Sentiment Analysis of Online Reviews: A Machine Learning Based Approach with TF-IDF Vectorization

Khalid Alemerien[1,*], Aram Al-Ghareeb[2]
and Malek Zakarya Alksasbeh[3]

[1]*Information Technology Department, Tafila Technical University Tafila, Jordan*
[2]*Deanship of Scientific Research and Graduate Studies, Tafila Technical University, Tafila, Jordan*
[3]*Faculty of Information Technology, Al-Hussein Bin Talal University, Jordan*
E-mail: Khalid.alemerien@ttu.edu.jo; 820220604002@stu.ttu.edu.jo; malksasbeh@ahu.edu.jo
Orcid: 0000-0003-4485-8298; 0009-0003-7877-1753
[*]*Corresponding Author*

Received 09 June 2024; Accepted 20 November 2024

Abstract

Nowadays, online reviews wield considerable influence over consumer decision-making processes. Surveys show 84% of people compare their trustworthiness to recommendations from personal connections in these online reviews. Online reviews of services or destinations can significantly benefit the tourism industry. Therefore, our primary intent of this study is to leverage Machine Learning (ML) and Natural Language Processing (NLP) for sentiment analysis of hotel reviews in Jordan in order to assist both hotel owners and tourists. In this study, we proposed a ML-based approach using Support Vector Machine (SVM) and TF-IDF to perform sentiment analysis of hotel reviews into positive or negative. In addition, our experiments were performed using our real dataset, "JOHotelRating", which was

Journal of Mobile Multimedia, Vol. 20_5, 1089–1116.
doi: 10.13052/jmm1550-4646.2055

gathered in the Jordanian context. In the feature extraction stage, we utilized the Term Frequency-Inverse Document Frequency (TF-IDF) method. In the machine learning (ML) classification phase, we utilized various algorithms such as Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), Bernoulli's Naïve Bayes (BNB), Decision Tree (DT), and Random Forest (RF). SVM with TF-IDF for feature extraction, emerged as the standout performer, achieving an impressive 97% accuracy in sentiment classification. Our proposed approach offers the hotel owners a time-saving method to identify positive and negative reviews, allow them to understand trends, and enhance the overall customer experience. On the tourist side, the study attempts to tackle the challenge of comprehending numerous reviews by providing sentiment analysis, ultimately aiding them in making better-informed decisions when selecting a hotel in Jordan.

Keywords: Machine learning (ML), natural language processing (NLP), sentiment analysis, online review, tourism, support vector machine (SVM).

1 Introduction

Throughout the day, millions of tourists post their opinions, recommendations, and observations regarding accommodations, services, and tourist destination on an assortment industry-standard websites [1]. Online reviews are crucial and wield significant influence; they represent user opinions about products or services and are often found on retail sites or in search engine results. These reviews discuss features, user preferences, and product accuracy. Ratings, indicated by stars or numbers, are frequently included to provide an average score based on multiple reviews [2]. Therefore, the significance of electronic word-of-mouth in tourism domain is well proven in the extant literature [3]. Research indicates that 91% of individuals often or sometimes read reviews regarding to service or product on the internet, and 84% trust these reviews as much as recommendations from friends or family. 68% of people make decisions based on their reading of one to six internet reviews [4].

However, a popular product or service may have thousands or even hundreds of reviews. Those who want to get the product or service may find it difficult to browse through the many reviews and determine whether or not to get it because of this large number. Additionally, it's difficult for the provider to track and manage all of the client feedback. Furthermore, some false information may appear early in the search results, which can make it

more difficult for users to obtain accurate and comprehensive information before making a decision [5].

The tourism business is one of several that has been propagated to adopt data-driven decision-making [6]. Especially tourists need to make a decision about their hotel stay based on the previous reviews. It is impressive to provide both tourists and service providers with intelligent solutions that may help them make timely and correct decisions [7]. This kind of solution is constructed based on the integration between machine learning and natural processing techniques. As a result, the tourism industry can effectively adapt these solutions to improve the quality of the services provided. One of the success factors of effective tourism digitization is ensuring consumer social engagement.

Understanding consumer social engagement can be achieved through the analysis of customer review data. Sentiment analysis is a tool used for this purpose, providing insights into how consumers feel about a company, product, or feature at a fundamental level. This analysis delves into the perceptions of products and brands, offering a detailed understanding of consumer sentiments at various levels [8]. Sentiments are typically categorized as positive, neutral, or negative. This process is considered a crucial step in comprehending customer experiences and opinions. The statistics mentioned that more than 60% of companies find sentiment analysis data in reviews extremely useful for ongoing customer service and sales. This underscores the practical impact of sentiment analysis. Furthermore, the statistics highlighted that a majority of businesses recognize the value of extracting meaningful insights from customer sentiments to achieve their main objectives, such as enhancing their customer service processes, refining sales strategies, and ultimately boosting the overall customer experience [9].

Sentiment analysis relies on two key technologies, including machine learning (ML) and natural language processing (NLP). Natural language processing enables computer applications to comprehend and respond to human language to acquire linguistic understanding. Meanwhile, machine learning is known for its efficiency in processing vast amounts of customer data and employing statistical models and algorithms. These computer technologies develop and learn on their own to determine textual sentiments through recognizing such patterns in data [10].

In this research, we will focus on conducting sentiment analysis for hotel reviews, with a specific emphasis on those originating in Jordan. The data indicates that there has been a consistent increase in the number of classified hotels over the years, with a reported count of 294,000 units in 2021, up from

285,000 units in 2020. According to Ministry of Tourism and Antiquities, Number of Classified Hotels in Jordan [11], the data is updated yearly, and the average count from December 1998 to 2021 is 246,000 units.

The subsequent sections of this paper are organized as follows: Section 2 delivers an overview of literature review. In Section 3, we outline the methodology, while Section 4 details the experimental setup and Section 5 presents the experimental results discussion. The paper concludes by summarizing and outlining prospects in Section 6.

2 Related Works

In this section, we provide an overview of the state-of-the-art studies that have been conducted using ML and DL techniques and lexicon analysis methods.

2.1 Machine Learning with TF-IDF

Puh and Bagić in [1] explored the application of machine and deep learning models for sentiment analysis on tourist reviews, using ML models (Naïve Bayes, SVM, CNN, LSTM, BiLSTM) and DL methods (BiLSTM) on the TripAdvisor Hotel Review Dataset. The results showed that deep learning models outperform traditional machine learning algorithms, with the top-performing BiLSTM achieving 72% accuracy for five classes and 89% for three classes. The study acknowledges the limitation of assuming the credibility of all reviews in the dataset and suggests future work involving data enlargement from diverse sources and incorporating different data types across various domains.

Wadhe and Suratkar in the study [12], used machine learning algorithms including, Naive Bayes, SVM, and RF, to collect data from platforms like Mouthshut and TripAdvisor. The dataset contains 3209 reviews from diverse tourism websites in CSV format, including text and associated ratings. Researchers refined their analysis with strategies like multilingual review classification, various feature selection methods, and exploring deep learning techniques. Ratings above 3 are labeled as positive $(+1)$, below 3 as negative (-1), and equal to 3 as neutral (0). They used TF-IDF Vectorization, which notably improved classification accuracy over CountVectorization in the review dataset. The composition of TF-IDF vectorization and RF achieved the highest accuracy at 86%.

As presented in [5], the focus is on online hotel reviews, proposing a supervised machine learning approach using unigram features and both

frequency and TF-IDF information for polarity classification. Employing SVM-based models, the study utilized a hotel-review corpus by Tan S.H., highlighting the superiority of TF-IDF over frequency. The SVM with unigram features and TF-IDF achieved an accuracy of 87.2%. The dataset comprises 4000 balanced reviews (positive and negative), divided into four folds for analysis. The authors plan to explore semi-supervised learning and introduce NLP techniques to enhance sentiment analysis.

Dharma and Saragih in the study [13], aimed to analyze hotel reviews using sentiment analysis, employing the Support Vector Machine (SVM) approach. Three feature extraction methods were compared: Bag of Words, TF-IDF, and an improved TF-IDF. These methods convert text data into numerical representations to capture the influence of words in the reviews. Furthermore, the study revealed that TF-IDF is the most effective, with an accuracy score of 71.75%, and concluded that TF-IDF is effective at capturing the sentiment of hotel reviews. This leads to the fact that the TF-IDF method can indicate the significant influence of certain words on expressed sentiments.

Srivastava et al. in the study [14], set out to perform a sentiment analysis of a textual dataset using various methods. The researchers investigated the effectiveness of both machine learning and lexicon-based approaches. SVM and Logistic Regression (LR) with Stochastic Gradient Descent (SGD) algorithms were utilized in the study. Moreover, the researchers employed lexicon-based models like AFINN and VADER. Term Frequency-Inverse Document Frequency (TF-IDF) and bag of words (BoW) methods are used to represent the textual features. The study found that TF-IDF with SVM achieved the highest accuracy at 96.3%, which outperformed the bag of words approach at 95.2%. Furthermore, the findings showed that the VADER method is more accurate at an accuracy of 88.7% than the AFINN method at an accuracy of 86.0%. This study underscores the effectiveness of combining machine learning techniques with appropriate text feature representations for sentiment analysis.

2.2 Machine Learning with Other Sentiment Lexicon Methods

Ye et al. in the study [15], carried out a study aimed at assisting potential tourists and the tourism industry in efficiently extracting valuable insights from hotel reviews. The study implemented a model that employed two supervised machine learning algorithms (NB and SVM) with the character-based N-gram model. The utilized dataset comprises 1191 labeled tuples from

the top seven popular travel destinations in the US and Europe on Yahoo's Travel. Each review was rated on a 5-scale by reviewers. The SVM and N-gram approach outperformed NB, all achieving accuracies of at least 80%. The study offered a number of recommendations, including investigating the applicability of classification methods in destinations beyond Western countries, examining readership for travel blogs, and conducting a longitudinal study to track changing consumer perceptions over time.

In the study [16], Rai and Ahirwal introduced a method that generates a tourism review sentiment lexicon, employing Part-Of-Speech (POS) tags and SentiwordNet within the lexicon creation process. Three classifiers, namely SVM, k-nearest Neighbors (KNN), and RF, are employed by utilizing Lexicon Features. Data acquisition involves the collection of 450 reviews for both positive and negative sentiments from diverse individuals and locations on TripAdvisor. The evaluation of the sentiment analysis framework with utilizing the proposed lexicon demonstrates its effectiveness in the context of tourism records. The obtained results revealed noteworthy accuracy, with KNN achieving approximately 80.57%, SVM achieving around 89.36%, and RF achieving a high accuracy of 93.6%.

Kulkarni et al. [17] focused on the core concept that positive reviews play a crucial role in attracting potential users, while negative reviews may deter them. Consequently, there is a paramount need to analyze these reviews for a more profound understanding of user experiences and to undertake initiatives for improvement. The study employed various machine learning algorithms, such as MNB, BNB, and RF, alongside deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The study used mixed user reviews of products and services, involving the Amazon Product Reviews dataset from Kaggle consisting of one million reviews as well as destination reviews from platforms like TripAdvisor and Google consisting of 30,000 reviews. Notably, the results demonstrated the superiority of the deep-learning algorithm in classifying user reviews. The CNN achieved an accuracy of 94.40%, and the RNN reached 94.56%, surpassing the accuracies of traditional machine learning algorithms (BNB (82.75%), MNB (84.48%), and RF (84.60%)). This may affect the generalization of its results in hotels reviews context.

As described in [18], sentiment classification in reviews employs the Multinomial Naïve Bayes classifier and Bag of Word (BoW) with an emphasis on preprocessing, feature extraction, and feature selection. The collected dataset was labeled with binary classes including, positive (1) and negative (0). The dataset was unbalanced with 3946 positive reviews and 1053

negative reviews. The best results, achieving a 91% average F1-Score with 10-fold cross-validation, are achieved through preprocessing and feature selection.

Li and Liu [19] focused on predicting hotel customer satisfaction through a dataset of ten thousand TripAdvisor reviews. It evaluates machine learning models – LR, NB, DT, RF, SVM, and Neural Network. The study, employing two 10-fold cross-validation experiments, trained models on review titles (average length: 25.7 bytes). Results showed promising 84% to 87% classification accuracies, with the SVM model reaching nearly 92% accuracy by including review contents and filtering irrelevant words. The paper concluded that classical models like SVM and Naive Bayesian perform well in terms of accuracy, while neural networks demand careful design and parameter tuning for optimal performance.

The research [20] aimed to find the best machine-learning method for analyzing sentiment in hotel reviews, specifically focusing on 5-star hotels in Istanbul. They manually collected 708 reviews and categorized them as positive or negative using four different machine learning techniques: LR, KNN, NB, and SVM. The results showed that LR achieved the highest accuracy at 92%, followed closely by the SVM at 90%. NB had an accuracy of 77%, while KNN lagged with 66%. These findings indicate that LR is the most effective method for sentiment analysis.

Table 1 summarizes the details of the proposed approaches that were presented in literature including, used dataset, rating method, machine learning approach, best result by each proposed model, and main objective of the research. This paper presents substantial contributions that extend previous research. Our research study distinguishes itself from other research works in the field due to its use of five different machine learning classifications (Multinomial NB, Bernoulli Naive Bayes, DT, RF, and SVM) along with the use of TF-IDF as a vectorization method in Natural Language Processing (NLP). The uniqueness of the research lies in the approach to rating classification. While other research works typically use two or three ML classifications with one or two rating scale scenarios, this research explores three kinds of rating scales with four scenarios. The dataset initially contains reviews with a 1–5 rating scale. The research starts by trying a 1–5 rating scale. Then, it minimizes the ratings to three sentiments: positive (1,2), neutral (3), and negative (4,5). Subsequently, it further minimizes the ratings to two sentiments: only positive and negative. In two sentiments, two scenarios are considered based on the key indicator "3." In the first scenario, rating 3 is considered negative, resulting in negative (1, 2, 3) and positive (4, 5)

Table 1 Summary of literature review

Study	Dataset	Rating	NLP methods	ML methods	Results
(Puh and Bagić 2023) [1]	TripAdvisor, 20491hotel reviews	5 and 3 ratings	TF-IDF and global vectors (GloVe)	NB, SVM, CNN, LSTM, and BiLSTM	BiLSTM achieved 72% accuracy for five classes and 89% for three classes
(Wadhe and Suratkar 2020) [12]	3209 reviews from Mouthshut and TripAdvisor	3 ratings	TF-IDF Vectorization and CountVectorization	NB, SVM, and RF	TF-IDF Vectorization and RF achieved the highest accuracy 86%
(Shi and Li 2011) [5]	4000 hotel reviews from hotel-review corpus by Tan S.H.	2 ratings	Unigram features and both frequency and TF-IDF	SVM	The SVM with unigram features and TF-IDF achieved an accuracy of 87.2%
(Dharma and Saragih 2022) [13]	311 hotel reviews from TripAdvisor	3 ratings	TF-IDF and BoW	SVM	SVM achieved an accuracy 71.75%
(Srivastava et al. 2022) [14]	20000 hotel reviews from TripAdvisor	3 ratings	TF-IDF, AFINN, VADER, and SGD optimizer	SVM and LR	SVM achieved an accuracy 96.3%
(İnan 2024) [20]	708 hotel reviews	2 ratings	N/A	SVM, LR, KNN, and NB	LR achieved the highest accuracy at 92%
(Ye et al. 2009) [15]	1191reviews	5 ratings	N-gram model	NB, SVM, and the SVM	The SVM and N-gram achieved accuracy greater than 80%
(Rai and Ahirwal 2018) [16]	450 positive and negative reviews on TripAdvisor	2 ratings	sentiment lexicon using Part-Of-Speech (POS) tags and SentiwordNet method	SVM, KNN, and RF	RF with accuracy 93.6%

(*Continued*)

Table 1 Continued

Study	Dataset	Rating	NLP methods	ML methods	Results
(Kulkarni et al. 2019) [17]	1 million product review from Amazon and 30000 reviews from Kaggle	2 ratings	N/A	Multinomial NB, RF, Bernoulli's NB, CNN, and RNN	RNN achieved the highest accuracy with 94.56%
(Farisi et al. 2019) [18]	4999 hotel reviews from finitis's Business	2 ratings	Bag of Words (BoW)	Multinomial Naïve Bayes	91% average F1-Score
(Li and Liu 2020) [19]	10000 hotel reviews from TripAdvisor	2 ratings	N/A	LR, NB, DT, RF, SVM, and Neural Network	The SVM achieved a 92% accuracy

sentiments. In the second scenario, rating 3 is considered positive, leading to negative (1,2) and positive (3,4,5) sentiments. Another contribution is the real dataset (JOHotelRating), which is derived from the Jordanian context.

To sum up, the research conducts four experimental scenarios, while others in the field typically use positive-negative or positive-neutral-negative scenarios. The ultimate goal is to explore different approaches to sentiment analysis and classification based on the rating scales provided in the dataset.

3 Methodology

Sentiment analysis has been conducted on the JOHotelRating dataset using a series of defined steps shown in Figure 1.

3.1 Dataset preparation

A two-step approach was taken to prepare the dataset for our research, ensuring its compatibility with machine learning approaches.

Step 1: Data acquisition

Tripadvisor.com was selected for data collection due to its popularity as a travel platform, providing reviews and information on hotels, restaurants, attractions, and more. Users can browse and share their insights through reviews, ratings, and photos. These insights can contribute diverse perspectives on travel destinations and services as shown in Tripadvisor [21]. Our

Figure 1 The proposed ML-based approach.

primary focus is on conducting a thorough sentiment review analysis tailored specifically to hotels in Jordan. In pursuit of this, we gathered a comprehensive dataset named JoHotelRating, drawing reviews from 22 distinct hotels in Jordan. The dataset is structured with three key columns: 'index,' 'review,' and 'rating,' resulting in a dataset comprising 33,575 rows. In three weeks of focused data collection, we ensured the completeness and accuracy of the JoHotelRating dataset, providing a valuable resource for evaluating sentiments in Jordanian hotel experiences. In Figure 2, the distribution of the hotel review ratings in JoHotelRating dataset is shown.

Step 2: Data preprocessing

In order to preprocess the textual dataset, we apply fundamental natural language processing (NLP) tasks using the NLTK library in Python. These tasks focus on converting reviews to lowercase, removing punctuation and numbers, and excluding stop words from the reviews. Additionally, the process

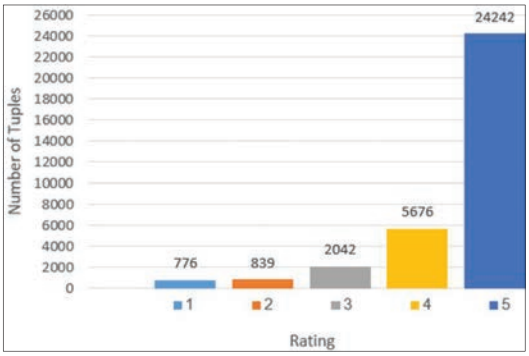


Figure 2 Distribution of the hotel review ratings in the JoHotelRating dataset.

involves applying tokenization to segment the reviews into individual words and implementing lemmatization to normalize the words. These procedures enhance the quality and uniformity of the textual dataset and prepare the data for subsequent analysis or modeling purposes. These procedures are executed through the NLTK library. NLTK text preprocessing is depicted in Figure 3 below.

3.2 Feature Extraction

In natural language processing (NLP), text data needs to be converted into a numerical representation for the application of machine learning algorithms [22]. TF-IDF (Term Frequency-Inverse Document Frequency) was applied.

Term Frequency (TF)

Term Frequency (TF) calculates the frequency of a term (t) within a document (D). TF indicates the significance of words within an individual document. As described in Equation (1), the Term Frequency (TF) is expressed as the count of term (t) occurrences in document (D) divided by the total number of words in document (D).

$$\boldsymbol{TF}(\boldsymbol{t}) = \frac{count(t)\ in\ D}{Total\ words\ in\ D} \tag{1}$$

Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) denotes the number of documents within a document set that include a specific term. IDF determines the significance of a word. For example, if the term 'X' is found in all documents in the set,

it is undoubtedly less informative. We can calculate the Inverse Document Frequency (IDF) using Equation (2).

$$IDF(w) = log\left(\frac{Total\ no.\ of\ docs}{No.\ of\ docs\ containing\ the\ term\ 't'}\right) \tag{2}$$

TF-IDF

From Equation (3), we can compute and express text as numbers in a meaningful way. The calculated values of TF and IDF in Equations (1) and (2) are the inputs for calculating the TF-IDF values as given in Equation (3).

$$TF - IDF = TF(t) * IDF\ (t) \tag{3}$$

3.3 Training the Model for Classification

The model was trained using 80% of the available dataset. In this study, classification algorithms such as linear support vector Machines, multinomial Naive Bayes, Bernoulli Naive Bayes, DT, and RF were applied to train the review dataset for research purposes. These algorithms are explained as follows:

3.3.1 Multinomial Naïve Bayes (MNB)

MNB is a text analysis algorithm that is particularly effective for problems with multiple classes. MNB is a popular classifier, which is commonly used for analyzing categorical text data in Natural Language Processing (NLP). The MNB algorithm works on a dataset that has a multinomial distribution. The MNB algorithm is performed through three steps, including using term frequencies as predictors, calculating the probability of each tag for a given sample, and outputting the tag with the highest probability [23, 24]. Equation (4) describes the mathematical basis of MNB, where A represents the hypothesis and B indicates the evidence [25].

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{4}$$

3.3.2 Bernoulli Naïve Bayes (BNB)

The BNB is an advancement version of the Naive Bayes (NB) Algorithm, works as similar as the Multinomial classifier. The BNB classifier is designed for classifying binary features to determine whether a particular term appears or does not appear in a particular document. Furthermore, BNB plays a

crucial role in text classification tasks [24, 26]. Equation (5) describes the conditional probability $P(x_i|y)$ that calculates the likelihood of x_i occurring, given the occurrence of event y. where i denotes the specific event and x_i takes on a binary value of either 0 or 1 [25].

$$P(x_i|y) = P(x_i = 1|y)x_i + (1 - P(x_i = 1|y))(1 - x_i) \tag{5}$$

3.3.3 Decision Tree (DT)

DT is a supervised learning technique that forms a tree structure where decisions depend on features in the dataset. Nodes represent these features, and branches show decision rules. DT helps to find an understandable and easier solution to interpret. The leaf nodes represent the outcomes and make them the ultimate results [27].

3.3.4 Random Forest (RF)

RF, a supervised learning algorithm, operates on the principle of ensemble learning, where multiple classifiers are combined to address complex problems and enhance model performance. Rather than relying on a single decision tree, Random Forest aggregates predictions from each tree and, by considering the majority of votes, makes the final output prediction [10].

3.3.5 Support Vector Machine (SVM)

The SVM is a supervised algorithm that aims to identify the optimal hyperplane (defined as the most effective decision boundary) within an N-dimensional space. This hyperplane is designed to segregate data points into distinct classes within the feature space. The objective is to maximize the margin between the nearest points of the different classes [28]. The Equation (6) defines the hyperplane that separates the input space can be represented as

$$W^T X_i + b = 0 \tag{6}$$

Where, w represents a weight vector perpendicular to the hyperplane. x is the training data and b is the bias term.

3.4 Testing and Visualization the Proposed ML-based Approach

Testing has been done on 20% of the dataset. The effectively trained model can categorize the sentiment of each review into two groups: positive and negative. The distribution of predicted sentiments in the dataset was visualized using Matplotlib's pyplot module. Sentiments were categorized as

positive and negative based on the model predictions. Valuable insights into the predicted sentiment distribution of the dataset were provided through this visualization, contributing to a comprehensive analysis of sentiment predictions.

3.5 Performance Evaluation

To evaluate the performance of classification models in our proposed ML-based approaches, various performance measures are employed, including accuracy score, precision, recall, and f1-score. According to [31], the precision score of each ML-based approach can be computed as shown in Equation (7). In addition, the value of F1-score is calculated as shown in Equation (8). Furthermore, the recall score is measured as shown in Equation (9) and accuracy score is derived based on Equation (10). We further explain that False Positive (FP) denotes that the ML algorithm predicts a positive result (hotel rating) incorrectly, while False Negative (FN) denotes that a ML algorithm predicts a negative result (hotel rating) incorrectly. Conversely, True Positive (TP) indicates correct positive results (hotel rating), and True Negative (TN) denotes correct negative results (hotel rating).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{8}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{9}$$

$$\text{F1} - \text{Score} = \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{10}$$

4 Experiments

In this section, the outcomes of four experiments conducted within the dataset, featuring different scenarios for the rating feature, will be discussed.

4.1 Experimental Setup

In conducting this experiment, the Python programming language was employed [32] on a Windows 10 Pro 64-bit operating system. The hardware

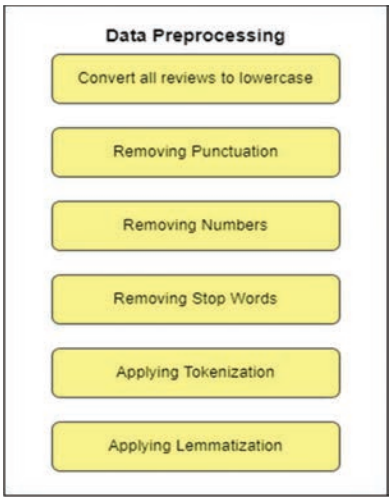


Figure 3 The NLTK text preprocessing.

and software requirements are detailed as follows:

Hardware Requirement:
Processor: Intel(R) Core(TM) i3-8130U CPU @ 2.20GHz (2.20 GHz)
Memory: 8 GB Random Access Memory (RAM)
Storage: 476GB Hard Disk

Software Requirement:
Python Environment: Utilizing Google Colab
Web Scraping Tool: Octoparse

4.2 Experimental Scenarios

The dataset initially comprises reviews with ratings ranging from 1 to 5. Initially, we trained a model with a five-class classification system, where each rating represented a distinct class. Subsequently, we explored different strategies for simplifying the classification task. Initially, we condensed the ratings into three classes: Negative (1, 2) and, Neutral (3), and Positive (4, 5). We trained the model with this modified three-class setup and recorded the corresponding accuracy.

Further simplification was attempted by reducing the classes to two: Negative (1, 2, 3) and Positive (4, 5). The model was trained and evaluated under this binary classification. Additionally, we experimented with an alternative

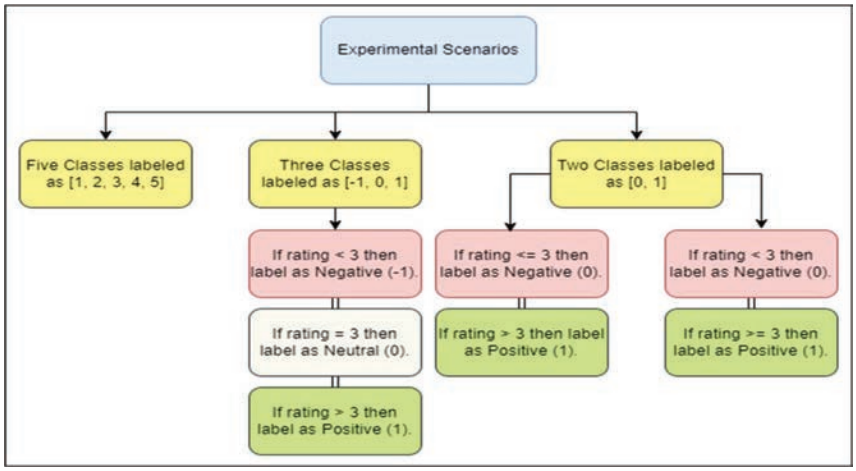


Figure 4 The four experimental scenarios.

binary classification, categorizing Negative (1, 2) and Positive (3, 4, 5). The four experimental scenarios we navigated are depicted in Figure 4.

4.3 Implementation

In the implementation phase, the data acquisition process focused on leveraging the popularity of TripAdvisor.com as a travel platform for collecting relevant data. A dedicated dataset, JoHotelRating, was meticulously curated, encompassing reviews from 22 hotels in Jordan with key columns including 'index,' 'review,' and 'rating,' totaling 33,578 rows. Subsequently, data preprocessing tasks were executed using the NLTK library in Python. This involved converting reviews to lowercase, removing punctuation and numbers, excluding stop words, tokenizing the text for segmentation into individual words, and lemmatizing to normalize words. The outcome of these preprocessing steps was an improvement in the quality and uniformity of the text data, laying a solid foundation for subsequent analysis and modeling.

TF-IDF was employed for feature extraction, transforming the textual data into a numerical representation that is well-suited for application in machine learning algorithms within the realm of natural language processing (NLP). The training of the model involved an 80-20 data split, with various classification algorithms like linear support vector machines, multinomial Naive Bayes, Bernoulli Naive Bayes, decision tree, and random forests applied for research purposes. Testing was conducted on the remaining 20%

of the data, leading to the development of an effectively trained model capable of categorizing sentiment into two groups: positive and negative. Both the classification and feature extraction stages were seamlessly executed using the Python scikit-learn (sklearn) library.

In experimental scenarios, the initial classification employed a five-class system, followed by simplified classifications condensing ratings into three and two classes for binary classification. The model's performance was evaluated using various parameters such as accuracy, recall, precision, and f1-score.

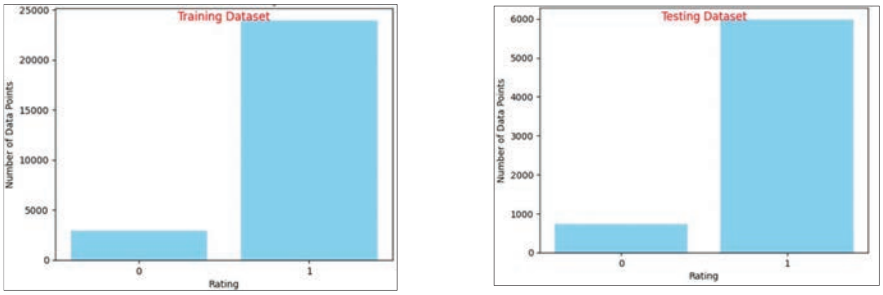
5 Results and Discussion

As mentioned earlier, we conducted four experimental scenarios, each with distinct parameters. In the first scenario, where a five-class system was employed, the results were as follows: MNB, DT, and RF exhibited identical accuracy rates of 73%, while BNB displayed a slight variation with 72%. Notably, SVM emerged as the top-performing model that achieves the highest accuracy (78%) among the other selected machine learning algorithms. In the second scenario, we utilized three rating classes, including 5 and 4 as positive, 3 as neutral, and 1 and 2 as negative. Notable improvements were observed compared to the first scenario. Both MNB and RF classifiers achieved an impressive accuracy of 89%, while BNB closely followed with an accuracy of 86%. DT exhibited notable performance and outperformed all other models, achieving a remarkable accuracy of 93%. This enhancement in accuracy underscores the effectiveness of the refined classification approach based on the three rating classes. In the last two experimental scenarios, we used a binary classification by simplifying the rating classes into two categories (Negative and Positive). In the first scenario, we consider (1, 2, 3) as a Negative and (4, 5) as a Positive. Subsequently, in the second scenario, we altered the classification to Negative (1, 2) and Positive (3, 4, 5) by changing the position of the rating 3 within the classes.

The outcomes differed significantly for both approaches. MNB achieved 89% accuracy with 3 considered as Negative, whereas it reached an improved accuracy of 95% when 3 was classified as Positive. Similarly, BNB displayed differences in accuracy, achieving 89% with 3 as Negative and 93% with 3 as Positive. Both RF and DT models achieved 91% accuracy in the first approach and improved to 95% when 3 was categorized as Positive. Notably, Support Vector Machine exhibited 95% accuracy with 3 as Negative and a higher accuracy of 97% with 3 as Positive. The impact of rating 3 on the results

Table 2 The accuracy scores achieved by each ML-based model for the four experimental scenarios

			Two Classes	
ML	Five Classes	Three Classes	3 as Negative	3 as Positive
MNB	73%	89%	89%	95%
BNB	72%	86%	88%	93%
DT	73%	90%	91%	95%
RF	73%	89%	91%	95%
SVM	78%	93%	95%	97%

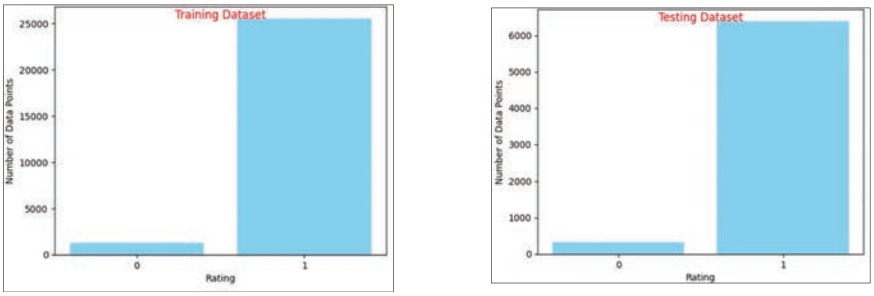


(a)Training dataset (b) Testing dataset

Figure 5 The distribution of testing and training datasets for the scenario where 3 is considered negative in (a) and (b).

prompted a decision to further investigate the binary classification with 3 considered as part of the Positive class. This recognition emphasizes the significance of thoughtful class definition in binary classification scenarios, as it can markedly influence the model's performance.

Table 2 provides a comprehensive summary of model accuracies in the four experimental scenarios. Significantly, the binary classification, designating 3 as positive, consistently achieved the highest accuracies across all scenarios. As a result, we will focus our research efforts on exploring this particular classification scenario. In Figure 5, the test dataset and training dataset are visualized for the scenario where 3 is classified as negative, while Figure 6 showcases the same datasets with 3 considered as positive. It is noteworthy that reviews with a rating of 3 are relatively scarce, yet their impact on accuracy is substantial when classified as positive. This observation underscores the sensitivity of model outcomes to the classification of this specific rating, emphasizing the importance of careful consideration in defining class labels for optimal model performance.



(a)Training dataset (b) Testing dataset

Figure 6 The distribution of testing and training datasets for the scenario where 3 is considered positive in (a) and (b).

Table 3 Results interpretation with rating 3 as negative

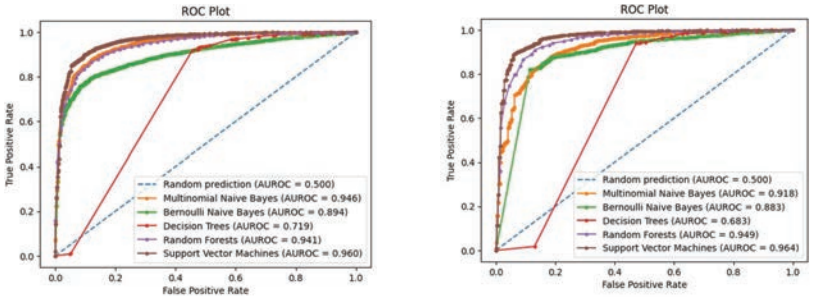
	Precision		Recall		F1-score		Accuracy
Rating	Negative	Positive	Negative	Positive	Negative	Positive	
MNB	1.00	0.89	0.03	1.00	0.06	0.94	**0.89**
BNB	0.45	0.94	0.54	0.92	0.49	0.93	**0.88**
DT	0.65	0.92	0.32	0.98	0.43	0.95	**0.91**
RF	0.95	0.90	0.14	1.00	0.24	0.95	**0.91**
SVM	0.83	0.96	0.63	0.98	0.72	0.97	**0.95**

Table 4 Results interpretation with rating 3 as Positive

	Precision		Recall		F1-score		Accuracy
Rating	Negative	Positive	Negative	Positive	Negative	Positive	
MNB	0.00	0.95	0.00	1.00	0.00	0.98	**0.95**
BNB	0.34	0.97	0.46	0.96	0.39	0.96	**0.93**
DT	0.53	0.97	0.30	0.99	0.39	0.98	**0.95**
RF	0.86	0.95	0.04	1.00	0.07	0.98	**0.95**
SVM	0.78	0.98	0.54	0.99	0.63	0.98	**0.97**

Tables 3 and 4 showcase classification algorithm performance metrics, including Precision, Recall, F1-score, and accuracy. Table 3 interprets results with a rating of 3 as negative, while Table 4 interprets results with a rating of 3 as positive. SVM achieved the highest accuracy at 97%, and DT, MNB, and RF shared an accuracy of 95%, each with distinct Precision, Recall, and F1-score outcomes. BNB achieved an accuracy of 93%.

In terms of precision, Table 3 depicts that the MNB and SVM classifiers have the highest scores: 100% on negative reviews and 96% on positive



(a) Rating 3 as a negative

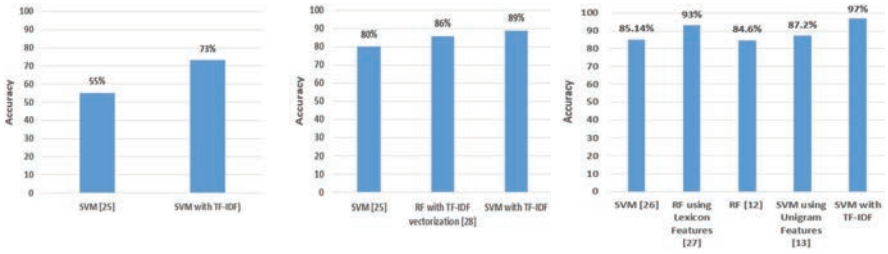
(b) Rating 3 as a positive

Figure 7 ROC curve of five machine learning algorithms when rating 3 as a negative (a) and rating 3 as a positive (b).

reviews, respectively. Furthermore, the results of MNB and RF classifiers in terms of recall metrics show that both of them have optimal values of 100% on positive reviews. The best F1-score achieved by SVM using the positive reviews. From Table 4, the SVM classifier achieved the highest results in terms of precision at 98% on the positive reviews. MNB and RF have an optimal recall score of 100%. The five classifiers have an F1-score of at least 96%. From Table 4, the results prove that considering 3 as positive affects significantly the performance of the proposed model.

We have comprehensively covered all scenarios, achieving the highest accuracy compared to other related works. Notably, our SVM model outperformed others in terms of accuracy. It's worth mentioning that some prior works focused on either 2-rating or 3-rating classification, while we explored a broader range. Additionally, a notable observation is that many related works gathered their datasets through self-construction. Figure 8 shows the ROC curves of five machine learning algorithms when rating 3 as a negative and rating 3 as a positive. The SVM shows superiority over other machine learning algorithms.

Figure 7 and Table 5 provide a comparative analysis of machine learning approaches with related works. In Figure 7, the accuracy score of our approach (SVM with TF-IDF) shows superiority compared to the study [25] in classifying hotel reviews into five rating classes. Moreover, our proposed approach achieved a higher accuracy score than the studies [25] and [28] in classifying hotel reviews into three rating classes. Our proposed approach shows significant enhancement in its performance in terms of accuracy scores in classifying hotel reviews into two rating classes compared to the state-of-the-art studies such as [12, 13, 26], and [27].



(a) Five classes (b) Three classes (c) Two classes

Figure 8 Illustration of a comparative analysis of machine learning approaches alongside related works.

Table 5 A comparative analysis of machine learning approaches with related works

Ref	ML Methods	Dataset	Five Classes	Three Classes	Two Classes
(Puh and Bagić 2023) [1]	NB, SVM, CNN, LSTM, and BiLSTM	TripAdvisor Hotel Review Dataset	SVM gives **55%**	SVM gives **80%**	N/A
(Ye et al. 2009) [15]	NB, SVM, and character-based N-gram mode	Self-constructed	N/A	N/A	SVM gives **85.14%**
(Rai and Ahirwal 2018) [16]	SVM, KNN, and RF by using Lexicon Features	Self-constructed	N/A	N/A	RF using Lexicon Features gives **93%.**
(Wadhe and Suratkar 2020) [12]	NB, SVM, and RF	Self-constructed	N/A	TFIDF Vectorization + RF gives **86%**	N/A
(Kulkarni et al. 2019) [17]	MNB, BNB, and RF	Self-constructed	N/A	N/A	RF gives **84.60%,**
(Shi and Li 2011) [5]	SVM with unigram feature and TF-IDF	hotel-review	N/A	N/A	SVM using unigram feature with (TF-IDF) gives **87.2%**
The proposed Model	**SVM, MNB, BNB, DT, and RF.**	**Self-constructed**	**SVM with (TF-IDF) gives 73%**	**SVM with (TF-IDF) gives 89%**	**SVM with (TF-IDF) gives 97%**

In Table 5, we compare the performance of our proposed approach to the studies that have the following criteria: The studies focused on hotel reviews and used two, three, or five rating classes. In addition, the studies were constructed using machine learning algorithms such as SVM, NB, KNN, and

RF. Furthermore, the datasets that were used in the previous studies have the same features that formulate our dataset. And finally, the studies evaluated the performance of their proposed approaches using the accuracy metric. For the three rating scenarios, our proposed approach shows superiority compared to the reported studies in Table 5.

To sum up, ML based solutions can significantly benefit the tourism domain, particularly in the analysis of hotel reviews. Therefore, we believe that the proposed machine learning based model for sentiment analysis in tourism context can be part of or leads to several effective practical applications such as, First, providing a review summarization [29], where machine learning models can be trained to summarize lengthy hotel reviews, providing a concise overview of customer feedback. This can be useful for both hotel management and potential guests who want quick insights into the overall sentiment. Second, developing chatbots for customer service, the integration between NPL and ML enables the development of chatbots that can engage with customers, answer queries, and provide information about hotel facilities, policies, and local attractions. Third, detecting the fraud reviews [30], where the ML algorithms can be employed effectively to detect fraudulent reviews or ratings. This ensures that the feedback influencing potential guests is genuine, leading to better decision-making [33]. Fourth, ML based solution can analyze historical booking data, seasonal trends, and external factors to predict future demand for hotel rooms. This may help to optimize pricing, allocate resources efficiently, and plan for peak periods. Finally, analyzing the competitors, where ML based solutions can help the hotels owners to understand their strengths and weaknesses in comparison to others in the tourism market.

The contribution of this study concentrates on sentiment analysis in the context of the tourism industry. The solutions introduced based on sentiment analysis using ML and NPL methods can serve the tourism field in many directions. For example, the proposed model can help hotel management engage with their customers more effectively through understanding sentiments. Hotels can respond to positive feedback and address concerns expressed in negative feedback. This leads to a positive relationship with customers. Furthermore, the proposed model can be applied to other services and products associated with the tourism industry to manage the business reputation, enhance the quality of services and products, analyze the market, and research new trends.

6 Conclusion

Our research focused on comprehending sentiments in Jordanian hotel reviews to assist both hotel owners and tourists. This involved considering various scenarios and utilizing Rating 3 as a key indicator. Employing machine learning and natural language processing techniques, specifically SVM with TF-IDF for feature extraction, resulted in an impressive 97% accuracy in sentiment classification for two classes (positive and negative). These findings contribute valuable insights into the effective analysis of sentiments in hotel reviews in Jordan, providing a foundation for enhancing the understanding of customer experiences in the hospitality industry. One of potential future directions is that focusing on the aspect-based sentiment analysis. Instead of providing an overall sentiment, ML based solutions can examine specific aspects mentioned in reviews, such as cleanliness, service, amenities, and location. This granular analysis assists hotel owners in pinpointing areas in need of enhancement.

Another potential avenue for future research involves employing the ML and deep learning techniques along with computer vision tools, to analyze images and videos associated with hotel reviews. This might entail locating particular amenities such as, pool or room view that mentioned in reviews in order to provide a more comprehensive understanding of customer experiences.

Conflict of Interest

The authors have no relevant financial or non-financial interests to disclose.

References

[1] Puh, K., and Bagić Babac, M. (2023). Predicting sentiment and rating of tourist reviews using machine learning. Journal of Hospitality and Tourism Insights, 6(3), 1188–1204. https://doi.org/10.1108/JHTI-02-2022-0078.

[2] Liu, Y., Ding, X., Chi, M., Wu, J., and Ma, L. (2024). Assessing the helpfulness of hotel reviews for information overload: A multi-view spatial feature approach. Information Technology & Tourism, 26(1), 59–87. https://doi.org/10.1007/s40558-023-00280-x.

[3] Chen, W. (2024). Exploring the Dynamics of Electronic Word-of-Mouth in Chinese Tourism: A Social Network Perspective. Journal of the Knowledge Economy, 1–23. https://doi.org/10.1007/s13132-024-01780-9.

[4] Akbar, A. R., Kalis, M. C. I., Afifah, N., Purmono, B. B., and Yakin, I. (2023). The Influence of Product Packaging Design and Online Customer Review on Brand Awareness and Their Impact on Online Purchase Intention. South Asian Res J Bus Manag, 5(1), 10–18.

[5] Shi, H. X., and Li, X. J. (2011, July). A sentiment analysis model for hotel reviews based on supervised learning. In 2011 International Conference on Machine Learning and Cybernetics (Vol. 3, pp. 950–954). IEEE.

[6] Rodrigues, V., Eusébio, C., and Breda, Z. (2023). Enhancing sustainable development through tourism digitalisation: a systematic literature review. Information Technology & Tourism, 25(1), 13–45. https://doi.org/10.1007/s40558-022-00241-w.

[7] Wang, W. (2023). Design of cloud computing database and tourism intelligent platform based on machine learning. Soft Computing, 1–9. https://doi.org/10.1007/s00500-023-08642-7.

[8] Hartmann, J., Heitmann, M., Siebert, C., and Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing, 40(1), 75–87.

[9] Taherdoost, H., and Madanchian, M. (2023). Artificial intelligence and sentiment analysis: A review in competitive research. Computers, 12(2), 37. https://doi.org/10.3390/computers12020037.

[10] Vargas-Calderón, V., Moros Ochoa, A., Castro Nieto, G. Y., and Camargo, J. E. (2021). Machine learning for assessing quality of service in the hospitality sector based on customer reviews. Information Technology & Tourism, 23, 351–379. https://doi.org/10.1007/s40558-021-00207-4.

[11] Ministry of Tourism and Antiquities. Number of Classified Hotels in Jordan: 1998-2021 (2023) CEIC Data. https://www.ceicdata.com/en/jordan/tourist-accommodation-establishments-statistics/number-of-classified-hotels.

[12] Wadhe, A. A., and Suratkar, S. S. (2020, February). Tourist place reviews sentiment classification using machine learning techniques. In 2020 international conference on Industry 4.0 Technology (I4Tech) (pp. 1–6). IEEE.

[13] Dharma, A. S., and Saragih, Y. G. R. (2022). Comparison of Feature Extraction Methods on Sentiment Analysis in Hotel Reviews. Sinkron: jurnal dan penelitian teknik informatika, 7(4), 2349–2354. https://doi.org/10.33395/sinkron.v7i4.11706.

[14] Srivastava, R., Bharti, P. K., and Verma, P. (2022). Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis. International Journal of Advanced Computer Science and Applications, 13(3). https://doi.org/10.14569/IJACSA.2022.0130312.

[15] Ye, Q., Zhang, Z., and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised Machine learning approaches. Expert systems with applications, 36(3), 6527–6535. https://doi.org/10.1016/j.eswa.2008.07.035.

[16] Rai, P., and Ahirwal, R. (2018). Tourism Review Sentiment Analysis using Lexicon Features and Machine Learning Approach. E ISSN, 2348-1269.

[17] Kulkarni, A., Barve, P., and Phade, A. (2019). A machine learning approach to building a tourism recommendation system using sentiment analysis. International Journal of Computer Applications, 178, 48–51.

[18] Farisi, A. A., Sibaroni, Y., and Al Faraby, S. (2019, March). Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012024). IOP Publishing.

[19] Li, X., and Liu, C. (2020, April). Comparison of Machine Learning Models for Sentimental Analysis of Hotel Reviews. In IOP Conference Series: Materials Science and Engineering (Vol. 806, No. 1, p. 012029). IOP Publishing.

[20] İnan, H. E. (2024). Comparison of Machine Learning Algorithms for Classification of Hotel Reviews: Sentiment Analysis of TripAdvisor Reviews. GSI Journals Serie A: Advancements in Tourism Recreation and Sports Sciences, 7(1), 111–122.

[21] Tripadvisor (2023). https://www.tripadvisor.com, last accessed in 15/11/2024.

[22] Xiang, Z., Du, Q., Ma, Y., and Fan, W. (2018). Assessing reliability of social media data: lessons from mining TripAdvisor hotel reviews. Information Technology & Tourism, 18, 43–59. https://doi.org/10.1007/s40558-017-0098-z.

[23] Sharupa, N. A., Rahman, M., Alvi, N., Raihan, M., Islam, A., and Raihan, T. (2020, July). Emotion detection of Twitter post using multinomial Naive Bayes. In 2020 11th International Conference on Computing,

Communication and Networking Technologies (ICCCNT) (pp. 1–6). IEEE.

[24] Sarang, P. (2023). Naive Bayes: A Supervised Learning Algorithm for Classification. In Thinking Data Science: A Data Science Practitioner's Guide (pp. 143–152). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-02363-7_7.

[25] Mahmoud Masadeh, Moustapha. A, Sharada B, Hanumanthappa J, Hemachandran K, Channabasava Chola and Abdullah Y. Muaad, Investigating the Impact of Preprocessing Techniques and Representation Models on Arabic Text Classification using Machine Learning, International Journal of Advanced Computer Science and Applications (IJACSA), 15(1), 2024. http://dx.doi.org/10.14569/IJACSA.2024.01501110.

[26] Tandon, V., and Mehra, R. (2023). An Integrated Approach For Analysing Sentiments On Social Media. Informatica, 47(2). https://doi.org/10.31449/inf.v47i2.4390.

[27] Priyanka, and Kumar, D. (2020). Decision tree classifier: a detailed survey. International Journal of Information and Decision Sciences, 12(3), 246–269. https://doi.org/10.1504/IJIDS.2020.108141.

[28] Chory, R. N., Nasrun, M., and Setianingsih, C. (2018, November). Sentiment analysis on user satisfaction level of mobile data services using Support Vector Machine (SVM) algorithm. In 2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS) (pp. 194–200). IEEE.

[29] Siautama, R., IA, A. C., and Suhartono, D. (2021). Extractive hotel review summarization based on TF/IDF and adjective – Noun pairing by considering annual sentiment trends. Procedia Computer Science, 179, 558–565. https://doi.org/10.1016/j.procs.2021.01.040.

[30] Cai, M., Du, Y., Tan, Y., and Lu, X. (2023). Aspect-based classification method for review spam detection. Multimedia Tools and Applications, 1–22. https://doi.org/10.1007/s11042-023-16293-x.

[31] Alemerien, K., Alsarayreh, S., and Altarawneh, E. (2024). Diagnosing Cardiovascular Diseases using Optimized Machine Learning Algorithms with GridSearchCV. Journal of Applied Data Sciences, 5(4), 1539–1552.

[32] Shrivastava, A. (2024). A Deep Learning model based on CNN using Keras and TensorFlow to determine real time melting point of chemical substances. ELCVIA Electronic Letters on Computer Vision and Image Analysis, 23(1), 47–67.

[33] Gupta, S., Singhal, N., Hundekari, S., Upreti, K., Gautam, A., Kumar, P., and Verma, R. (2024). Aspect Based Feature Extraction in Sentiment Analysis using Bi-GRU-LSTM Model. Journal of Mobile Multimedia, 20(4), 935–960.

Biographies

Khalid Alemerien is an associate professor of Software Engineering/Computer Science at Information Technology Department in the college of ICT, Tafila Technical University (TTU), Jordan. Dr. Alemerien was awarded his Masters and Ph.D. in Software Engineering/Computer Science from North Dakota State University, USA in 2013 and 2014, respectively. His research interests focus on Software Engineering, Usable Security and Privacy, Machine Learning and Deep Learning, Informatics, E-learning, Privacy and Security in IoT-based Systems. Dr. Alemerien has published numerous research papers in prestigious journals and conference proceedings.

Aram Al-Ghareeb holds a bachelor degree in computer science. She is currently pursuing a master's in Cyber Physical Systems (CPSs) at Tafila

Technical University, Jordan. Her research interests focus on CPSs, IoT based Systems, Machine Learning, and Computer Education.

Malek Zakarya Alksasbeh is currently a Full Professor in the Faculty of Information Technology at Al-Hussein Bin Talal University, Ma'an, Jordan. He received his B.S. degree in Computer science from Mu'tah University, Jordan, in 2005, and the M.S. and Ph.D. degrees in Information Technology from the University Utara Malaysia (UUM), Malaysia, in 2008 and 2012, respectively. His research interests are in the areas of Smart Systems, Information Retrieval, Deep learning, and Instructional Technology.