10.48047/jocaaa.2024.32.01.82

Enhanced Hybrid Feature Selection Technique for Sentiment Classification

Pankaj Kumar Gautam[1](0000-0003-2918-8855), Akhilesh A. Waoo[2*](0000-0001-6788-710X)

[1, 2] *Department of Computer Science, Faculty of Computer Application & Information Technology and Sciences, AKS University Satna, M.P., INDIA 485001*

*** Corresponding Author-** akhileshwaoo@gmail.com

Abstract

Redundant, irrelevant, and high-dimensional features give inaccurate results in SA. Feature selection (FS) is an important step in the SA preprocessing process that improves classification accuracy, precision, recall, and F1 score by reducing feature dimensionality and overcoming underfitting and overfitting by selecting relevant features. Additionally, FS significantly reduces computing time, complexity, and storage. This paper uses a hybrid method with the one-way ANOVA filter and recursive feature elimination wrapper to handle the high-dimensional Amazon review dataset problem. This method improves the performance of Amazon reviews regarding classification accuracy, precision, recall, and F1 score. Four feature selection techniques, such as one-way ANOVA with RFE, Mutual Information (MI), Genetic Algorithm (GA), and Principal Component Analysis PCA with six classifiers (K-Nearest Neighbor (KNN), Naïve Bayes (NB), Linear Support Vector Classifier (Linear SVC), Decision Tree (DT), Random Forest Classifier (RFC), and Logistic Regression (LR), are used in this study. The Amazon review dataset was used in this experiment. One-way ANOVA with RFE has reduced the 50% feature set to obtain a higher accuracy of 0.951 with the KNN classifier, a higher precision of 0.935 with the KNN classifier, a higher recall with the KNN classifier, and a higher F1-score with the KNN classifier. Other classifiers also get desirable accuracy, precision, recall, and F1-score.

Keywords: feature selection; sentiment analysis; pre-processing; hybrid feature selection method; one-way ANOVA; recursive feature elimination (RFE).

Introduction

People provide reviews or feedback digitally for any product on an online platform or social media site. This huge amount of data is fruitful for business analytics, but this data must be processed properly [8]. This approach is known as sentiment classification. Sentiment classification, also called opinion mining, is a process where the extraction and comprehension of opinion, emotion, and feedback from the text document are represented in positive, negative, or neutral form [10, 13]. Reviews, ratings, comments, and customer feedback are all forms of electronic word-of-mouth (e-WOM). Currently, e-WOM sentiment classification is widely used. It is among the key elements influencing how customers behave [9]. It has a great impact on the product purchase decision and business strategy. However, this data is in a huge volume and an unstructured format, which includes unwanted words like symbols, stop-words, and sometimes URLs. The manual search and evaluation of this data resulted in a significant overload [8]. Consequently, it is important to draw attention to the sentiment classification for product reviews [9].

Sentiment classification, or sentiment analysis (SA), is carried out on three levels: sentence, document, and aspect level [8]. In this study, we concentrate on sentence-level SA, utilizing tokenizers to process each sentence in the review and translate it into words. In SA, the important steps are text pre-processing, feature extraction, and feature selection. Feature extraction creates a feature vector for machine-understanding form and feature selection; it extracts and selects the most important features. Text pre-processing, feature extraction, and selection are essential steps in textual SA [9].

The state-of-the-art methods of SA are primarily supervised and lexicon-based [10]. In the lexicon-based method, the overall sentiment of the given text is obtained by finding and mapping each word to its category, which is positive, negative, or neutral, or by calculating the numerical score of each word. On the other hand, the supervised method depends on learning the classification model. This model learns from the training set's labeled data to find the testing data category [1]. However, feature selection

10.48047/jocaaa.2024.32.01.82

techniques in both methods find the most important feature for classification.

In feature extraction, a feature vector is created for machine comprehension, while in feature selection, the most crucial features are extracted and chosen for SA. Specifically, feature selection selects the subset of relevant and most significant features, whereas feature extraction concentrates on finding and extracting emotion words and phrases [10]. Feature extraction and selection are crucial to reduce model complexity and storage needs and enhance performance [1].

In machine learning (ML) systems, there are three methods for selecting features: filter, wrapper, and hybrid. The filter method uses feature relevance statistics such as mutual information, one-way ANOVA, and Chi-Square to determine feature importance. Contrarily, the wrapper is a classifier-dependent method that chooses features according to the ML's predicted classifier performance on a specific subset [9]. The learning algorithm process involved in feature selection is the reason it takes time. The hybrid feature selection methodology combines any filter and wrapper feature selection approaches to overcome the limitations of the filter and wrapper methods. The best hybrid feature selection strategy integrates the advantages and disadvantages of the wrapper and filter feature selection strategies. However, the hybrid feature selection strategy produced the best classification results. Moreover, the performance of ML is highly dependent on the size of its features. As a result, the ideal plan for overcoming the shortcomings of the filter and wrapper approaches is hybrid feature selection [9]. A large dataset's worth of redundant and irrelevant features can drastically lower machine learning models' effectiveness and seriously impair their performance [8]. They could also lead to issues with overfitting, underfitting, and the curse of dimensionality [4].

This paper presents a hybrid feature selection method based on the one-way ANOVA filter method and the recursive feature elimination wrapper method to overcome the limitations of both methods. The one-way ANOVA is employed as an FS filter method to assess a feature's influence on a target class, and the RFE will rank the features by selecting the significant features. This hybrid method is compared with the other three state-of-the-art feature selection methods.

The structure of this article is as follows: Section 1: Introduction; Section 2: Review of Literature; Section 3: Methodology; Section 4: Word Cloud; Section 5: Feature Extraction and Feature Selection; Section 6: Classifiers and Comparison of Evaluation Metrics for Various State-of-the-Art Techniques; Section 7: Results and Discussion; Section 8: Conclusions with some findings.

Review of Literature

Alassaf and Qamar developed one-way ANOVA as a filter feature selection method to improve classification performance and handle the curse of dimensionality for classifying Qassim University, Saudi Arabia, tweets in the Twitter dataset. TF-IDF is used as a feature extraction method, and classifiers SVM, k-NN, LR, MLP, and NB are used in this study. SVM obtained good performance with an F1-score of 88%.

Yuce, Nielsen, and Wargocki used the Taguchi method, Grey Relational Analysis (GRA), and ANOVA to find optimization in CFD analyses of ventilation performance in construction.

Kausar, Fageeri, and Soosaimanickam proposed a bag-of-words feature extraction approach to extract features from Amazon reviews. The word cloud visualization technique represents the words in the form of the frequency with which they appear in the review text. Logistic regression and decision tree machine learning classifiers were used to get an optimum accuracy of 99% with decision trees.

J and U suggested the Pearson correlation coefficient-based Harris Hawks optimization-based Recurrent Neural Network-Long Short Term Memory (PCCHHO-RNN-LSTM) algorithm to minimize the data's dimensionality. MATLAB software with an RNN-LSTM classifier used for Amazon data got 95.8% accuracy.

Khare et al. employ support vector machine (SVM) classifiers and the TF-IDF feature extraction method to remove sarcastic tweets from Twitter data from the Lok Sabha election.

Chakraborty et. al. and Justin et al. suggested traditional methods. The AdaBoost, Random Forest Classifier (RF), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), deep learning methods, Convolutional Neural Network, and Long Short-Term Memory Network (LSTM) were used to obtain the polarity of Bengali textual Facebook comments and posts. LSTM achieved an improved accuracy of 96.95%. In the classical approach, RF and SVM achieve maximum accuracy of 78.37% and 78.23%, respectively [6,34].

Methodology

Fig. 1 illustrates the process of this study, which begins with data collection and ends with the classification model's evaluation.

10.48047/jocaaa.2024.32.01.82

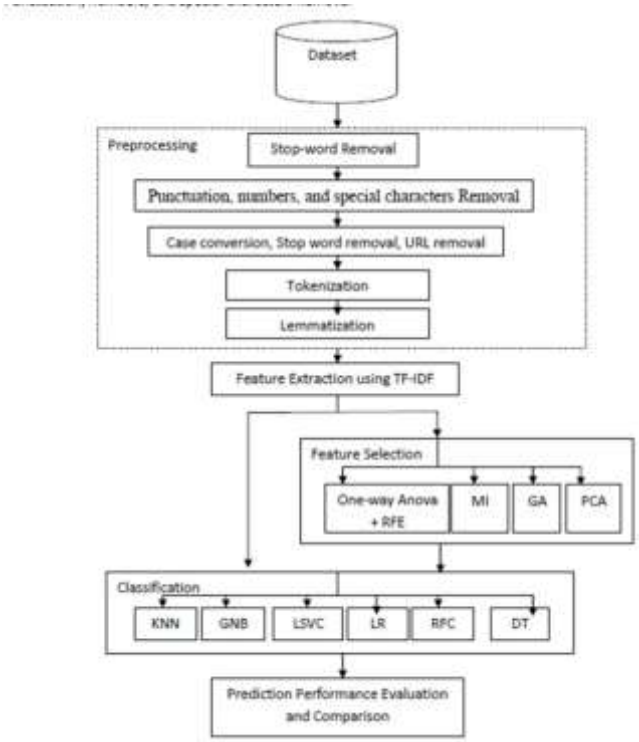


Fig. 1 The proposed hybrid framework for sentiment analysis

Data Collection

The Amazon dataset used in this experiment was obtained from https://kaggle.com in CSV format. This dataset has 568454 user ratings and reviews related to food (see Fig. 3). The score column is used to express product ratings. The dataset includes one more column of sentiment to represent the review's sentiment as positive, negative, or neutral. We have assigned positive reviews for ratings 4 and 5, negative reviews for ratings 1 and 2, and neutral reviews for rating 3. Due to computation constraints, we have taken 10,000 reviews from this dataset, of which 7616 are positive, 1522 are negative, and 862 are neutral (see Fig. 2; Table 1)

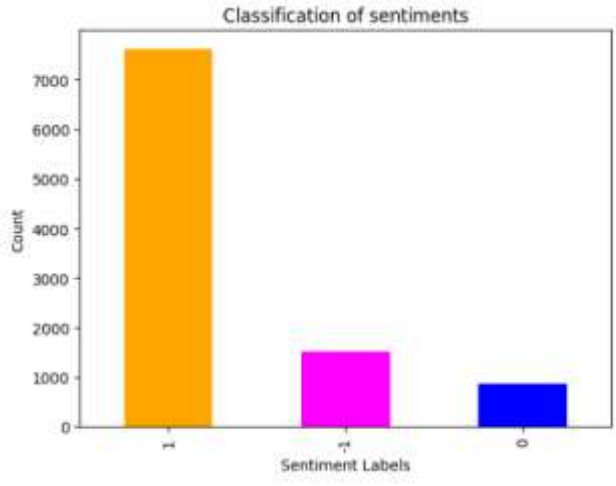
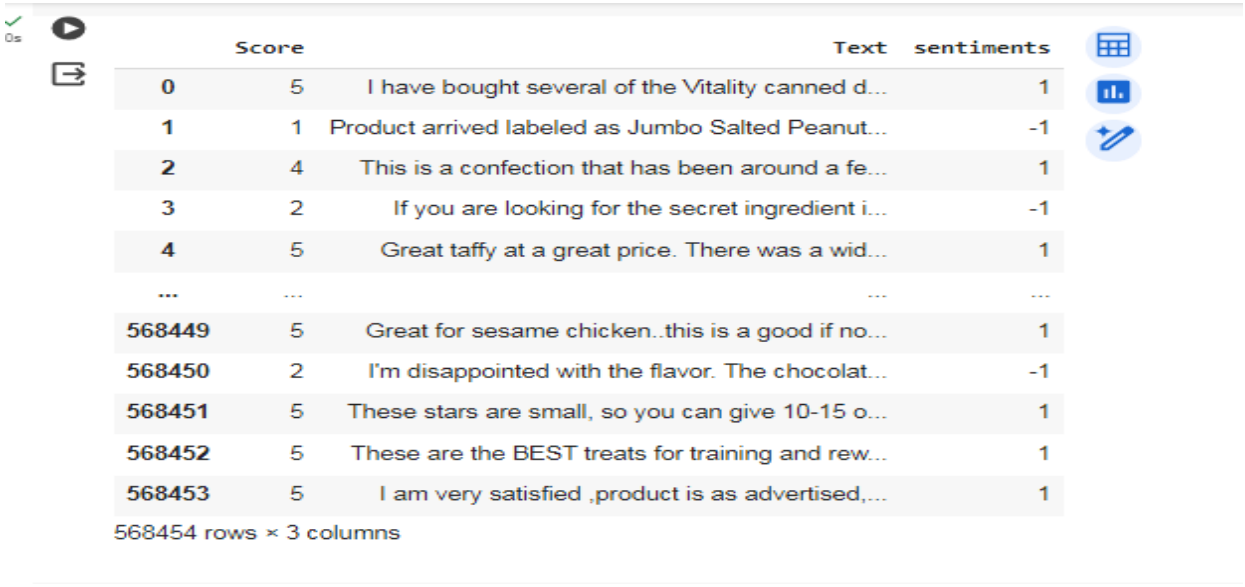


Fig. 2 Bar plot of sentiment counts

10.48047/jocaaa.2024.32.01.82

Table 1: Description of the Dataset

Dataset	Reviews	Class	Category	No of Reviews
Amazon	10000	Positive Negative Neutral	Food products	Positive: 7616 Negative: 1522 Neutral: 862



	Score	Text	sentiments
0	5	I have bought several of the Vitality canned d...	1
1	1	Product arrived labeled as Jumbo Salted Peanut...	-1
2	4	This is a confection that has been around a fe...	1
3	2	If you are looking for the secret ingredient i...	-1
4	5	Great taffy at a great price. There was a wid...	1
...	...	...	...
568449	5	Great for sesame chicken..this is a good if no...	1
568450	2	I'm disappointed with the flavor. The chocolat...	-1
568451	5	These stars are small, so you can give 10-15 o...	1
568452	5	These are the BEST treats for training and rew...	1
568453	5	I am very satisfied ,product is as advertised,...	1

568454 rows × 3 columns

Fig. 3 Screenshot of the Amazon Reviews Dataset

Preprocessing

Data gathered from social media platforms such as Amazon is typically unstructured. The collected data may contain grammatical, typographical, and noisy data (see Table 2). This means the text must be cleaned and pre-processed before any analysis. Preprocessing seeks to decrease the dimensionality of the analysis and the input data. Some redundant words should be eliminated because they do not affect the text's polarity. Case conversion, HTML tag removal, null value removal, punctuation mark removal, URL removal, number removal, stop word removal, slang word removal, tokenization, and lemmatization are included in data preprocessing approaches [1].

Stop-word Removal: The goal of this work is to exclude words or tokens from the text document that are frequently called "functional words" because they are meaningless words like "this," "is," "but," "and," etc. Removal of stop-words decreases computational complexity and does not impact classification because they appear repeatedly in a text. Stop-words in the English language are employed in this study [6, 12].

Punctuation, numbers, and special characters Removal: The text's sentiment is not significantly affected by using numbers, punctuation, or special characters. Therefore, they must be removed. The dimensions of the data decrease when these terms are removed [6, 11].

Case conversion: Uppercase and lowercase words have the same sentiments; therefore, words and tokens must change from uppercase to lowercase before processing [6].

URL removal: Since Uniform Resource Locators (URLs) are not important in determining the sentiment of a text, they should be removed from user-provided text reviews [3, 6].

Tokenization: Strings, or statements, are divided into tokens, which are collections of words, phrases, symbols, and other keywords. For example, the phrase "Ram is a good player" breaks into the tokens "Ram," "is," "good," and "player" [4].

Lemmatization: Writing words in a text does not follow a set format. As a result, even though a word in a review paragraph is written differently, it still has the same meaning. Lemmatization, which converts each of these terms into a base form found in a dictionary, can assist in eliminating this requirement. We call this form a lemma. WordNetLemmatizer was employed in this investigation for lemmatization [7, 12].

10.48047/jocaaa.2024.32.01.82

Table 1 Review before and after preprocessing

Before pre-processing	After pre-processing
"I have bought several of the Vitality canned dog food products and found them all good quality. The product looks more like a stew than processed meat and smells better. My Labrador is finicky, and she appreciates this product better than most."	['bought',' several', 'vitality', 'canned', 'dog ', 'food ', 'products ',' found', 'good', 'quality', 'product', 'looks',' like', 'stew ', 'processed ', 'meat ', 'smells ', 'better ', 'labrador', 'finicky ',' appreciates ', 'product ',' better']

Word Cloud

The dataset contains many common words. The word cloud visualization approach visually shows these most commonly used terms. This study describes three visualizations: positive sentiment visualization, negative sentiment visualization, and neutral sentiment visualization. Fig. 4 shows that all positive words appear frequently in their corresponding word cloud [14].

Fig. 4 shows the most common words used in positive reviews.

Feature Extraction

Representation is an important stage in sentiment classification. Raw data generally contains noise, which needs to be filtered using various pre-processing techniques. After preprocessing, the data is transformed into a term-document matrix, or TDM, which counts the occurrences of every single word. Two feature extraction techniques supported by the TDM are TF-IDF and bag-of-words. A word's TF-IDF score is determined by multiplying its TF and IDF; this score can be found by multiplying these two parts. The terms that appear most frequently in the review dataset are assigned a higher weight in the TF score than others. The IDF scaling factor gives more importance to the least common words in the sample. This score is lower for common and unusual words than for other word categories. If we exclude terms with low TF-IDF values, we can eliminate them. [5, 15]

TF-IDF: This technique, which considers both term frequency (TF) and inverse document frequency (IDF), can extract information. Every phrase and word has a unique TF and IDF mark assigned to it. A term's TF*IDF weight is calculated by adding its entire TF and IDF product scores [5].

Hybrid Feature Selection

In sentiment analysis, the primary goal of feature selection is to identify the optimal collection of features that enables the construction of practical models. Creating an effective prediction model is the main objective of most applications. Finding the factors that improve this accurate prediction is another, and occasionally even more crucial, task. This involves narrowing down the large set of observed variables to those containing information rather than noise [33]. Several feature selection techniques have been presented to report the set of genuinely relevant

10.48047/jocaaa.2024.32.01.82

variables, each based on a different set of concepts and methods. A hybrid feature selection is used in this study. [30] A hybrid feature selection was used in this investigation. The first method is the one-way ANOVA filter method based on the F-value as a score or weight of the feature. Following this feature selection procedure, the output is sent to the wrapper-based attribute reduction technique, RFE. To assess the suggested hybrid feature selection framework, the resulting feature subsets are sent to the classification algorithms [33].

Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is one of the wrapper feature selection methods. It is an attribute selection method based on brute force, which looks for a subset of features. All features in the training dataset are used at first, and it successfully removes the weakest features until the target number of features is left. Features in RFE are categorized based on the characteristics of the model. With recurrent feature deletion, RFE eliminates dependencies and co-linearity from a model by removing a specified number of features every cycle [30].

The Recursive Feature Elimination Algorithm is described as follows [30]:

Algorithm. Recursive Feature Elimination
Tune/Train the model on the training set using all predictors
Calculate model performance
Calculate variable importance or rankings
for each subset size S_i, *i=1,2,.. S* do
Keep the S_i's most important variables
Tune/Train the model on the training set using S_i, predictors
Calculate model performance
end for
Calculate the performance profile over the S_i
Determine the appropriate number of predictors
Use the model corresponding to the optimal S_i

One-way ANOVA

In the feature space, one-way ANOVA has demonstrated its efficacy in resolving high-dimensional problems. To find the features most relevant to a machine learning task, a one-way ANOVA is employed as an FS filter method to assess a feature's influence on a target class. If the variance, the average of the squared deviations from the mean, differs between classes, then the feature impacts the target class. Other features are completely ignored when measuring the association between each feature and the target variable separately. Consequently, regardless of the correlation between features, a one-way ANOVA selects the feature based on the variance difference across classes [33].

$$F - value = \frac{MS_B}{MS_W} \quad \text{eq. (1)}$$

Where MS_B is the between-group variability, defined as:

$$MS_B = \frac{\Sigma_i n_i(\bar{x_i}-\bar{x})}{g-1} \quad \text{eq. (2)}$$

Where
n_i= number of observations in the *ith* group
$\bar{x_i}$=sample mean in the *ith* group
$\bar{x}$= group mean of the data
MS_W= within-group variability, defined as:

$$MS_W = \frac{\Sigma_{IJ} n_i(x_{ij}-\bar{x_i})2}{n-g} \quad \text{eq. (3)}$$

Where
x_{ij}=*jth* sample in the *ith* group,
g= number of groups.

Only m% of the top-scoring features will be used for the training and testing phases. We have taken 10%, 30%, and 50% top-scoring F-value features.

INPUT: A pair (E, Y), where E is the features obtained by TF-IDF and Y represents the class of each feature. Also, m% denotes the percentage of the selected features [1].

OUTPUT: A subset of features obtained based on the F-value.

Begin

1. n_classes = count(y)
2. For each E_j€ (E, Y)
3. n_sample_per_classes = Count (Yi)
4.n_samples = Count ((E, Y))
5. d_{ebug} = n_classes-1// Degrees of freedom between classes
6. dawg = n_samples-1 //Degrees of freedom within classes
7. ss_all_features = sum (square (E))
8. sum_all_features = sum (E)
9. square_of_sum_all_features = square (sum_all_features)
10. $SS_{tota\ l}$= $ss_{all\_features}$-$\frac{\text{square\_of\_sum\_all\_features}}{\text{n\_samples}}$
11. SS_{bg} = 0
12. For each Y_i € Y do
13. SS_{bg}= SS_{bg} + $\frac{\text{square(sum(Eyi))}}{\text{count(E, Yi)}}$
14. End For

10.48047/jocaaa.2024.32.01.82

15. SSbg = SSbg - $-\frac{\text{square\_of\_sum\_all\_features}}{\text{n\_samples}}$
16. $SS_{wg} = SS_{total} - SS_{bg}$
17. $M_{sb} = \frac{SSbg}{dfbg}$ // Variance between classes
18. $M_{wb} = \frac{SSwg}{dfwg}$// Variance within classes
19. F- value = $\frac{Mwg}{Mwg}$//Score of the feature
20. End For
21. Ascending Order (Based on value)
21. FS=Select (The highest %m of E based on the F-value)
22. Return (FS)

End

Mutual Information (MI)

In feature selection theory, features that provide meaningful information about a class are considered relevant, while features that don't offer much information about the output class are deemed irrelevant. Finding the informative features that provide the most information about the output class is, therefore, necessary. For this reason, two methods were proposed to quantify the information conveyed between two random variables: information entropy and MI. Assume that there are two random variables, M = {m1, m2,…, mk} and N = {n1, n2,…, nk}, where k is the total number of samples. The quantity of knowledge known as MI is shared by variables N and M. MI is calculated using Eq. 4 [16].

$$I(M,N) = \Sigma_{m\in M}\Sigma_{n\in N} pd(m.n) log \frac{pd(m,n)}{pd(m)p(n)} \quad \text{eq. (4)}$$

Where,

PD(m,n) = joint probability distribution

PD(m) = probability distribution of M

Pd (n) = the probability distribution of N.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique to reduce dimension by converting higher-dimensional data sets into lower-dimensional ones. It does this by converting an interrelated feature data set into a new set of uncorrelated features, known as principal components (PCs), of which the first few hold most of the variation in the entire data set. The PCs linearly combine the actual features, with the first PC having the greatest variation and the second PC being the most orthogonal to the first PC in terms of variance across the other PCs. In that order, the subsequent PCs come.

The fundamental premise of PCA is that while coordinates with smaller variations may be a source of noise that needs to be suppressed or ignored, coordinates with larger variations show the divergence between sample points. When two dimensions correlate, it indicates irrelevant information that won't be displayed. For this reason, PCA demands that the coordinates that come after be orthogonal to those that came before. Scaling has an impact on PCA. [18]

Genetic Algorithm(GA)

A probabilistic optimization technique called genetic algorithms was developed from the natural evolution of concepts [20]. In genetic algorithms, each chromosome stands for a point in the search space and could be a solution or response to the problem being investigated. The chromosomes (possible answers) consist of a set number of genes. A population is a collection of chromosomes. A fitness function for the problem must be developed before employing genetic algorithms to solve any given issue.

In an optimization issue, the genetic algorithm generates n candidates or solutions, and X chooses the best candidate from the set. Certain features of each candidate solution are changed and mutated to create the best candidate solutions. The genetic algorithm looks for local optimal solutions to find the best global solution. [20]

Classifier

This study uses six classical algorithms (KNN, GNB, LSVC, DT, RFC, and LR) for prediction.

Decision Tree (DT)

A supervised machine learning model decision tree is utilized for both classification and regression. In a decision tree, the training data is regarded as the root node. An organized depiction of these mapping interactions is a decision tree. A tree can consist of one leaf node that is linked to a more comprehensive structure with a root node connected to two or more subtrees, or a more specialized category. A test node can forecast the result based on how an instance's properties are configured. Every possible result has a sub-tree linked to it [21].

Linear Support Vector Classifier (Linear SVC)

10.48047/jocaaa.2024.32.01.82

Support vector machine, or SVM, is a machine learning technique that can be applied to regression, outlier identification, and classification. Its linear form is called linear SVC [7]. A support vector classifier (linear SVC) uses the problem's dimensions to identify which line or hyperplane best fits the data. Dimension is another word for feature number. Because SVC ignores all outliers and just chooses the best hyperplane to distinguish between the classes, it is chosen [21].

Logistic regression (LR)

A supervised machine-learning technique called logistic regression was developed to aid in classification problems. The problem is called a classification learning problem when the target variable is categorical. Logistic regression maps a function from the attributes of the dataset to the targets to forecast the likelihood that a new example will be among the target classes. Sometimes this algorithm is known by its other name, Maximum Entropy. The logistic regression classification algorithm is one of the mathematical creations belonging to the family of extended linear models. The odds of a trial's outcomes are described using the modeling technique known as logistic regression. A parameter grid with several values for the hyperparameters "C," "penalty," and "solver" to be employed in a logistic regression model is used for hyperparameter tuning [17, 19].

K-Nearest Neighbor (KNN)

Machine learning may be done very well with the K-Nearest Neighbor (KNN) method, which is easy to use. The application of it is advantageous for both regression analysis and classification.
That said, classification prediction is where it is most frequently applied. Making predictions for fresh data based on its resemblance to previously taught data, the K-Nearest Neighbor method clusters data into coherent clusters or subsets. Using the class with which it shares the closest neighbors, the input is classified into the appropriate category [17].

Naïve Bayes (NB)

Naïve Bayes is a straightforward classification that finds the probability from the sum of values and the frequency of available data. With the supposition that class one is independent (not interdependent), the probabilities of each class are calculated using the Bayes theorem, which forms the basis of Naïve Bayes. An additional definition is Naïve Bayes, a technique that uses past performance to forecast future opportunities. When used on huge datasets, Naïve Bayes has superior computing time and accuracy levels [17].

Here is how the Bayes' theories typically look:

$$P\left(\frac{H}{X}\right) = P(H).P(H)/P(X) \qquad \text{eq. (5)}$$

Random Forest Classifier (RFC)

A machine learning algorithm called Random Forest Classifier (RFC) can do both regression and classification. Several decision trees that are randomly chosen from a portion of the training data make up this approach. The votes cast by various decision trees accumulate to provide the classification results.

The Random Forest method frequently makes use of a technique called bootstrap aggregation, or bagging. By using the bagging technique, the calculated prediction function's variance is decreased. This approach is thought to be efficient for high-variation data and low-bias processes, including tree-based techniques. When it comes to classification, every tree will yield a prediction result, which will subsequently be determined by taking the average of the results; that is, a majority vote will be cast [17].

Evaluation metrics

Precision (P), recall (R), accuracy, and F1 Score (F) metrics are used to assess the performance of the system [3] [22] [23].

TP (true positive): depicts correctly classified positive samples
FN (false negative): depicts incorrectly classified positive samples
FP (false positive): depicts incorrectly classified negative samples.
TN (true negative): depicts correctly classified negative samples [2].

Precision is the proportion of true positives to the sum of true positives and false positives.

$$\text{PRE} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{eq. (6)}$$

Recall is the proportion of true positives to the sum of true positives and false negatives [33].

$$\text{REC} = \frac{\text{TP}}{\text{TP+FN}} \qquad \text{eq. (7)}$$

Classification accuracy is the proportion of true positives (TP) and true negatives (TN) to the total number of instances [13].

10.48047/jocaaa.2024.32.01.82

$$\mathbf{ACC} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \quad \text{eq. (8)}$$

The F1 score is the harmony mean of precision and recall. This performance metric solves defects in precision and recall in multi-class imbalanced data [13].

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad \text{eq. (9)}$$

Results and discussion

With the aid of auxiliary libraries, this study used Python 3 and Google Colab to carry out pre-processing, ML modeling, visualization, and data purification. To create training and testing sets for sentiment classification, supervised machine learning was utilized. Once features were extracted and selected from the training and testing data, a classifier model (K-Nearest Neighbor (KNN), Naïve Bayes (NB), Linear Support Vector Classifier (Linear SVC), Decision Tree (DT), Random Forest Classifier (RFC), and Logistic Regression (LR)) was fed the features. Four feature selection techniques, such as one-way ANOVA with RFE, Mutual Information (MI), Genetic Algorithm (GA), and Principal Component Analysis (PCA), with classifiers KNN, GNB, LSVC, LR, RFC, and DT, are used. Here, with one-way ANOVA, this study has used 10%, 30%, and 50% features, and with mutual information, 10%, 20%, 30%, and 50% features were used. Table 3 depicts that one-way ANOVA with 50% feature + RFE and KNN has achieved the highest precision with 0.935 compared to the remaining feature selection methods and classifiers. Here, the base condition represents that no feature selection technique was used. Fig. 5 displays this scenario. Table 4 depicts that a one-way ANOVA with 50% feature + RFE and KNN achieved the highest recall with 0.951, followed by the remaining feature selection methods and classifiers. Fig. 6 displays this condition. Table 5 depicts that one-way ANOVA with 50% feature + RFE and KNN has achieved the highest accuracy with 0.951 compared to the remaining feature selection methods and classifiers. Fig. 7 displays this condition. Table 6 depicts that one-way ANOVA with 50% feature + RFE and KNN has achieved the highest F1-score with 0.928, followed by the remaining feature selection methods and classifiers. Fig. 8 displays this scenario. Table 7 displays that one-way Anova with 10% feature + RFE has decreased 90% features, one-way Anova with 30% feature + RFE has decreased 70% features, and one-way Anova with 50% feature + RFE has reduced 50% features. Mutual information with a 10% feature has reduced the feature set to 90%; mutual information with a 20% feature has reduced the feature set to 80%. Mutual information with a 30% feature has a 70% feature set; mutual information with a 50% feature has a 50% feature set. Genetic algorithms have selected only one feature, while principal components have selected only five features.

Table 3: Precision comparison of six different classifiers using various feature selection techniques on the Amazon dataset

Feature Selection Methods	KNN (%)	GNB (%)	LSVC (%)	LR (%)	RFC (%)	DT (%)
BASE (without feature selection)	0.713	0.660	0.816	0.783	0.824	0.728
One-way ANOVA with 10% feature + RFE	0.575	0.626	0.796	0.643	0.748	0.642
One-way ANOVA with 30% feature + RFE	0.726	**0.728**	**0.824**	**0.842**	**0.824**	0.725
One-way ANOVA with 50% feature + RFE	**0.935**	0.668	0.762	0.824	0.815	0.726
Mutual Information with 10% feature	0.670	0.620	0.798	0.796	0.820	0.739
Mutual Information with 20% feature	0.638	0.640	0.797	0.793	0.824	0.739
Mutual Information with 30% feature	0.679	0.669	0.809	0.794	**0.825**	**0.740**
Mutual Information with 50% feature	0.724	0.651	0.809	0.797	0.825	0.749
Genetic Algorithm	0.664	0.024	0.575	0.575	0.575	0.575
Principal Component Analysis	0.629	0.586	0.575	0.575	0.663	0.656

10.48047/jocaaa.2024.32.01.82

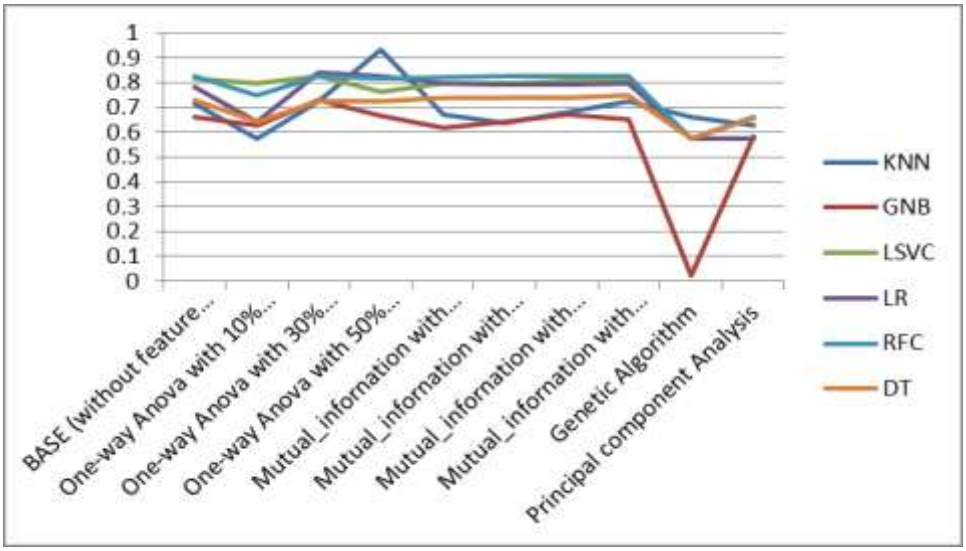


Fig. 5 A precision comparison of six different classifiers using various feature selection techniques on the Amazon dataset

Table 4: Recall the comparison of six different classifiers using various feature selection techniques on the Amazon dataset.

Feature Selection Methods	KNN (%)	GNB (%)	LSVC (%)	LR (%)	RFC (%)	DT (%)
BASE (without feature selection)	0.763	0.561	0.836	0.815	0.802	0.738
One-way ANOVA with 10% feature + RFE	0.758	**0.740**	0.818	0.734	0.865	0.801
One-way ANOVA with 30% feature + RFE	0.809	0.737	**0.846**	0.594	**0.885**	**0.809**
One-way ANOVA with 50% feature + RFE	**0.951**	0.702	0.795	0.594	0.879	**0.809**
Mutual information with 10% feature	0.758	0.175	0.824	**0.823**	0.811	0.742
Mutual information with 20% feature	0.757	0.232	0.823	0.821	0.810	0.748
Mutual information with 30% feature	0.759	0.305	0.832	0.822	0.806	0.750
Mutual information with 50% feature	0.758	0.384	0.832	0.823	0.804	0.759
Genetic Algorithm	0.760	0.155	0.758	0.758	0.758	0.758
Principal Component Analysis	0.734	0.733	0.758	0.758	0.723	0.712

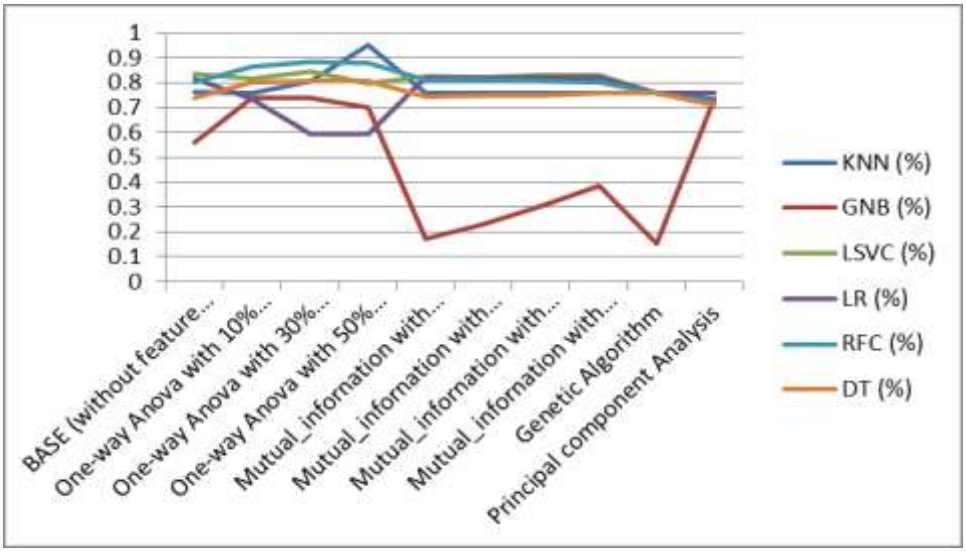


Fig. 6: Recall the comparison of six different classifiers using various feature selection techniques on the Amazon dataset.

10.48047/jocaaa.2024.32.01.82

Table 5: Accuracy comparison of six different classifiers using various feature selection techniques on the Amazon dataset

Feature Selection Methods	KNN (%)	GNB (%)	LSVC (%)	LR (%)	RFC (%)	DT (%)
BASE (without feature selection)	0.763	0.561	0.836	0.815	0.802	0.738
One-way ANOVA with 10% feature + RFE	0.758	**0.740**	0.818	0.734	0.865	0.801
One-way ANOVA with 30% feature + RFE	0.809	0.737	**0.846**	0.594	**0.885**	**0.809**
One-way ANOVA with 50% feature + RFE	**0.951**	0.702	0.795	0.594	0.879	**0.809**
Mutual information with 10% feature	0.758	0.175	0.824	**0.823**	0.811	0.742
Mutual information with 20% feature	0.757	0.232	0.823	0.821	0.810	0.748
Mutual information with 30% feature	0.759	0.305	0.832	0.822	0.806	0.750
Mutual information with 50% feature	0.758	0.384	0.832	**0.823**	0.804	0.759
Genetic Algorithm	0.760	0.155	0.758	0.758	0.758	0.758
Principal Component Analysis	0.734	0.733	0.758	0.758	0.723	0.712

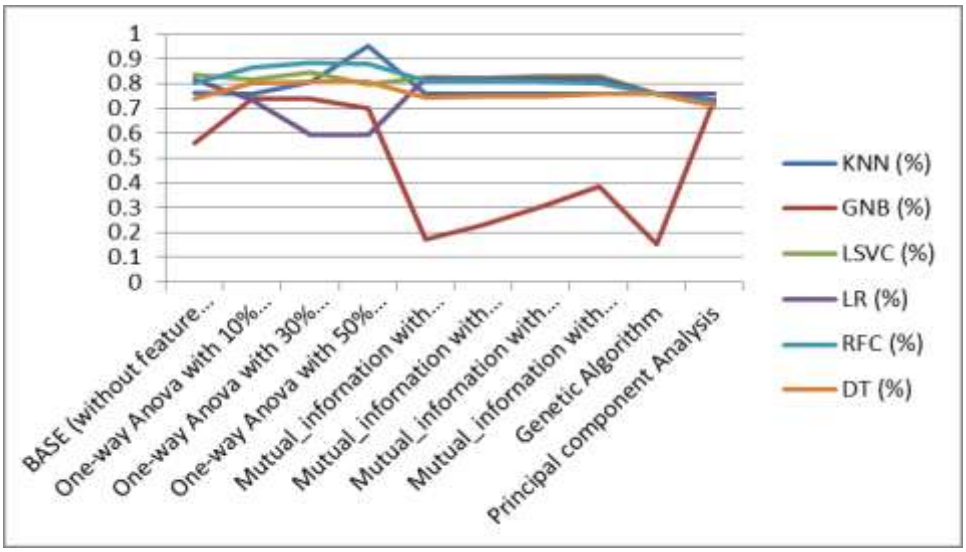


Fig.7 Accuracy comparison of six different classifiers using various feature selection techniques on the Amazon dataset

Table 6: F1-score comparison of six different classifiers using various feature selection techniques on the Amazon dataset

Feature Selection Methods	KNN (%)	GNB (%)	LSVC (%)	LR (%)	RFC (%)	DT (%)
BASE (without feature selection)	0.723	0.600	**0.820**	0.774	0.748	0.733
One-way ANOVA with 10% feature + RFE	0.654	0.657	0.775	0.649	0.802	0.713
One-way ANOVA with 30% feature + RFE	0.724	**0.729**	0.813	0.650	**0.832**	0.724
One-way ANOVA with 50% feature + RFE	**0.928**	0.680	0.744	0.650	0.823	0.724
Mutual information with 10% feature	0.661	0.193	0.804	**0.788**	0.767	0.740
Mutual information with 20% feature	0.656	0.275	0.805	0.785	0.763	0.743
Mutual information with 30% feature	0.656	0.370	0.813	0.786	0.757	0.745
Mutual information with 50% feature	0.657	0.454	0.813	**0.788**	0.751	**0.754**
Genetic Algorithm	0.672	0.042	0.654	0.654	0.654	0.654
Principal Component Analysis	0.662	0.650	0.654	0.654	0.682	0.676

10.48047/jocaaa.2024.32.01.82

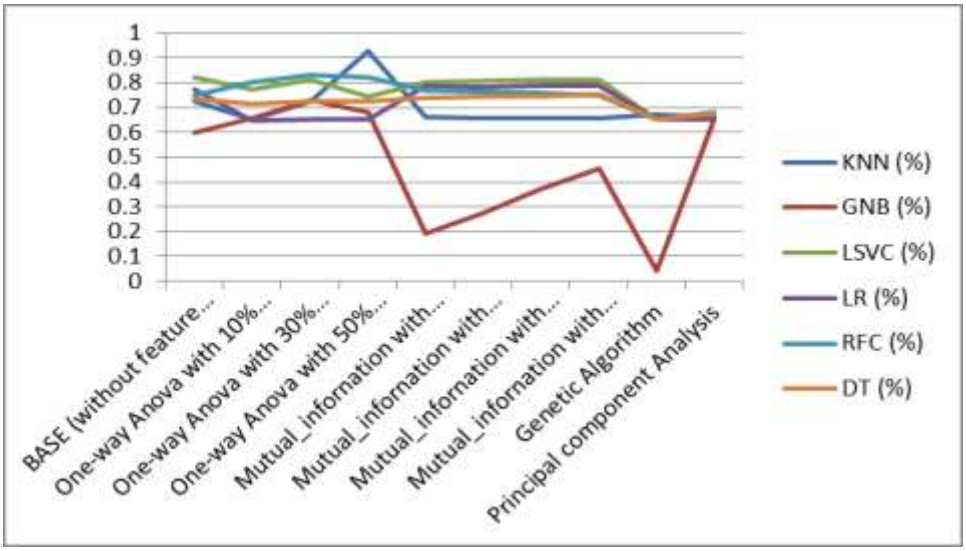


Fig. 8 F1-score comparison of six different classifiers using various feature selection techniques on the Amazon dataset

Table 7 Feature size comparison

Feature selection technique	No of features
Base Experiment	19336
One-way ANOVA with 10% feature + RFE	1933
One-way ANOVA with 30% feature + RFE	5799
One-way ANOVA with 50% feature + RFE	9665
Mutual information with 10% feature	1933
Mutual information with 20% feature	3866
Mutual information with 30% feature	5799
Mutual information with 50% feature	9665
Genetic Algorithm	1
Principal Component Analysis	5

Conclusion

Hybrid feature selection is an important step in the SA pre-processing process that has improved classification accuracy, precision, recall, and F1-score by reducing feature dimensionality and removing irrelevant and redundant features. By selecting only relevant features of the study, it also helps to overcome underfitting and overfitting problems. Additionally, FS reduces computing time, complexity, and storage.

This study illustrates that each feature selection strategy has reduced feature set comparison with the base experiment hybrid feature selection method with one-way ANOVA, and RFE has reduced the 50% feature set to obtain a higher accuracy of 0.951 with the KNN classifier, higher precision with 0.935 with the KNN classifier, higher recall with the KNN classifier, and a higher F1-score with the KNN classifier. Other classifiers also get desirable levels of accuracy, precision, recall, and F1-score, which indicate that overfitting and underfitting have been removed. The reduced feature set of 50% with higher accuracy in precision, recall, accuracy, and F1-score indicates that this study has reduced the feature set to its optimum level by removing irrelevant and redundant features. It also represents that dimensionality has been reduced to its optimum level. In the future, we can deduce the possibility of this hybrid feature selection approach from other feature selection methods, datasets, or classifiers.

10.48047/jocaa.2024.32.01.82

Reference

1. Alassaf M, Ali MQ. Improving sentiment analysis of Arabic tweets by one-way ANOVA. Journal of King Saud University Computer and Information Sciences. 2022; 34, no. 6: 2849–2859.
2. Yuce BE, Peter VN, Pawel W. The use of Taguchi, ANOVA, and GRA methods to optimize CFD analyses of ventilation performance in buildings. Building and Environment. 2022; 225: 109587.
3. Kausar MA, Sallam OF, Arockiasamy S. Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews. Engineering, Technology & Applied Science Research. 2023; 13, no. 3: 10849–10855.
4. Sangeetha JUK. Sentiment analysis of Amazon user reviews using a hybrid approach. Measurement: Sensors. 2023; 27: 100790.
5. Khare A, Amisha G, Sudhakar S, Shiv P. Sentiment Analysis and Sarcasm Detection in Indian General Election Tweets. In Research Advances in Intelligent Computing, CRC Press. 2023; pp. 253-268.
6. Chakraborty P, Farah N, Humayra AC. Sentiment analysis of Bengali Facebook data using classical and deep learning approaches. In Innovation in Electrical Power Engineering, Communication, and Computing Technology: Proceedings of the Second IEPCCT; Springer Singapore. 2022; pp. 209–218.
7. Duma RA, Zhendong N, Ally SN, Jude TK, Abdulganiyu AY. A Deep Hybrid Model for Fake Review Detection by Jointly Leveraging Review Text, Overall Ratings, and Aspect Ratings. Soft Computing. 2023; 27, no. 10: 6281–6296.
8. Khan J, Aftab A, Youngmoon L. Intelligent hybrid feature selection for textual sentiment classification: IEEE Access. 2021; 9: 140590-140608.
9. Nafis NSM, Suryanti A. An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification: IEEE Access. 2021, 9, 52177-52192.
10. Rijal S, Pandu AC, Eka MSSC, Putri HP, Andiyan A, Robbi R. Integrating Information Gain Methods for Feature Selection in Distance Education Sentiment Analysis during COVID-19, TEM Journal. 2023; 12, no. 1.
11. Khaliq A, Sarbast S, Darwesh AM. Sentiment Analysis Using Hybrid Feature Selection Techniques. UHD Journal of Science and Technology. 2020; 4, no. 1: 29–40.
12. Qaisar SM. Sentiment analysis of IMDb movie reviews using long short-term memory." In 2020, the 2nd International Conference on Computer and Information Sciences (ICCIS): IEEE. 2020; pp. 1-4.
13. Hasan G, Mamlook A, Emhamed R, Samara G, Nasayreh A et al. Arabic sentiment analysis of Monkeypox using deep neural networks and optimized hyperparameters of machine learning algorithms. Social Network Analysis and Mining. 2024; 14, no. 1: 1–18.
14. Gupta K, Nasmin J, Neda A. A Combined Approach to Sentimental Analysis Using Machine Learning Techniques. Revue d' Intelligence Artificielle. 2023; 37, no. 1 ().
15. Hilmi KM, Maple C, Watson T. Hybrid feature selection technique for the intrusion detection system. International Journal of High Performance Computing and Networking. 2019; 13, no. 2: 232–240.
16. Edi S, Rifai AP, Risnumawan A, Saputra M. A comparison of text weighting schemes on sentiment analysis of government policies: a case study of the replacement of national examinations. Multimedia Tools and Applications. 2022; 81, no. 5: 6413–6431.
17. Kwame AE, Nyame G, Qin Z, Addo PC, Gyamfi EO et al. Stock market prediction with a Gaussian naïve Bayes machine learning algorithm. Informatica. 2021; 45, no. 2.
18. Ibrahim HF, Ingosan JS, Arden NA. Oyam, Hu Y. Classifying wastes using random forests, Gaussian Naïve Bayes, support vector machines, and multilayer perceptron. In IOP Conference Series: Materials Science and Engineering: IOP Publishing. 2020, vol. 803, no. 1, p. 012017.
19. Hossein K, Moghadam BN, Abdollahi J, Branch A. Maximizing the Impact on Social Networks Using the Combination of PSO and GA Algorithms. Future Generation in Distributed Systems. 2023; 5: 1–13.
20. Muhammad DM, Khan SS, Khan M, Ghaffar MB, Khan B et al. Sentiment Analysis Based on Performance of Linear Support Vector Machine and Multinomial Naïve Bayes Using Movie Reviews with Baseline Techniques. Journal on Big Data. 2023; 5.
21. Khare A, Amisha G, Sudhakar S, Prakash S. Sentiment Analysis and Sarcasm Detection in Indian General Election Tweets. In Research Advances in Intelligent Computing: CRC Press. 2023, pp. 253-268.
22. Sharma S. Waoo AA. Customer Behavior Analysis in E-Commerce using Machine Learning Approach: A Survey. IJSRCSEIT. 2023; 9, pp.163-170.
23. Sharma S, Waoo AA. Decision Tree Machine Learning Approach for Customer Behavior Analysis on Online Product.
24. Warner E, Lee J, Hsu W, Syeda-Mahmood T, Kahn JCE et al. Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects. International Journal of Computer Vision. 2024; 1-17.
25. ACCENTURE GSL. TARGET IDENTIFICATION USING BIG DATA AND MACHINE LEARNING.INDIA. 2024; Patent 512287, filed 6 Feb 2018, and issued 19 Feb 2024.
26. Hadi M, Koblas M, Shoaraee S. SENTIMENT ANALYSIS. 2021; US 2020/0065383 A1, filed 24 Aug 2018, and issued 23 Mar 2021.

10.48047/jocaaa.2024.32.01.82

27. Zadeh AL, Tadayon S, Tadayon B. System and method for extremely efficient image and pattern recognition and artificial intelligence platform. 2024; US 11914674 B2, filed 6 Dec 2021, and issued 27 Feb 2024.
28. Adewole KS, Balogun AO, Raheem MO, Jimoh MK, Jimoh RG, et al. Hybrid feature selection framework for sentiment analysis on large corpora. Jordanian Journal of Computers and Information Technology. 2021; 7(2).
29. Andrian BW, Tobing FAT, Pane IZ, Kusnadi A. Implementation Of Naïve Bayes Algorithm In Sentiment Analysis Of Twitter Social Media Users Regarding Their Interest To Pay The Tax. International Journal of Science, Technology & Management. 2023; 4(6), 1733-1742.
30. Alsemaree O, Alam AS, Gill SS, Uhlig S. Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions. Heliyon. 2024; 10(9).
31. Parlak B, Uysal AK. A novel filter feature selection method for text classification: Extensive Feature Selector. Journal of Information Science. 2023; 49(1): 59-78. https://doi.org/10.1177/0165551521991037
32. Nguyen TT, Tran GT. Vietnamese sentiment analysis for hotel review based on overfitting training and ensemble learning. In Proceedings of the 10th International Symposium on Information and Communication Technology. 2019; pp. 147-153.
33. Dange S, Nitnaware P. A Novel Machine learning and Internet of Things (IoT) Based Dual Monitoring System for Proactive Botnet Attack Prevention. SN COMPUT. SCI.2025; 6, 708. https://doi.org/10.1007/s42979-025-04253-1
34. Swapnil Justin, Aaradhya Waoo, Akhilesh A. Waoo, Performance Analysis of Activation Functions In Neural Networks, ShodhKosh: Journal of Visual and Performing Arts, January 2024 5(1), 588–598, DOI: 10.29121/shodhkosh.v5.i1.2024.1908.

Statements and Declarations

Declaration of conflicting interest: The authors declare no conflicts of interest regarding the publication of this manuscript.

Funding: The authors did not receive any funding for this manuscript.

Ethical approval and informed consent statements: Not applicable

Data availability statement: The dataset used in this study is available on https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews/.