# Trust Based Predictive Analysis on E-Commerce Applications

**Ms. G.V. Rajya Lakshmi[1], Mr. A.V.N. Pavan Kumar[2], Ms. S. Mounica[2], Mr. E. Harish[2], Mr. M. Naresh[2]**

[1]Assistant Professor, Department of CSE, Lakireddy Bali Reddy College of Engineering (A), Mylavaram, Krishna (Dt.), Andhra Pradesh, India – 521230
[2]B. Tech Project Students, Department of CSE, Lakireddy Bali Reddy College of Engineering (A), Mylavaram, Krishna (Dt.), Andhra Pradesh, India – 521230

**Article History**: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

**Abstract:**
One of the problems that online businesses platforms facing is gaining a greater understanding of the customer's thoughts and sentiments on the products. This problem can be solved by using sentiment analysis to derive additional insights from consumer feedback. Customer feedback provides a useful platform to discover a huge range of customer- initiated reactions to the product(s) that they have purchased. Text analytics provide businesses a more holistic picture of customer's satisfaction or dissatisfaction. It is insufficient to rely on customer's ratings alone to find out their experiences. This is because reviews can provide important feedback to businesses. In this paper, for text classification, we proposed the Naive Bayes algorithm. Despite its simplicity, it always outperforms much more complex solutions. The naive Bayes algorithm will be used in this paper to predict the mood of feedback. In the proposed plan, Natural language processing is used to extract features from the text of the feedback to train the algorithm.

**Keywords:** User trust, Naïve Bayes, Natural Language Processing

## 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. The 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, information about the product is necessary to determine customer requirements. Thus, the machine learning algorithms can be used to find the relevant information about 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. has been done based on recommendation analysis which will classify product recommendation of review data set.

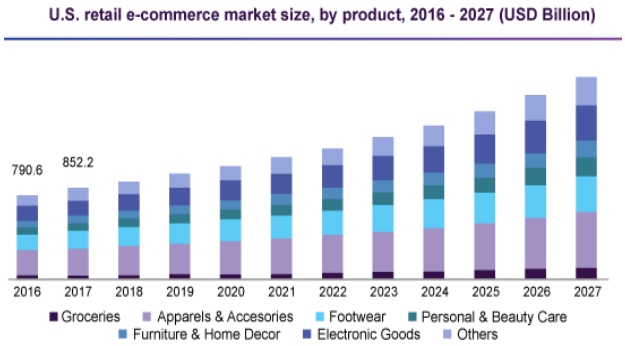



Fig. 1: US retail E-Commerce growth

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 data set on 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 Naive Bayes [10], random forest, to study the impact of various parameters. Experiments have been conducted to identify new insights into the effect of various attributes for recommendation analysis.

## 2. Related Work

In paper [3], the authors evaluated the performance of various algorithms on 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 could be much more effective when used as a recommendation tool.

Authors in [4] used classifiers like Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Multinomial Naive Bayes (MNB) on data. Results have shown that Naive Bayes gives the highest accuracy. They have concentrated on online reviews such as review texts.

In paper [5], the Amazon review dataset is considered for the work, including the review data of Laptops, Tablets, TVs, Cameras, and video surveillance. Machine learning classifiers are used to Sort user reviews into positive and negative categories. This paper concludes that the best method for classifying Product Reviews is to use Naive Bayes. Authors have not focused on advanced machine learning techniques to improve classification accuracy.

Authors in [6] proposed a new framework to analyze customers' sentiments using different data mining techniques. The proposed framework is based on data collection, Pre-Processing [9] 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 [7], The writers devised a new method for categorizing feedback on a scale of 1 to 5 based on the sentiments expressed in the words. Food reviews are analyzed using rating scores combined with existing text analyzing processes. They used groups of words to decide. Better results are produced using score rating in this proposed method. Only fewer features have been considered for analysis.

Authors in [8] 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. They identified that the Decision tree performed well compared to other algorithms also, negligible false-positive was obtained, which is essential for any classifier.

Authors went for traditional machine learning approaches and ignored benefits of modern machine learning. A new classifier system proposed by the author in [8] analyzed 400 Thai customer reviews about hotels from a website to categorize the customer comments. This classifier model has projected good probability results that utilized naïve Bayes and decision tree techniques and concluded that the Naive Bayes method produced the best results. The Authors have not put the focus on sentiment polarity word extraction using data Pre-Processing.

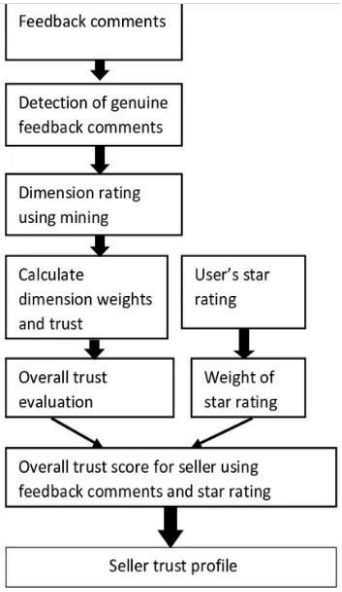
## 3. Proposed System

The proposed system, Naive Bayes algorithm is used to classify recommendations from e-commerce review data.

**Collection of Dataset:**

E-Commerce reviews dataset collected from Online Source in json format. There are 100000 rows in the dataset. Each row corresponds to a customer review which includes the following attributes '{Review, ReviewText}' in Key-value pairs.

**Workflow of Proposed System:**



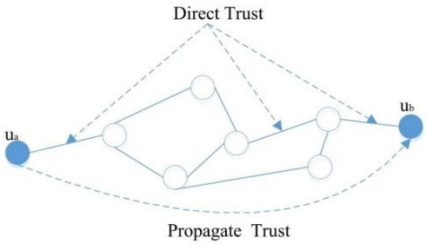

**User Reviews:**

Which have collected from e-commerce websites are given as input. The datasets include the reviews the products that users post, nearly 100000 reviews will be taken as the input source to the proposed system.

**Pre-Processing:**

All the data cleaning tasks like lowercase conversion, remove whitespaces, special characters, punctuation, numbers, stop words have been performed. Then tokenization and lemmatization have been done. Finally, a collection of words and TF-IDF has been applied.

**Defining Trust:**




**Random Forest Algorithm:**

The Random Forest (RF) classifiers are suitable when dealing with text classification. An RF classifier consists of a set of base classifiers and, it is trained with the help of random subsets of features. A vast number of relatively uncorrelated models (trees) operating as a committee provides better performance than any individual constituent models.

In our project, Accuracy measure is around 81%.

**Naive Bayes Algorithm:**

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

The Bayes theorem is the foundation of the Naive Bayes algorithm, which is a kind of supervised machine learning algorithm. This type of algorithm supports high dimensional training dataset. It is an effective classification algorithm that supports building machine learning models that can help for better predictions. It finds the classifier based on probability, which is also known as Bayes Rule.

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

Given that B has occurred, we can calculate the probability of A occurring. The proof is B, and the hypothesis is A. The predictors/features are assumed to be independent in this case.

In our project, Accuracy measure is around 85%.

**Support Vector Machine:**

The "Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used to solve problems such as classification and regression. It is, however, often used in the solution of classification problems.

In our project, Accuracy measure is around 88%.

**Logistic Regression:**

Another method for generating multivariable composites that can be used to distinguish two or more classes is to use logistic regression. In this sense, logistic regression has many advantages over other statistical approaches because it lacks the restrictive statistical assumptions of DFA.

In our project, Accuracy measure is around 88%.

**4. Methodology**

As previously mentioned, we used a data set with a total of 100000 rows. The analysis text is then pre-processed with the NLTK kit, which eliminates any unnecessary terms (like nouns, pronouns, etc.). The Training section will then be split into five (5) sections using the K-fold split technique. We use four classifiers for each break, and the results are presented in the form of a confusion matrix for each classifier, followed by a Classification Report which includes the metrics (like Accuracy, precision, recall, F1-score).

Finally, it compares the Classification Reports for all the classifiers used in the training project. During this time, rules are extracted in pickled form, and graphs in the form of heat maps are produced.

Following are the procedure for the project:

**Step -1:** Our project will give the output in two ways.

a) It will analyze the dataset given and returns the metrics for the classification algorithms.

b) It will ask the user to enter a review and give the result, whether it is positive or negative.

**Step -2**: In the first way, it fetches the data in the form of a dictionary and converts it into a data frame by using pandas.

**Step -3:** Then next, pre-process Step in this We are cleaning the dataset by removing the unwanted data in the reviews with the help of the NLTK tool.

**Step -4:** Then apply the cross-validation strategy by using the k-fold splits process. In this project, we are doing it with five (5) splits.

**Step -5:** For each split in the k-fold splits process, we are applying the four classification algorithms, then train the dataset and generates the rules on that dataset and, also generates some heats maps for the top features in the metrics for each classification algorithms.

**Step -6:** For each classifier, we are generating the following metrics:

a. Confusion matrix
b. Precision
c. Recall
d. F1-Score
e. Classification Accuracy

**Step -7:** Finally, it will generate a comparison of different metrics for all the four classifiers.

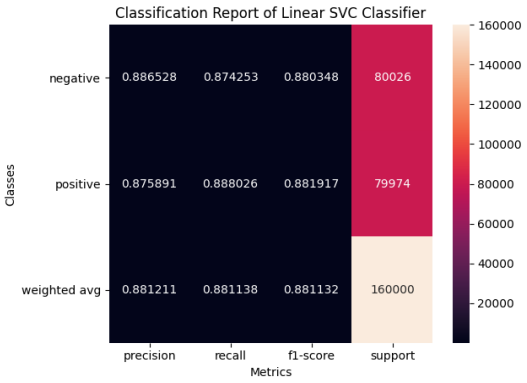



Fig. Heat Maps

```
Comparison of different metrics for the various Classifiers used:

              Classifier  Accuracy  Precision    Recall  F1-score   ROC AUC
0        Multinomial NB  0.853113   0.855563  0.849551  0.852547  0.853111
1   Logistic Regression  0.880494   0.875409  0.887176  0.881253  0.880496
2            Linear SVC  0.881138   0.875891  0.888026  0.881917  0.881140
3         Random Forest  0.816612   0.795492  0.852189  0.822865  0.816624
```

## 5. Performance Metrics

Several quality control methods, including Precision, Recall, F1-Score, and Classification accuracy, will be used to consider the metrics for the proposed work.

**Precision:** The fraction of retrieved words and texts that are compatible with the find is known as precision.

**Recall:** It is the percentage of information (such as words and texts) that is effectively saved that is consistent with the queries.

**F1-Score:** F-Score is used to calculate both recall and precision. The provided equation can be used to predict it.

$$F = (2 * \text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

**Classification Accuracy:** The percentage of correctly analyzed information (such as words and texts) is calculated using the following equation, which can be estimated.

## 6. Result Analysis

Our project will produce results in two ways: the first time we run the code with the data set, it will break the data set into train data and test data and then train the algorithm with train it generates some rules, it will produce precision, recall, F1-Score, and help for positive and negative classification on the test data. In other words, it prompts the user to enter review, it classifies as Positive or negative.

```
Enter your review:
This product is good compared to other

Predicted sentiment: Positive

Evaluation metrics of Classifier Random Forest:
Confusion Matrix:
[[62505 17521]
 [11821 68153]]

Classification Report:
              precision    recall  f1-score   support

    negative   0.840957  0.781059  0.809902     80026
    positive   0.795492  0.852189  0.822865     79974

    accuracy                       0.816612    160000
   macro avg   0.818225  0.816624  0.816384    160000
weighted avg   0.818232  0.816612  0.816382    160000
```

## 7. Conclusion

In this project, the Dataset has taken from Internet Source which is a json format contains two fields and 100000 rows. After preprocessing the machine learning algorithms are applied to classify recommendation that is recommended or not recommended. The naive is proposed to obtain better accuracy results. This paper concludes the accuracy of the data collected, in another way it will gives that the user entered review is positive or negative. In the Future, all product reviews 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.

**Acknowledgements**

We, authors of this research article whole heartedly thank Head of the Department, Department of CSE, Lakireddy Bali Reddy College of Engineering (A), Mylavaram and "Research Center of LBRCE", recognized by JNTUK, Kakinada for providing us the infrastructure facilities during the progress of work.

**References (APA)**

[1] Shaozhongzhang , "Mining Users Trust From e-commerce".

[2] Sharma, G., Bajpai, N., Kulshreshtha, K., Tripathi, V. and Dubey. P, "Foresight foronline shopping behavior: a study of attribution for "what next syndrome"", Foresight, Vol. 21 no.2, pp. 285-317, 2019

[3] Z. Zhang and H. Liu, "Application and researchof improved probability matrix factorization techniques in collaborative filtering", International Journal of Control & Automation, Vol. 7, no. 8, pp. 79–92,2014.

[4] Z. Zhang and H. Liu, "Social recommendation model combining trust propagation and sequential behaviors", Applied Intelligence, Vol. 43, no. 3, pp. 695–706,2015.

[5] Z.-J. Zhang and H. Liu, "Research on contextawareness mobile SNS recommendation algorithm", Pattern RecognitionandArtificial Intelligence, Vol. 28, no. 5, pp. 404– 410,2015.

[6] AndreeaSalinca "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] Basha, S., & Subrahmanyam, M. (2019), Research of various data mining techniques for IoT applications, International Journal of Recent Technology and Engineering, 8(2 Special Issue 11), 1083-1086.

[10] Basha, S., Bhavani, S., & Subrahmanyam, M. (2020), Self-regulated deepfakes detection mechanism using cart, Journal of Advanced Research in Dynamical and Control Systems, 12(2), 923-930.