66.6271%
RRSE	105.4194 %	68.0246%
TPR	0.782	0.963
FPR	0.172	0.021
Precision	0.812	0.966
F-Measure	0.791	0.963
ROC Area	0.892	0.982

Table 2: Performance analysis of the CART classification techniques for the original dataset and pre-processed dataset

Performance Metrics	Type of Dataset	
	Original Dataset	**Pre-Processed Dataset**
Accuracy	82.0253 %	98 %
Kappa Statistic	0.625	0.9609
MAE	0.0879	0.0296
RMSE	0.2691	0.145
RAE	45.8877 %	8.7118 %
RRSE	87.3506 %	35.2013 %
TPR	0.820	0.980
FPR	0.181	0.016
Precision	0.813	0.980
F-Measure	0.813	0.980
ROC Area	0.808	1.0

Table 3: Performance analysis of the ID3 classification techniques for the original dataset and pre-processed dataset

Performance Metrics	Type of Dataset	
	Original Dataset	**Pre-Processed Dataset**
Accuracy	74.4304 %	90.5%
Kappa Statistic	0.4341	0.8067
MAE	0.1023	0.0633
RMSE	0.3198	0.2517
RAE	53.0184 %	18.6164 %
RRSE	103.813 %	61.0936 %
TPR	0.744	0.905
FPR	0.327	0.108
Precision	0.73	0.908
F-Measure	0.7165	0.892
ROC Area	0.705	0.898

Figure 1 represents the performance analysis of the J48, ID3, and CART classification techniques. From figure 1, it is evident that the CART classification offers maximum accuracy for a pre-processed dataset.

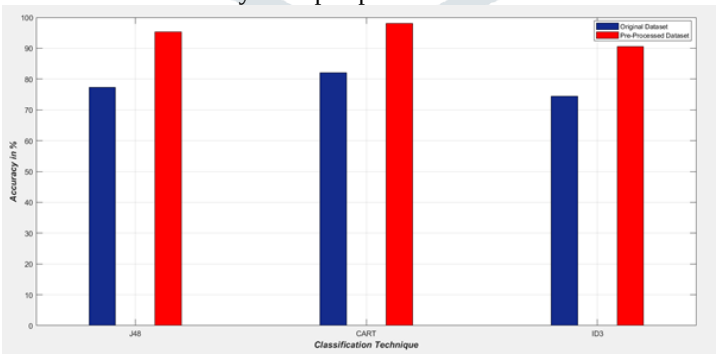

Figure 1: Performance Analysis of the J48, ID3, and CART classification technique on original dataset and pre-processed dataset

Figure 2 represents the Performance analysis on the Kappa Statistic value of the J48, CART, and ID3 classification methods for the original dataset and Pre-Processed dataset. From figure 2, it is strong that the pre-processed dataset gives maximum Kappa Statistic value with CART classification technique than the other methods with J48 and ID3 classification.

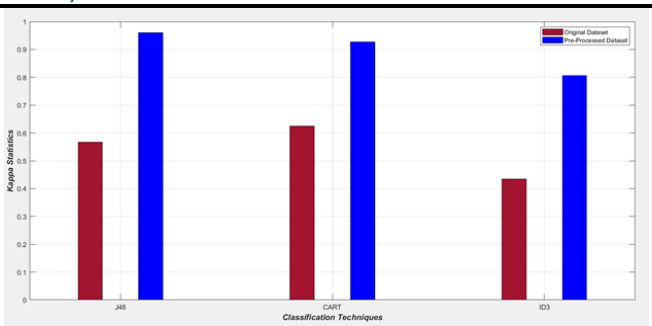

Figure 2: Performance analysis on the Kappa Statistic value of the J48, CART, and ID3 classification techniques for the original dataset and Pre-Processed dataset

Figure 3 depicts the Performance analysis on the Mean Absolute Error (MAE) of the J48, CART, and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 3, it is clear that the pre-processed dataset gives minimum MAE value with CART classification technique than the other methods with J48 and ID3 classification.

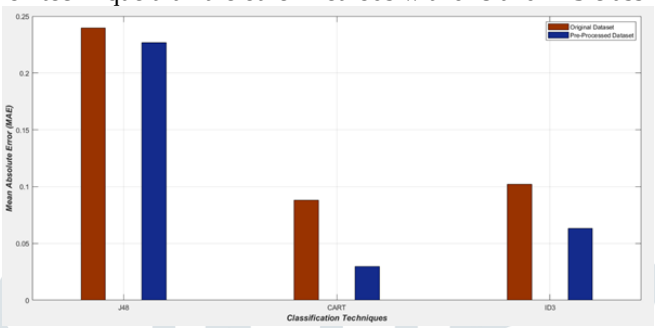

Figure 3: Performance analysis on the Mean Absolute Error (MAE) of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

Figure 4 depicts the Performance analysis on the Root Mean Squared Error (RMSE) of the J48, CART and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 4, it is clear that the pre-processed dataset gives minimum RMSE value with CART classification technique than the other methods with J48 and ID3 classification.

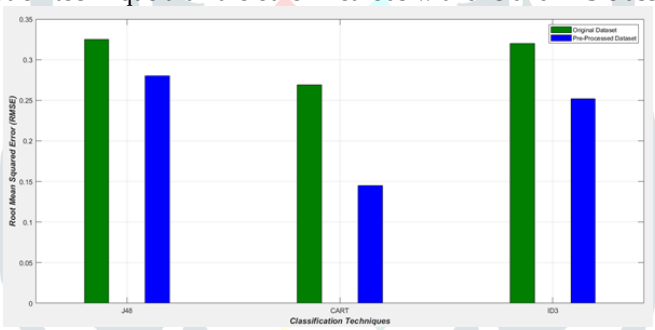

Figure 4: Performance analysis on the Root Mean Squared Error (RMSE) of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

Figure 5 presents the Performance analysis on the Relative Absolute Error (RAE) of the J48, CART, and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 5, it is clear that the pre-processed dataset gives minimum RAE value with ID3 classification technique than the other methods with J48 and CART classification.

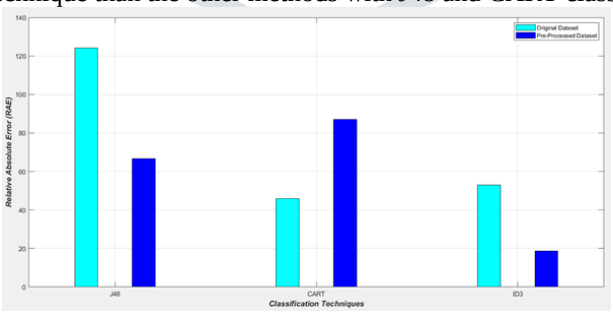

Figure 5: Performance analysis on the Relative Absolute Error (RAE) of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

Figure 6 presents the Performance analysis on the Root Relative Squared Error (RRSE) of the J48, CART and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 6, it is clear that the pre-processed dataset gives minimum RRSE value with CART classification technique than the other methods with J48 and ID3 classification.

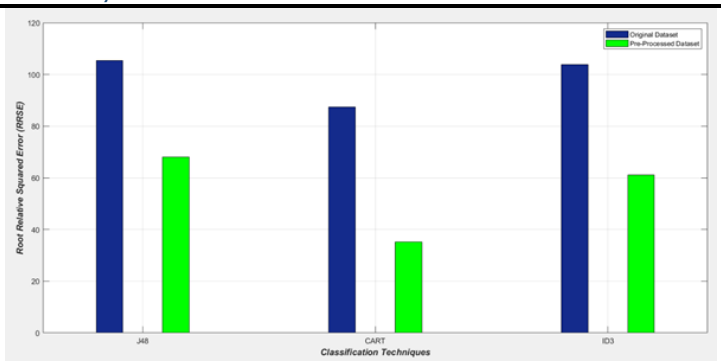

Figure 6: Performance analysis on the Root Relative Squared Error (RRSE) of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

Figure 7 presents the Performance analysis on the True Positive Rate (TPR) of the J48, CART, and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 7, it is clear that the pre-processed dataset gives maximum TPR value with CART classification technique than the other methods with J48 and ID3 classification.

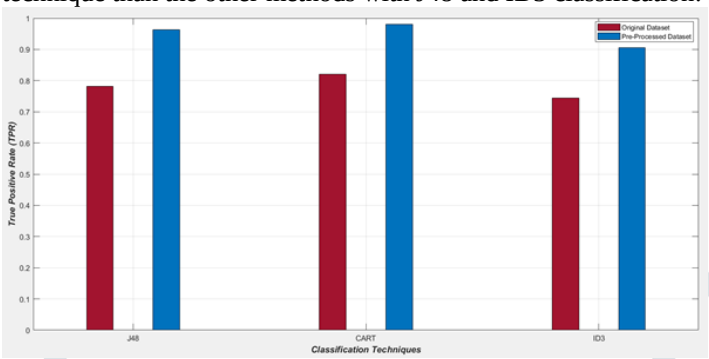

Figure 7: Performance analysis on the True Positive Rate (TPR) of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

Figure 8 presents the Performance analysis on the False Positive Rate (FPR) of the J48, CART, and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 8, it is clear that the pre-processed dataset gives minimum FPR value with CART classification technique than the other methods with J48 and ID3 classification.

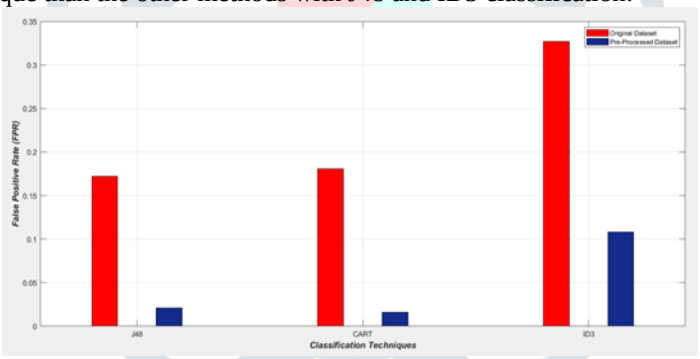

Figure 8: Performance analysis on the False Positive Rate (FPR) of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

Figure 9 presents the Performance analysis on the Precision of the J48, CART, and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 9, it is clear that the pre-processed dataset gives maximum Precision value with CART classification technique than the other methods with J48 and ID3 classification.

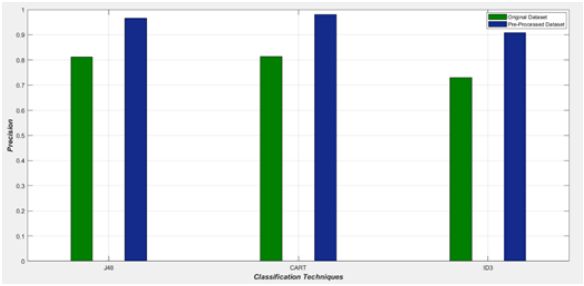

Figure 9: Performance analysis on the Precision of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

Figure 10 presents the Performance analysis on the F-Measure of the J48, CART, and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 10, it is clear that the pre-processed dataset gives maximum F-Measure value with CART classification technique than the other methods with J48 and ID3 classification.

Figure 11 presents the Performance analysis on the Receiver Operating Curve (ROC) area of the J48, CART and ID3 classification techniques for the original dataset and Pre-Processed dataset. From figure 11, it is clear that the pre-processed dataset gives maximum ROC value with CART classification technique than the other methods with J48 and ID3 classification.

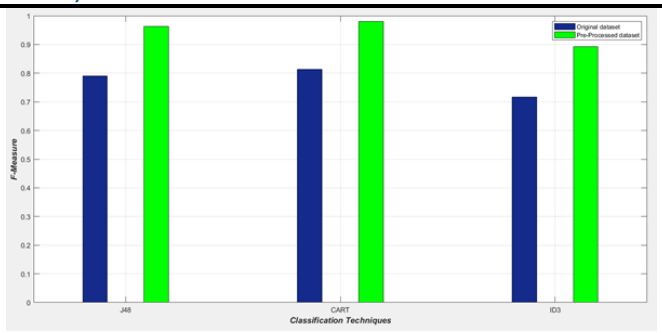

Figure 10: Performance Analysis on the F-Measure of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

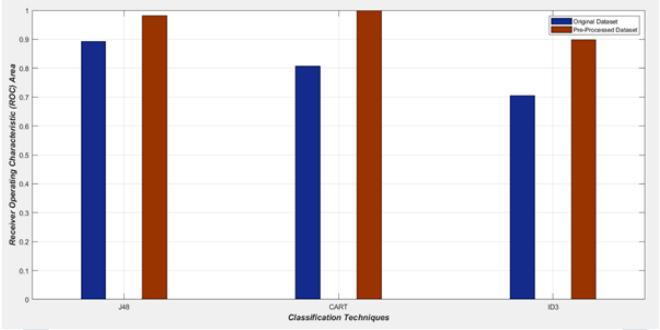

Figure 11: Performance Analysis on the Receiver Operating Characteristic (ROC) area of the J48, CART, and ID3 classification techniques for Original dataset and pre-processed dataset

VI. CONCLUSION

This project provides a new insight on how text mining leading to sentiments analysis can improve commonly used five-star rating method. Along with that we also understood basics text parsing, text topic and concept link. I took the instance of Echo Dot (2nd Generation) – Smart Speaker with Alexa – Black and analyzed users' review to extract significant sentiments for this product. As a result, I was able to capture sentiments about what users really like and what not which is further integrated with five-star rating system. This project can be further used to make better recommendations. When a user attempts to search for a product and service, the recommendation prompts with the top rated products. If we provide user, products sentiments along with the users rating, it will be easier for user to decide what is the best option according the requirement and thus solves the problem of missing important information and bimodal distribution.

Most of reviews have lengthy comments without ratings. It is a time-consuming factor to read all the reviews and come to a conclusion. In the future, the model predicts the opinion from the reviews without ratings. In future, the logistic regression and Naïve Bayes predictions of opinions are much similar than the multinomial and Bernouli classifiers are utilized.

REFERENCES

[1] J., Dean, & S., Ghemawat (2010). MapReduce: a flexible data processing tool. Communications of the ACM, 53(1), 72-77.
[2] J., Mehine (2011). Raamistiku Apache Pig kasutamine suuremahulises andmeanalüüsis (Doctoral dissertation, Tartu Ülikool).
[3] B., Jopson (2011). Amazon urges California referendum on online tax. The Financial Times, 4.
[4] J., McAuley, R. Pandey & J. Leskovec (2015, August). Inferring networks of substitutable and complementary products. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.
[5] J., McAuley, C., Targett, Q., Shi, & A., Van Den Hengel (2015, August). Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 43-52). ACM.
[6] J., McAuley, & A. Yang, (2016, April). Addressing complex and subjective product-related queries with customer reviews. In Proceedings of the 25th International Conference on World Wide Web (pp. 625-635). International World Wide Web Conferences Steering Committee.
[7] S., Mohanty, K., NathRout, S., Barik, & S.K., Das. A Study on Evolution of Data in Traditional RDBMS to Big Data Analytics.
[8] S., Singh, V., Mandal, & S., Srivastava. The Big Data Analytics with Hadoop.
[9] Apache Hadoop, http://hadoop.apache.org
[10] R., Shobana, D., Saranya. Hadoop on Big Data Analysis. International Journal of Advanced Research Trends in Engineering and Technology
[11] S., Dhawan, & S., Rathee (2013). Big data analytics using Hadoop components like pig and hive. American International Journal of Research in Science, Technology, Engineering & Mathematics, 88, 13-131.
[12] Pig Latin Reference Manual 2. https://pig.apache.org/docs/r0.8.1/piglatin_ref2.html
[13] S., Rathi. A brief Study of Big Data Analytics using Apache Pig and Hadoop Distributed File System
[14] E. L., Lydia, & M. B., Swarup. Analysis of Big data through Hadoop Ecosystem Components like Flume, MapReduce, Pig and Hive.
[15] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh. "Sentiment analysis on product reviews using machine learning techniques." *Cognitive Informatics and Soft Computing*. Springer, Singapore, 2019. 639-647.
[16] Bose, Rajesh, et al. "Sentiment Analysis on Online Product Reviews." *Information and Communication Technology for Sustainable Development*. Springer, Singapore, 2020. 559-569.
[17] Khan, Aurangzeb, et al. "Sentiment Classification of User Reviews Using Supervised Learning Techniques with Comparative Opinion Mining Perspective." *Science and Information Conference*. Springer, Cham, 2019.

[18] Pandey, Avinash Chandra, Saksham Raj Seth, and Mahima Varshney. "Sarcasm Detection of Amazon Alexa Sample Set." *Advances in Signal Processing and Communication*. Springer, Singapore, 2019. 559-564
[19] Devi, DV Nagarjuna, et al. "Assay: Hybrid Approach for Sentiment Analysis." *Information and Communication Technology for Intelligent Systems*. Springer, Singapore, 2019. 309-318.
[20] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems*, Preprint: 1-18.
[21] M. Durairaj, T S Poornappriya, "Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM.", *International Journal of Engineering & Technology*, 7 (4) (2018) 5856-5861.
[22] M. Durairaj, T. S. Poornappriya, "Importance of MapReduce for Big Data Applications: A Survey", *Asian Journal of Computer Science and Technology*, Vol.7 No.1, 2018, pp. 112-118.
[23] M. Lalli, V.Palanisamy,(2016), "Filtering Framework for Intrusion Detection Rule Schema in Mobile Ad Hoc Networks", International Journal of Control Theory and Applications –(IJCTA),9(27), pp. 195-201, ISSN: 0974-5572
[24] M. Lalli, V.Palanisamy,(2017), "Detection of Intruding Nodes in Manet Using Hybrid Feature Selection and Classification Techniques", Kasmera Journal, ISSN: 0075-5222, 45(1) (SCIE)(Impact Factor:0.071).
[25] M. Lalli, V.Palanisamy, (Sep 2014), "A Novel Intrusion Detection Model for Mobile Adhoc Networks using CP-KNN", International Journal of Computer Networks & Communications- (IJCNC), Vol.6, No.5, ISSN:0974-9322.
[26] M. Lalli, "Statistical Analysis on the KDD CUP Dataset for Detecting Intruding Nodes in MANET", *Journal of Applied Science and Computations*, Volume VI, Issue VI, JUNE/2019, 1795-1813.
[27] M. Lalli, "Intrusion Detection Rule Structure Generation Method for Mobile Ad Hoc Network", *Journal of Emerging Technologies and Innovative Research*, June 2019, Volume 6, Issue 6, 835-843.