

# Sentiment Exploring on Feedback of E-commerce Data Using Machine Learning Algorithms



Amrithkala M. Shetty, Mohammed Fadhel Aljunid, and D. H. Manjaiah

**Abstract** In today's fast-growing Internet world, customer ratings and reviews play an essential role in online buying on e-commerce websites such as Amazon, Flipkart, and others. Sentiment analysis is crucial for increasing customer satisfaction on e-commerce sites since it contains a lot of consumer feedback. In this work, we have used Amazon Women's E-Commerce Clothing Reviews dataset. We have used CountVectorizer and TF-IDF and trained the data on five machine learning (ML) classifiers, namely logistic regression (LR), multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), support vector machine (SVM), random forest (RF), and AdaBoosting (AB). When comparing the ML model's accuracy scores concerning the CountVectorizer, it was discovered that MNB and LR models had the highest accuracy of 0.94, while RF had the lowest accuracy of 0.90. SVM achieved the maximum accuracy of 0.94 using the TF-IDF approach, and MNB achieved the lowest accuracy of 0.89. The accuracy, precision, recall, F1-score, and AUC-ROC curve help us to determine the performance of the ML algorithms. To examine the dataset's attributes and comprehend the relationships between the variables, many statistical techniques were applied.

**Keywords** Sentiment analysis · CountVectorizer · TF-IDF · Logistic regression · Multinomial Naive Bayes · Bernoulli Naive Bayes · Support vector machine · Random forest · AdaBoosting

---

Mohammed Fadhel Aljunid and D.H. Manjaiah: These authors contributed equally to this work.

---

A. M. Shetty (✉) · D. H. Manjaiah  
CS Department, Mangalore University, Mangaluru, Karnataka 574199, India  
e-mail: [amrithkalas@gmail.com](mailto:amrithkalas@gmail.com)

M. F. Aljunid  
Computer and Informatics Center, Thamar university, Thamar, Yemen  
e-mail: [ngm505@yahoo.com](mailto:ngm505@yahoo.com)

# 1 Introduction

Online shopping has grown significantly and becomes quite popular in recent years as a result of the widespread use of the internet and the quick development of mobile devices. Customers like internet shopping because it allows them to take decisions based on product ratings, reviews, and technical details [1]. Rating and reviews play an important role in understanding the sentiment of the customer. A numerical rating scale from 1 for the worst to 5 for the best is available on Amazon.com.

Not only the customer but the manufacturer also is informed of the product's advantages and disadvantages. The manufacturer can design new business strategies and take the necessary actions to improve the product based on user input. Categorizing reviews as positive, negative, or neutral is the core task of sentiment analysis. The sentiment is an essential component of human communication. It is a subset of natural language processing (NLP). Aspect level, phrase level, sentence level, and document level are the four levels of sentiment analysis [2]. Opinions are frequently disguised in lengthy forum messages and blogs. Finding relevant sources, extracting linked sentences with viewpoints, reading them, summarizing them, and organizing them into useful forms are tough for a human reader. Therefore, automated opinion discovery and summarization methods are needed. This requirement has led to the development of sentiment analysis, often known as opinion mining. Because of its tremendous value in practical applications, there has been an incredible development in both academic research and industrial needs [3]. Amazon is a well-known e-commerce website with thousands of items available for purchase online. Users can leave reviews and rate the product to convey their thoughts about it. These reviews are important; sentiment analysis is performed on them. In this paper, we analyze the Amazon Women's Clothing E-Commerce dataset. First, preprocessing techniques are used to clean the text in the reviews. The next step is feature extraction, which converts the text into vectors so that ML methods can be applied to them. CountVectorizer and TF-IDF are feature extraction techniques used in the study. Each of these techniques is applied to ML algorithms such as LR, MNB, BNB, SVM, RF, and AB. The main objective of this study is to assess how well various machine learning (ML) algorithms perform while using the CountVectorizer and TF-IDF vectorization techniques. In addition to this, statistical analysis was also performed.

The remainder of the paper is structured as follows: We provide a summary of the research on sentiment analysis on e-commerce datasets in Sect. 2. The methodology, which includes data preprocessing stages, data annotation, data analysis, and data visualizations, as well as vectorizations and ML models is introduced in Sect. 3. In Sect. 4, a thorough experimental study is offered. Sections 5 then wrap up with the results, conclusions, and future work.

## 2 Related Work

The study of sentiment in reviews of commercial products has received a lot of attention. E-commerce sites like Amazon, Flipkart, Snapdeal, and others have gained prominence in recent years.

In this section, we evaluate earlier research on the Amazon E-Commerce Women's Clothes Review dataset.

Three tree-based classification methods, including the classification tree, RF, Gradient Boosting, and XG boost, are utilized in the [4] study to classify the reviews of apparel e-commerce sites. The Gradient Boosting technique shows steady results, but XGBoost exhibits overfitting if the number of trees is too high. Classification tree is strong at detecting bad reviews but poor at detecting positive ratings.

In [5], the primary goal is to establish a relationship between review attributes and product recommendations using natural language processing. This study was done to gain a deeper understanding of consumer psychology and sentiment in the e-commerce transaction industry. In this case, five ML algorithms—LR, SVM, RF, XGBoost, and light GBM—were used. The dataset of reviews of women's clothes on Amazon was used to apply the algorithms. The maximum AUC value and accuracy were provided by the light GBM algorithm. The RBF Kernel SVM algorithm, however, had the worst results, with an accuracy score of 81%.

In [6], the reviews from various e-commerce websites are examined in this study to develop a robust and proposed recommendation system. The major objective is to enhance customer service. The study is done on a women's apparel review dataset. The dataset in this case is subjected to univariate and multivariate analysis. For recommendation and sentiment classification, a bi-directional recurrent neural network with a long short-term memory unit was employed.

In [7], the many labels' customer review categorization task aims to pinpoint the many perspectives that users have regarding the products they are purchasing. The reviews from a Turkish e-commerce website were scraped and divided into three groups: Electronics, Woman's clothing, Home, and life. Here, rather than analyzing reviews based on polarity, we analyze reviews based on aspects. In this dataset, many algorithms and seven evaluation metrics were used. To perform this study, a new corpus was developed and the review evaluation problem was transformed into a multi-class and labeled topic modeling problem.

Wassan et al. [8] presented a deep learning method "Deep Sentiment Classifier" (DSC) to assess whether or not to recommend a product. In this study, descriptive statistics, multivariate distribution, univariate distribution, and multivariate analysis are all used in the statistical analysis. To analyze the classification of sentiments, they used statistical analysis, deep learning, and machine learning analysis. In this study, they used two different kinds of datasets, including reviews of women's clothing and reviews of movies from the IMDB dataset. With an F1-score of 93.52% in the dataset for women's clothing's recommended classification, DSC did well in the emotional classification. Additionally, they evaluated DSC on the IMDB datasets, which obtained an average F1-score of 88.01% for the data mining task. The

study's conclusions demonstrated the model DSC's successful performance when the dataset's samples were distributed both uniformly and unevenly.

Several deep learning models, including convolutional neural networks, recurrent neural networks, and bi-directional long short-term memory, were evaluated in [9] using various word embedding techniques, including the Bi-directional Encoder Representations from Transformers (BERT) and its variants, FastText, and Word2Vec. Data augmentation was carried out using the Easy Data Augmentation approach, yielding two datasets (original versus augmented). Many machine learning techniques, including Naive Bayes, LR, and SVM, among others, were also used to compare the model's performance. The outcomes show that the enhanced dataset improves prediction when compared to the original dataset. Word2Vec was demonstrated to provide superior results to FastText for the context-free embeddings, despite the little variances. Similarly, RoBERTa outperformed BERT and ALBERT in terms of results. Additionally, our findings show that ensemble models outperform individual models and machine learning models in terms of performance.

Sentiment analysis is carried out on the Women's e-clothing review dataset from the Kaggle data repository [10]. Two artificial neural networks—CNN and LSTM—are used to classify reviews as positive or negative. Performance measurement parameters including accuracy, recall, specificity, F1-score, and ROC curve are used to evaluate the models. With classification accuracy of 91.69%, specificity of 92.81%, the sensitivity of 76.95%, and F1-score of 56.67%, LSTM outperforms CNN. The research found that the suggested LSTM model is reliable, accurate, and efficient.

### 3 Methodology

The methodology used in this study is comprised seven phases. It is described as follows: Data collection is the initial stage, during which Amazon women's clothes reviews are obtained from the Kaggle repository. Data preprocessing is the second phase, where the missing values are handled and text preprocessing is done. The third phase is data annotation. The fourth phase is data visualization. In the fifth phases, feature engineering techniques like CountVectorizer and TF-IDF are used. Data prediction is followed by the use of machine learning algorithms in the sixth step. For each classifier, a confusion matrix is built. The performance of the trained models is evaluated at the final stage. The performance of ML algorithms is also analyzed using ROC.



Fig. 1 Missing value

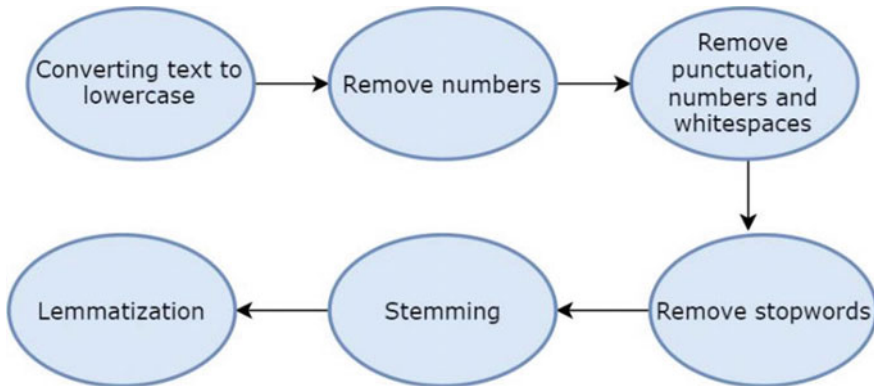
### 3.1 Data Preprocessing

#### 3.1.1 Data Observation and Missing Value Handling

We collect data from the Kaggle “Women’s Clothing E-Commerce” dataset, which includes customer reviews. This feedback is actual business information. As a result, all data are anonymized to safeguard people’s privacy, and its source is hidden from view. First, we do data preparation and load several modules, such as Numpy, pandas. Reading the data reveals that there are 23,486 rows and 10 columns of variable information. Customer reviews and information are included in each row of the dataset, including the clothing ID, age, title, review text, rating, recommendation IND, positive feedback count, department name, department name, and category name. Next, we look for outliers and missing numbers. It displays the distribution of missing values which is shown in Fig. 1.

#### 3.1.2 Text Preprocessing

We undertake text preprocessing, a crucial stage in natural language processing, after the analysis of the numerical data. We consistently convert the text to lowercase and eliminate numerals, punctuation, stop words, and whitespace to keep the data structured and prepared for the next phase. The words are then stemmed to turn the words into their root form (for example, the words amaze, amazing, amazed are all turned into amaze). Figure 2 shows the process of text preprocessing.



**Fig. 2** Text preprocessing

### **3.2** *Data Annotation*

An essential step in the test preparation process is data annotation, which is particularly useful in the supervised learning mode since it may aid in identifying recurring patterns in the data. As a result, we must use clean, labeled data to train our models, and the completeness and consistency of the labeling process will have a significant impact on the development of machine learning models. Based on the review text's rating of the product by the consumer, tags are included in this job. Scores of 3 or more are regarded as positive and are represented by the number 1. Scores under 3 are viewed as negative and are represented as 0.

### **3.3** *Data Analysis and Visualization*

Our ability to further analyze problems and patterns is aided by visualization, which provides us with clear, understandable information about the data. As a result, many visual charts will be utilized in this part to illustrate various data circumstances.

#### **3.3.1** *Age, Rating, and Positive Feedback Distribution Analysis*

According to Fig. 3, the age distribution, score distribution, and positive feedback count of each age group are represented by displot, box plot, and join the plot, respectively. This online store's primary customers are between the ages of 30 and 40, and it has a reasonably high user rating of 5 compared to other age groups. Additionally, the box plot's first, tertile, and median are very close, suggesting that the proportion is generally age-independent.

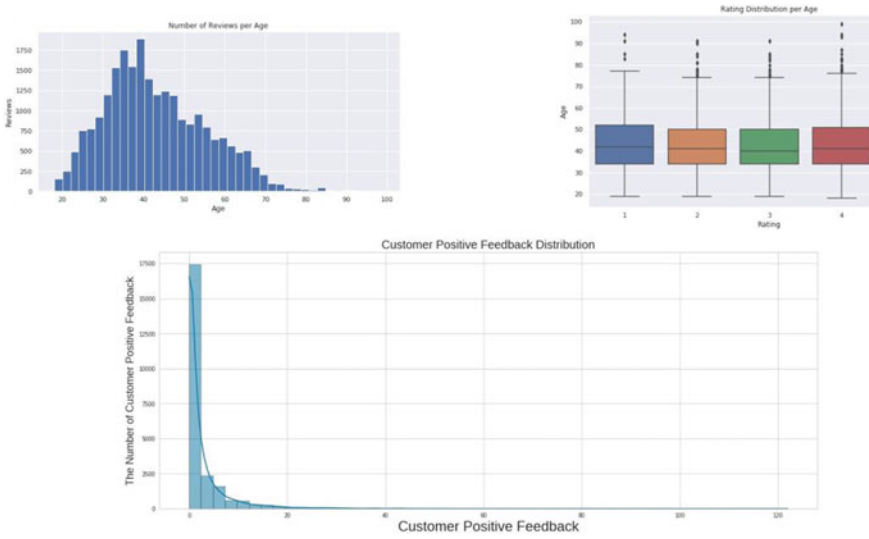


Fig. 3 Age, rating, and positive feedback distribution

Customers are generally satisfied and the site has received positive feedback overall.

### 3.3.2 Division, Class, and Department Category Analysis

The bar plot in Fig. 4 demonstrates that there are three major divisions: General, General Petite, and Intimates. The general category has received the most reviews, while intimates have received less. Figure 4 shows six departments. In comparison to other departments, the tops' department has the highest frequency distribution.

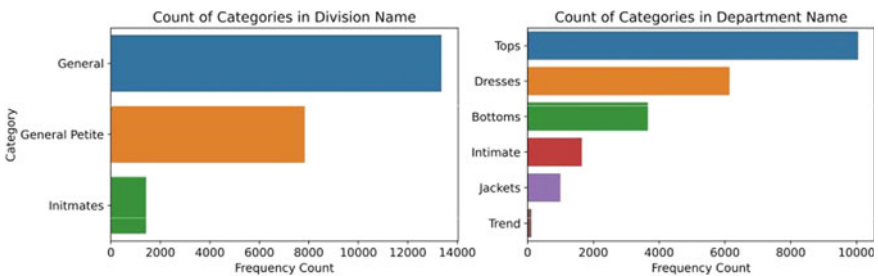


Fig. 4 Division and department category analysis

### 3.3.3 Analysis of Star Products

Dresses, knits, and blouses are the top three most reviewed apparel types. Out of the 20 classes in Fig. 5, the class named Dresses received the most reviews. It also shows the top 60 most reviewed clothing IDs. As shown in Fig. 6, three dress IDs from 1078, 862, and 1094 have obtained more ratings compared to others. They also observed that these products mainly were regular.

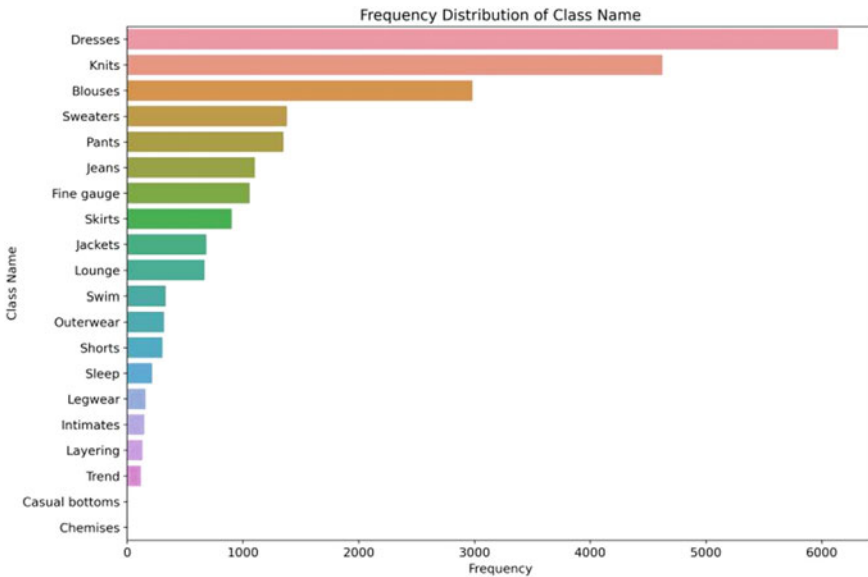


Fig. 5 Frequency distribution of class name

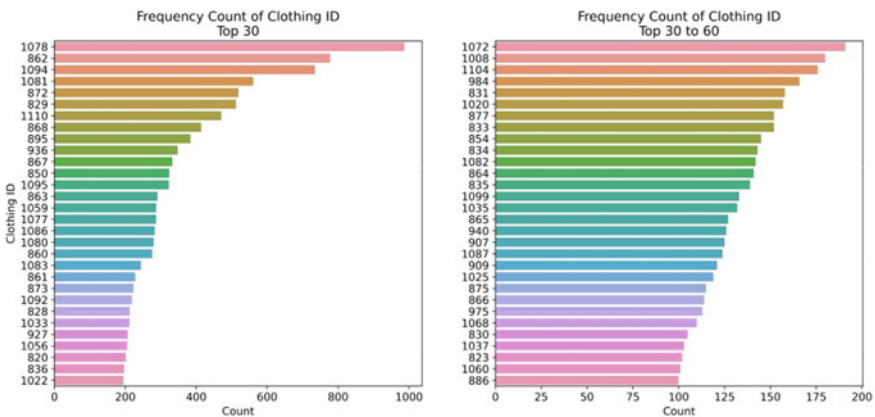


Fig. 6 Frequency count of top 60 clothing ID



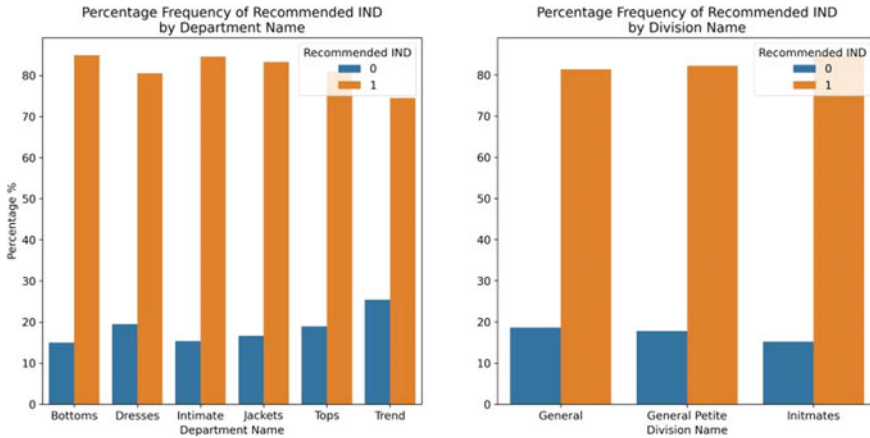


Fig. 7 Percentage frequency of recommended IND by department name and by division name

### 3.3.4 Analysis of Division, Class, and Department Categories Based on Recommendations

Recommended IND is a two-valued variable with values 0 or 1. A rating of 1 indicates that the customer recommends the product, whereas a rating of 0 means that they do not. Figure 7 shows the percentage of frequency of recommended IND by department name and division name. The bar plots demonstrate that customers greatly recommend bottoms, and per class, we can observe that the intimate's division is the most well-liked one.

### 3.3.5 Correlation Analysis

The correlation between two numerical variables can be better understood using a correlation matrix. It aids in our comprehension of their inter-relationship. A correlation matrix is employed to comprehend the link between the variables. In this Fig. 8, we see a strong correlation between recommendations and ratings at a value of 0.8.

### 3.3.6 Analysis of Average Ratings, Ratings, and Reviews

In Fig. 9, the bar plot illustrates the relationship between recommendation and rating, i.e., when a review includes a recommendation, the rating is high, and otherwise, it is halved. Positive reviews outnumber negative and neutral reviews in Fig. 10, as can be observed. Customers can rate products on a scale of 1–5, and it is obvious from the bar plot that the highest rating is 5. And it is also clear from the graph that a rating of 5 is gained by positive reviews. The final bar plot demonstrates that compared

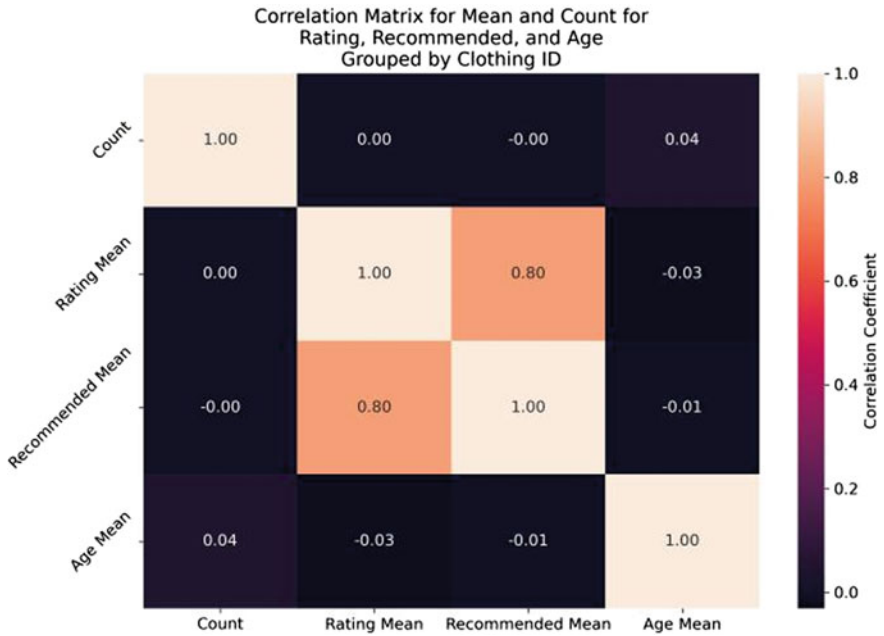


Fig. 8 Correlation matrix

to negative and neutral sentiment scores, positive reviews have received the highest ratings.

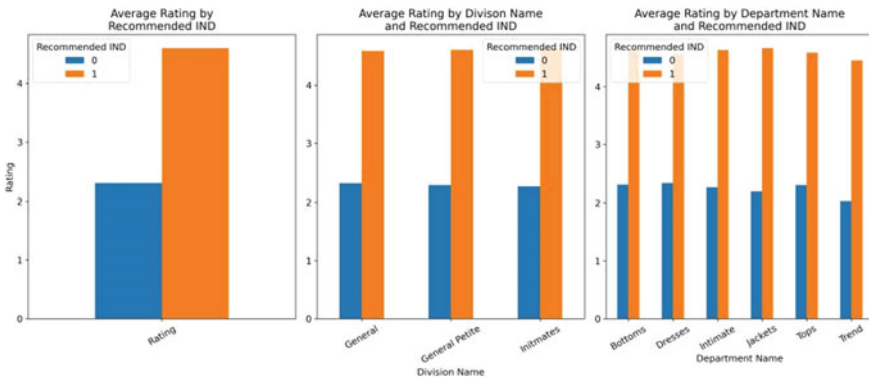


Fig. 9 Analysis of average ratings

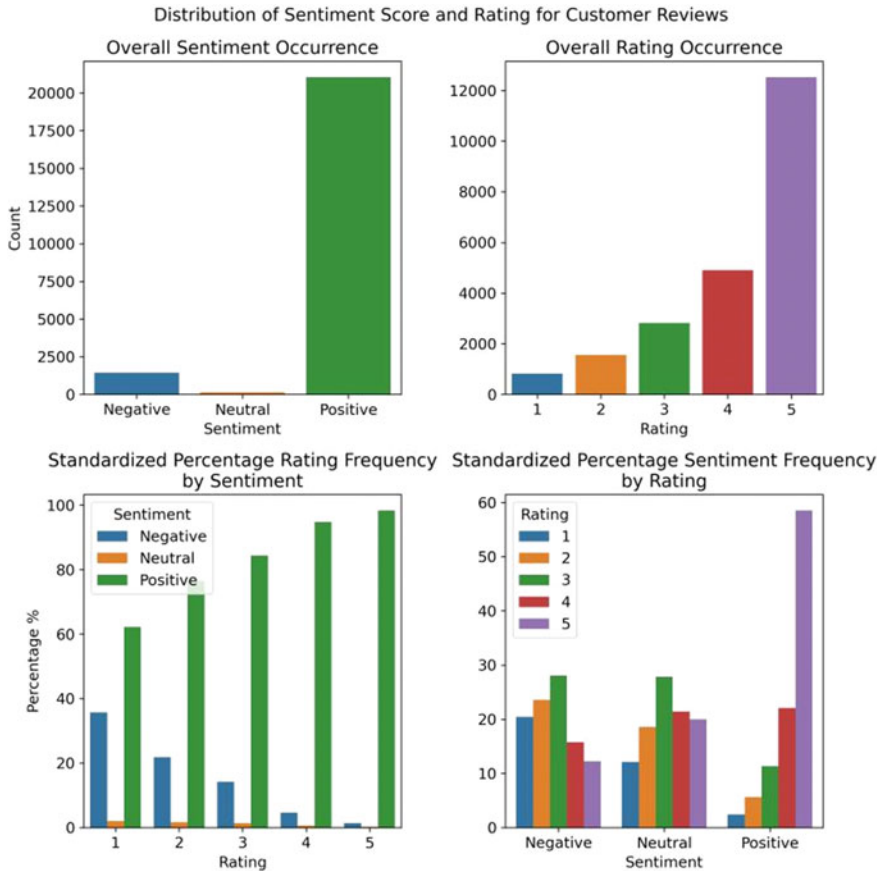


Fig. 10 Distribution of sentiment score and rating for customer reviews

### 3.3.7 Analyzing Keywords

Word cloud is the visual portrayal of text. Word clouds are used in text analytics and text summarization. It is constructed using keywords present in the text. They sum up single words without understanding their linguistic relationships or meanings. They have no or very little interaction capabilities and are utilized statistically [11]. Figure 11 displays the keywords that appear the most in the review title. With the exception of the word “Flaws,” the majority of the words in the cloud convey positivity. In other words, there could be keywords that can be used in place of the negative word indicators but did not make it into the word cloud. Figure 12 displays the phrases that appear often in reviews with high ratings. Positive sentiment is shown by the majority of the terms in this word cloud, and it is obvious that positive reviews have received high ratings. The bar plot in Fig. 14 shows the occurrence score of the frequent words used in high-rated comments. It is observed that dress, love, and size

# WordCloud for Titles



Fig. 11 Word cloud for titles

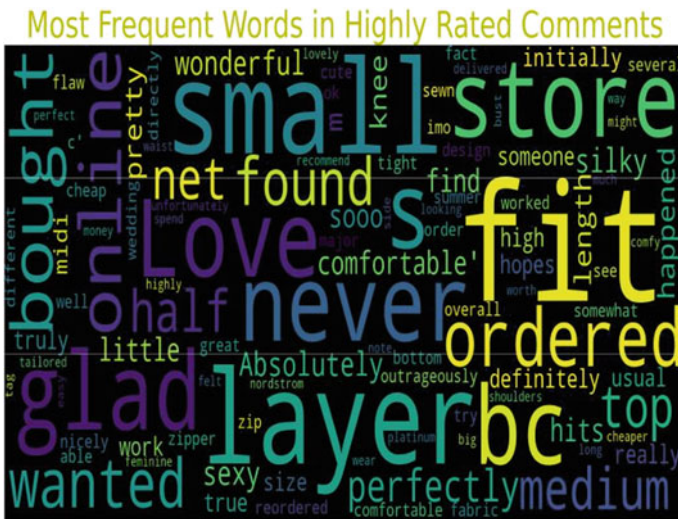


Fig. 12 Most frequent words in highly rated comments

are the most frequently used words. Figure 13 displays the words that appear often in reviews with negative ratings. Unfavorable reviews receive low ratings since the majority of the words in this word cloud express negative sentiment.

### 3.4 Bag of Words (BOW)

The Bag-of-Words (BOW) model is an easy model used for text depiction. Depending on how frequently a word appears in the documents or sentences, it generates a matrix. The significant shortcomings of this model, however, are that the sentence or document structure, grammar as well as semantic meaning between the words is

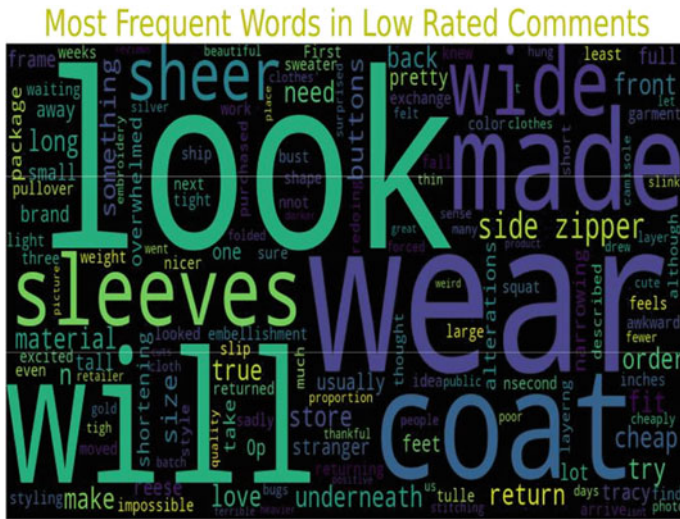


Fig. 13 Most frequent words in low rated comments

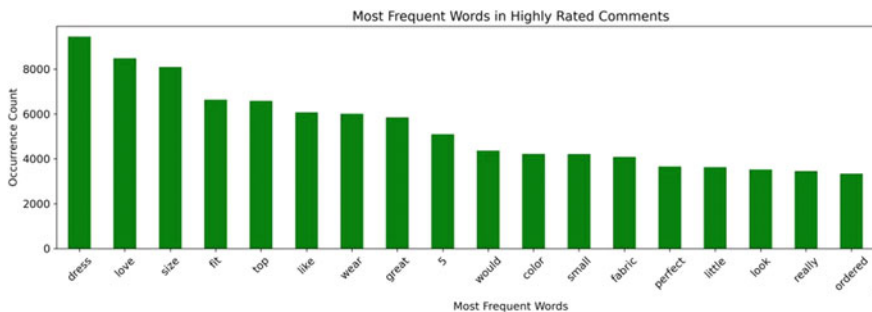


Fig. 14 Positive words

not retained [12]. In another word, it assumes that we do not consider the contextual relationship between words in the text but only consider the weight of all words. The weight is related to the frequency of words appearing in the text. The BOW model will first perform word segmentation. Afterword segmentation, by counting the number of times each word appears in the text, we can get the word-based features of the text. If these words and the corresponding word frequencies of each text sample are put together, which is what we often call vectorization.

Words are converted into numbers by vectorizations so that ML techniques can be applied to them and the computer can analyze them. We explore the CountVectorizer and the TF-IDF approach in this paper.

**Table 1** CountVectorization matrix

Sentences	Boy (Feature1)	Good (Feature2)	Ram (Feature3)	Shyam (Feature4)
Sentence1	1	1	1	0
Sentence2	1	1	0	1

### 3.4.1 CountVectorizer

CountVectorizer is a vectorization method that tracks a word's frequency across the document. It applies to real-world datasets. Text is converted into a vector by indicating a word's presence with a 1 and its absence with a 0.

An example of CountVectorizer is shown in Table 1.

Sentence1 = Ram is a good boy.

Sentence2= Shyam is a good boy.

The matrix above contains two rows and two columns. It is useful when the dataset contains multiple texts and each word has to be converted to a vector. The semantic information is lost while using CountVectorizer, which is a drawback.

### 3.4.2 TF-IDF

Vectors are numerical values that can be processed by any machine learning algorithm. Term Frequency–Inverse Document Frequency is abbreviated as TF-IDF. A word's TF-IDF score is estimated by multiplying two factors: Term Frequency and Inverse Document Frequency. Term Frequency refers to the frequency of the term appearing in the document. Words with a high TF score have importance in documents [13]. The IDF of a term represents the number of times the word appears in multiple documents.<sup>1</sup> Words having a high DF value are unimportant since they appear in all documents [13]. The equation for Term Frequency is given by Eq. 1.

$$TF = \frac{\text{Number of times keyword is found in a document}}{\text{Number of words in a document}}. \quad (1)$$

The equation for Inverse Document Frequency is given by Eq. 2.

$$IDF = \log \left\{ \frac{\text{Number of documents}}{\text{Number of documents containing the keyword}} \right\}. \quad (2)$$

Therefore, by combining Eqs. (1) and (2), we get

$$TF - IDF = \text{Term Frequency (TF)} * \text{Inverse Document Frequency (IDF)}.$$

<sup>1</sup> <https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/>.

### 3.5 Machine Learning Algorithms

Machine learning (ML) is made up of many algorithms whose primary goal is to address data problems. There is no one-size-fits-all algorithm that is best for solving an issue [14]. The data scientist must select the best algorithm for his data problem. Here is a brief overview of some of the ML algorithms that has been employed in our study.

#### 3.5.1 Logistic Regression (LR)

Logistic regression, often known as the logit model, is a statistical model used in categorization and predictive analytics. Using previous observations of a data collection, it predicts a binary result, such as yes or no.<sup>2</sup> LR is used when there is one and more than one independent variable. Using machine learning to predict a person's probability of contracting COVID-19 infection is an example of logistic regression. Because there are only two possible answers to this query—yes, they are infected, or no, they are not—this classification is known as binary.

#### 3.5.2 Multinomial and Bernoulli Naive Bayes (MNB and BNB)

The Naive Bayes classification technique is formulated on the Bayes theorem. There are three types of this algorithm: Gaussian, multinomial, and Bernoulli. It is a probabilistic classifier that is utilized for both binary and multi-class classifications. When the training data are small and the input variables are categorical, it works well. It is based on probability models with strong independence assumptions. Each pair of characteristics being classified by this algorithm is distinct from one another.<sup>3</sup> Multinomial Naive Bayes handles discrete values. The concept of term frequency, or how frequently a word appears in a document, is used by the MNB classifier. In contrast to MNB, BNB classifier does not provide information on term frequency and instead relies on the binary concept of whether a phrase appears in a text or not [15].

#### 3.5.3 Support Vector Machine (SVM)

SVM is a supervised machine learning approach that can perform regression analysis as well as classification. Between the two classes lies a decision boundary called a hyperplane. The number of attributes affects the hyperplane's dimension. The hyperplane is a line when there are two features. A two-dimensional plane is a

---

<sup>2</sup> <https://www.ibm.com/in-en/topics/logistic-regression>.

<sup>3</sup> <https://www.geeksforgeeks.org/naive-bayes-classifiers/>.

hyperplane if there are three features.<sup>4</sup> It is capable of doing both linear and nonlinear classifications. It can handle semi-structured and structured data, but its performance decreases when the dataset is largely due to the increased training time [16]. When there are many observations and variables, SVM can be applied.

### 3.5.4 Random Forest (RF)

Multiple decision trees make up the random forest algorithm. We call it a Random Forest because we employ random choice of data and features to create a forest of decision trees (many trees).<sup>5</sup> Each decision tree's predictions are used by RF to create the final output, which is based on the results of the majority of prediction votes. The RF's performance improves as the number of decision trees increases.<sup>6</sup> It is based on ensemble learning. Ensemble learning is a machine learning approach that improves performance by mixing multiple models. Bagging, stacking, and boosting are the three main forms of ensemble learning.

### 3.5.5 AdaBoosting (AB)

The full form of AdaBoost is adaptive boosting, and it is usually used for binary classification. Boosting is an ensemble technique for constructing a strong classifier from weak classifiers. AB consumes less memory and has fewer computational needs [17]. AdaBoost is extensively utilized in a variety of applications, such as credit risk assessment, facial recognition, and image categorization. It is also applied in binary classification scenarios, where it is required to distinguish between two categories.

## 4 Experimental Evaluation

The computing environment, which includes the hardware, software, and datasets used, is described in this section. This section discusses in detail the dataset and performance evaluation metrics.

### 4.1 Computing Environment

We implemented our experimentation of the models on Ubuntu 16.4 operating system running on Intel® Core™ i 52400 CPU 3.10 GHz 4 processors and a hard disk of

---

<sup>4</sup> <https://www.geeksforgeeks.org/support-vector-machine-algorithm/>.

<sup>5</sup> <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.

<sup>6</sup> <https://www.javatpoint.com/machine-learning-random-forest-algorithm>.



500 GB. The machine learning models were implemented in Python 3 under Google Colab.

## 4.2 Dataset

The dataset chosen for the study should be large and enriched to get accurate results. Women's clothing E-Commerce dataset from the Kaggle repository is used for the study. There are 23,486 rows in this dataset, along with ten feature variables. Each row represents a customer review and contains ten columns, namely Clothing ID, Age, Title, Review Text, Rating, Recommended IND, Positive Feedback Count, Division Name, Department Name, and Class Name.

### 4.2.1 Performance Evaluation Metrics

The performance of the machine learning algorithms has been accessed in terms of accuracy, precision, recall, and F-measure which have been measured from the confusion matrix.

Confusion matrix (CM) is the visual depiction of statistical results, where the output can be divided into two or more classes. It is called a confusion matrix or error matrix. The classifier displays statistics about the actual and predicted levels for each review in the text. Most supervised machine learning algorithms' performance is assessed using it. It represents the false positive, true positive, true negative, and false negative values for the classifier. It is extremely useful for measuring accuracy, recall, precision, specificity, and most importantly area under the curve (AUC)-receiver operating characteristic curve (ROC) [18]. For evaluating the effectiveness of the machine learning algorithm, the ROC curve and AUC value are also provided.

The performance metrics are discussed below:

- Accuracy: The ratio of precise predictions to all predictions is known as accuracy. The metric is only valid if each category has an equal number of observations. The more accurate it is, the better the model [19]. This metric's complement is termed error. It is shown in Eq. 3.

$$\text{Accuracy (\%)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (3)$$

- Precision: It is used to determine when a category set occurrence is classified as a part of the category set. Greater precision denotes that the positive recognition is more precise [19]. It is shown in Eq. 4.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (4)$$

- Recall (sensitivity): It is defined as the ratio of samples that are correctly predicted as positive to all positive samples as shown in Eq. 5.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (5)$$

- F1-score (F-measure): By calculating the harmonic mean of these two metrics, as indicated in Eq. 6, it is possible to quantify both precision and recall. The outcome is a number between 0 and 1.

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (6)$$

### 4.3 Results

In this work, the feature extraction methods such as TF-IDF and CountVectorizer are employed to convert input text into vectors. Feature extraction methods are applied to various machine learning models such as LR, MNB, BNB, SVM, RF, and AB. We used accuracy, precision, recall, and F1-score as performance evaluation metrics. We evaluate the model's capacity to generate the best categorization results using these metrics.

Tables 2 and 3 show the performance accuracy of machine learning classifiers using CountVectorizer and TF-IDF, respectively. In the case of the CountVectorization feature extraction technique, MNB and LR attain the best accuracy of 0.94 and other models such as BNB, SVM, RF, and AB attain accuracy of 0.92, 0.93, 0.90, and 0.91, respectively. The low accuracy of 0.90 is attained by RF.

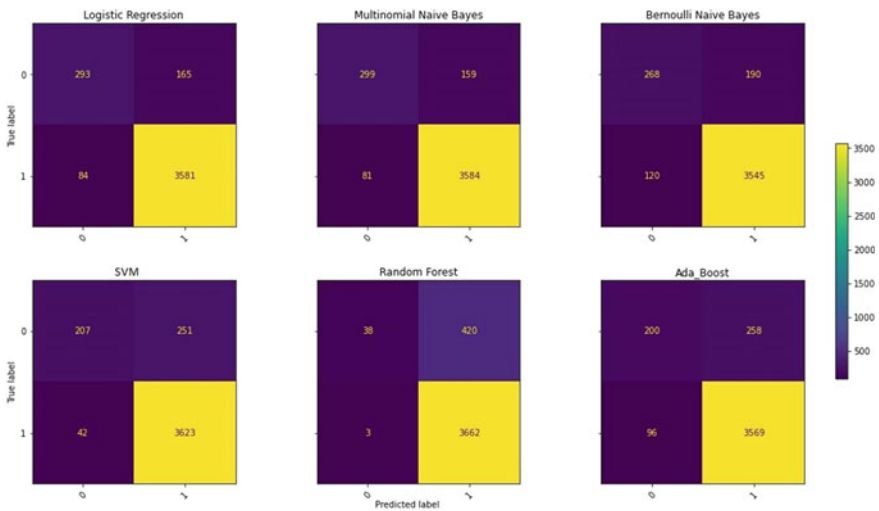
While using the TF-IDF feature extraction technique, SVM outperforms other machine learning models such as LR, MNB, BNB, RF, and AB with the highest accuracy of 0.94 with 0.94 precision, 0.99 recall, and 0.96 F1-score. The lowest accuracy of 0.89 is attained by MNB.

**Table 2** Experimental results on the women's clothes reviews' dataset with different machine learning models based on the CountVectorizer mechanism

	Accuracy	Precision	Recall	F1-score
LR	0.94	0.96	0.98	0.97
MNB	0.94	0.96	0.98	0.97
BNB	0.92	0.95	0.97	0.96
SVM	0.93	0.94	0.99	0.96
RF	0.90	0.90	1.00	0.95
AB	0.91	0.93	0.97	0.95

**Table 3** Experimental results on Women’s Clothes Reviews dataset with different ML models based on TF-IDF vectorizer mechanism

	Accuracy	Precision	Recall	F1-score
LR	0.93	0.94	0.99	0.97
MNB	0.89	0.89	1.00	0.94
BNB	0.92	0.95	0.97	0.96
SVM	0.94	0.94	0.99	0.96
RF	0.90	0.90	1.00	0.95
AB	0.91	0.93	0.97	0.95

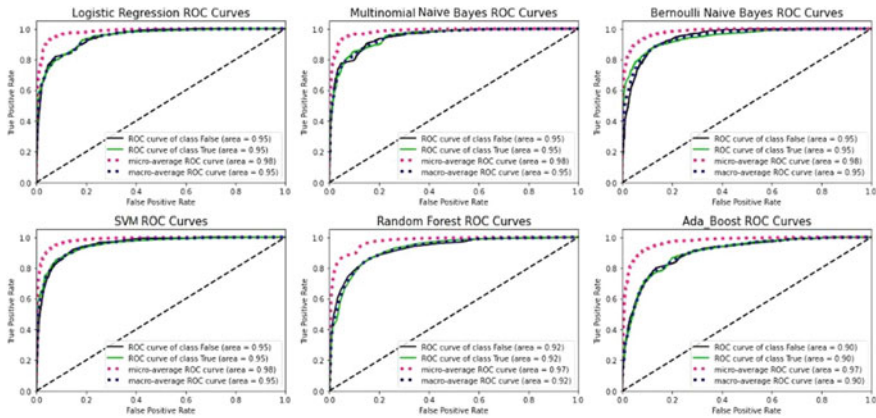


**Fig. 15** Confusion matrix of the proposed models based on CountVectorization mechanism

The results are shown in Figs. 15 and 17, which illustrate the confusion matrices. In the form of a confusion matrix, the performance outcomes of several machine learning approaches have been displayed.

### 4.3.1 Confusion Matrix of Machine Learning Models Based on the CountVectorizer Technique

Figure 15 shows the predictions of LR, MNB, BNB, SVM, RF, and AB classifiers. The maximum accuracy of 0.94 is provided by the LR model, which makes 3581 correct predictions and 293 wrong ones. The MNB model, which has high accuracy comparable to that of LR, makes 3584 correct predictions and 299 incorrect ones. The BNB model predicts 3545 correctly and 268 incorrectly. About 3623 accurate predictions and 207 incorrect predictions are provided by the SVM. With the lowest



**Fig. 16** ROC curves for the proposed models based on CountVectorization mechanism

accuracy, RF makes 3662 accurate predictions and 38 incorrect ones. The AB model yields 3569 right predictions and 200 incorrect predictions.

#### 4.3.2 Confusion Matrix of Machine Learning Models Based on TF-IDF Technique

Figure 17 shows the predictions of LR, MNB, BNB, SVM, RF, and AB classifiers. About 3623 correct predictions and 221 incorrect predictions are produced by LR. MNB gives 3665 correct predictions and 1 wrong prediction with the lowest accuracy of 0.89. The BNB model predicts 3545 correctly and 268 incorrectly. With a high accuracy of 0.94, the SVM makes 3620 right predictions and 238 incorrect ones. The RF model yields 3657 accurate predictions and 46 inaccurate ones. The AB model makes 3561 correct estimates and 187 incorrect ones.

For an in-depth study, ROC curves for the proposed models using CountVectorizer and TF-IDF are shown in Figs. 16 and 18. Receiver operating characteristics (ROCs) are a graphical tool used to evaluate classifiers. X-axis is the False-Positive Rate (FPR) and Y-axis is the True-Positive Rate (TPR). The top left corner of the plot is considered the ideal point. ROC leads to another metric called area under the curve (AUC). To consider a model to be good, the AUC value must be higher. The trade-off between TPR and FPR for a binary classifier is visually represented by the ROC curve and the entire classifier performance is summarized by the AUC using a single numerical value. The micro average and macro average are also calculated. Micro average is the sum of True Positive and False Positive, whereas macro average is the average curves across all classes.<sup>7</sup>

<sup>7</sup> <https://www.scikit-yb.org/en/latest/api/classifier/rocauc.html#multi-class-rocauc-curves>.

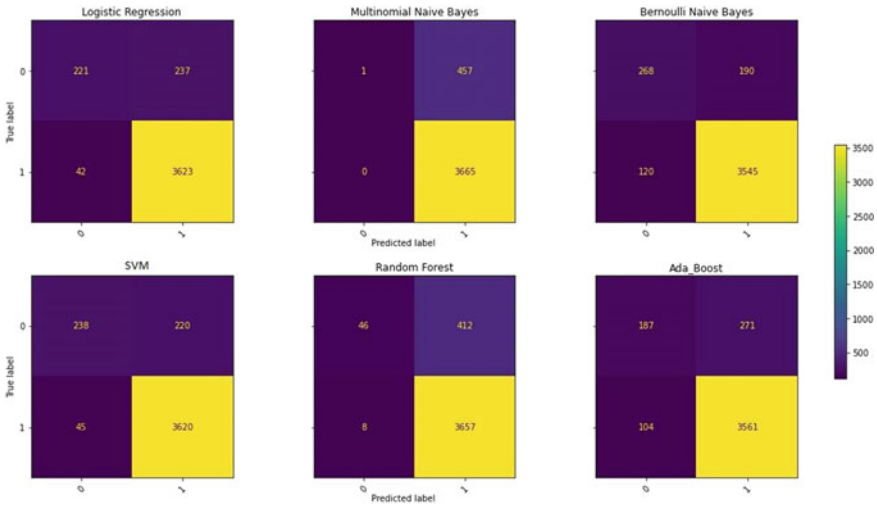


Fig. 17 Confusion matrix of the proposed models based on TF-IDF vectorization mechanism

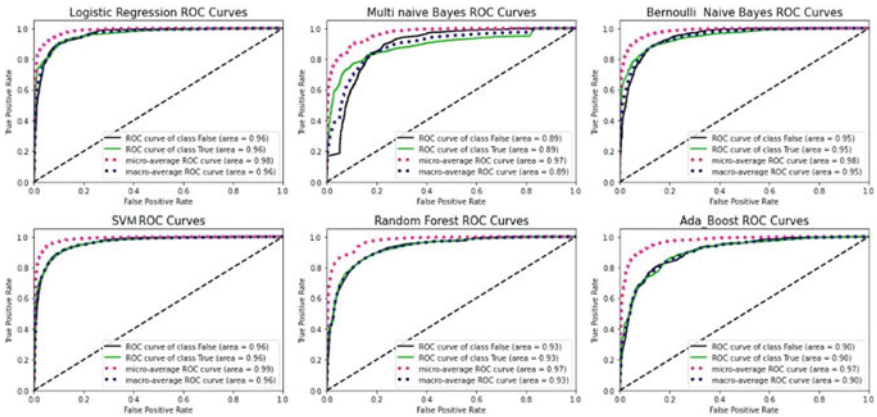


Fig. 18 ROC curves for the proposed models based on TF-IDF vectorization mechanism

### 5 Conclusion and Future Work

In this paper, we investigated the performances of six ML algorithms such as LR, MNB, BNB, SVM, RF, and AB on Amazon E-Commerce clothing review dataset. The study was conducted on two vectorization techniques, CountVectorizer and TF-IDF. In the study, evaluation metrics such as accuracy, precision, recall, F1-score, and ROC are used. While comparing the accuracies of the machine learning models with respect to CountVectorizer, it was observed that MNB and LR models attained the highest accuracy of 0.94 and the lowest accuracy of 0.90 was attained by RF.

In the case of TF-IDF, it is observed that SVM attained highest accuracy of 0.94, and the lowest accuracy of 0.89 was attained by MNB. Only two feature extraction strategies were utilized here; many other techniques may be employed, and they can all be analyzed using various ML models. Additionally, deep learning techniques can also be used. This idea may be used to test various classification strategies, and it can be used to test various datasets and feature extraction techniques.

## References

1. Katthik RV, Sannasi G (2021) A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce. *Appl Soft Comput* 108:107396 (2021). <https://doi.org/10.1016/j.asoc.2021.107396>
2. Wankhade M, Rao ACS, Kulkarni C (2022) A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* 1–50
3. Liu B et al (2010) Sentiment analysis and subjectivity. In: *Handbook of natural language processing*, vol 2, pp 627–666
4. Kubrusly J, Neves AL, Marques TL (2022) A statistical analysis of textual e-commerce reviews using tree-based methods. *Open J Stat* 12(03):357–372
5. Lin X (2020) Sentiment analysis of e-commerce customer reviews based on natural language processing. In: *Proceedings of the 2020 2nd international conference on big data and artificial intelligence*. ACM, New York, NY, USA
6. Agarap AF (2018) Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (RNN). [arXiv:1805.03687](https://arxiv.org/abs/1805.03687)
7. Deniz E, Erbay H, Coşar M (2022) Multi-label classification of e-commerce customer reviews via machine learning. *Axioms* 11(9). <https://doi.org/10.3390/axioms11090436>
8. Wassan S, Shen T, Xi C, Gulati K, Vasan D, Suhail B (2022) Customer experience towards the product during a coronavirus outbreak. *Behav Neurol* 2022:4279346. <https://doi.org/10.1155/2022/4279346>
9. Balakrishnan V, Shi Z, Law CL, Lim R, Teh LL, Fan Y (2022) A deep learning approach in predicting products' sentiment ratings: a comparative analysis. *J Supercomput* 78(5):7206–7226. <https://doi.org/10.1007/s11227-021-04169-6>
10. Nawaz Z, Zhao C, Nawaz F, Safeer AA, Irshad W (2021) Role of artificial neural networks techniques in development of market intelligence: a study of sentiment analysis of eWOM of a women's clothing company. *J Theor Appl Electron Commerce Res* 16(5):1862–1876. <https://doi.org/10.3390/jtaer16050104>
11. Kabir AI, Ahmed K, Karim R (2020) Word cloud and sentiment analysis of amazon earphones reviews with r programming language. *Informatica Economica* 24:55–71. <https://doi.org/10.24818/issn14531305/24.4.2020.05>
12. Kumar JA, Trueman TE, Cambria E (2022) Gender-based multi-aspect sentiment detection using multilabel learning. *Inf Sci* 606:453–468. <https://doi.org/10.1016/j.ins.2022.05.057>
13. Kim S-W, Gil J-M (2019) Research paper classification systems based on TF-IDF and LDA schemes. *HCIS* 9(1):1–21
14. Mahesh B (2019) *Machine learning algorithms—A review*
15. Singh G, Kumar B, Gaur L, Tyagi A (2019) Comparison between multi-nomial and Bernoulli naïve Bayes for text classification. In: *2019 International conference on automation, computational and technology management (ICACTM)*, pp 593–596. <https://doi.org/10.1109/ICACTM.2019.8776800>
16. Ray S (2019) A quick review of machine learning algorithms. In: *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, pp 35–39

17. Aljedaani W, Rustam F, Ludi S, Ouni A, Mkaouer MW (2021) Learning sentiment analysis for accessibility user reviews. In: 2021 36th IEEE/ACM international conference on automated software engineering workshops (ASEW). IEEE, pp 239–246
18. Behera RK, Jena M, Rath SK, Misra S (2021) Co-lstm: Convolutional LSTM model for sentiment analysis in social big data. *Inf Process Manage* 58(1):102435
19. Li H, Chen Q, Zhong Z, Gong R, Han G (2022) E-word of mouth sentiment analysis for user behavior studies. *Inf Process Manage* 59(1):102784. <https://doi.org/10.1016/j.ipm.2021.102784>