KSÜ Mühendislik Bilimleri Dergisi, 27(3), 2024
Araştırma Makalesi

KSU J Eng Sci, 27(3), 2024
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 16.01.2024
Kabul Tarihi :31.07.2024

Received Date : 16.01.2024
Accepted Date : 31.07.2024

CLASSIFICATION OF CUSTOMER SENTIMENTS BASED ON ONLINE REVIEWS: COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

ÇEVRİMİÇİ DEĞERLENDİRMELER ÜZERİNDEN MÜŞTERİ DUYGULARININ SINIFLANDIRILMASI: MAKİNE ÖĞRENMESİ VE DERİN ÖĞRENME ALGORİTMALARININ KARŞILAŞTIRMALI ANALİZİ

Vahid SINAP[1*] (ORCID: 0000-0002-8734-9509)

[1] Ufuk University, Department of Management Information Systems, Ankara, Türkiye

*Sorumlu Yazar / Corresponding Author: Vahid SINAP, vahidsinap@gmail.com

ABSTRACT

E-commerce's transformation of consumer behavior has increased the importance of understanding customer emotions, especially in the transition from traditional retail models to online platforms. The proliferation of online shopping has fundamentally changed not only shopping habits but also consumer interactions and purchase decisions. This research aims to compare and analyze the performance of various text mining and machine learning algorithms in the context of sentiment analysis and online review data. For this purpose, analyses were performed with a total of five supervised classification algorithms including Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, AdaBoost, and a deep learning model, CNN Model. The dataset used in the study includes customer reviews obtained from a women's clothing e-commerce platform. The missing data were completed by pre-processing the dataset. Count Vectorizer and TF-IDF vectorization were performed to transform the textual data. In addition, various text preprocessing steps were applied. According to the findings obtained from the research, AdaBoost and Naive Bayes algorithms were the most effective algorithms in terms of classifying customer sentiments. No significant difference was detected in terms of the vectorization method used. Although the CNN Model showed high performance, the generalizability of the model was considered low because overfitting was detected during the training of the model.

Keywords: Natural language processing, sentiment analysis, text mining, machine learning, deep learning

ÖZET

Geleneksel perakende modellerinden çevrimiçi platformlara geçişte e-ticaretin tüketici davranışlarını dönüştürücü etkisi müşteri duygularını anlamanın önemini artırmıştır. Bu araştırma, çeşitli metin madenciliği ve makine öğrenmesi algoritmalarının duygu analizi ve çevrimiçi değerlendirme verileri bağlamında performanslarını karşılaştırmayı amaçlamaktadır. Bu amaç doğrultusunda Lojistik Regresyon, Naive Bayes, Destek Vektör Makinesi, Rastgele Orman ve AdaBoost olmak üzere toplam beş denetimli sınıflandırma algoritması ve bir derin öğrenme modeli olan CNN Model ile analizler gerçekleştirilmiştir. Çalışmada kullanılan veri seti, bir kadın giyim e-ticaret platformundan elde edilen müşteri değerlendirmelerini içermektedir. Veri setinde ön işlemeler gerçekleştirilerek eksik veriler tamamlanmıştır. Count Vectorizer ve TF-IDF vektörizasyonları yapılarak metinsel verilerin dönüşümü sağlanmıştır. Bunlara ek olarak çeşitli metin ön işleme adımları uygulanmıştır. Araştırmadan elde edilen bulgulara göre müşteri duygularını sınıflandırma bağlamında en etkili algoritmalar AdaBoost ve Naive Bayes algoritmaları olmuştur. Kullanılan vektörizasyon yöntemi açısından önemli bir farklılık tespit edilmemiştir. CNN Model yüksek performans gösterse de modelin eğitimi sırasında aşırı öğrenme tespit edildiği için modelin genellenebilirliği düşük kabul edilmiştir.

Anahtar Kelimeler: Doğal dil işleme, duygu analizi, metin madenciliği, makine öğrenmesi, derin öğrenme

ToCite: SINAP, V., (2024). CLASSIFICATION OF CUSTOMER SENTIMENTS BASED ON ONLINE REVIEWS: COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING ALGORITHMS. *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, 27(3), 779-799.

INTRODUCTION

The widespread use of e-commerce has led to a transformation in consumer behavior. Online commerce, which replaces traditional retail shopping models, has the potential to offer consumers a wider range of products and a more comprehensive shopping experience. This transformation has an impact on many aspects from shopping habits to preferences. Consumers now prefer online platforms more than physical store visits when purchasing products and services (Shanthi & Desti, 2015), and this situation fundamentally changes the shape and dynamics of shopping. This online shopping trend enables consumers to examine product reviews on various platforms and interact with other consumers through these reviews before purchasing products (Rosário and Raimundo, 2021). In this context, it has placed a significant emphasis on the review of products on online platforms and the exchange of reviews. While consumers guide other consumers by sharing their experiences, companies have the opportunity to improve their products and services by evaluating customer reviews (Singh et al., 2017). This interaction has become a powerful factor in shaping consumers' purchasing decisions (Mariani and Borghi, 2021). This shift can be attributed to significant advances in technology and the increased connectivity and social interaction offered by modern digital environments. Effective handling of customer experience and sentiments is important for achieving long-term business success in the ever-changing digital commerce space (Kamal and Himel, 2023). Therefore, understanding the dynamics of customer sentiments in the context of online reviews plays an important role in shaping brand success and marketing strategies.

According to the diffusion of innovations theory, the introduction of new products is closely related to social dynamics (Wani and Ali, 2015). Early adopters have a significant influence in shaping mainstream markets through positive word-of-mouth publicity and fostering a socially favorable environment for the introduction of innovative products (Carrigan et al., 2011). Analyzing how this theory applies to the current environment of online reviews and consumer behavior is important for understanding the interaction between customer sentiment and purchase decisions. Online reviews provide a platform for consumers to evaluate new products and share their experiences (Singh et al., 2017). This platform serves as a tool especially for early adopters, as this group is often the first to discover and start using innovative products. By sharing their positive experiences, these users can leave a positive impact on other consumers on social networks and online platforms. This shows how positive word-of-mouth promotion, one of the basic principles of diffusion of innovations, plays a role in the online environment (Hennig-Thurau et al., 2015).

Diffusion of innovations states that the social environment plays a critical role in the process of product adoption by consumers (Chandrasekaran and Tellis, 2017). In this process, consumers, especially early adopters, learn about the product through social interactions, recommendations, and online reviews. Therefore, online reviews are an important source for understanding social interactions as part of innovation diffusion and the factors influencing the product adoption process.

As e-commerce expands with the proliferation of products and rapid technological advances, customers struggle with the challenge of navigating a diverse marketplace (Chawla and Kumar, 2022). Understanding and effectively managing consumer sentiment in this evolving digital environment is an important issue that needs to be addressed. Traditional metrics such as online review scores and volume, while informative, fail to capture the subtle nuances of customer emotions (Lian et al., 2023). Firstly, online review scores are important in terms of expressing customer satisfaction; however, these reviews are often left by extremely satisfied or dissatisfied consumers (Han and Anderson, 2020). Therefore, average online review scores may contain excessive bias. This bias becomes more pronounced for popular brands such as Apple and Samsung, where online review ratings tend to cluster at the extremes, impeding marketers from acquiring a comprehensive understanding of overall customer sentiment (Kapoor and Banerjee, 2021). Secondly, although the volume of online reviews usually indicates how much attention a product has received, the number of reviews is not directly related to the future success of a product (Zhuang et al., 2018). The number of reviews about the product can increase the popularity of a product. However, this does not mean that the product will be received positively or will sell more in the future (Racherla and Friske, 2012). Therefore, the volume of online reviews does not provide much information on customer opinions. Thus, the mentioned measure may be insufficient to ensure success in the diffusion of innovation. These limitations show that traditional measures are incomplete in capturing all aspects of customer sentiment. To overcome these limitations, advanced analytical approaches including text mining, natural language processing (NLP) and machine learning (ML) are emerging as powerful tools to unravel the complexity of consumer sentiments (Hartmann and Netzer, 2023). However, the application of these advanced techniques presents its own challenges and complexities that require critical scrutiny

to guarantee a comprehensive understanding of the complex interplay between consumer sentiments and e-commerce dynamics.

To effectively apply techniques such as text mining, NLP, and ML, large and qualified datasets are needed. There should be enough in-depth reviews on a particular product or service category (Zhang et al., 2019). In addition, to successfully apply these techniques, people who are skilled in data science, analytics, and programming, which is a field that requires expertise, are needed. Besides, techniques such as text mining and NLP need to be appropriately adapted to the complexity and diversity of language to fully recognize sentimental nuances (Mohammad, 2016). Different forms of expression, wordplay, and emotional tones across languages are important factors that need to be considered for these techniques to work correctly.

Text mining, NLP, and ML also have certain differences and limitations of use. Text mining is considered effective in analyzing large text data. It also has the ability to identify patterns and extract basic statistical information. However, it may be limited in complex tasks such as sentiment analysis and may have difficulty in fully understanding emotional nuances and subtexts of language (Qaiser and Ali, 2018). NLP is designed to understand more complex aspects of language. It has the capabilities of inferring meaning, understanding emotional tones, and analyzing word context (Guo, 2022). However, it may require more training and data to adapt to the complexity of language. Performance may be degraded, especially when multiple languages are used. Although NLP is more advanced than text mining, it is sometimes insufficient to fully understand and decode complex emotional content. ML is a powerful tool for analyzing large datasets and learning patterns (Alexopoulou et al., 2017). However, it needs a large and representative dataset to obtain accurate results. It can also be susceptible to problems such as overlearning or lack of data training. All three methods come with their own advantages and limitations. Therefore, choosing the most appropriate one in the context of an application depends on the complexity of the subject matter and the desired results.

The main objective of this research is to compare and analyze the performance of various text mining and ML algorithms. By identifying the different advantages, disadvantages, and limitations of these algorithms, the research aims to evaluate their effectiveness, especially on sentiment analysis and online review data. In line with the results obtained, it is aimed to determine the best performing method and to understand the effectiveness of these algorithms in specific usage scenarios. In addition, the research is expected to be a guide to better understand the competition between algorithms used in sentiment analysis and similar tasks and to select the most appropriate algorithm in the context of the application.

Related Work

In the field of text mining and sentiment analysis, there are important studies that deal with various aspects of the subject. These studies are categorized into four main areas: fundamental sentiment analysis methods including aspect-based sentiment analysis and supervised learning approaches, advanced NLP and deep learning models, comparative analyses and techniques such as sentiment analysis on movie reviews and comparative studies on sentiment analysis, and methods for handling big data and specialized datasets including social media sentiment analysis and ML techniques for online customer reviews.

Various studies have explored fundamental sentiment analysis methods across different domains. Li et al. (2023) employed aspect-based sentiment analysis (ABSA) to predict restaurant survival, emphasizing its use in categorizing reviews by specific aspects like location and service. Hossain and Rahman (2023) utilized ML techniques, including AFINN and VADER algorithms, to classify sentiment in insurance reviews, highlighting logistic regression's effectiveness. Zhang et al. (2021) proposed a hybrid approach integrating word embedding and dependency parsing for hotel review sentiment analysis. In an evaluation, the proposed hybrid approach was found to be superior to individual techniques in terms of sentiment classification performance. Rain (2013) assessed sentiment limitations using Likert scales in Amazon product reviews. According to the results of the study, it was found that this type of evaluation tool allows limited expression of sentiments, leads to inadequate expression of language and cultural differences, and the review may go beyond the characteristics of the product. Hu et al. (2010) applied SentiWordNet for sentiment analysis of product features. Angulakshmi and ManickaChezian (2014) outlined opinion mining tools for sentiment classification.

In NLP and deep learning models for sentiment analysis, Patel et al. (2023) explored the Airline reviews dataset using ML algorithms like Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT), with Google's

BERT demonstrating superior performance across metrics. Barik et al. (2023) integrated LSTM with grey wolf optimization (DGWO) and BERT for word embedding, achieving high accuracy in product review sentiment analysis. Zhao et al. (2021) introduced LSIBA-ENN, an optimized ML algorithm, showing enhanced recall rates in sentiment analysis. Poomka et al. (2021) investigated customer satisfaction through sentiment analysis of Amazon book reviews, highlighting LSTM and gated iterative units' effectiveness. Pradhan et al. (2016) conducted a comprehensive study on supervised learning approaches, emphasizing sensitivity and recall metrics. Alantari et al. (2022) compared methods for linking consumer ratings with text-based reviews, highlighting pre-trained neural networks' predictive accuracy over topic models. Demirbilek and Demirbilek (2023) conducted a sentiment analysis on Google comments about a state university in Central Anatolia. They used machine learning methods like Logistic Regression (LR), Gaussian Naive Bayes, SVM, and CatBoost, as well as LSTM and AWS Comprehend. All methods achieved over 80% success, with AWS Comprehend excelling in all metrics except sensitivity, demonstrating the effectiveness of these techniques for sentiment analysis.

In comparative analyses and techniques, Tran et al. (2022) performed sentiment analysis on movie reviews using RF, NB, SVM, blending, voting, and RNN techniques, proposing frameworks that outperform Stanford CoreNLP in predictive accuracy, particularly with voting and RNN models. Dey et al. (2020) compared SVM and NB for sentiment analysis of Amazon product reviews, with SVM demonstrating consistent performance. Ramadhan et al.'s (2023) study compared NB and SVM classifiers for sentiment analysis on wireless headphone reviews from Tokopedia, highlighting the NB classifier's superior accuracy, recall, and F-measure, and the SVM classifier's high precision.

Advancements in big data analytics have facilitated innovative approaches to sentiment analysis. Researchers like Biradar et al. (2022) and Obiedat et al. (2022) have developed methods leveraging advanced algorithms to extract insights from real-time social media and review datasets. Biradar et al. (2022) focused on developing big data technologies to process real-time social media data for sentiment analysis, emphasizing dataset preprocessing, domain-specific clustering, and feature extraction using n-gram models and TF-IDF vectors. Their approach integrates unsupervised clustering for domain-specific insights and supervised ML for efficient Twitter data processing, achieving a reported 80% accuracy and enhancing computational efficiency compared to traditional methods. Obiedat et al. (2022) introduced a hybrid method combining SVM with Particle Swarm Optimization (PSO) and various oversampling techniques to address data imbalance in sentiment analysis of restaurant reviews from Jeeran, a platform featuring Arabic reviews. The study highlights the PSO-SVM approach's effectiveness in optimizing feature weights and improving classification accuracy, F-score, G-mean, and Area Under the Curve (AUC) across different dataset versions, underscoring its superiority over alternative classification methods. These studies exemplify advancements in applying advanced algorithms to big data for enhancing sentiment analysis and data-driven decision-making processes.

This study offers a different perspective from previous studies by evaluating the performance of sentiment analysis and classification algorithms in the e-commerce domain in depth. In particular, the focus on customer reviews obtained from women's clothing e-commerce platforms is a unique feature of the study. Unlike sentiment analysis studies in the existing literature, which are usually conducted on general product categories, this research targets a specific niche and addresses industry-specific challenges and opportunities. Moreover, the effective completion of missing values and the methodologically innovative implementation of text preprocessing steps enhance the capacity of this study to produce well-founded conclusions. In addition, the research evaluates in detail the performance of various classification algorithms in sentiment analysis as well as the deep learning model CNN Model and demonstrates how effective these models are in e-commerce customer reviews. Besides, the problems inherent to ML techniques that occur during the procedure followed in building the model are reported in the research, and suggestions are made to solve the problems instead of being solved in the background. The unique contribution of this study is to guide e-commerce companies to understand customer feedback more effectively and successfully integrate this understanding into their strategic decision-making processes. Therefore, it constitutes an important resource for decision makers, marketers, and researchers in the industry.

MATERIAL AND METHOD

In this section, explanations of the ML algorithms used in the research, performance criteria used in the comparison of algorithms, features of the dataset, and information about the data preparation process are given.

Algorithms Utilized

ML is a sub-branch of artificial intelligence and involves computers making intelligent decisions by learning from data. ML is divided into various methods according to the nature of the data and the purposes of the analysis. In this context, three main categories stand out: (1) Supervised Learning, (2) Unsupervised Learning, and (3) Reinforcement Learning (Sarker, 2021). Supervised learning is an approach in which an ML model tries to understand the relationship between a given input and a target output. This approach is divided into two main subcategories: classification and regression. Classification is used to determine whether an input belongs to a certain category or not. Classification algorithms learn patterns in the dataset and use these patterns to classify new and unknown data into specific categories. In the context of customer sentiment analysis, supervised classification algorithms play an important role in analyzing customer reviews quickly and effectively. These algorithms can analyze customer reviews and categorize them into positive and negative sentiments. The algorithms evaluate customer expressions based on certain features and classify sentiment based on the learned patterns. The main advantages of supervised classification algorithms are fast execution, the ability to process large datasets, and their learning capabilities. However, they also have some challenges, such as the need for correctly labeled datasets and the need to be trained to perform well in a given context (Zhou et al., 2017).

In this research, a total of five supervised classification algorithms, namely LR, NB, SVM, RF, and AdaBoost, were used to examine customers' reviews and perform sentiment classification. In addition, Convolutional Neural Networks (CNN) Model, which is a deep learning (DL) model, was also used in the analysis.

Logistic Regression

LR is a statistical modeling technique used to estimate the probability that a dependent variable belongs to one of two categories. It is especially used in classification problems. It is named after the logit function in mathematics (Das, 2021). Basically, LR is an extended form of linear regression. However, the output of LR is probability values, which are converted into log odds ratio using a logit transformation. This process is used to express probabilities between 0 and 1. The formula for LR is given in Equation 1.

$$P(Y=1) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k)}} \quad (1)$$

When the formula in Equation 1 is analyzed:

- $P(Y=1)$ is the probability that the dependent variable belongs to category 1.
- e is the Euler number.
- b_0, b_1, b_2, ... b_k are the learned coefficients of the model.
- X_1, X_2, ... X_k are the independent variables.

Naive Bayes

NB is a probabilistic ML algorithm used especially in text classification problems. It is based on Bayes' Theorem and takes its name from the "naive" (pure, simple) assumption in this theorem. The assumption suggests that each feature in the model is independent of the others (Berrar, 2018). The formula of NB is given in Equation 2.

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)} \quad (2)$$

In the formula in Equation 2:

- $P(C \mid X)$ denotes the probability of belonging to a given class (C),
- $P(X \mid C)$, the probability of observed traits (X) belonging to class (C),
- $P(C)$ denotes the probability of belonging to class (C) in general.
- $P(X)$ is the overall probability of the observed features.

Support Vector Machine

SVM is an algorithm used in ML for classification and regression problems. Basically, it works by creating a hyperplane to separate data points into two classes. In creating this hyperplane, SVM focuses on maximizing the margin between classes (Pisner and Schnyer, 2020). SVM is successful with linearly separable datasets but can be

extended for non-linearly separable datasets by using kernel functions. Kernel functions provide linear separability by transforming the data into another space (Cervantes et al., 2020). In this way, SVM can also be used for multi-class problems and regression problems. The SVM decision function is given in Equation 3.

$$f(x) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) \tag{3}$$

Looking at Equation 3, $f(x)$ is the function representing the classification decision. **w** is the weight vector, important parameters that determine the hyperplane. **x** is the weight vector, important parameters that determine the hyperplane. b is the bias term. If $f(x)$ is positive or negative, it is assigned to a class. This decision function is used to classify the data points, allowing to evaluate the performance of the learned model. The SVM learning process involves optimizing this decision function and finding the optimal hyperplane.

Random Forest

RF is an ensemble-based learning model. It is a powerful tool for classification and regression tasks. The algorithm builds a more generalizable model by assembling a set of decision trees. Each decision tree independently performs classification or regression based on features. These trees are usually deeply split and focused on different subgroups. RF combines the predictions of these trees into a more reliable and higher performing model (Dogru and Subasi, 2018). The "randomness" feature of the algorithm involves training each tree on a different subset, which increases its generalization ability and provides resistance to overfitting (Sylvester et al., 2018). Although it does not have a specific formulaic equation, it is a model in which predictions are generated by a combination of partitions and decision mechanisms within each tree.

AdaBoost

AdaBoost is an ensemble learning algorithm that combines weak learners to form a strong learner. It can be used in classification or regression tasks and is specifically designed to improve performance in successive iterative stages. In each iteration, a new weak learner is added, focusing on examples that were misclassified in previous stages of the model. In this way, the weight of the incorrectly predicted examples is increased, and the model adapts to better classify these examples (Li et al., 2008). The basic formula of AdaBoost includes a combination where the weight of each weak learner is related to the prediction accuracy. This relationship is expressed using the weighted error rate. AdaBoost is combined with simple models such as decision trees and applies an adaptive weighting strategy to increase the power of each learner. The mathematical formula of the algorithm is expressed in Equation 4.

$$F(x) = \sum_{t=1}^{T} \alpha_t f_t(x) \tag{4}$$

In this formula:

- $F(x)$ is the total learners,
- T denotes each weak learner iteration of AdaBoost.
- α_t is the coefficient assigned weight to the t-th weakest learner.
- $f_t(x)$ is the prediction function of the t-th weakest learner.

In each iteration, the AdaBoost algorithm gives weight to data points, taking into account the errors of previous iterations. It emphasizes previously misclassified examples more and trains the next weaker learner by combining these weighted learners together. The coefficients (α_t) determine the importance of each weak learner and give more weight to correct classification.

Deep Learning based CNN Model

DL is a sub-branch of ML used to solve complex problems and automate learning processes. DL models are usually multilayer neural networks and therefore have a "sequential" structure. CNN Model is a structure that combines the layers of a neural network in a sequential manner (Zhou and Troyanskaya, 2015). The CNN Model can include different types of layers such as convolutional layers, fully connected layers, and activation layers. The characteristic of this model is that the layers are connected sequentially and these connections are optimized in a learning process (Li et al., 2018). DL is often applied in areas such as visual and NLP, as it can learn complex patterns in large datasets.

Performance Metrics

Evaluating the effectiveness of ML models is a critical element for their successful design and implementation. The evaluation of models is important to understand how well the algorithms are compatible with real-world data. In this context, performance metrics are used to evaluate the accuracy, precision, sensitivity, and overall performance of the model. Metrics derived from basic measures such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) help to understand the classification capabilities and overall performance of the model (Carvalho et al., 2019). TP represents the number of times the model correctly identifies a positive situation. TN represents the number of times the model correctly identifies a negative case. FP represents the number of times the model incorrectly identifies a negative state as positive. FN refers to the number of times the model incorrectly identifies a positive state as negative. Furthermore, tools such as AUC scores, confusion matrices, and F1-Score combine different metrics to provide a more comprehensive assessment. The correct interpretation of these metrics is critical for understanding the reliability and overall effectiveness of the model.

Accuracy: It refers to the proportion of instances correctly predicted by a classification model. It is calculated by dividing the number of correctly predicted samples by the total number of samples. Its formula is given in Equation 5.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

Precision: Measures the probability that samples predicted as positive are actually positive. It focuses on reducing FP predictions. Its formula is given in Equation 6.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall: Measures the proportion of truly positive samples that are correctly predicted. It focuses on reducing false negative predictions. Its formula is given in Equation 7.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

F1-Score: Provides a balanced performance measure by combining precision and recall metrics. It uses the harmonic mean between the two. Its formula is given in Equation 8.

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \tag{8}$$

Receiver Operating Characteristic Curve (ROC): An important tool used to evaluate the performance of classification models. The curve shows the relationship between two key metrics, sensitivity (True Positive Rate) and specificity (True Negative Rate). Sensitivity represents the true positive rate, i.e. the proportion of true positives out of total positive cases, while specificity represents the true negative rate, i.e. the proportion of true negatives out of total negative cases (Carter et al., 2016). The curve is plotted using the metrics False Positive Rate (FPR), shown in Equation 9 on the horizontal axis, and True Positive Rate (TPR), formulated in Equation 10, on the vertical axis. In an ideal model, the ROC curve passes through the upper left corner. In this case, the model provides high sensitivity and high specificity.

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

AUC (Area Under the Curve): Refers to the area under the ROC curve. For an ideal classifier, the AUC will approach 1, while for a randomly estimated model, the AUC will be 0.5. AUC is the model's ability to discriminate between the positive and negative classes.

Dataset

The dataset used in the research consists of customer reviews from a women's clothing e-commerce platform (Brooks, 2018). The dataset consists of 23.486 records and 10 columns. Each row represents one customer review. Each entry contains a written comment and additional customer information. The data is anonymized to protect confidentiality. Company references in the review text and body have been replaced with the term "retailer". The features present in the dataset and their descriptions are given below.

- Clothing ID: Integer categorical variable that refers to the specific product being reviewed.
- Age: Age of the reviewer, positive integer variable.
- Title: Text variable containing the title of the review.
- Review Text: Text variable containing the text of the review.
- Rating: Product rating given by the customer, positive ordinal integer variable ranging from 1 worst to 5 bests.
- Recommended IND: Binary variable indicating whether the customer recommends the product; 1 recommended, 0 not recommended.
- Positive Feedback Count: Positive integer indicating the number of times other customers found this review positive.
- Division Name: Categorical name indicating the general division of the product.
- Department Name: Categorical name indicating the department name of the product.
- Class Name: Categorical name indicating the class name of the product.

There are missing values in the dataset. There are 3.810 missing values in the "Title" column, 845 in the "Review Text" column, and 14 missing values in each of the "Division Name", "Department Name", "Class Name" columns (Figure 1). The presence of missing values can make it difficult to fully utilize the dataset and affect the analysis results. In addition, it can be more difficult to detect missing values in textual features, especially if an empty string is entered as a textual expression. In this case, strategies for dealing with missing values should be identified and their potential impact on the analysis should be evaluated.

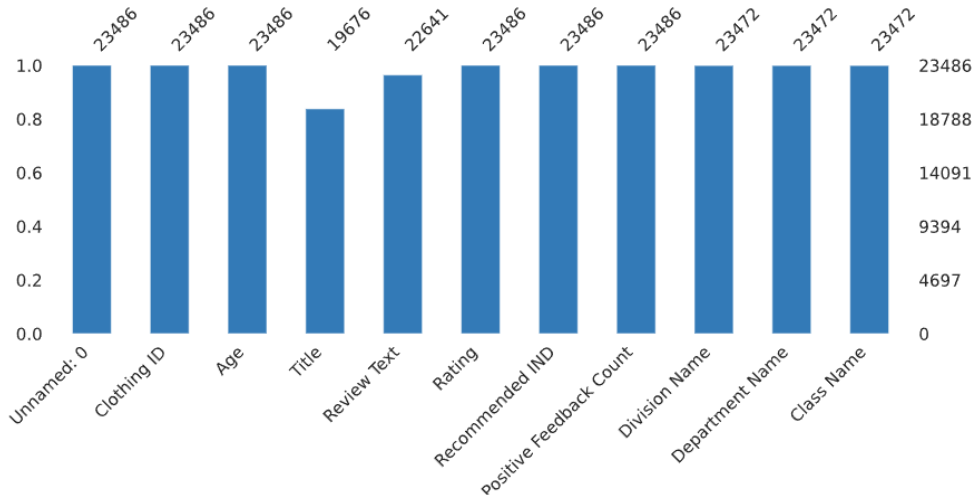

Figure 1. Missing Values

Data Preparation

Data preprocessing steps were performed to eliminate the deficiencies in the dataset and to obtain more reliable results in the analysis. First, missing 'Title' values were changed to 'No Title' to ensure consistency in the analysis and to avoid problems caused by missing values. This ensures that the missing values in the 'Title' column are filled with a meaningful value. Subsequently, rows with missing 'Review Text' values were removed to ensure that the underlying data used in text-based analyses is complete. To increase the accuracy of the analysis and add credibility to the results, a dataset free of missing 'Review Text' values was obtained. Finally, duplicate values were deleted to remove unnecessary repetitions in the dataset and reduce inconsistencies in the analysis. These processes contribute to making the dataset more organized and homogeneous.

Text Vectorization

ML algorithms take numerical feature vectors as input. Therefore, when working with textual data, it is necessary to convert each data into a numeric vector (Weiss et al., 2010). This process is referred to as “vectorization” in the literature (Jararweh et al., 2019). Two different vectorization methods were used in this study. The first one is the "Count Vectorizer" vectorization approach where each text is represented as a vector of word counts.

Count Vectorizer is a tool for converting text data into a matrix of token counts, a numerical representation. In the research, the text collection was converted into a predefined set of tokens (words), and a matrix was obtained for each token, showing how many times it occurs in a given document (Turki and Roy, 2022). According to the Count Vectorizer's logic, each token forms a column, and each document (text sample) forms a row. Each cell of the matrix indicates the number of times a particular word occurs in a given document. This allows text data to be represented in a numerical format.

The second text vectorization method used in the research is called TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is a measure used to determine the importance of a word in a document (Aizawa, 2003). This measure combines the frequency of a word in a document and the overall frequency of that word in all documents. Term Frequency (TF) refers to the frequency of a word in a document. That is, it shows how many times a word occurs in a document. However, it only measures the absolute frequency of the word in the document and therefore this value does not take into account the relative importance of a word within a document (Bafna et al., 2016).

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \tag{11}$$

Inverse Document Frequency (IDF) is a measure of the overall importance of a word. If a word is rare in general, its IDF value will be high. This means that a particular word is less common in the overall collection and therefore has more weight (Robertson, 2004).

$$IDF(t,D) = \log\left(\frac{Total\ number\ of\ documents\ in\ the\ corpus}{Number\ of\ documents\ containing\ term\ t\ +\ 1}\right) + 1 \tag{12}$$

TF-IDF vectorization creates a vector containing the TF-IDF scores of all words in a document. This provides a numerical representation of the importance of each word in the document, which can then be given as input to ML algorithms.

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D) \tag{13}$$

Text Preprocessing

Text is the most unstructured form of data available, so it contains various types of noise. This means that the data cannot be directly analyzed without any preprocessing. The process of cleaning and standardizing text, removing noise, and making it ready for analysis is called "text preprocessing" (Denny and Spirling, 2018). The following steps were followed for text preprocessing in the research.

Tokenization: Tokenization is recognized as an important pre-processing step in text mining and NLP. In this step, textual data is broken down into smaller and more meaningful chunks. It often plays an important role in

understanding the structural properties of language, as sentences are the basic structural units that often carry meaning. Brackets such as punctuation marks, spaces, or special characters are used to separate the text into sentences and words. The words and terms identified in each sentence are treated as tokens (Vijayarani and Janani, 2016). That is, each word or term in the text is referred to as a token and analyzed using these tokens.

Noise Reduction: This step was performed by identifying any piece of text in the data that is incompatible with the context and the final output. Such noise elements include language stop words, i.e. words that are commonly used in a language but do not contribute to the meaning of the text (is, am, the, of, in, etc.). Furthermore, this step is concerned with removing all types of noise in the text, including URLs or links, capitalization, punctuation, and industry-specific words (Xiong et al., 2006). In this way, noise elements within the text are reduced and the text is more suitable for analysis.

Lexical Normalization: The lexical normalization step was carried out to clean up the multiple representations of a single word, in particular word variations such as "run", "runner", "ran", "runs" and "running". Such variations mean semantically different things but are similar in context (Han and Baldwin, 2011). This step performs the function of transforming the differences of each word into a normalizing form, also known as a "lemma". Lemmatization is one of two methods of lexical normalization, stemming or lemmatization (Badaro et al., 2014). For this case, lemmatization is preferable because lemmatization performs a stemming operation by returning the stem form of each word, which is not only limited to removing suffixes but also performs stemming.

These steps are carried out to better understand the text and make it coherent and suitable for analysis. Tokenization breaks the text into more manageable chunks; Noise Reduction removes redundant information for analysis; Lexical Normalization regularizes word variations. These steps improve accuracy and reliability in the analysis process while making the text ready for analysis.

Model Setups

In this study, a total of five supervised classification algorithms (LR, NB, SVM, RF, AdaBoost) and a DL model (CNN) were used to perform sentiment analysis on customer reviews. While creating the models, the dataset was divided into two parts, 80% for training and 20% for testing. Since text data is usually high dimensional and irregular, large datasets are recommended for text mining and NLP applications. In the analyses, 10-fold cross-validation was employed. The reason for using the 10-fold cross-validation method is that the dataset used in the research is considered a relatively small dataset, especially for text mining applications. By employing 10-fold cross-validation in the analyses, it is ensured that the model is trained and evaluated on different subsets of the data, providing a more robust estimation of the model's performance. This method allows the model to learn from a variety of data splits and helps in assessing its generalization ability more effectively.

In all algorithms used, the random state was set to 42. Hyperparameter tuning was performed using Grid Search to optimize the performance of each model. In the LR algorithm, the parameter C was optimized, and the best value found was 0.1. C is known as the regularization term and controls the amount of regularization; smaller values of C mean more regularization. The max_iter parameter was set to 1000, which means that the optimizer algorithm will perform a maximum of 1000 iterations. The class_weight parameter was set to "balanced", so that automatic weights are used to address the imbalance between classes. For the RF algorithm, Grid Search determined that a total of 140 decision trees (n_estimators=140) and a maximum depth of 15 (max_depth=15) yielded the best performance. The n_jobs parameter is set to -1 so that all processor cores are used for training the trees. The class_weight parameter is set to "balanced" so that automatic weights are used to handle imbalance between classes. The kernel used in the SVM was optimized to be the Radial Basis Function (RBF), and the C parameter was set to 1.5. In SVM optimization, the C parameter indicates the extent to which misclassification of each training sample is avoided. For the AdaBoost algorithm, Grid Search found that using 420 weak learners and a learning rate of 0.05 provided the best results. For the Naive Bayes algorithm, the var_smoothing parameter was optimized to 1e-9. The specific hyperparameters for each algorithm are summarized in Table 1.

Table 1. Optimized Hyperparameters for Each Algorithm

Algorithm	Parameter	Value
LR	C	0.1
	max_iter	1000
	class_weight	balanced
RF	n_estimators	140
	max_depth	15
	n_jobs	-1
	class_weight	balanced
SVM	kernel	RBF
	C	1.5
AdaBoost	n_estimators	420
	learning_rate	0.05
NB	alpha	0.5
	var_smoothing	1e-9

The study was developed in Python programming language and implemented using Jupyter Notebook IDE. For text mining and sentiment analysis processes, NLTK (Natural Language Toolkit) and Scikit-learn libraries were used. In addition, Matplotlib and Seaborn libraries were preferred for data visualization and analysis. Pandas and NumPy libraries were used for data manipulation and analysis.

EXPERIMENTAL STUDY AND FINDINGS

The findings are presented in three sections. In the first section, confusion matrices, ROC curves and AUC scores of five supervised classification algorithms, LR, NB, SVM, RF, AdaBoost, are reported. In the second section, AUC and Validation AUC Score and Loss and Validation Loss Score graphs of CNN Model are shown. In the third section, the performance of the models created in the research is compared.

Evaluation of Supervised Classification Models

The models were first analyzed using confusion matrices. The results obtained through the confusion matrices reflect the ability of each model to accurately classify customer sentiments. When Figure 2 and Figure 3 are examined, it is seen that the NB model stands out with high TN and TP values in both text vectorization methods, but the FP value is lower in the Count Vectorizer method. This shows that NB is successful in accurately distinguishing between negative and positive emotions, especially when the Count Vectorizer method is used. The other models obtained similar results in both vectorization methods. In addition to NB, LR and SVM models also provided high accuracy. However, the FP value of the AdaBoost model was found to be slightly high. This indicates that the model tends to classify negative comments as positive. In other words, the AdaBoost model indicates that it may tend to misinterpret sentiments in situations compared to other models. However, confusion matrices alone are not sufficient to make such an inference.

ROC curve analysis was performed to better understand the performance of the models. The AUC scores of each model were used to measure the classification abilities of the models. These measurements were performed and reported separately for Count Vectorizer and TF-IDF methods. ROC curves visualize the balance between sensitivity and specificity to assess the classification ability of the models. AUC score evaluates the overall classification performance of the model by measuring the area under the ROC curve.

When Figure 4 and Figure 5 are analyzed, LR and SVC models stand out as the highest performing models with an AUC score of 0.93 when the TF-IDF method is used. When Count Vectorizer is used, LR's AUC score decreases to 0.92, while RF's remains the same. This shows that the models have a high ability to distinguish between positive and negative classes and that there is no significant difference in terms of the vectorization method used. RF and NB models also showed strong performance with AUC scores of 0.90 and 0.91 respectively. The ROC curves of these models indicate a balanced performance between sensitivity and specificity. The AdaBoost model performed slightly lower than the other models with an AUC score of 0.89.

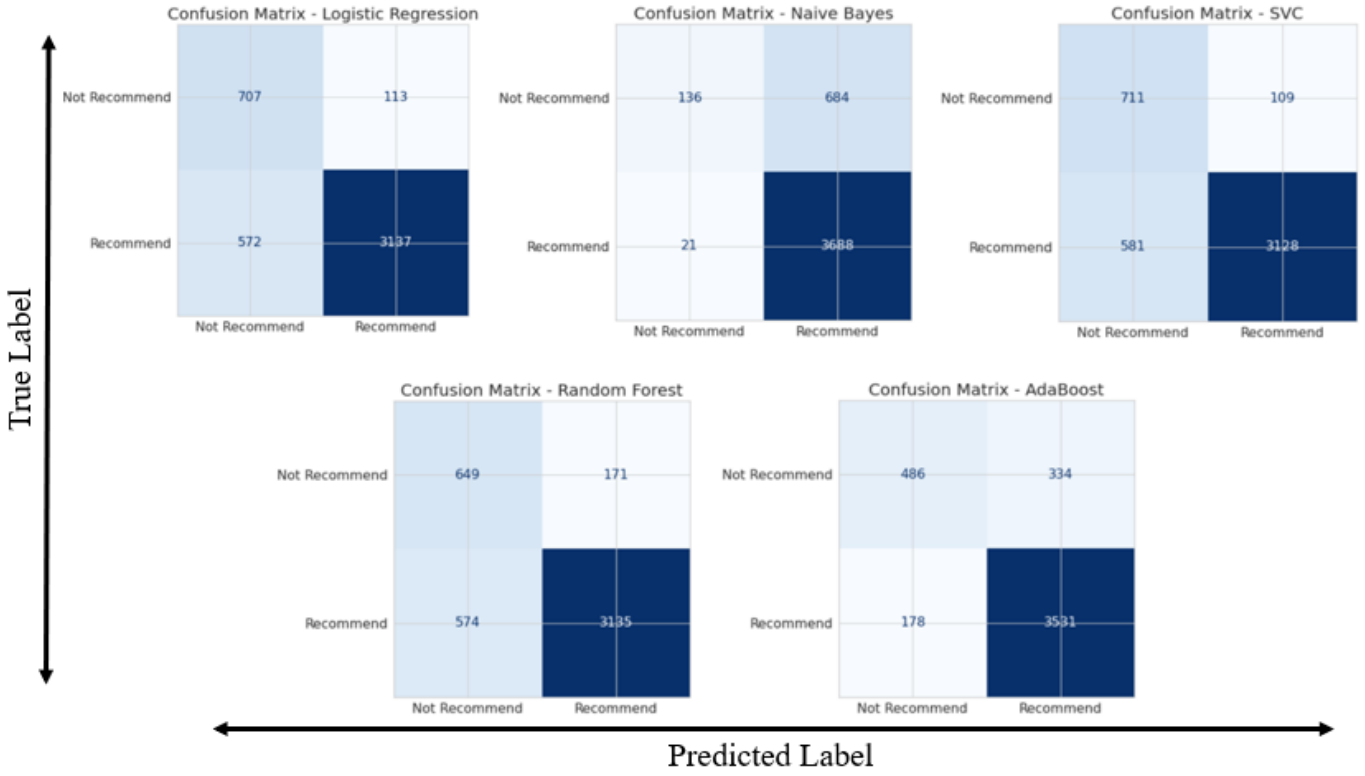

Figure 2. Confusion Matrices (Based on TF-IDF)

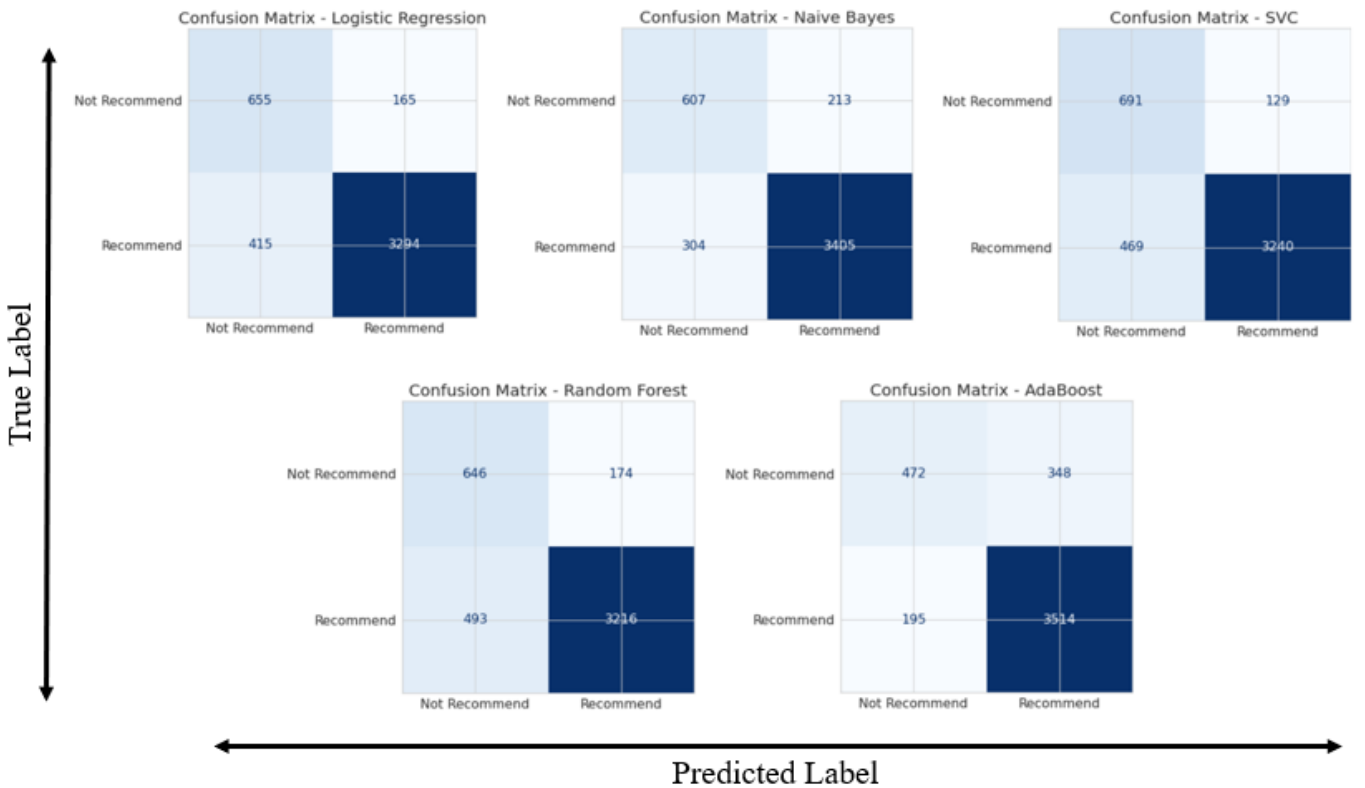

Figure 3. Confusion Matrices (Based on Count Vectorizer)

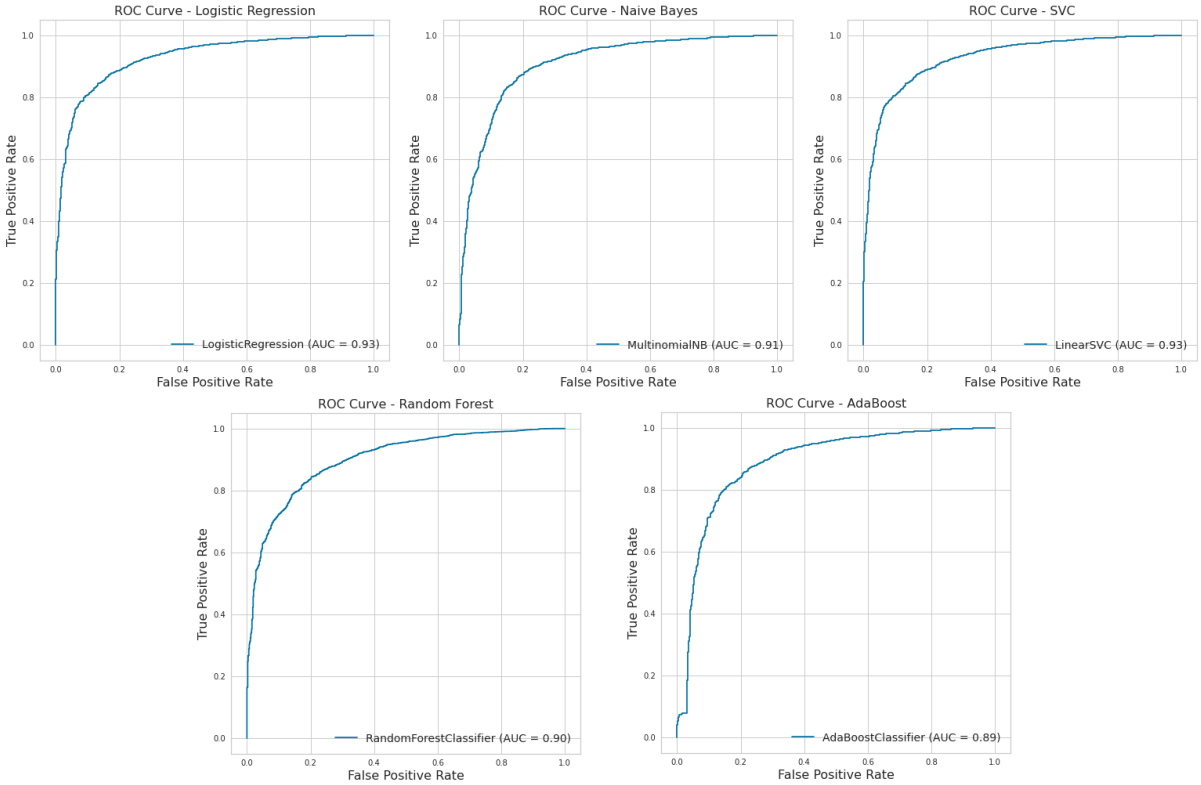

Figure 4. ROC Curves and AUC Scores (Based on TF-IDF)

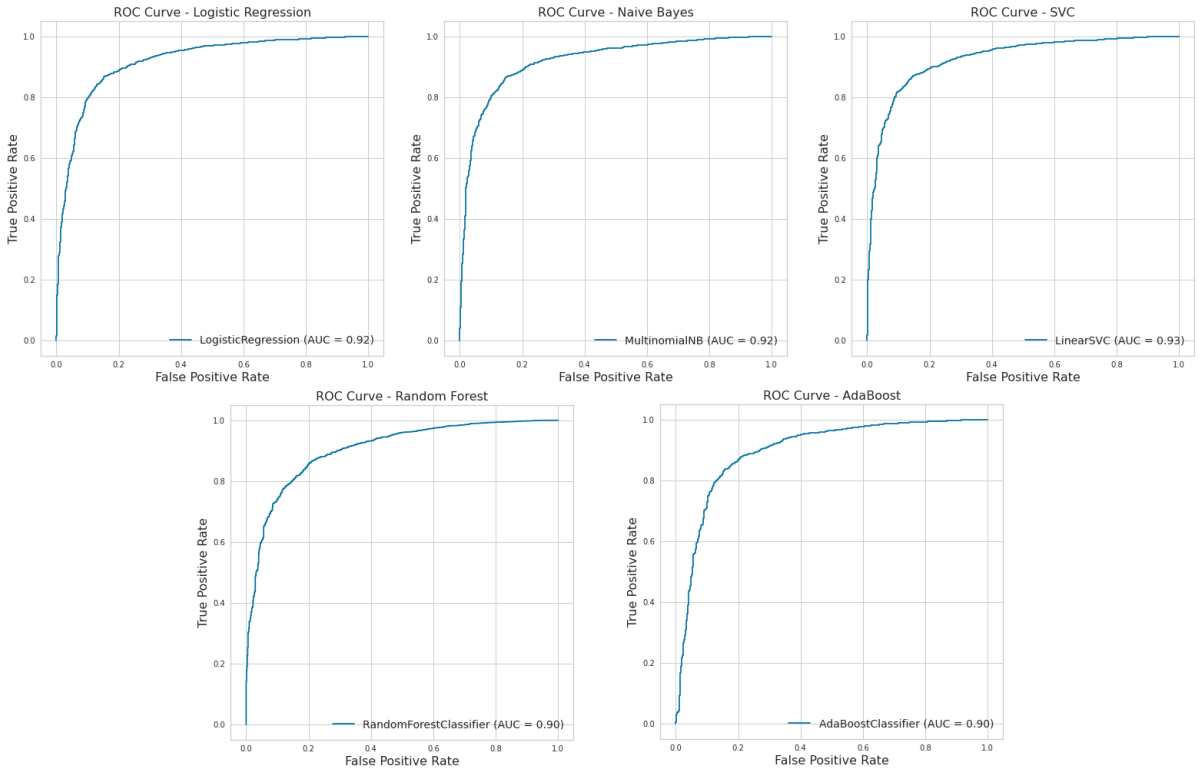

Figure 5. ROC Curves and AUC Scores (Based on Count Vectorizer)

Evaluation of CNN Model

The first analysis of the CNN Model focuses on the AUC scores of the model during the training process and its performance on the validation set. The first graph in Figure 6 shows the model's AUC scores for the training set and validation set as the number of epochs (number of training rounds) increases. Initially, we observe that the AUC

score obtained by the model on the training set has a high initial value (0.70) and this score is consistently maintained as the epochs progress. However, the AUC score for the validation set drops below the score obtained on the training set after approximately epoch 35. This may indicate that the model exhibits an overfitting tendency during the training process. The model is highly adapted to the training set, achieving high performance on the training data, but tends to underperform against new and unseen data. The validation AUC score stabilizes at 0.90 at the end of the training process, while the overall AUC score remains at 0.95. This may indicate that the model's overall performance decreases when the patterns it initially learned during the training process are applied to the validation set. Although the overall performance of the model is high, the overlearning tendency and the poor performance on the validation set suggest that the generalization ability of the model may be limited.

The second graph in Figure 6 focuses on the model's loss values during the training process and the loss values on the validation set (val_loss). The graph shows the loss values of the model over epochs. When the graph is analyzed, it is observed that the loss value of the model on the training set starts from 0.8 at the beginning and decreases below 0.3 as the epochs progress. This means that the model learns better and better during the training process and is able to identify patterns in the training set more effectively. On the other hand, the loss value (val_loss) on the validation set stabilizes around 0.35 after approximately epoch 35. However, the loss value on the training set dropped below val_loss. This suggests that the model may have overfitted the training set and underperformed on the validation set. The fact that the loss value of the model on the training set is lower than that on the validation set indicates the tendency of the model to overfit during the learning process.

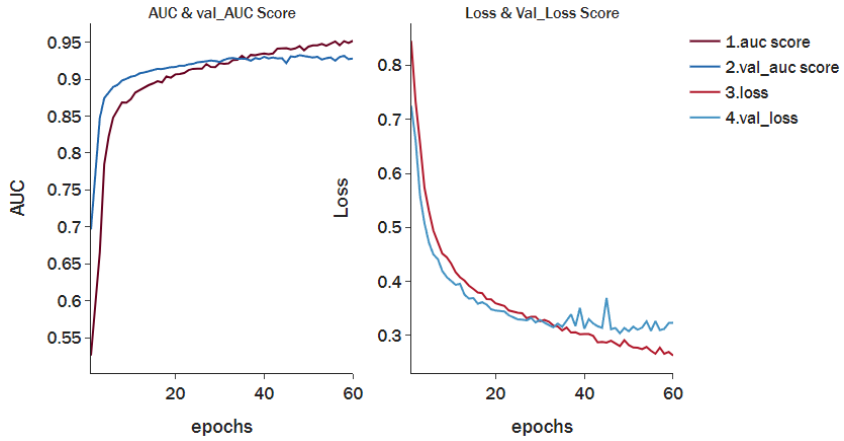

Figure 6. Training and Validation Performance

Comparison of Model Performances

Table 2 shows the performance comparisons of the models used in the study. Sentiment analysis performed with NB, AdaBoost, LR, SVM, RF, and DL models were compared using two different text vectorization methods, Count Vectorizer and TF-IDF. While the NB model achieves 90.3% accuracy when trained with Count Vectorizer, this rate drops to 88.1% when TF-IDF is used. In the AdaBoost model, on the contrary, a lower accuracy is obtained when trained with Count Vectorizer. The LR model performs similarly with both vectorization methods. The model with the highest accuracy rate is AdaBoost model. When trained with AdaBoost TF-IDF (F1-Score: 0.955), achieved 92.4% accuracy.

NB shows consistent performance with both Count Vectorizer and TF-IDF, achieving high precision (94.3% and 93.5%) and recall (91.2%). AdaBoost performs well across both vectorization methods but notably excels with TF-IDF, achieving high precision (94.2%) and recall (97.2%). SVM also demonstrates strong performance, especially with Count Vectorizer, where it achieves high precision (95.8%) and decent recall (87.2%). RF and LR exhibit competitive results, with RF showing slightly higher precision but lower recall compared to LR across both vectorizers. DL, although not vectorized in the traditional sense, achieves competitive overall metrics. In summary, while each model exhibits strengths across various metrics and vectorization techniques, AdaBoost stands out for its robust performance. With particularly high precision and recall AdaBoost proves effective in classification tasks, especially suited for applications where balancing precision and recall is crucial, such as analyzing customer reviews. The choice of model and vectorization method should be guided by specific objectives, ensuring alignment with the desired performance metrics and the characteristics of the dataset at hand.

Table 2. Model Performance Comparison

Model	Text Vectorization	Accuracy	Precision	Recall	F1-Score
NB	Count Vectorizer	0.903	0.943	0.912	0.927
	TF-IDF	0.881	0.935	0.912	0.923
AdaBoost	Count Vectorizer	0.915	0.935	0.960	0.940
	TF-IDF	**0.924**	0.942	**0.972**	**0.955**
LR	Count Vectorizer	0.875	0.949	0.878	0.915
	TF-IDF	0.862	**0.962**	0.859	0.912
SVM	Count Vectorizer	0.898	0.958	0.872	0.913
	TF-IDF	0.861	**0.963**	0.839	0.897
RF	Count Vectorizer	0.891	0.948	0.871	0.904
	TF-IDF	0.876	0.948	0.846	0.894
DL	-	0.879	**0.962**	0.929	0.929

DISCUSSION

In this section, an analysis is presented on how the characteristics of the dataset and the selection of machine learning algorithms impact the performance of sentiment analysis models, with particular attention to preprocessing steps and vectorization methods.

The dataset used in this research, features a mix of numerical, categorical, and textual data, requiring careful preprocessing to ensure compatibility with various machine learning algorithms. The review text and titles are critical features for sentiment analysis, and their quality and variability can affect NLP model performance. Preprocessing steps such as tokenization and stop-word removal are essential for standardizing the text data. Different vectorization methods, like Count Vectorizer and TF-IDF, impact model performance. For example, NB performed better with the Count Vectorizer method due to its frequency-based approach, while LR and SVM benefited from TF-IDF's nuanced feature weighting. Algorithms have varying sensitivities to dataset characteristics; NB is efficient with high-dimensional data, while SVM and LR handle overlapping classes well. Ensemble methods like AdaBoost may offer higher accuracy but can be sensitive to noisy data, as indicated by slightly higher false positive rates. The anonymization of data can affect the context and sentiment in reviews. Although necessary for privacy, such modifications may strip nuances that aid sentiment detection, emphasizing the need for thorough preprocessing.

The high performance of the AdaBoost model in the text classification task has been emphasized in several studies in the literature and is also supported by this study (Feng et al., 2017). In particular, the high precision achieved by the model when trained with both Count Vectorizer and TF-IDF reflects its ability to accurately classify customer sentiments. This is of great importance in emotion-driven applications such as customer sentiment analysis, where positive and negative reviews can be accurately identified. Other studies in the literature have often attributed AdaBoost's success to the ensemble nature of the model and its ability to evaluate the interactions between features (Wyner et al., 2017). These features of the model allow it to handle the complexity of text datasets by effectively combining different types of features and performing more robustly (Xia et al., 2011). In addition to this success, the interpretability advantages of the AdaBoost model have also been highlighted in the literature. Since AdaBoost builds a strong model by combining weak learners (usually decision trees), it is easier to understand the contribution of each learner and explain why the model makes a particular prediction (Li et al., 2005). This feature contributes to the preference for AdaBoost, especially in industrial applications, when the understandability of the model's decisions is an important factor.

The findings of the study show that the AdaBoost model has a low AUC score and a high FP value compared to the other models. On the other hand, the AdaBoost model exhibited the highest accuracy and recall values. This may suggest that the model has difficulty with a particular class or tends to misclassify it. However, the high precision and recall demonstrate the model's ability to correctly identify positive reviews. Therefore, it is important to combine these results with other contextual factors to develop further understanding of how to use the model in a specific application context and which performance metrics are prioritized.

Due to their performance advantages on large text datasets, some sources in the literature suggest that DL models are preferable (Kowsari et al., 2019; Minaee et al., 2021). These models are known for their complex learning structures and deep feature discovery capabilities. DL methods are designed to understand hidden patterns and relationships more effectively in text data (Rygielski et al., 2002). Furthermore, these advantages often require larger datasets and high computational power, which can result in cost and infrastructure challenges. Training DL models can often take

a long time and require large amounts of computational resources (Najafabadi et al., 2015). This can be a significant constraint, especially for organizations with a limited budget or infrastructure.

In this study, it was observed that the DL model achieves high accuracy due to its complex learning structures. However, despite these advantages, it can be more costly in terms of training time and computational power, which is one of the factors affecting the model selection decision. Model selection should be carefully evaluated based on various factors such as application context, dataset features, and available infrastructure. DL model with high overall performance performs poorly on the validation set due to the tendency to overfit during the training process. This suggests that the model tends to overfit the training data and may have less generalization ability to new data. To increase the generalization ability of the model, overfitting control techniques and more diverse datasets can be used to make the model form a more general representation.

Table 3 illustrates various studies in the literature that have investigated the dataset used in this study, detailing their objectives, methodologies, and findings. Due to the focus of other studies on specific tasks and metrics for evaluating their models, comparisons with this study may be limited in terms of the range of text mining and machine learning algorithms evaluated and the metrics used. This study, in comparison, comprehensively evaluates a variety of models including NB, AdaBoost, LR, SVM, RF, and DL using both Count Vectorizer and TF-IDF. When examining Table 3, it is evident that Inácio and Oliveira (2024) achieved an increase in the F1 score from 53.2% to 68.7% for the BERTimbau-large model, showing variability across different models. Agarap (2018) obtained F1 scores of 88% for recommendation classification and 93% for sentiment analysis. In contrast, this study achieved a significantly higher F1 score of 95.5% with AdaBoost using TF-IDF, demonstrating superior performance in both precision (94.2%) and recall (97.2%). This highlights a substantial improvement in F1 score compared to Inacio and Oliveira's (2024) as well as Agarap's (2018) results, underscoring the effectiveness of AdaBoost, particularly when leveraged with TF-IDF, in sentiment analysis tasks. Georgescu and Kinnunen (2020) achieved 93% accuracy using LR and NB, with a kappa statistic of 79, providing recommendations for customer profiling and business strategies. However, since metrics such as recall, precision, and F1-score were not reported, relying solely on accuracy may not sufficiently demonstrate the overall effectiveness of their models. Additionally, the absence of cross-validation in their study suggests that performance values calculated using only train-test split validation method are likely to be overestimated compared to more robust cross-validation methods, which generally yield more accurate performance estimates. The primary goal of this study is to compare various methods rather than solely aiming for high scores. Nevertheless, it has achieved consistent and generalizable results.

Table 3. Comparative Review of Studies Utilizing the Same Dataset

Study	Objective	Methods	Key Findings
Inácio and Oliveira (2024)	Use of multimodal transformers, proposal of a new corpus	Combining LLMs and numerical features, feature pooling method	Increased F1-score from 53.2% to 68.7% for BERTimbau-large, variability across different models
Maronikolakis and Schütze (2021)	Multiple domain adaptation, low-resource usage	Multidomain models	Performance advantages of multidomain models over single-domain models, resource efficiency
Agarap (2018)	Model improvement, hyperparameter tuning	Bidirectional RNN-LSTM model	F1-scores of 88% for recommendation classification, 93% for sentiment analysis
Georgescu and Kinnunen (2020)	Analysis of customer reviews on fashion items	LR, NB	93% accuracy, 79 kappa statistic, recommendations for customer profiling and business strategies
Cloutier and Japkowicz (2023)	Effectiveness of LLM-augmented oversampling on binary and multiclass classification	Various resampling methods including LLM-augmented techniques	LLM-augmented methods show mixed results in binary classification tasks, significant improvements in multiclass tasks, particularly achieving highest performance on small multiclass tasks with improvements ranging from 85% to 115%
This study	Evaluate effectiveness of various algorithms and text vectorization methods	Evaluation of NB, AdaBoost, LR, SVM, RF, and DL models using Count Vectorizer and TF-IDF	AdaBoost achieves 92.4% accuracy with TF-IDF (F1-Score: 0.955), excelling in precision (94.2%) and recall (97.2%).

CONCLUSION

This research aims to compare the performance of various supervised classification and DL models for sentiment analysis on customer reviews obtained from a women's clothing e-commerce platform.

The data preparation phase of the study constitutes one of the most important parts of the research. First, missing values issues were addressed with various strategies. Significant dataset deficiencies that could cause problems in text-based analysis were cleaned. Then, Count Vectorizer and TF-IDF methods were used for text vectorization. Count Vectorizer is a tool for converting text data into a matrix of token counts, a numerical representation. Each token creates a matrix of the number of times it occurs in a given document. TF-IDF is a measure used to determine the importance of a word in a document. This method combines the frequency of the word in the document and its frequency in the overall collection. Text preprocessing steps of tokenization, noise reduction and lexical normalization are then applied. In the tokenization step, texts are broken down into more manageable chunks by segmenting them into words. The noise reduction step was used to remove redundant information, including stop words of the language. Lexical normalization was performed to regularize word variations and to obtain the root form of each word. These steps were important to prepare the text for analysis and improve model performance.

Five different supervised classification models (LR, NB, SVM, RF, AdaBoost) and CNN Model, were trained on the dataset and their performances were compared. The results show that there are some differences between the models. AdaBoost and NB models are prominent algorithms in the context of customer sentiment analysis with their high accuracy. SVM and LR models also achieved successful results but fell behind in terms of accuracy. An overfitting problem was detected in the CNN Model. Overfitting means that the model is overly adapted to the dataset used in the training process. In this case, the model may perform well on the training data, but its ability to generalize to new and unseen data may suffer. Especially in text-based models, noise, omissions in the dataset or the complexity of the relationships between features can increase the risk of overfitting. Depending on the size and complexity of the dataset, an effective approach may be to add regularization terms to control the tendency of overfitting or to prefer simpler models over more complex models. In addition, no significant performance differences were observed in terms of vectorization methods applied to the models in the study.

The overall results of the study show that the application context and performance metrics should be considered before choosing between different models. Especially in text-based analysis, data preprocessing steps and text preprocessing methods are critical factors affecting model performance. Future work could focus on testing model performance on larger and diverse datasets and more in-depth analysis to determine which model is more effective in specific application scenarios. Such research in the field of sentiment analysis can help e-commerce platforms evaluate customer reviews more effectively and improve the user experience.

REFERENCES

Agarap, A. F. (2018). Statistical analysis on E-commerce reviews, with sentiment classification using bidirectional recurrent neural network (RNN). *arXiv preprint arXiv:1805.03687*.

Aizawa, A. (2003). An information-theoretic perspective of TF-IDF measures. *Information Processing & Management*, *39*(1), 45-65.

Alantari, H. J., Currim, I. S., Deng, Y., & Singh, S. (2022). An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing*, *39*(1), 1-19.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*(S1), 180-208.

Angulakshmi, G., & ManickaChezian, R. (2014). An analysis on opinion mining: techniques and tools. *International Journal of Advanced Research in Computer and Communication Engineering*, *3*(7), 2319-5940.

Badaro, G., Baly, R., Hajj, H., Habash, N., & El-Hajj, W. (2014, October). A large scale Arabic sentiment lexicon for Arabic opinion mining. In *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)* (pp. 165-173).

Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 61-66). IEEE.

Barik, K., Misra, S., Ray, A. K., & Bokolo, A. (2023). LSTM-DGWO-Based sentiment analysis framework for analyzing online customer reviews. *Computational Intelligence and Neuroscience*, *2023*.

Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *403*, 412.

Biradar, S. H., Gorabal, J. V., & Gupta, G. (2022). Machine learning tool for exploring sentiment analysis on twitter data. *Materials Today: Proceedings*, *56*, 1927-1934.
Brooks, N. (2018). Women's E-Commerce Clothing Reviews. Kaggle. https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

Carrigan, M., Moraes, C., & Leek, S. (2011). Fostering responsible communities: A community social marketing approach to sustainable living. *Journal of Business Ethics*, *100*, 515-534.

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, *159*(6), 1638-1645.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215.

Chandrasekaran, D., & Tellis, G. J. (2017). A critical review of marketing research on diffusion of new products. *Review of Marketing Research*, *3*, 39-80.

Chawla, N., & Kumar, B. (2022). E-commerce and consumer protection in India: The emerging trend. *Journal of Business Ethics*, *180*(2), 581-604.

Cloutier, N. A., & Japkowicz, N. (2023, December). Fine-tuned generative LLM oversampling can improve performance over traditional techniques on multiclass imbalanced text classification. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 5181-5186). IEEE.

Das, A. (2021). Logistic regression. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 1-2). Cham: Springer International Publishing.

Demirbilek, M., & Demirbilek, S. Ö. (2023). Sentiment analysis based on google comments with machine learning methods and Amazon Comprehend: The case of a university in Central Anatolia. *Journal of University Research*, *6*(4), 452-461.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168-189.

Dey, S., Wasif, S., Tonmoy, D. S., Sultana, S., Sarkar, J., & Dey, M. (2020, February). A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. In *2020 International Conference on Contemporary Computing and Applications (IC3A)* (pp. 217-220). IEEE.

Dogru, N., & Subasi, A. (2018, February). Traffic accident detection using random forest classifier. In *2018 15th learning and technology conference (L&T)* (pp. 40-45). IEEE.

Feng, X., Liang, Y., Shi, X., Xu, D., Wang, X., & Guan, R. (2017). Overfitting reduction of text classification based on AdaBELM. *Entropy*, *19*(7), 330.

Georgescu, I., & Kinnunen, J. (2020). Consumer recommendation dynamics in online retail business under logistic regression and naïve Bayes analyses. In *Proceedings of the International Conference on Applied Statistics* (Vol. 2, No. 1, pp. 120-128).

Guo, J. (2022). Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, *31*(1), 113-126.

Han, B., & Baldwin, T. (2011, June). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 368-378).

Han, S., & Anderson, C. K. (2020). Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly*, *61*(2), 142-153.

Hartmann, J., & Netzer, O. (2023). Natural language processing in marketing. In *Artificial Intelligence in Marketing* (Vol. 20, pp. 191-215). Emerald Publishing Limited.

Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, *43*, 375-394.

Hossain, M. S., & Rahman, M. F. (2023). Customer sentiment analysis and prediction of insurance products' reviews using machine learning approaches. *FIIB Business Review*, *12*(4), 386-402.

Hu, W., Gong, Z., & Guo, J. (2010, November). Mining product features from online reviews. In *2010 IEEE 7th International Conference on E-Business Engineering* (pp. 24-29). IEEE.

Inácio, M., & Oliveira, H. G. (2024, March). Exploring multimodal models for humor recognition in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese* (pp. 568-574).

Jararweh, Y., Al-Ayyoub, M., Fakirah, M., Alawneh, L., & Gupta, B. B. (2019). Improving the performance of the needleman-wunsch algorithm using parallelization and vectorization techniques. *Multimedia Tools and Applications*, *78*, 3961-3977.

Kamal, M., & Himel, A. S. (2023). Redefining Modern Marketing: An analysis of AI and NLP's influence on consumer engagement, strategy, and beyond. *Eigenpub Review of Science and Technology*, *7*(1), 203-223.

Kapoor, S., & Banerjee, S. (2021). On the relationship between brand scandal and consumer attitudes: A literature review and research agenda. *International Journal of Consumer Studies*, *45*(5), 1047-1078.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Li, C., Zhang, Z., Lee, W. S., & Lee, G. H. (2018). Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5226-5234).

Li, H., Bruce, X. B., Li, G., & Gao, H. (2023). Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews. *Tourism Management*, *96*, 104707.

Li, X., Wang, L., & Sung, E. (2005, July). A study of AdaBoost with SVM based weak learners. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (Vol. 1, pp. 196-201). IEEE.

Li, X., Wang, L., & Sung, E. (2008). AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence*, *21*(5), 785-795.

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: speech, text, and face. *Entropy*, *25*(10), 1440.

Maronikolakis, A., & Schütze, H. (2021, April). Multidomain pretrained language models for green NLP. In *Proceedings of the Second Workshop on Domain Adaptation for NLP* (pp. 1-8).

Mariani, M., & Borghi, M. (2021). Are environmental-related online reviews more helpful? A big data analytics approach. *International Journal of Contemporary Hospitality Management*, *33*(6), 2065-2090.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, *54*(3), 1-40.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement* (pp. 201-237). Woodhead Publishing.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1-21.

Obiedat, R., Qaddoura, R., Ala'M, A. Z., Al-Qaisi, L., Harfoushi, O., Alrefai, M. A., & Faris, H. (2022). Sentiment analysis of customers' reviews using a hybrid evolutionary svm-based approach in an imbalanced data distribution. *IEEE Access*, 10, 22260-22273.

Patel, A., Oza, P., & Agrawal, S. (2023). Sentiment analysis of customer feedback and reviews for airline services using language representation model. *Procedia Computer Science*, *218*, 2459-2467.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101-121). Academic Press.

Pradhan, V. M., Vala, J., & Balani, P. (2016). A survey on sentiment analysis algorithms for opinion mining. *International Journal of Computer Applications*, *133*(9), 7-11.

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, *181*(1), 25-29.

Racherla, P., & Friske, W. (2012). Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications*, *11*(6), 548-559.

Rain, C. (2013). Sentiment analysis in Amazon reviews using probabilistic machine learning. *Swarthmore College*, *42*.

Ramadhan, F. A., Ruslan, R. R. P., & Zahra, A. (2023). Sentiment analysis of e-commerce product reviews for content interaction using machine learning. *Cakrawala Repositori IMWI*, *6*(1), 207-220.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503-520.

Rosário, A., & Raimundo, R. (2021). Consumer marketing strategy and e-commerce in the last decade: a literature review. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(7), 3003-3024.

Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, *24*(4), 483-502.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160.

Shanthi, R., & Desti, K. (2015). Consumers' perception on online shopping. *Journal of Marketing and Consumer Research*, *13*, 14-21.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, *70*, 346-355.

Sylvester, E. V., Bentzen, P., Bradbury, I. R., Clément, M., Pearce, J., Horne, J., & Beiko, R. G. (2018). Applications of random forest feature selection for fine-scale genetic population assignment. *Evolutionary Applications*, *11*(2), 153-165.

Tran, D. D., Nguyen, T. T. S., & Dao, T. H. C. (2022). Sentiment analysis of movie reviews using machine learning techniques. In *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London*, Volume 1 (pp. 361-369). Springer Singapore.

Turki, T., & Roy, S. S. (2022). Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer. *Applied Sciences*, *12*(13), 6611.

Vijayarani, S., & Janani, R. (2016). Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, *3*(1), 37-47.

Wani, T. A., & Ali, S. W. (2015). Innovation diffusion theory. *Journal of General Management Research*, *3*(2), 101-118.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media.

Wyner, A. J., Olson, M., Bleich, J., & Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, *18*(1), 1558-1590.

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, *181*(6), 1138-1152.

Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering*, *18*(3), 304-319.

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, *99*, 238-248.

Zhang, J., Lu, X., & Liu, D. (2021). Deriving customer preferences for hotels based on aspect-level sentiment analysis of online reviews. *Electronic Commerce Research and Applications*, *49*, 101094.

Zhao, H., Liu, Z., Yao, X., & Yang, Q. (2021). A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. *Information Processing & Management*, *58*(5), 102656.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12*(10), 931-934.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350-361.

Zhuang, M., Cui, G., & Peng, L. (2018). Manufactured opinions: The effect of manipulating online product reviews. *Journal of Business Research*, *87*, 24-35.