Libyan Journal of Medical and Applied Sciences LJMAS
Online ISSN: 3006-1113
Volume 3, Issue 3, 2025, Page No: 29-38
Website: https://ljmas.com/index.php/journal/index

User Comment Classification Using Linguistic and Sentiment-Based Features

Mona Ali Mohammed [1 *], Reem.Abdalhadi Alsunousi [2]
[1,2] Faculty of Science, Omar Al-Mukhtar University, Al Bayda, Libya
[*]*Corresponding author: mona.boshebh@omu.edu.ly*

Received: June 05, 2025 Accepted: July 10, 2025 Published: July 16, 2025

Cite this article as: M, A, Mohammed., R, A, Alsunousi. (2025). User Comment Classification Using Linguistic and Sentiment-Based Features. Libyan Journal of Medical and Applied Sciences (LJMAS). 2025;1(1):29-38.

Abstract:
The online e-commerce market is growing and becoming increasingly competitive. There are many of the data that businesses provide includes client input, such as product and service reviews. However, customer reviews have a crucial role in the business development and have been valuable sources for marketing intelligence. This paper focuses on explore the effectiveness of using feature extraction during the data preprocessing stage to enhance the performance of learning algorithms. In particular, the experiments were ongoing to investigate the impact of using feature extraction in the classification outcomes with several machine learning models. Three new columns extracted from the features were utilized with five classification algorithms during data preprocessing to classify the sentiment of Amazon reviews. The results are referring to advantages of using the feature extraction which helps making accurate models. However, Random Forest classifier achieved the best performance among other techniques across both experiments. Addition to that, for Naive Bayes classifier there is no improvement in the model accuracy.

Keywords: Feature Extraction, Classification Task, Classification Datasets, Supervised Learning Models.

تصنيف تعليقات المستخدمين باستخدام الخصائص اللغوية والعاطفية

منى علي محمد[1*]، ريم عبدالهادي السنوسي[2]
[2،1]قسم الحاسوب، كلية العلوم، جامعة عمر المختار, البيضاء، ليبيا

الملخص
يشهد سوق التجارة الإلكترونية عبر الإنترنت نموًا متزايدً .تتضمن العديد من البيانات التي تقدمها الشركات مدخلات العملاء، مثل تقييمات المنتجات والخدمات .ومع ذلك، فإن تقييمات العملاء لها تأثير قوي على تطوير الأعمال وكانت مصادر قيّمة لذكاء التسويق .تركز هذه الورقة على استكشاف فعالية استخدام استخراج الميزات كخطوة معالجة مسبقة للبيانات لتحسين أداء خوارزميات التعلم .وعلى وجه الخصوص، كانت التجارب جارية للتحقيق في تأثير استخدام استخراج الميزات في نتائج التصنيف باستخدام العديد من نماذج التعلم الآلي .تم استخدام ثلاثة أعمدة جديدة مستخرجة من الميزات لخمس خوارزميات تصنيف في مرحلة المعالجة المسبقة للبيانات للتنبؤ بتعليقات تقييمات أمازون .تشير النتائج إلى فائدة استخدام استخراج الميزات التي تساعد في إنشاء نماذج دقيقة .على اية حال، يتمتع مصنف الغابة العشوائية بأفضل أداء بين التقنيات الأخرى في التجربتين .بالإضافة إلى ذلك، بالنسبة لمصنف بايز الساذج، لا يوجد تحسن في دقة النموذج..

الكلمات المفتاحية: استخراج الميزات، مهمة التصنيف، مجموعات بيانات التصنيف، نماذج التعلم الخاضع للإشراف.

Introduction

Customer reviews are efficient in the e-commerce area so, businesses collect customer inputs (product and service reviews, for example). However, customer reviews have a crucial role in the business development and have been valuable sources for marketing intelligence[1]. It is playing a crucial role in influencing purchasing decisions. it helps the It assists customers who wish to look up product reviews before making a purchase. Moreover, it helps companies those want to observe the public's reaction to improve upon their existing service or the product they are selling. However, machine learning (ML) algorithms have become extremely useful tools for efficient analysis and classification due to the large numbers of data[1][2][3]. Sentiment analysis of these reviews presents challenges due to noisy data, subjective language, and class imbalance.

Over time, ML algorithms have become essential tools in natural language processing (NLP) tasks. one of the most popular tasks in ML is text classification. The purpose of classification is to use historical data to infer the class of future data objects. The method involves establishing a set of models that make it possible to recognize and identify different types of data [4][5]. However, Sentiment analysis and prediction are common applications

of ML algorithms [6][4]. The explosion of online shopping, especially on sites like Amazon, has produced a huge database of user-generated evaluations. Review categorization is referred to as a process of automatically assigning newly submitted textual reviews based on patterns learned by machine learning so the Amazon Customer Reviews Datasets offers a thorough resource for ML model evaluation and training[5]. As no approach is flawless for every data set, classification algorithms have been designed in the literature, to achieve the goal of creating a high-quality dataset before developing ML models. This comes in the context of providing the optimal representation of the data for the model [7][2] [8].

A feature refers to a unique measurable attribute or characteristic of a data point that is fed into a ML algorithm. Features can be numerical, categorical, or text-based, and they represent various dimensions of the data that are related to the given problem. However, it also identified as a variable or attribute [5]. However, in the pre-processing level the reliability of the dataset's features may be enhanced. Pre-processing data is crucial in order to decrease the influence of data outliers or distortion and raise the forecasting performance of the models to generalize to unseen data [3] [6] [9]. The procedure of converting raw data into features that are appropriate for ML models, is known as feature engineering. It is the process of choosing, extracting, and adjusting the most significant features from the existing data to develop more reliable and efficient ML models. One of the Feature engineering methods is a feature extraction [10][11][12]. Feature extraction is the procedure for creating new features from available ones to propose more relevant information to the ML models. This is done by modifying, combining, or summarizing existing features [13][14][15].

The study investigates the effectiveness of using feature extraction during the data preprocessing phase to enhance the performance of learning algorithms. In particular, the experiments were ongoing to investigate the impact of using Feature Extraction in the classification outcomes with several ML models.

To examine the effectiveness of using feature extraction, the accuracy of ML models evaluated after applying feature extraction as a data pre-processing step. However, common classification algorithms were used in the experiments. The classifiers used in this study include logistic regression (LR), Naïve Bayes (NB), Decision Tree (DT), Support Vector Machines (SVM) and random forest (RF).

Related works

The authors in [16] purposed to examine the use of feature extraction, NLP, and applying deep learning and curriculum learning to false news identification curriculum learning, and deep learning for the fake news recognition. The authors presented an enhanced two-phase process model built upon a combination of these strategies, The document provides a variety of results include accuracy percentages, area under the curve (AUC) percentages, and comparisons of performance metrics for different methods and datasets. The writers of [2] considered feature extraction as an essential stage in the text classification process, intending reduce error rates and enhance accuracy of classification as a crucial step of customer sentiment analysis for hotel services, and the impact of these design factors on the predictive accuracy. The researchers address the results of the study, where the DT algorithm provided optimal performance measured by accuracy, recall, and precision especially on a large feature set. The authors also outline that the SVM and NB classification algorithms are recommended in the case of small- and medium-scale feature sets due to their strong predictive accuracy and F-measure, together with user-friendly interpretation. The researchers in [11] proposed a model using deep learning models, NLP strategies, statistical descriptors, and feature extraction for detecting fake news using deep learning, feature extraction, NLP, and statistical descriptors. The provided model contains two stages: the features extraction from the text and title of news samples, and a hybrid method to classify news data. In addition, the report focuses on the importance of extracting key and useful features from the content of the datasets. The study explored the feature extraction process in-depth. It offers two new features named coherence and cohesion, as well as other key features, that were derived from news samples.

The researchers in [5] have designed a strong sentiment analysis model qualified of classifying customer feedback into three sentiment categories which are positive, negative, and neutral. The research included the collection and organization of a dataset of product reviews on Amazon. The study involved the application of NLP approaches including feature extraction and the incorporation of extra meta-data, including product classification or customer reviews. Potential integration of other meta-information, like product classification, was mentioned in the paper or consumer feedback to extract features. It reflects the consideration of diverse data sources for feature extraction in the sentiment analysis of Amazon client reviews. The BERT algorithm has provided the best performance among others, obtaining an accuracy rate of 89%. The study conducted by [17] examines a hybrid network English word segmentation processing method using BI-GRU and CRF models. The researchers discuss feature extraction in the context of deep learning and its application to multi-modal feature extraction. It proposes a multi-modal neural network with a multilayer sub-neural network for each mode to transform features from various modes to the same-modal features. These methods aim to outline the practical challenge of structural differences among several data modalities and enhance the efficiency of feature extraction. The systematic literature review by[18] intended to detect, analyze, and evaluate every research finding that is available to answer targeted research questions in the context of feature extraction from Text-based requirements for reuse in software product lines.

The authors pointed out the importance of the feature extraction process for the reuse of natural language requirements in software product lines. Furthermore, the review highlighted the necessity for more research work. Other study by [19] aimed to anticipate customer responses to the question of whether they would suggest the company to friends or family based on the analysis of transcriptions of their phone conversations with the service center. The study examined both traditional and deep learning-based text feature extraction techniques, using models trained on 20,000 transcripts and pre-trained models. It highlighted the role of domain-specific training and full transcript analysis for efficient classification of customer service interactions. Another study [20] the authors target to emphasize the value of efficient text preprocessing and feature extraction techniques in NLP, especially as part of text classification tasks and information retrieval systems. Additionally, the document discusses the influence of text preprocessing on feature extraction and selection, along with the effectiveness of TF-IDF in penalizing frequent but less valuable words within the document structure. In the research [15]. The authors showed that the objective of the Label-Sentence Bi-Attention Fusion Network (LSBAFN) model is to boost multi-label text classification by effectively capturing multi-level information and categorize content of documents. The LSBAFN model obtains this by integrating multi-level feature extraction mechanisms including local sentence and global, label-driven features. to extract multiple text features at different granularities. The research paper [14] highlights feature extraction for automated tweet classification. It emphasized the use of NLP techniques and hybrid methods to identify characteristics in tweets for classification opinion polarities and topic categories. The authors suggest adding language features in a ML process to achieve better performance, as standalone traditional NLP techniques were not sufficient to retrieve the required information. The paper also spotlights the use of decomposition of hashtags and other NLP features to feed classifiers with richer features, and the significance of reducing the risk of overfitting by eliminating features that have little or no impact on the results. Additionally, the researchers experimented with multiple combinations of NLP features and syntactic analysis to predict opinion polarities, and examined the difficulties related to annotation and feature analysis. The article [12] investigates feature extraction as a method to automated feature representation learning from big data using deep learning. It highlighted standard methods used in extracting text features such as mapping, filtration, and clustering methods.

Material and Methods

1- Supervised Machine Learning Algorithms

This study leverages five widely used supervised ML algorithms for text classification tasks: NB, LR, SVM, DT, and RF. Supervised ML operates on labeled datasets, enabling models to learn the mapping between input features and target outputs, particularly when predicting discrete class labels. These algorithms were selected due to their robustness, interpretability, and proven performance in tasks involving high-dimensional input features, particularly in text-based applications.

1.1 Naïve Bayes Classifier

Naïve Bayes (NB) is a statistical classifier that estimates the likelihood of a given class by applying Bayes' Theorem to the noticed characteristics. This approach is especially advantageous in high-dimensional data environments, such as text classification, due to its computational efficiency and interpretability [21]. Bayes' Theorem can be mathematically formulated as follows:

$$P(C_K \backslash X) = \frac{P(X \backslash C_K).P(C_K)}{P(X)}$$

Where:
$P(C_K \backslash X)$: The posterior probability of class C_K given the features X.
$P(X \backslash C_K)$: The likelihood of observing features X given class C_K.
$P(C_K)$: The prior probability of class C_K.
$P(X)$: Evidence (normalization constant) [22].
Since the evidence $P(X)$:is constant for all classes during comparison, classification is typically done using:

$$\hat{C} = \text{argmaxy } C_K [\, P(C_K). \prod_{i=1}^{n} p(x_i \backslash C_K)$$

This formulation relies on the assumption that all input features x_i are conditionally independent of one another given the class label C_K:

$$P(X \backslash C_k) = \prod_{i=1}^{n} p(x_i \backslash C_K)$$

which is the core simplifying assumption behind the "naïve" nature of the model [23]. Despite its simplicity, this assumption often yields accurate results in real-world applications.
Advantages of NB include:

- Minimal training data requirements.
- Efficiency in computation and storage.
- Effective even with irrelevant features or class imbalance, particularly in text mining [24].

Disadvantages:

- The assumption of conditional independence rarely holds true in real data.
- Poor performance when features are highly correlated [21].

1.2 Support Vector Machine

It is a type of supervised ML model essentially designed for classifying tasks, in some cases, extended to regression problems. Its fundamental mechanism involves constructing a decision boundary—known as a hyperplane—that best separates different classes by maximizing the boundary among them in the characteristic space [25].
For binary classification, the SVM optimization objective can be formulated as

$$\underset{w,b}{\text{minimize}} \frac{1}{2} \|w\|^2$$

Subject to the constraint:
$y_i(w^T x_i + b) \geq 1 \; for \; i = 1,2,3,\cdots,m$
Where:
$\boldsymbol{w}$**:** is the weight vector that defining the orientation of the decision boundary.
$\boldsymbol{b}$**:** is the bias term shifting the hyperplane from the origin.
$\boldsymbol{x}_i$: represents the feature vector of the i^{th} training instance.
$\boldsymbol{y}_i$: $\in\{-1,+1\}$ is the class label for the i^{th} instance.
$\boldsymbol{m}$**:** is the total number of training examples.
This formulation ensures that every training instance is accurately classified and positioned beyond the boundary of the defined margin. The constraint $y_i(w^T x_i + b) \geq 1$ ensures proper separation.
SVM has demonstrated outstanding performance in text classification domains, such as spam filtering and document categorization, because of its capability and robustness in managing large-dimensional data [26].

1.3 Logistic Regression (LR)

It is a commonly applied statistical approach used to address binary categorization problems. It estimates the likelihood of a discrete class result by transforming a weighted sum of the input features through a sigmoid activation function [27]
The probability is computed using the following equation:

Where:

- $\boldsymbol{w}$: is the vector of weight.

$$P(y = 1 \mid x) = \sigma(w^\top x + b) = \frac{1}{1 + e^{-(w^\top x + b)}}$$

- $\boldsymbol{x}$: is the feature vector input.
- $\boldsymbol{b}$: is a term for bias (intercept).
- $\boldsymbol{\sigma(z)}$: is the sigmoid activation function,
- $\boldsymbol{P(y = 1 \mid x)}$: indicates the likelihood that a given input is associated with the positive category.

This formulation transforms output from a linear function to a probabilistic domain between zero and one; this makes it particularly suitable for tasks involving classifications of binary. The sigmoid presents non-linearity, enabling a model to handle complex feature interactions while maintaining a linear decision boundary. Although it is a linear classifier, LR is effective, interpretable, and widely used in practical applications due to its simplicity and robustness [28].

1.4 Decision Tree (DT)

Decision Trees (DT) are predictive models that follow a hierarchical structure of decisions, where the input space is repeatedly divided into smaller, non-overlapping regions based on the values of specific features. They are widely used in both classification and regression tasks due to their simplicity, interpretability, and efficiency [29][30]. Each tree is composed of:

- Specific attributes are estimated by internal decision points in relation to threshold values.
- Each branch signifies a possible result of a decision made at a node.

– Terminal nodes provide the predicted output class or value.

The tree-building process begins at a root node and moves in a top-down manner, applying a greedy algorithm that iteratively splits the dataset into increasingly uniform subsets. The partitioning proceeds till certain termination criteria are met, such as a defined tree depth, a lower number of samples in a node, or the absence of further information gain, or when further splits yield no meaningful gain in information [31]

To choose each node's optimal feature and split point, the algorithm typically uses impurity-based criteria such as:

Gini Index:

$$\boldsymbol{Gini(t) = 1 - \sum_{i=1}^{C} p(i \mid t)^2}$$

Entropy:

$$\boldsymbol{Entropy(t) = \sum_{i=1}^{C} -p(i \mid t) \log_2 p(i \mid t)}$$

Where:

- $\boldsymbol{C}$: is the number of classes.
- $\boldsymbol{p(i \mid t)}$: is the proportion of samples of class iii at node t.

A split is chosen when it results in a maximum decrease of impurity, alternatively referred to as information gain. If no such split improves the impurity, the node is declared a leaf node[32].

Although DTs are efficient and easy to interpret, they tend to overfit the training data, particularly when allowed to grow without constraint. To address this issue, post-processing techniques like pruning or advanced ensemble methods [33] [34].

1.5 Random Forest (RF)

It is a group-based learning algorithm that builds a set of DTs and combines their predictions to increase the accuracy of regression or classification. It is known for its high performance, resistance to overfitting, and robustness to noise and data imbalance[35].

RF operates using two fundamental basics:

1. Bootstrap Sampling (Bagging): Every DT in a forest is trained using a randomly selected portion, with replacement, of the main set of data.
2. Random Feature Selection: At every node divide, a random subset of features is selected, and the best split is chosen only from this subset. These mechanisms introduce diversity among trees, which reduces the model's variance and avoids overfitting.[32]

For classification, the model prediction is based on majority voting across all trees:

$$y\hat{} = mode(h1(x), h2(x), \dots, hT(x))$$

At regression tasks, the output is computed by averaging the predictions produced by all individual trees in the forest:

$$y\hat{} = \frac{1}{T} \sum_{i=1}^{T} hi(x)$$

$\boldsymbol{hi(x)}$: is the prediction from the $\boldsymbol{i^{th}}$ tree**.**

$\boldsymbol{T}$: is the ensemble's total number of trees.

RF are capable of handling both numerical and categorical variables, scale well to high-dimensional data, and provide measures of feature importance. Due to their generalization ability, RF models are used in diverse domains such as text classification, bioinformatics, and financial modeling [36][37].

Despite their advantages, for large-scale issues, RF models can be computationally costly. Research has proposed several improvements to make them more efficient in handling large datasets[38].

2- Performance Evaluation Criteria

Assessing classification models' performance is essential for choosing the best algorithm and making sure the model operates consistently on unknown data. One of the most common tools for evaluating classification performance is the confusion matrix, which summarizes prediction results and provides the foundation for calculating key performance metrics [39] [40]. Table 1 illustrates the general structure of the confusion matrix using four key components:

A: True Positives (TP) correctly classified positive instances

B: False Positives (FP) incorrectly classified negative instances as positive

C: False Negatives (FN) incorrectly classified positive instances as negative
D: True Negatives (TN) correctly classified negative instances

Table 1. Classification models evaluation (Confusion Matrix)

Confusion Matrix		**Target**			
		Positive	Negative		
Model	Positive	A	B	Positive Predictive Value	a/(a+b)
	Negative	C	D	Negative Predictive Value	d/(c+d)
		Sensitivity	Specificity	**Accuracy**=(a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

The following metrics are obtained from the confusion matrix:

Accuracy: Indicates the percentage of cases that are accurately classified.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision (Positive Predictive Value): Measures how many of the predicted positives are truly positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall (Sensitivity): Indicates the proportion of true positives that were accurately predicted.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity: Indicates the percentage of accurately determined true negatives.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

F1 Score: The precision and recall harmonic means, especially useful in imbalanced datasets.

$$\text{F1} - \text{score} = 2. \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Particularly in domains where the cost of false positives and false negatives varies substantially, these metrics provide a fair assessment of model performance.[40] .

3- The Data set used

In this study, ML models were trained and evaluated utilizing a well-structured dataset. ML models are expected to learn patterns from labeled training data and generalize to unseen data during testing. The dataset was split into two main subsets, training data and testing data, which allows for evaluating the model's performance without bias from prior exposure [41] [31]. Factors such as data quality and the method of splitting significantly affect the reliability and effectiveness of ML algorithms [41] [42]. Validation data were additionally used to assess the model's ability to make accurate predictions before deploying it in real-world scenarios version [43][26].

The dataset employed in this research consists of approximately 3,150 Amazon Alexa product reviews, specifically focusing on Alexa Echo devices [44]. Each of the entries represents aggregate information about 6 features. The features can be summarized as follows: rating, date, variation, verified reviews, and feedback. The target variable is binary, indicating 0 for negative feedback or 1 for positive feedback. However, dividing the data set to separate training and testing portions, the training data (75% of the data set) were employed by the model to acquire knowledge, and the testing data (25% of the data set) were used by the model to estimate unobserved data, which will assess model performance.

Experiments

1- Pre-processing of the datasets

The dataset comprises diverse opinions and expressions, as different users convey their views in various ways. The data used for this study is pre-labeled, containing both negative and positive sentiments, which simplifies the analysis process. However, raw textual data with polarity often contains inconsistencies and redundant elements. Since data quality directly affects model performance, pre-processing is applied to enhance its reliability. This step includes removing duplicated terms, eliminating unnecessary punctuation, and refining the structure to improve overall data efficiency.

2- Feature Extraction

In practical applications, not every feature contributes meaningfully to identifying or classifying data instances. Including too many irrelevant or redundant features can negatively affect model accuracy and efficiency. Therefore, selecting the most informative features is essential to enhance the overall effectiveness of the feature set in machine learning tasks. Moreover, an overly large feature space can increase computational complexity, slow down the training process, and require additional memory resources, which may degrade the performance of the learning algorithm [45] [46]. In this study we utilized the generated features to develop new features as following:

Polarity: is an important concept in NLP that assists machines to understand the sentiment and emotions conveyed in human language. The value ranges from -1 to 1, with -1 indicating negative feelings and +1 representing positive feelings.

Subjectivity: is an important concept in NLP that assists machines to understand the sentiment and emotions conveyed in human language. Its value lies between 0 and 1, capturing subjective views and personal judgments.

Capital letter count column: indicates the number of capital letters in each review. Capital letters are often used by people to emphasize certain words or phrases.

3- Text representation

because ML running mathematical operations and algorithms, there are various methods to represent text data.

In this paper, the used dataset has features stored as categorical values. In order to make it valid for use as a model input categorical values should be converted into a numerical form therefore, we will employ Term Frequency-Inverse Document Frequency (TF-IDF) which is the representation of the raw text into numerical format vector or matrix representations[10]. It is typical algorithm to convert text into a meaningful representation of numbers for predictive tasks.

4- Training and classification

The experiments, have been conducted ten classification tasks. After that compared the results to demonstrate the performance of the model will change if feature extraction is used as a preprocessing step.

Experiment 1: examine the classifiers and evaluate the results using classification algorithms: NB, LR, SVM, DT and RF. In all of these, models were trained and then calculated the prediction performance. then a confusion matrix will be used to assess the estimated target value on the various models.

Experiment2: examined the classifiers and evaluate the results after applied of feature extraction as a preprocessing step. Then the confusion matrix will be used to evaluate the target values predicted by the models. The approximated target value by models will be evaluated using a confusion matrix. The block diagram of experiments is presented in Figure1.

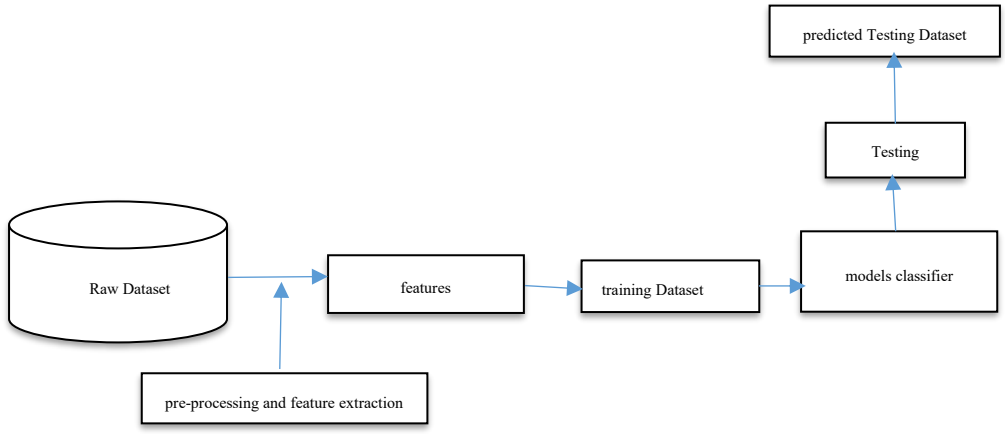

Figure1. Block diagram illustrating the conducted experiments.

Results and discussion

The experiments were carried out on ten classification tasks to identify the feedback label. However, all experimental results will be provided with a brief discussion in order to illustrate how the models perform may differ after using feature extraction in the data pre-processing stage. A summary of the models' accuracy is presented in Table 2

Table 2: Accuracy results for the evaluated models

Model	Before using Feature Extraction (%)	After using Feature Extraction (%)
Logistic Regression	90.7	91.6
Support Vector Machine	90.0	92.4
Naïve Bayes	90.7	90.8
Decision Tree	89.8	92.5
Random Forest	91.7	94.1

By comparing the accuracy of the models, we came to the conclusion that there is an improvement in the classifiers accuracy after applied feature extraction during data pre-processing step as shown in figure 6. However, the classifiers have performed the best performance while using RF.

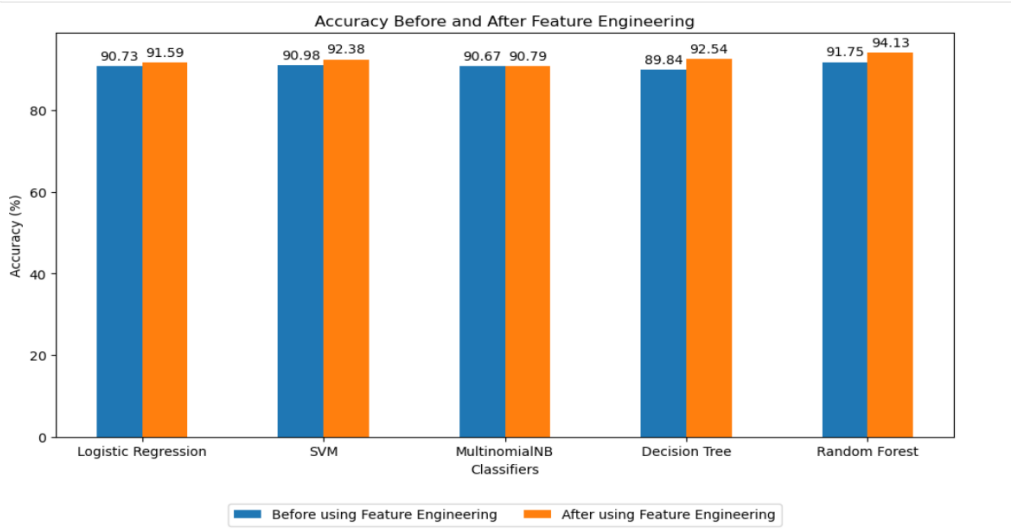

Figure2. Comparison of the models' accuracy

Conclusion

This work examined the effectiveness of using feature extraction during the data preprocessing stage regarding the accuracy of widely used supervised learning models. An extra three new columns have been extracted from the features were applied for five classification algorithms during the data pre-processing stage to predict the feedback of amazon reviews. The results are highlight the benefits of using feature extraction which helps making accurate models. However, RF classifier achieved the best performance among other techniques across both experiments. Addition to that, for NB classifier there is no improvement in the model accuracy.

Disclaimer

The article has not been previously presented or published, and is not part of a thesis project.

Conflict of Interest

There are no financial, personal, or professional conflicts of interest to declare.

References

[1] F. Huseynov and Y. C. Güler, "The Impact of Online Consumer Reviews on Online Hotel Booking Intention," *Journal of Business Research - Turk*, vol. 13, no. 3, pp. 2634–2652, Sep. 2021, doi: 10.20491/isarder.2021.1282.

[2] B. N.-A. A. Intelligence and undefined 2021, "Classification of customer reviews using machine learning algorithms," *Taylor & FrancisB NooriApplied Artificial Intelligence, 2021•Taylor & Francis*, vol. 35, no. 8, pp. 567–588, 2021, doi: 10.1080/08839514.2021.1922843.

[3] N. Sunil and F. Shirazi, "Customer Review Classification Using Machine Learning and Deep Learning Techniques," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14025 LNCS, pp. 581–597, 2023, doi: 10.1007/978-3-031-35915-6_42.

[4] A. Manhar, S. Hariramani, S. Wadhwani, and A. Manahar, "Sentimental Analysis on social media," 2022, [Online]. Available: https://www.researchgate.net/publication/362580636

[5] H. Ali, E. Hashmi, S. Yayilgan Yildirim, and S. Shaikh, "Analyzing Amazon Products Sentiment: A Comparative Study of Machine and Deep Learning, and Transformer-Based Techniques," *Electronics (Switzerland)*, vol. 13, no. 7, Apr. 2024, doi: 10.3390/electronics13071305.

[6] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst Appl*, vol. 57, pp. 117–126, Sep. 2016, doi: 10.1016/J.ESWA.2016.03.028.

[7] I. Ba'abbad, T. Althubiti, A. Alharbi, K. Alfarsi, and S. Rasheed, "A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications," *Journal of Data Analysis and Information Processing*, vol. 09, no. 03, pp. 162–174, 2021, doi: 10.4236/jdaip.2021.93011.

[8] D. Antypas, A. Ushio, J. Camacho-Collados, L. Neves, V. Silva, and F. Barbieri, "Twitter Topic Classification," Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.09824

[9] J. Szymański, "Comparative analysis of text representation methods using classification," Feb. 17, 2014. doi: 10.1080/01969722.2014.874828.

[10] M. I. Syafaah and L. Lestandy, "Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory)," 2022. [Online]. Available: https://atapdata.ai/dataset/192/HIMPUNAN_DATA_E

[11] M. Madani, H. Motameni, and H. Mohamadi, "Fake news detection using deep learning integrating feature extraction, natural language processing, and statistical descriptors," *SECURITY AND PRIVACY*, vol. 5, no. 6, Nov. 2022, doi: 10.1002/spy2.264.

[12] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," Dec. 01, 2017, *Springer International Publishing*. doi: 10.1186/s13638-017-0993-1.

[13] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The Impact of Features Extraction on the Sentiment Analysis," *Procedia Comput Sci*, vol. 152, pp. 341–348, Jan. 2019, doi: 10.1016/J.PROCS.2019.05.008.

[14] A. Stavrianou, C. Brun, T. Silander, and C. Roux, "NLP-based Feature Extraction for Automated Tweet Classification." [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[15] A. Li and L. Zhang, "Multi-Label Text Classification Based on Label-Sentence Bi-Attention Fusion Network with Multi-Level Feature Extraction," *Electronics (Switzerland)*, vol. 14, no. 1, Jan. 2025, doi: 10.3390/electronics14010185.

[16] M. Madani, H. Motameni, and R. Roshani, "Fake News Detection Using Feature Extraction, Natural Language Processing, Curriculum Learning, and Deep Learning," *Int J Inf Technol Decis Mak*, vol. 23, no. 3, pp. 1063–1098, May 2024, doi: 10.1142/S0219622023500347.

[17] D. Wang, J. Su, and H. Yu, "Feature extraction and analysis of natural language processing for deep learning english language," *IEEE Access*, vol. 8, pp. 46335–46345, 2020, doi: 10.1109/ACCESS.2020.2974101.

[18] N. H. Bakar, Z. M. Kasirun, and N. Salleh, "Feature extraction approaches from natural language requirements for reuse in software product lines: A systematic literature review," *Journal of Systems and Software*, vol. 106, pp. 132–149, Aug. 2015, doi: 10.1016/j.jss.2015.05.006.

[19] A. Kelm, P. Plebanski, and R. A. Klopotek, "Impact of Deep Learning-Based Text Feature Extraction Methods on Binary Classification Quality of Customer Service Call Transcripts," in *2024 IEEE 17th International Scientific Conference on Informatics, INFORMATICS 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 138–145. doi: 10.1109/Informatics62280.2024.10900923.

[20] A. Tabassum and R. R. Patil, "A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing," *International Research Journal of Engineering and Technology*, 2020, [Online]. Available: www.irjet.net

[21] H. Zhang, "The Optimality of Naive Bayes." [Online]. Available: www.aaai.org

[22] P.-Nin. Tan, Michael. Steinbach, and Vipin. Kumar, *Introduction to data mining*. Pearson, 2018.

[23] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers."

[24] Juan Ramos, "Using TF-IDF to determine word relevance in document queries," *First Instructional Conference on Machine Learning.*, 2003.

[25] D. M. S. Derek A. Pisner, *Support vector machine. Machine Learning.* . 2020.
[26] H. Elaidi, Y. Elhaddar, … Z. B.-… on I. S., and undefined 2018, "An idea of a clustering algorithm using support vector machines based on binary decision tree," *ieeexplore.ieee.orgH Elaidi, Y Elhaddar, Z Benabbou, H Abbar2018 International Conference on Intelligent Systems and Computer, 2018•ieeexplore.ieee.org*, Accessed: Jul. 31, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8354024/
[27] M., & J. K. Kuhn, *Applied Predictive Modeling. Springer.* 2013.
[28] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R Second Edition," 2021.
[29] M. Kuhn and K. Johnson, "Applied predictive modeling," *Applied Predictive Modeling*, pp. 1–600, Jan. 2013, doi: 10.1007/978-1-4614-6849-3/COVER.
[30] K. Priya, M. Kypa, … M. R.-2020 4th I., and undefined 2020, "A novel approach to predict diabetes by using Naive Bayes classifier," *ieeexplore.ieee.orgKL Priya, MSCR Kypa, MMS Reddy, GRM Reddy2020 4th International Conference on Trends in Electronics and, 2020•ieeexplore.ieee.org*, Accessed: Jul. 31, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9142959/
[31] T. Ma, K. Yamamori, A. T.-2020 I. 9th Global, and undefined 2020, "A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification," *ieeexplore.ieee.orgTM Ma, K Yamamori, A Thida2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), 2020•ieeexplore.ieee.org*, Accessed: Jul. 31, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9291921/
[32] T. M. D. H. G.-M. M. N. H. A. Esmaily H, "A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes," *J Res Health Sci*, no. 18(2): 412, 2018.
[33] S. R., & L. D.] Safavian, "A survey of decision tree classifier methodology,". *IEEE Transactions on Systems, Man, and Cybernetics, 21(3), DOI: 10.1109/21.97458*, pp. 660-674., 1991.
[34] L. Rokach and O. Maimon, "DATA MINING WITH DECISION TREES." [Online]. Available: http://www.worldscientific.com/series/smpai
[35] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.
[36] H., L. C., & L. Z. (2022). Wang, "A novel hybrid machine learning model for credit scoring based on random forest and logistic regression.,". *Expert Systems with Applications, 202, 117068.*, no. DOI: 10.1016/j.eswa.2022.117068, 2022.
[37] C. ZhangY. Ma, "Ensemble machine learning: Methods and applications," *Springer.*, no. DOI: 10.1007/978-3-030-10475-7, 2019.
[38] H., & M. Y. He, "Improvements of random forest algorithm for classification tasks in large datasets.," *Big Data Research, 30, 100322.,* no. DOI: 10.1016/j.bdr.2022.100322, 2022.
[39] M., & L. G. Sokolova, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45(4), no. DOI: 10.1016/j.ipm.2009.03.002, pp. 427–437.
[40] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7.
[41] Ü. Çavuşoğlu, "A new hybrid approach for intrusion detection using machine learning methods," *Applied Intelligence*, vol. 49, no. 7, pp. 2735–2761, Jul. 2019, doi: 10.1007/S10489-018-01408-X.
[42] P. Wang, Y. Zhang, W. J.-2021 I. 4th Advanced, and undefined 2021, "Application of K-Nearest neighbor (knn) algorithm for human action recognition," *ieeexplore.ieee.orgP Wang, Y Zhang, W Jiang2021 IEEE 4th Advanced Information Management, Communicates, 2021•ieeexplore.ieee.org*, Accessed: Jul. 31, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9482165/
[43] M. M. Abualhaj, A. A. Abu-Shareha, M. O. Hiari, Y. Alrabanah, M. Al-Zyoud, and M. A. Alsharaiah, "A Paradigm for DoS Attack Disclosure using Machine Learning Techniques," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, p. 2022, Accessed: Jul. 31, 2024. [Online]. Available: www.ijacsa.thesai.org
[44] A. Saha, "Alexa Reviews Dataset," https://www.kaggle.com/datasets/sid321axn/amazon-alexa-reviews.
[45] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, Aug. 2016, doi: 10.1109/TEVC.2015.2504420.
[46] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," 2022, *Frontiers Media SA*. doi: 10.3389/fbinf.2022.927312.