# INFLUENCE OF CUSTOMER REVIEW RATING IN E-COMMERCE

**[1]R.Amutha, [2]D. Vimal Kumar**

[1]Research Scholar, Department of Computer Science.Nehru Arts and Science College, Coimbatore, India.
[1]Assistant Professor,Department of Computer Science,PSG College of Arts & Science, Coimbatore, India.
E-Mail:amutharajan@gmail.com
[2]Associate Professor, Department of Computer Science, Nehru Arts and Science College, Coimbatore, India.
E-Mail:drvimalcs@gmail.com

**Abstract-** - Recommendation analysis is widely used in many e-commerce companies to provide specialized offers and services to their consumers. Also, classification algorithms are playing a vital role in this type of recommendation related applications. These algorithms are used to detect patterns, behaviors for better decision-making. At times, it is essential to integrate the features of many algorithms to get an optimal solution to a given problem. So this paper proposes a new hybrid ensemble algorithm to analyze recommendation in Women's E-Commerce Clothing data. Dataset has taken from Kaggle which contains age, review, rating, department, positive feedback count and clothing recommendation. The proposed algorithm is compared with existing algorithms like naïve bayes, random forest, hard voting and soft voting. Results show that the hybrid ensemble out performs traditional algorithms.

**Keywords-** Women Clothing, Recommendation Analysis, Hybrid Ensemble, Clothing Review, Hard Voting, Soft Voting.

## 1. INTRODUCTION

A recommendation classification system has been demanding research for social media and e-commerce related companies. Recommendation analysis is used to provide the most relevant and accurate products to the customers by analyzing useful patterns from the large database. Recommendation analysis system discovers hidden patterns in the data set by analyzing customer choices and gives the outcomes that correlate the customer needs and interests. Amazon Flipkart, Facebook, etc., have been using a recommendation analysis system for suggesting products that consumers like very much. To understand the customer issues and to improve the company profits, recommendation analysis system emerges [1].

The popularity of online shopping increases nowadays. Analysis of the product based on customer review plays an important role in business. To analyze the popularity of online shopping products, the information about the product is necessary to determine customer requirements. Thus the machine learning algorithms can be used to find the relevant information about particular products, positive and negative reviews and suggestions for selecting the product to buy. Ratings and reviews are the most important factors of e-shopping. It involves the study of customer interest and product recommendations. In the modern E-commerce era, customers need a shopping assistant, which will suggest many interesting products according to customer interests [2]. Fig 1 shows retail e-commerce year wise in US.

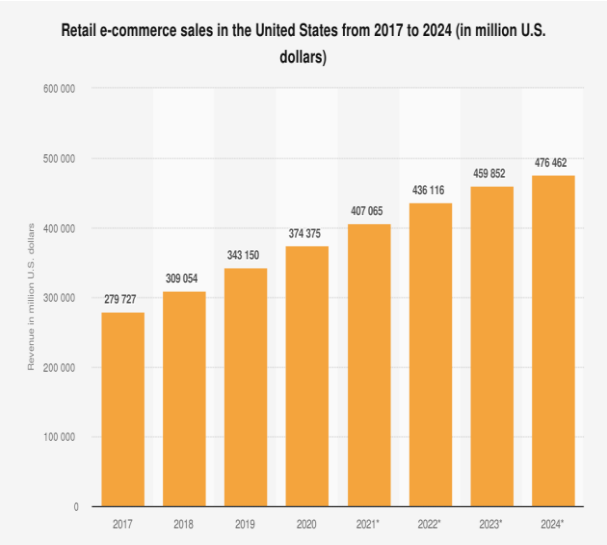



**Fig 1.** Retail e-commerce year wise in US

The recommendation system should provide the best way for decision making about the product. The system gives good idea about the particular product. The system will help to classify the products based on user recommendations. Different classification algorithms can be used to classify the product to purchase based on customer reviews. Based on the previous reviews the system recommends the customer to purchase the cloths. Personalized recommendation algorithm integrating trust relationship and time series [3], context information [4], user preference [5] is necessary nowadays.

Recommender analysis is applied to a variety of user data from customer reviews to social network posts. None of the literature work has been done based on recommendation analysis which will classify product recommendation of women's clothing review data. Review plays an important role in deciding purchasing strategies for different products. In this study, explores the impact of age, review, rating, department, positive feedback count and clothing recommendation, as this can help customers to get a better understanding about the women cloth purchasing.

The paper analyzes the categorized data for a recommendation based on all features. The new hybrid ensemble is proposed to obtain better recommendation accuracy. Finally, recommendation classification analysis is done using different Machine Learning (ML) approaches including naïve bayes, random forest, hard voting and soft voting to study the impact of various parameters. Experiments have been conducted to identify new insights into the effect of various attributes for recommendation analysis.

This paper is structured as follows. Our proposed work described in section 2. The recommended classifications are explained in section 3. Section 4 illustrates the results of the different algorithms; and finally, Conclusion of the proposed work and set of future work presented in section 5.

## 2 LITERATURE REVIEW

In paper [6], the authors evaluated the performance of various algorithms on women e-commerce data. They focused on good accuracy and good generalization results on several datasets, by using machine learning techniques. Logistic Regression worked best on the dataset compared to Bernoulli Naïve Bayes and Multinomial Naïve Bayes. However, the authors stated in future work that sentiment analysis can be much more effective when it is used as a recommendation tool.

Authors in [7] used classifiers like Support Vector Machine, Logistic Regression, Random Forest and Naive Bayes on women clothing data. Results have shown that Naive Bayes gives the highest accuracy. They have concentrated on online women clothing reviews features such as rating, class name, age and review texts.

In paper [8], Amazon review dataset is considered for the work which includes the review data of Laptops, tablets, TVs, Camera and video surveillance. Machine learning algorithms are used to classify review categories such as positive or negative. This paper concludes that Naïve Bayes gives best results to classify the Products Reviews. Authors have not focused on advanced machine learning techniques in order to improve classification accuracy.

Authors in [9] proposed a new framework to analyze customers' sentiments using different data mining techniques. The proposed framework is based on data collection, preprocessing and feature extraction and feature engineering. The results have shown that TF-IDF is the best measure in comparison with others. Authors have not used large data to evaluate the performance of algorithms.

In paper [10], authors introduced a new system which helps to classify the reviews on rating scale between 1 and 5 based on the sentiments in the words. Food reviews are analyzed using rating score combined with existing text analyzing process. They used groups of words to make a decision. Better results are produced using score rating in this proposed method. Only fewer features have been considered for analysis.

Authors in [11] compared different data mining classification techniques such as Decision Tree, Random Forest, Multilayer Perception, Radial Basis Function, Ada Boost,Sequential Minimal Optimization, Naive Bayes and Decision Stump and identified that decision tree performed well compared to other algorithms. Also, negligible false-positive obtained which is very essential for any classifier. Authors went for traditional machine learning techniques and ignored the benefits of modern machine learning.

A new classifier system proposed by the author in [12], analyzed 400 Thai customer reviews about hotels from a website to categorize the customer comments. This classifier model has projected good probability results which utilized naïve bayes and decision tree techniques and concluded that the Naive Bayes gives the better result. Authors have not put the focus on sentiment polarity word extraction using data preprocessing.

## 3 PROPOSED SYSTEM

In the proposed system, the new hybrid ensemble algorithm is introduced to classify recommendation from women e-commerce clothing review data.

### 3.1 Collection of Dataset

Women's Clothing E-Commerce dataset collected from Kaggle. There are 23486 rows and 10 feature variables in the dataset. Each row corresponds to a customer review which includes the following attributes.

**Table 1** Attribute Descriptions

| |
|---|
| **Clothing ID**: This refers the specific piece being reviewed. It is an Integer Categorical variable.<br>**Age:** Reviewer's age represented in Positive Integer.<br><br>**Title:** It contains the information about the title of the review. This feature represented in string format.<br>**Review Text:** String variable for the review body.<br>**Rating:** This feature holds the Positive Ordinal Integer value of the product score granted by the customer rating from 1 Worst, to 5 Best.<br>**Recommended IND:** This value represents the Product recommendation. (Where 1 represents the product is recommended, 0 is not recommended).<br>**Positive Feedback Count**: number of other customers who found this review as positive is represented in the positive count.<br>**Division Name:** Product high level division in this column. This feature holds the Categorical value.<br>**Department Name:** This feature represents the department name of the product.<br>**Class Name:** Product class name holds the Categorical value. |

The table 1 explains the range, data type and other details of all the attributes. Recommended IND is taken as a class label in this paper.

### 3.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is a smart analytical system, to obtain insights and hidden patterns without applying any technique. It is easy to understand data and attribute characteristics with the help of EDA.

The figure 2 representsa number of reviews in each department. Tops department has the highest number of reviews. Intimate and Jackets has less than 10% of reviews, while bottoms have 16% and other dresses have 27%.

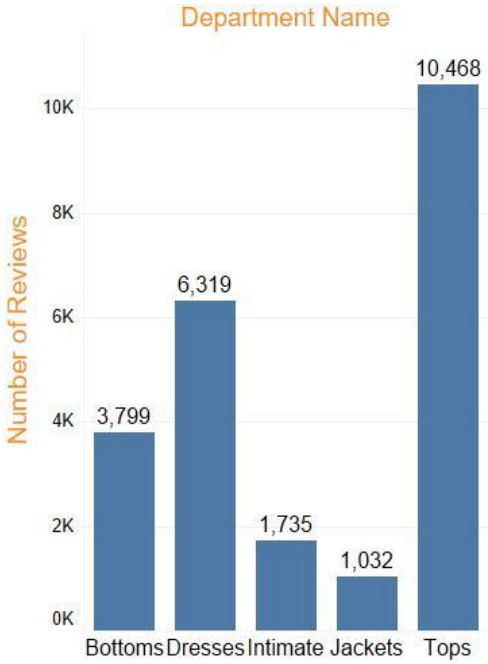



**Fig. 2 Departmentwise Reviews**

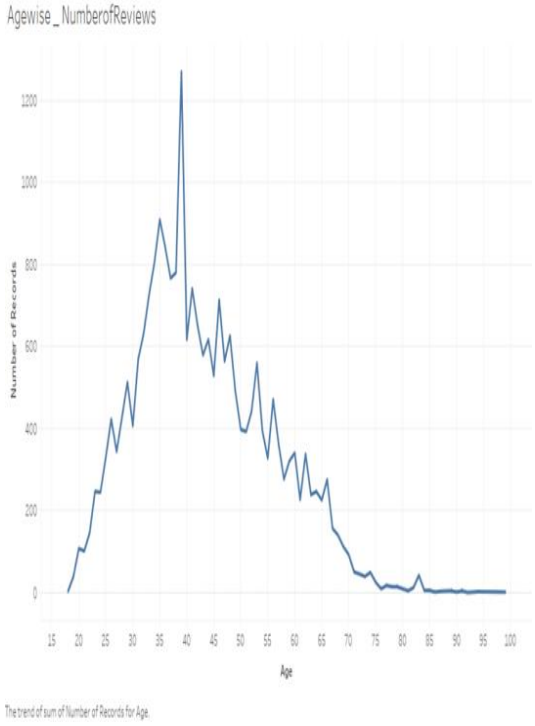



**Fig. 3 Age wise Reviews**

The figure 3 illustrates a number of reviews based on age. From the figure, it is clear that middle-age woman written more reviews. Age range 30 to 50, purchases more than other age groups. 35 to 40 age group women have better purchasing history.

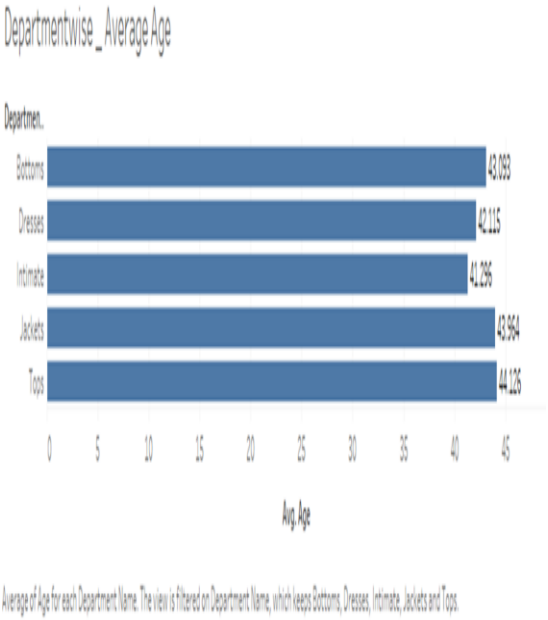



**Fig. 4 Department wise Average Age**

The figure 4 represents the average age based on department wise. The average age of all departments between 40 to 45. This shows the domination of the middle age group.

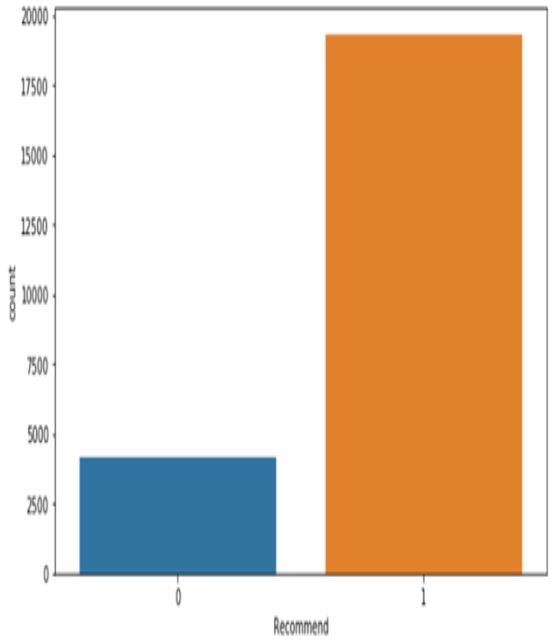



**Fig. 5 Class Label**

The above figure 5 illustratesthe class label recommendation. 19314 records belong to the recommendation, while 4172 records belong to not recommended. Here class 0 indicates not recommended and class 1 represents recommended.

### 3.3 Preprocessing

All the data cleaning tasks like lowercase conversion, remove whitespaces, special characters, punctuation, numbers, stop words has been performed. Then tokenization and lemmatization have been done. Finally, bag of words and TF-IDF (Term Frequency–Inverse Document Frequency) has been applied.

### 3.4 Machine Learning

Different machine learning algorithms like naïve bayes, random forest, hard voting and soft voting has been applied to classify the recommendations.

a. **3.4.1 Naïve Bayes Algorithm**
b.

The Naive Bayes classifier is a single classifier that classifies data based on probabilities of events. For text classification, it performs well.

Naïve bayes is a type of supervised learning algorithm and it is based on Bayes theorem. This type of algorithm supports high dimensional training dataset. It is an effective classification algorithm which supports to build a machine learning models that can help for better predictions. It finds the classifier based on the probability which is also known as Bayes Rule.

$P(A|B)=(P(B \mid A)P(A))/(P(B))$

We can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent

There are three types of Naives Bays classifier

**i) Multinomial Naïve Bayes**

This is used for document classification problem. This classifier uses the features are the frequency of the words occurred in the document.

**ii) Bernoulli Naive Bayes**

Here the predictors are Boolean type of variables. The class variable are take up the Boolean values.

**iii) Gaussian Naive Bayes**

Gaussian distribution applied when the predictors are continuous value.

c. **3.4.2 Random Forest Algorithm**
d.

The Random Forest (RF) classifiers are suitable when dealing with the text classification. An RF classifier consists of a set of base classifiers and it is trained with the help of random subsets of features.

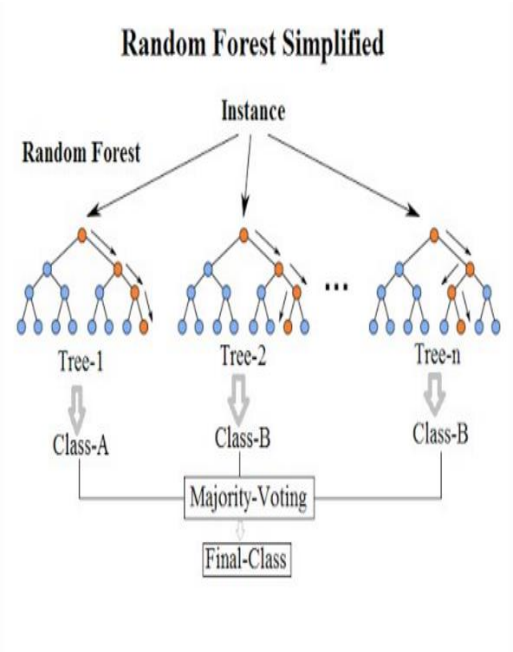



**Fig. 6** Random Forest Classification

A huge number of relatively uncorrelated models (trees) operating as a committee is provided a better performance than any of the individual constituent models. Fig. 6 shows Random Forest Classification.

### 3.4.3 Hard Voting

In hard voting, each base will vote for a class, and which has the majority will win. Here, the classified class label of multiple classifiers is the mode of the distribution of individually classified labels. Fig. 7 shows process of hard voting.

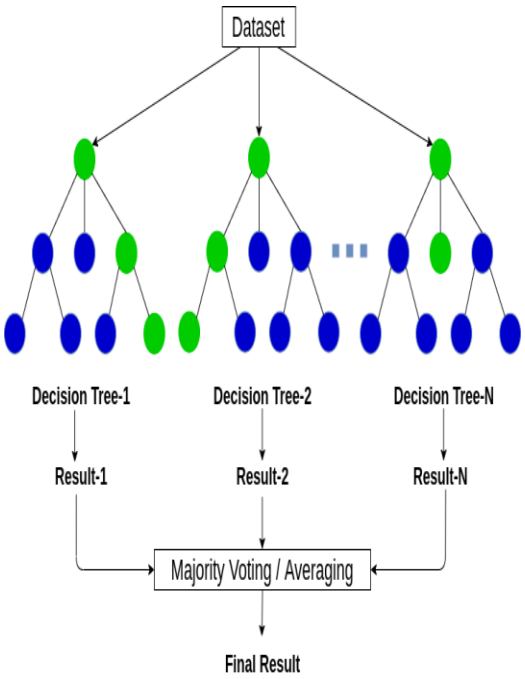



**Fig. 7** Hard Voting

### 3.4.4 Soft Voting

In soft voting, each base classifier gives a probability value that a particular record belongs to a specific class. The classifications are weighted by the classifier's importance and summed up. Then the class label with the greatest sum of weighted probabilities will win the vote. Fig. 8 shows process of Soft Voting.

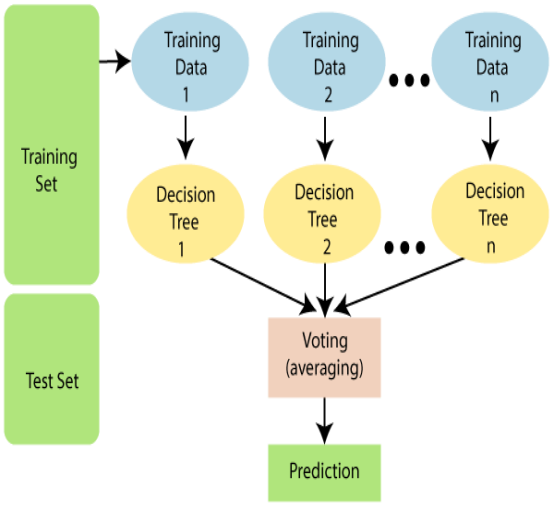



**Fig. 8** Soft Voting

### 3.4.5 Hybrid Ensemble

Hybrid Ensemble combines various classifiers to obtain better accuracy results. Features and outputs of base classifiers are integrated into this technique. Fig. 9 shows process of Hybrid Voting.

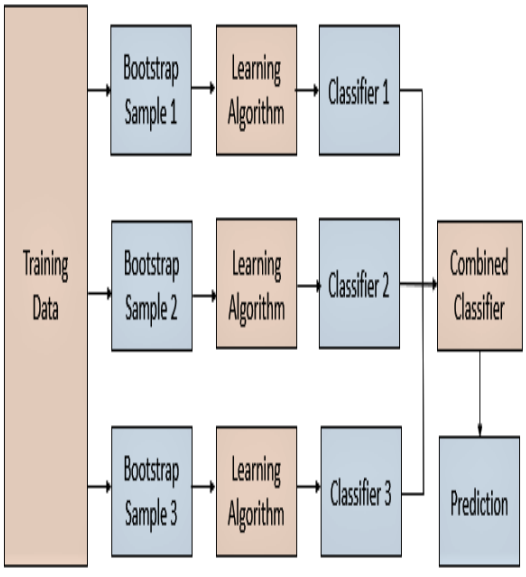



**Fig. 9** Hybrid Voting

**Hybrid Ensemble Algorithm**

**Algorithm**

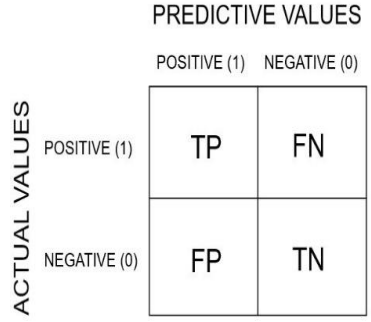
1. Apply base classifiers like random forest, naïve bayes, hand voting and soft voting.
2. Get the output of above base classifiers in one data frame with class label and create this as model.
3. Use that model to get output on test data.

## 4 EXPERIMENTAL RESULTS

### 4.1 Performance Metrics

In this paper, the recommendation is classified. So the performance metrics like accuracy, precision, recall and F-measure are considered.

**Table 2** Performance Metris

| | |
|---|---|
| Confusion Matrix =   | |
| $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$ | |
| $\text{Precision} = \frac{TP}{TP + FP}$ | |
| $\text{Recall} = \frac{TP}{TP + FN}$ | |
| $\text{F} - \text{Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$ | |

Performance measurements used in this paper are given in table 2. One common measure in the literature is accuracy. It is defined as correct classified records divided by the total number of records. The true positives (TP) and true negatives (TN) are correctly classified records. A false positive (FP) and false negative (FN) occurs when the outcome is incorrectly predicted.

### 4.2 Results

**Table 3** Results

| Algorithm | Accuracy | Precision | Recall | F-Measure |
|---|---|---|---|---|
| Random Forest | 86.73 | 0.86 | 0.87 | 0.85 |
| Naïve Bayes | 86.06 | 0.86 | 0.86 | 0.83 |
| Hard Voting | 87.41 | 0.87 | 0.87 | 0.85 |
| Soft Voting | 86.04 | 0.85 | 0.86 | 0.84 |
| Hybrid Ensemble | 88.70 | 0.88 | 0.89 | 0.88 |

The table 3 provides the results of different classifiers for the parameters accuracy, precision, recall and f-measure. In all the parameters, the new model performed well. Secondly, hard voting gives better results.

## 5 CONCLUSION

In this paper, Dataset has taken from Kaggle which contains 10 attributes. After preprocessing the machine learning algorithms are applied to classify recommendation that is recommended or not recommended. The new hybrid ensemble is proposed to obtain better accuracy results. Results of the proposed algorithms are compared with existing algorithms like naïve bayes, random forest, hard voting and soft voting. This paper concludes that hybrid ensemble gives the best results to classify the recommendations. In future, all product review can be considered to analyze sentiments and to recommend products. Also, improvement of deep learning can be considered with an ensemble to obtain better accuracy results.

## References

[1] Sharma, G., Bajpai, N., Kulshreshtha, K., Tripathi, V. and Dubey. P, "Foresight for online shopping behavior: a study of attribution for "what next syndrome"", Foresight, Vol. 21 no. 2, pp. 285-317, 2019.
[2] Z. Zhang and H. Liu, "Application and research of improved probability matrix factorization techniques in collaborative filtering", International Journal of Control & Automation, Vol. 7, no. 8, pp. 79–92, 2014.
[3] Z. Zhang and H. Liu, "Social recommendation model combining trust propagation and sequential behaviors", Applied Intelligence, Vol. 43, no. 3, pp. 695–706, 2015.

[4] Z.-J. Zhang and H. Liu, "Research on context-awareness mobile SNS recommendation algorithm", Pattern Recognition and Artificial Intelligence, Vol. 28, no. 5, pp. 404–410, 2015.
[5] Y.-M. Li, C.-L. Chou, and L.-F. Lin, "A social recommender mechanism for location-based group commerce," Information Sciences, Vol. 274, pp. 125–142, 2014.
[6] Andreea Salinca "Business Reviews Classification Using Sentiment Analysis", 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2015.
[7] Elshrif Elmurngi , Abdelouahed Gherbi,"Unfair reviews detection on Amazon reviews using sentiment analysis with supervised learning techniques", : Approaches based on sentiment analysis and supervised learning techniques for robust reputation systems in the e-commerce environment,2018
[8] Jagdale, R. S., Shirsat, V. S. and Deshmukh, S. N, "Sentiment analysis on product reviews using machine learning techniques", Cognitive Informatics and Soft Computing, pp. 639–647, 2019.
[9] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journa l,Vol. 5, no.4, pp.1093-1113,2014
[10] Sasikala P, L.Mary Immaculate Sheela, "Sentiment Analysis of Online Food Reviews using Customer Ratings", International Journal of Pure and Applied Mathematics, Vol. 119, no. 15, pp. 3509-3514, 2018.
[11] Mani, B. Suri& M. Kumar, "Performance Evaluation of Data Mining Techniques", Information and Communication Technology for Sustainable Development", Vol. 9, pp. 375-383, 2017.
[12] Songpan. W, "The Analysis and Prediction of Customer Review Rating Using Opinion Mining", IEEE 15th International Conference on Software Engineering Research, Management and Applications, 2017.