RESEARCH ARTICLE

Detecting Sentiment Analysis Correlation Between Hotel Employee and Customer Reviews Using Machine Learning Algorithms

Berat Bakan[1], **Abdullah Ammar Karcioglu**[1,*] **and Esra Katırcıoğlu**[2]

[1] Department of Software Engineering, Ataturk University, Erzurum 25240, Türkiye
[2] Department of Tourism Management, Ondokuz Mayıs University, Samsun 55270, Türkiye

Abstract

Analyzing heterogeneous online reviews from multiple stakeholder groups represents an emerging challenge in AI-driven service intelligence. This study proposes a two-stage sentiment correlation detection framework and applies it to customer and employee reviews of the Istanbul Marriott Şişli hotel. During dataset construction, a semi-supervised domain-specific blacklisting approach was developed alongside standard preprocessing steps to improve sentiment signal quality. In the first phase, customer and employee reviews were treated as separate datasets, and 5-fold cross-validation was applied using TF-IDF, BOW, and Word2Vec representations with multiple classifiers, achieving 99.8% and 93.8% F1-scores respectively. In the second phase, correlation analysis on the combined dataset yielded a 97.1% F1-score. In the final phase, LDA-based topic modeling identified three topic clusters, with the topic-based model achieving a 97.5% F1-score. The findings demonstrate that customer and employee perceptions converge at the macro level but diverge on a topic-by-topic basis. The proposed framework offers a generalizable approach for multi-source sentiment correlation analysis in domain-specific AI applications.

Keywords: sentiment analysis, multi-source correlation detection, semi-supervised preprocessing, topic modeling, service intelligence.

1 Introduction

The automated analysis of heterogeneous user-generated text from multiple stakeholder groups represents an emerging challenge in natural language processing and AI-driven decision support systems. In the tourism and hospitality sector, online user reviews are considered a strategic data source for evaluating service quality and customer satisfaction. Sentiment analysis studies, particularly those focusing on hotel reviews, have been extensively covered in the literature [1]. In this context, analyzing online reviews using machine

Submitted: 11 March 2026
Revised: 11 May 2026
Accepted: 01 July 2026
Published: 22 September 2026

Vol. 3, **No.** 3, 2026.
10.62762/TETAI.2026.991738

***Corresponding author:**
✉ Abdullah Ammar Karcioglu
ammar.karcioglu@atauni.edu.tr

Citation
Bakan, B., Karcioglu, A. A., & Katırcıoğlu, E. (2026). Detecting Sentiment Analysis Correlation Between Hotel Employee and Customer Reviews Using Machine Learning Algorithms. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 3(3), 188–201.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

learning methods has become a widespread research area for understanding consumer behavior and decision-making processes [2]. However, these reviews contain heterogeneous linguistic features such as spelling errors, repetitive expressions, emoji usage, and domain-specific terminology. This situation has complicated the analysis process in the texts [3]. Systematically analyzing large volumes of unstructured text data requires automated sentiment analysis approaches [4].

Sentiment analysis is defined as a method that aims to classify the emotional orientation of texts through natural language processing and machine learning techniques [5], highlighting the decisive role of linguistic variations and contextual features on model performance [6]. Although customer review-based studies are widely available in the literature [2], it has been noted that research that addresses employee feedback within the scope of sentiment analysis is limited [7]. In this study, considering this gap, an integrated framework is proposed that combines a semi-supervised domain-specific blacklisting approach for preprocessing with a two-stage sentiment correlation detection model, contributing to the emerging area of multi-source AI-driven opinion analysis in service domains. The study is structured to include an introduction, literature review, methodology, findings, and conclusion sections.

2 Related Work

The increase in user reviews shared on online platforms has made sentiment analysis research a significant area of study, particularly in the tourism and hospitality sector. Research examining the impact of online reviews on decision-making processes has shown that this data directly affects both consumer behavior and the strategic positioning of businesses [1, 2]. In this context, systematic analysis of large volumes of unstructured text data is critical for evaluating service quality and understanding customer experience.

In early applications, classical machine learning algorithms such as TF-IDF and Bag of Words representations, SVM, Logistic Regression (LR), and Random Forest were used. AUC=0.89 was obtained in Booking.com data [8], 87–92% accuracy in TripAdvisor analyses [9], and 85% accuracy in the Indonesian sample [10]. Thus, it has been shown that classical methods offer interpretable and low-cost solutions.

CNN, LSTM, and GRU models were integrated with Word2Vec and similar embedding techniques to enhance contextual representation. Accuracy of 98.08% [11] was reported in Traveloka data, deep learning-based improvements in Turkish hotel reviews [12], and 97.9% F1-score [13] in systems integrated into topic modeling. Aspect-based LSTM approaches achieved an F1-score of 0.75 on hotel review summarization [14], and Multilevel sentiment analysis using CNN and LSTM achieved F1-scores of 0.95 and 0.87 at document and aspect levels, respectively [15]. Transformer-based models provided significant performance improvements; 0.93 F1-score [16] was reported in BERT-based analyses, 87.2% F1-score in machine and deep learning comparison on tourist reviews [17], and 99.4% accuracy in Turkish big language model applications [18].

Integrating sentiment analysis with topic modeling has enabled the separation of service components [19]. 94.2% F1-score [20] was obtained in service area-based classifications, holistic evaluations in multi-criteria decision-making integration with sentiment analysis [21], systematic sentiment distributions in theme and aspect-based analyses [22] and 97% accuracy in employee review sentiment classification using hybrid voting classifiers with TF-IDF and BoW [23].

Although customer reviews have been extensively studied in the literature, sentiment analysis studies on employee feedback are limited. In an employee-focused study, a 93.7% F1-score was reported, and the correlation between organizational commitment and sentiment orientation was analyzed [7]. These assessments show that holistic studies that address customer and employee data in a comparative and unified sentiment space are limited; the present study aims to fill this gap. From an AI methodology perspective, this work introduces a structured pipeline for multi-source sentiment correlation that can be extended to other service domains and stakeholder configurations. Studies carried out in this field are shown comparatively in Table 1 in terms of dataset type, preprocessing and feature extraction techniques used, classification algorithms, and performance metrics.

3 Materials and Methods

In this study, a multi-stage methodological framework is presented for the joint sentiment analysis of heterogeneous review sources and the identification of cross-source perceptual correlations. The framework

Table 1. Detailed literature comparison.

Study	Dataset	Feature Extraction	Model / Method	Results
[7]	Employee review	NLP	Machine Learning	F1-Score = 0.937
[8]	Booking.com(TR)	TF-IDF	Random Forest	AUC = 0.89
[9]	TripAdvisor	BoW, TF-IDF	SVM, LR	Accuracy = 0.87–0.92
[10]	Indonesian Hotel	TextBlob+TF-IDF	SVM	Accuracy = 0.85
[14]	Traveloka Hotel	LSTM	Aspect-based LSTM	F1-Score = 0.753
[11]	Traveloka	Word2Vec(Skip-gram)	CNN	Accuracy = 0.9808
[12]	Turkish hotel	Word2Vec	GRU	F1-Score = 0.92
[13]	Antalya (15k+)	Topic Modeling + DL	Deep Learning	F1-Score = 0.979
[15]	Indonesian Hotel	CNN + LSTM	Multilevel DL	F1-Score = 0.95 (doc); 0.87 (aspect)
[16]	715 hotel review	BERT	Transformer	F1-Score = 0.93
[17]	Tourist reviews	TF-IDF, BoW	NB, SVM, CNN, LSTM	F1-Score = 0.872
[18]	Turkish Dataset	LLM (TURNA)	LLMs	Accuracy = 0.994
[20]	Antalya Hotel	Text Mining	Classifications	F1-Score = 0.942
[22]	Hotel reviews	Tema + Aspect	Integrated analysis	F-Score = 0.87
[23]	Employee reviews	TF-IDF, BoW, GloVe	RV-SGDC (LR+SVM+SGD)	Accuracy = 0.97
Our Study	Şişli Marriott Hotel	N-gram+Word2Vec	Linear SVC, LR, Perceptron, Extra Trees, XGBoost	F1-Score 0.998 (single); 0.971 (combined); 0.975 (topic)

is instantiated on hotel customer and employee reviews, though its design is generalizable to other multi-stakeholder text analysis scenarios. The proposed method includes domain-based preprocessing of raw text data, comparative evaluation of different text representation approaches, and systematic testing of machine learning models that allow for both direct and topic-based correlation detection. This section describes the data structure, preprocessing process, vectorization methods used, classification algorithms, correlation detection approaches, and evaluation criteria in detail.

3.1 Dataset

The datasets used in this study consist of hotel customer reviews and hotel employee reviews obtained from online platforms. Both datasets include user ratings from 1 to 5 and corresponding textual comments. The sentiment analysis problem was addressed using binary classification. In this context, reviews with scores of 1 and 2 were labeled negative, reviews with scores of 4 and 5 were labeled positive, and reviews with a score of 3, which could be considered neutral, were excluded from the analysis. This approach was preferred to reduce ambiguity between sentiment classes and to increase the discriminative power of the classification models. After label assignment, the customer review dataset comprised 2,891 positive and 461 negative samples, while the employee review dataset comprised 2,614 positive and 592 negative samples. After determining the sentiment labels of employee and customer reviews, they were analyzed both separately and as a combined dataset. This made it possible to compare the model performances under different scenarios. The characteristics of the datasets are shown in Table 2.

Table 2. Dataset statistics and composition.

Dataset	# of Reviews	Contents
Customer Reviews	3,352	Reviews, Score
Employee Reviews	3,206	
Total		**6,558**

3.2 The Preprocessing Process

Raw text data is noisy, disordered, and unstructured, and directly introducing this data to the model can hinder the learning process. Therefore, a multi-stage preprocessing process was applied to make the dataset suitable for model training and to obtain more reliable results. The preprocessing stages are shown in Figure 1.

First, all texts were converted to lowercase, and punctuation and special characters were removed to minimize stylistic differences. Then, instead of deleting emojis and similar emotional expressions in comments,

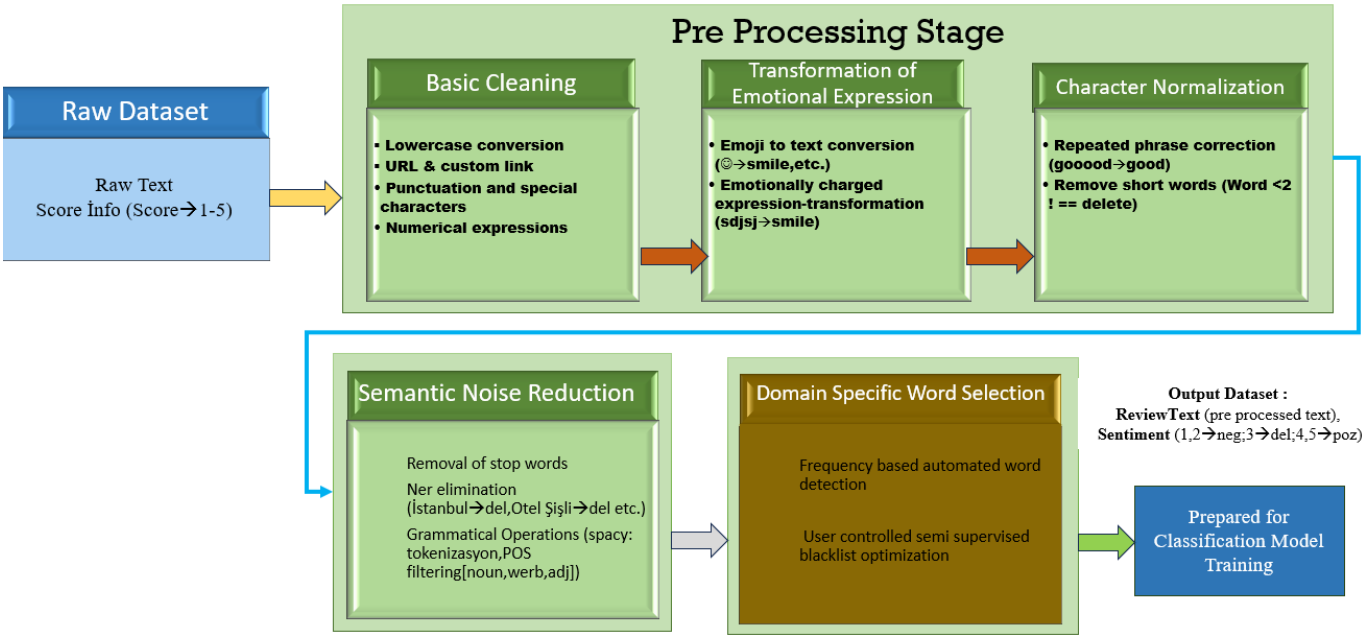


Figure 1. Multi-stage preprocessing pipeline for heterogeneous hotel reviews.

they were converted into textual expressions while preserving their semantic equivalents. This prevented the loss of cues that contribute to sentiment analysis. Next, repetitive characters used for emphasis were normalized (e.g., "goooood" → "good"). This process aims to prevent the learning of different spellings with the same meaning as separate words. Next, English stop words with limited semantic contribution were removed from the dataset, allowing the model to focus on distinctive words. Based on domain knowledge, terms frequently used in the hospitality industry but lacking sentiment value were identified and excluded using a domain-knowledge-guided filtering approach, in which candidate neutral terms were first extracted automatically by frequency analysis and then validated manually by domain experts — a process combining automatic candidate generation with human verification. Finally, named entities such as hotel and city names were removed. This process resulted in a more consistent and semantically richer text representation for the models.

3.3 Methods

A multi-stage method has been developed to examine the perceptual correlation between customer and employee reviews using a data-driven approach. The proposed method includes text preprocessing, application of different text representation approaches, sentiment classification using machine learning algorithms, and evaluation of the results from both general and subject-based correlation analysis perspectives.

Unlike classical sentiment analysis, this method is designed to reveal not only the accuracy of individual classifications but also the perceptual overlap between two groups of reviews. Accordingly, texts were preprocessed, the sentiment labeling process was completed, and numerical representations were obtained using various vectorization methods. Algorithms were trained on these representations, and their performance was evaluated using 5-fold cross-validation to ensure generalizability and the results were analyzed and evaluated.

Correlation identification was carried out in two stages: In the first stage, general sentiment distributions were examined using a combined dataset; in the second stage, sentiment correlations were analyzed on a topic basis by semantically separating interpretations through topic modeling. This approach holistically reveals both the general trends of perceptual correlations at the macro level and the topic-specific differences at the micro level. The general framework of this study is shown in Figure 2.

3.3.1 Text Vectorization Methods

In this study, the texts obtained after preprocessing steps were converted into numerical features for use in classification algorithms. Since text data cannot be directly fed as input to machine learning models, the documents need to be represented in a vector space. For this purpose, three different representation methods were used: Term Frequency–Inverse Document Frequency (TF-IDF), Bag of Words (BOW),

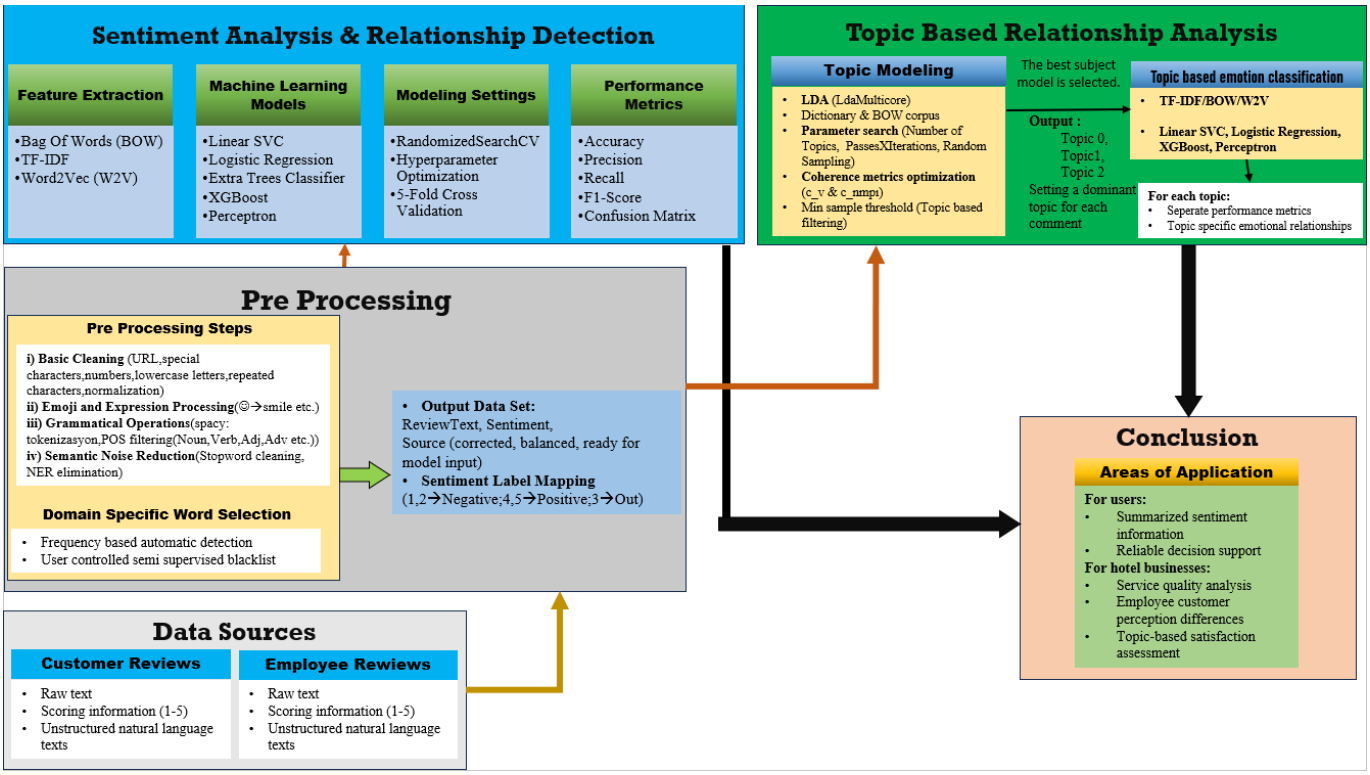


Figure 2. Overall framework of the two-stage sentiment correlation detection approach.

and Word2Vec.

TF-IDF is a statistical method that converts text data into numerical data and assigns weights to words. This method considers the importance of a word within the document and its distinctiveness in the dataset [24]. The TF-IDF method consists of two main components. Term Frequency (TF) indicates the frequency of a word's occurrence in the document. The TF value is given in Formula 1. Inverse Document Frequency (IDF) measures how rare the word is in the dataset and is given in Formula 2. N represents the total number of documents, and n_t represents the number of documents containing the word t. In this method, the weight of a word increases as its frequency in the document increases, but its weight decreases if the word appears in many documents. The TF-IDF value is obtained by multiplying the TF and IDF values and is given in Formula 3.

BOW is a simple and widely used method for representing texts based on word frequency [25]. In this approach, the order or grammatical structure of words is not considered. In the first step of the BOW approach, a dictionary is created from all the unique words in the dataset and is given in Formula 4. V represents all the unique words in the dataset. The document vector consists of numerical values indicating whether the words are present in the document or how many times they appear and is given in Formula 5. While BOW offers a simple and low-cost representation, it cannot model semantic correlations between words.

Word2Vec is an embedding method that aims to capture semantic correlations by representing words with high-dimensional vectors [26, 27]. It has two basic architectures: Continuous Bag-of-Words (CBOW) and Skip-Gram. In this study, CBOW was preferred because it is more efficient in small datasets. In the CBOW model, the vector of the target word is calculated by averaging the words in its context and is given in Formula 6. The C value represents the words in the context of the target word. In the Skip-Gram model, the aim is to predict the words around a given word, and this process is given in Formula 7. The Word2Vec model was trained with a vector size of 100 and a window width of 5; the minimum word frequency threshold was set to 1 to include low-frequency words as well, and document-level representations were created by averaging the relevant word vectors. This approach, unlike TF-IDF and BOW, preserves semantic correlations between words.

$$TF(t,d) = \frac{\text{The \# of instances of word 't' in document } d}{\text{Total \# of words in document } d} \quad (1)$$

$$IDF(t, D) = \log\left(\frac{N}{n_t}\right) \tag{2}$$

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times IDF(t, D) \tag{3}$$

$$V = \{w_1, w_2, w_3, \ldots, w_{|V|}\} \tag{4}$$

$$\mathbf{w} = [0, 0, \ldots, 1, 0, \ldots, 0] \tag{5}$$

$$\mathbf{v}_w = \frac{1}{|C|}\sum_{c \in C} \mathbf{v}_c \tag{6}$$

$$P(w_0|w_1) = \frac{e^{\mathbf{w}_0 \cdot \mathbf{w}_1}}{\sum_{w \in V} e^{\mathbf{w} \cdot \mathbf{w}_1}} \tag{7}$$

3.3.2 *Machine Learning Algorithms*

In this study, emotion classification is addressed within the framework of supervised learning. Machine learning algorithms from different paradigms were used for emotion prediction in text-based data. The selected models include linear classifiers, ensemble methods, and gradient boosting-based approaches. This diversity allows for different perspectives on the structural and semantic features of the data.

Hyperparameter optimization of all models was performed using RandomizedSearchCV, and their performance was evaluated with 5-fold cross-validation. Texts were fed into the model through numerical representations of words or word groups; the algorithms learned emotion classes from these representations.

a) Linear Support Vector Classifier (Linear SVC)

Linear SVC is a powerful supervised learning algorithm that performs the separation between classes using the maximum margin principle. It is widely used in sentiment analysis studies, especially because of its effectiveness in high-dimensional and sparse text data [28, 29]. Linear SVC learns the most suitable linear decision boundary separating classes according to the optimization problem presented in formula 8. w is the weight vector, b is the bias, x_i is the feature vector of the i-th instance, y_i is the actual class, ξ_i is the relaxation variable, and C is the regulation parameter. The model makes class predictions through the decision function and produces both efficient and reliable results in text-based classifications.

$$\min_{w,b,\{\xi_i\}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, ..., n. \tag{8}$$

b) Perceptron

Perceptron is a simple artificial neural network classifier that works on linearly separable datasets [30]. The Perceptron learning rule is expressed as shown in Formula 9. w_t represents the current weight vector, w_{t+1} the updated weights, η the learning rate, y_i the actual class label, and x_i the input feature vector. The algorithm offers satisfactory performance, especially when sufficient linear decision bounds exist.

$$w_{t+1} = w_t + \eta \cdot y_i \cdot x_i \tag{9}$$

c) Logistic Regression (LR)

LR is a linear model widely used in binary classifications that converts the linear combination of input features into probability through a sigmoid function [31] and is given in Formula 10. $w \cdot x + b$ represents the linear combination of word representations, and the sigmoid function converts this value into a probability in the range $[0, 1]$. The model estimates the probability of the text belonging to the emotion class through word representations and prevents overfitting with L1/L2 regularization.

$$P(y = 1|x) = \frac{1}{1 + e^{-(w \cdot x + b)}} \tag{10}$$

d) Extra Trees Classifier

Extra Trees is a randomized classifier created by the ensemble method of decision trees [32]. The final estimate is calculated by majority voting of the trees. The majority voting is calculated as presented in Formula 11. $T_M(x)$ represents the class estimate produced by the m-th decision tree for the input sample x, and M represents the total number of trees. Gini impurity is used in each node split. The Gini calculation is given in formula 12. This approach is an effective and generalizable method in sentiment classification due to its high randomness and capacity to model nonlinear correlations.

$$\hat{y} = \text{mode}\left(T_1(x), T_2(x), ..., T_M(x)\right) \tag{11}$$

$$Gini = 1 - \sum_{k=1}^{K} p_k^2 \tag{12}$$

e) Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting-based, scalable, and high-performance algorithm [33]. The model minimizes errors with successively weak learners (decision trees) and is given in Formula 13. $\sum_i l(y_i, \hat{y}_i)$ is the loss function, and $\Omega(f_k)$ represents the regulation term penalizing the model complexity. Thus, it learns complex interactions in text data effectively through parallel computation and regularization.

$$L = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{13}$$

3.3.3 Correlation Detection Methods

In this study, a two-stage correlation detection approach is proposed as a novel contribution to multi-source sentiment analysis. In this framework, perceptual correlation between customer and employee reviews is operationally defined as the degree to which the two review sources share a common sentiment structure — quantified by the stability of classification performance when both sources are jointly represented in a single sentiment space. This approach extends classical single-source sentiment classification by examining perceptual alignment between distinct reviewer groups at both the global and topic-specific levels. The proposed approaches aim to analyze comments not only at the emotional level but also by considering their semantic context. In the first method, an analysis based on direct emotion classification was performed on a combined dataset, while in the second method, a more detailed correlation detection process based on topic modeling was applied.

a) Correlation Detection in Combined Dataset

In this method, customer and employee reviews are combined into a single dataset. While preserving the source information of each comment, the classification process is carried out using common sentiment labels. Thus, the positions of the two groups in the same sentiment space are examined comparatively. Following preprocessing and labeling stages, the combined dataset is vectorized using TF-IDF, BOW, and Word2Vec methods. Linear SVC, LR, Perceptron, and XGBoost models are trained on the resulting representations; performance evaluation is carried out using 5-fold cross-validation. Accuracy, precision, sensitivity, and F1 score are used for evaluation.

b) Topic-Based Approach

A second method was applied to deepen the overall sentiment findings obtained from the combined dataset. This approach examined how the correlation between customer and employee comments differed not only at the general sentiment level but also within the context of specific topics. Accordingly, topic modeling was applied to the comments before sentiment classification.

The Latent Dirichlet Allocation (LDA) method was used in the topic modeling process [34]. LDA assumes that each document is a probabilistic combination of multiple latent topics, and each topic is represented by a specific word distribution. This correlation is shown in Formula 14, where z_k represents latent topics and K represents the total number of topics. This model allows for the identification of topic structures in texts that are not explicitly defined but are semantically consistent.

$$P(w|d) = \sum_{k=1}^{K} P(w|z_k)P(z_k|d) \tag{14}$$

The number of topics and model parameters were determined using the C_v and C_{NPMI} coherence metrics, which measure topic consistency. Different parameter combinations were evaluated using a random search method, and the model producing the highest coherence value was selected. This ensured the statistical and semantic consistency of the topic headings.

With the selected LDA model, a dominant topic was assigned to each comment, and the dataset was divided into topic-based subsets. Each subset was structured to include customer and employee comments related to the topic. This structure allowed for the comparison of sentiment expressions within the same topic context. Separate sentiment classification models were trained for each topic; the sensitivity of the topics was analyzed using different text representations. Performance evaluation was performed using 5-fold cross-validation; accuracy, precision, sensitivity, and F1 score measures were used.

3.3.4 Performance Evaluation Metrics

In this study, the performance of sentiment classification models was evaluated using the

Table 3. Sentiment analysis performance on customer reviews.

Validation	Model	Vectorization	Accuracy	F1-Score	Precision	Recall
5-Fold	**ExtraTreesClassifier**	**W2V**	0.998	**0.998**	0.998	0.998
5-Fold	ExtraTreesClassifier	TF-IDF	0.982	0.982	0.983	0.982
5-Fold	ExtraTreesClassifier	BOW	0.964	0.964	0.966	0.964
5-Fold	LinearSVC	TF-IDF	0.998	0.998	0.998	0.998
5-Fold	LinearSVC	BOW	0.998	0.998	0.998	0.998
5-Fold	LinearSVC	W2V	0.893	0.893	0.894	0.893
5-Fold	LR	TF-IDF	0.998	0.998	0.998	0.998
5-Fold	LR	BOW	0.998	0.998	0.998	0.998
5-Fold	LR	W2V	0.892	0.892	0.892	0.892
5-Fold	Perceptron	TF-IDF	0.997	0.997	0.997	0.997
5-Fold	Perceptron	BOW	0.997	0.997	0.997	0.997
5-Fold	Perceptron	W2V	0.88	0.898	0.9	0.9
5-Fold	XGBClassifier	BOW	0.998	0.998	0.998	0.998
5-Fold	XGBClassifier	W2V	0.998	0.998	0.998	0.998
5-Fold	XGBClassifier	TF-IDF	0.997	0.997	0.997	0.997

accuracy, precision, recall, and F1 score measures commonly used in the literature [35]. The calculations were based on the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values, which are components of the confusion matrix. Accuracy represents the correct prediction rate of the model among all samples and is given in Formula 15. Precision shows how many of the samples classified as positive by the model are actually positive and is given in Formula 16. Recall expresses how many of the truly positive samples were correctly identified and is given in Formula 17. The F1 score is the harmonic mean of the precision and recall values and is given in Formula 18. The F1 score provides a more balanced evaluation, especially in cases where the class distribution may be unbalanced.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{18}$$

4 Experimental Results

This section evaluates the proposed framework across three experimental scenarios — single-source classification, combined-source correlation detection, and topic-based correlation analysis — to assess both the performance and the generalizability of the AI-driven pipeline. Analyses were conducted separately on customer and employee comments, and together within the framework of a two-stage method developed to examine the correlation between the two data sources. To reliably measure the generalizability of the models, 5-fold cross-validation was applied in all experiments, and results were reported using accuracy, precision, recall, and F1-score metrics.

Pre-processed text data was used in the experiments. BOW, TF-IDF, and Word2Vec (W2V) methods were preferred for the numerical representation of the texts. Extra Trees Classifier, Linear SVC, LR, Perceptron, and XGBoost models were evaluated in the classification phase. Hyperparameter optimization was performed for all models using the RandomizedSearchCV method.

4.1 The Results of Customer Review Sentiment Analysis

The results of the customer review sentiment analysis are shown in Table 3. The accuracy, precision, sensitivity, and F1 score values obtained by different

Table 4. Sentiment analysis performance on employee reviews.

Validation	Model	Vectorization	Accuracy	F1-Score	Precision	Recall
5-Fold	ExtraTreesClassifier	TF-IDF	0.92	0.924	0.93	0.92
5-Fold	ExtraTreesClassifier	BOW	0.91	0.913	0.92	0.91
5-Fold	ExtraTreesClassifier	W2V	0.9	0.896	0.9	0.89
5-Fold	**LinearSVC**	**TF-IDF**	0.938	**0.938**	0.938	0.938
5-Fold	LinearSVC	W2V	0.926	0.926	0.927	0.926
5-Fold	LinearSVC	BOW	0.922	0.922	0.922	0.922
5-Fold	LR	TF-IDF	0.931	0.931	0.931	0.931
5-Fold	LR	BOW	0.926	0.926	0.927	0.926
5-Fold	LR	W2V	0.89	0.897	0.9	0.89
5-Fold	Perceptron	TF-IDF	0.893	0.893	0.894	0.893
5-Fold	Perceptron	BOW	0.869	0.869	0.869	0.869
5-Fold	Perceptron	W2V	0.85	0.86	0.88	0.84
5-Fold	XGBClassifier	BOW	0.873	0.873	0.873	0.873
5-Fold	XGBClassifier	W2V	0.84	0.84	0.84	0.84
5-Fold	XGBClassifier	TF-IDF	0.839	0.839	0.839	0.839

text representation methods (BOW, TF-IDF, W2V) and classification models using 5-fold cross-validation are presented.

As shown in Table 3, Linear SVC, LR, and XGBClassifier models exhibit high and stable performance when used with TF-IDF and BOW representations. It is noteworthy that in these combinations, accuracy and F1 scores reached 99%, and consistency was achieved between metrics. The ExtraTreesClassifier model also showed high success with all vectorization methods, producing particularly strong results with the W2V representation.

The overall findings indicate that customer reviews contain distinct and distinctive word patterns in terms of sentiment analysis; frequency-based representations (BOW and TF-IDF) offer an effective structure when used with linear classifiers. The high-performance values obtained demonstrate that the sentiment orientation of customer reviews can be reliably learned by the models.

4.2 The Results of Employee Review Sentiment Analysis

The results of employee review sentiment analysis are shown in Table 4. The performance values obtained with 5-fold cross-validation of different text representations and classification models in the experiments are shown.

The highest performance was obtained when the Linear SVC model was used in combination with the TF-IDF representation. In this combination, accuracy and F1 score reached 93.8%. Similarly, the LR+TF-IDF combination also produced consistent and high results. These findings show that TF-IDF representations offer an effective representation in employee comments when used with linear classifiers. The ExtraTreesClassifier model exhibited balanced performance with all vectorization methods, while the perceptron and XGBClassifier models produced high results, especially with TF-IDF and BOW representations. Overall, it was observed that the choice of model and vectorization method was decisive for performance in employee comments, and that TF-IDF-based approaches provided strong and consistent results.

4.3 The Results of Correlation Detection

In the first approach to correlation detection, the results of the 5-fold cross-validation of the first method applied to a single dataset created by combining customer and employee reviews are shown. Thus, it aims to evaluate the representability of two different types of reviews in a common sentiment space and the sustainability of the classification performance under a combined structure.

Table 5. Correlation detection results on the combined dataset.

Validation	Model	Vectorization	Accuracy	F1-Score	Precision	Recall
5-Fold	ExtraTreesClassifier	BOW	0.91	0.913	0.92	0.91
5-Fold	ExtraTreesClassifier	W2V	0.888	0.865	0.901	0.888
5-Fold	ExtraTreesClassifier	TF-IDF	0.855	0.855	0.855	0.855
5-Fold	**Linear SVC**	**TF-IDF**	0.971	**0.971**	0.972	0.971
5-Fold	Linear SVC	BOW	0.95	0.95	0.951	0.95
5-Fold	Linear SVC	W2V	0.938	0.938	0.938	0.938
5-Fold	LR	TF-IDF	0.964	0.964	0.966	0.964
5-Fold	LR	BOW	0.954	0.954	0.955	0.954
5-Fold	LR	W2V	0.85	0.86	0.88	0.84
5-Fold	Perceptron	TF-IDF	0.962	0.962	0.963	0.962
5-Fold	Perceptron	BOW	0.942	0.942	0.942	0.942
5-Fold	Perceptron	W2V	0.926	0.926	0.927	0.926
5-Fold	XGB Classifier	BOW	0.903	0.903	0.903	0.903
5-Fold	XGB Classifier	TF-IDF	0.89	0.897	0.9	0.89
5-Fold	XGB Classifier	W2V	0.865	0.865	0.865	0.865

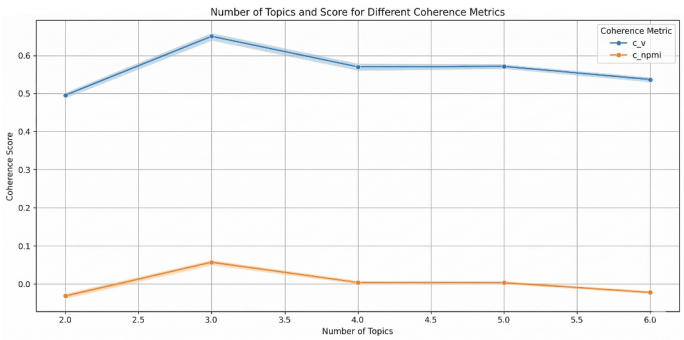


Figure 3. The results of different coherence metrics by number of topics.

As shown in Table 5, the highest performance was obtained when the Linear SVC model was used in conjunction with the TF-IDF representation. Accuracy, F1-score, precision, and sensitivity values reached approximately 97%. Similarly, LR and Perceptron models also yielded high and balanced results with the TF-IDF representation. The BOW representation showed strong performance, especially when used with linear classifiers. The Word2Vec representation also achieved high results when used with different models. Overall, it was observed that frequency and weight-based representations provided distinctive features in the combined data structure.

The findings demonstrate that customer and employee reviews contain significantly overlapping linguistic patterns in terms of sentiment orientation, and that both data sources can be successfully represented with a single model. High accuracy and F1-scores reveal that the initial approach to correlation detection in the combined dataset is valid and effective.

The results confirm that high classification performance is achieved in the combined dataset. In particular, the high accuracy and F1 values obtained with word-weighted representations indicate that customer and employee comments significantly overlap in terms of sentiment expression. This finding reveals a significant parallelism between employee perception and customer experience at the general sentiment level.

Furthermore, the combined dataset approach enables a quantitative comparison of the sentiment distributions of the two groups at the macro level. However, it does not provide a detailed explanation of the thematic dimensions through which the correlation is shaped. Therefore, the analysis is deepened in the following section using a topic-based approach.

4.4 The Results of Topic-Based Correlation Detection

In the second method for correlation detection, Latent Dirichlet Allocation (LDA) was applied to the combined dataset, separating the reviews into topic

Table 6. Topic-based sentiment analysis performance.

Validation	Model	Vectorization	Accuracy	F1-Score	Precision	Recall
5-Fold	Linear SVC - Topic 0	W2V	0.869	0.869	0.869	0.869
5-Fold	Linear SVC - Topic 0	BOW	0.86	0.868	0.899	0.86
5-Fold	Linear SVC - Topic 0	TF-IDF	0.872	0.854	0.886	0.872
5-Fold	Linear SVC - Topic 1	W2V	0.922	0.922	0.922	0.922
5-Fold	Linear SVC - Topic 1	TF-IDF	0.908	0.907	0.916	0.908
5-Fold	Linear SVC - Topic 1	BOW	0.887	0.886	0.892	0.887
5-Fold	Linear SVC - Topic 2	TF-IDF	0.963	0.962	0.965	0.963
5-Fold	Linear SVC - Topic 2	BOW	0.963	0.962	0.965	0.963
5-Fold	Linear SVC - Topic 2	W2V	0.893	0.893	0.894	0.893
5-Fold	LR - Topic 0	BOW	0.867	0.874	0.895	0.867
5-Fold	LR - Topic 0	TF-IDF	0.872	0.854	0.886	0.872
5-Fold	LR - Topic 0	W2V	0.84	0.84	0.84	0.84
5-Fold	LR - Topic 1	TF-IDF	0.901	0.9	0.912	0.901
5-Fold	LR - Topic 1	BOW	0.899	0.899	0.902	0.899
5-Fold	LR - Topic 1	W2V	0.855	0.855	0.855	0.855
5-Fold	LR - Topic 2	TF-IDF	0.963	0.962	0.965	0.963
5-Fold	LR - Topic 2	BOW	0.963	0.962	0.965	0.963
5-Fold	LR - Topic 2	W2V	0.962	0.962	0.963	0.962
5-Fold	Perceptron - Topic 0	BOW	0.889	0.893	0.906	0.889
5-Fold	Perceptron - Topic 0	TF-IDF	0.862	0.846	0.866	0.862
5-Fold	Perceptron - Topic 0	W2V	0.841	0.843	0.846	0.841
5-Fold	Perceptron - Topic 1	TF-IDF	0.903	0.903	0.905	0.903
5-Fold	Perceptron - Topic 1	BOW	0.868	0.868	0.869	0.868
5-Fold	Perceptron - Topic 1	W2V	0.852	0.86	0.89	0.852
5-Fold	Perceptron - Topic 2	TF-IDF	0.963	0.961	0.964	0.963
5-Fold	Perceptron - Topic 2	BOW	0.957	0.956	0.956	0.957
5-Fold	Perceptron - Topic 2	W2V	0.931	0.931	0.931	0.931
5-Fold	XGBoost Classifier - Topic 0	TF-IDF	0.908	0.907	0.916	0.908
5-Fold	XGBoost Classifier - Topic 0	BOW	0.867	0.873	0.89	0.867
5-Fold	XGBoost Classifier - Topic 0	W2V	0.853	0.859	0.877	0.853
5-Fold	**XGBoost Classifier - Topic 1**	**W2V**	0.976	**0.975**	0.976	0.976
5-Fold	XGBoost Classifier - Topic 1	TF-IDF	0.973	0.972	0.972	0.973
5-Fold	XGBoost Classifier - Topic 1	BOW	0.879	0.878	0.892	0.879
5-Fold	XGBoost Classifier - Topic 2	W2V	0.963	0.962	0.964	0.963
5-Fold	XGBoost Classifier - Topic 2	BOW	0.958	0.956	0.96	0.958
5-Fold	XGBoost Classifier - Topic 2	TF-IDF	0.888	0.865	0.901	0.888

headings. The aim was to reveal how the correlation between customer and employee perceptions is structured not only at the general sentiment level but also on a topic-by-topic basis. Different numbers of topics were tested for LDA, and the coherence metrics (c_v and c_npmi) were selected as the most suitable method. As shown in Figure 3, three topic headings were selected as the most suitable values in terms of semantic consistency and model simplicity, and the analyses were performed according to these values.

The 5-fold cross-validation results of the emotion

classification models trained for each topic obtained with LDA are shown in Table 6. For each topic, Linear SVC, LR, Perceptron, and XGBoost classifiers were evaluated separately with BOW, TF-IDF, and Word2Vec.

Table 6 shows that the models trained under Topic 2 generally exhibited the highest and most consistent performance. Specifically, the Linear SVC, LR, and Perceptron models achieved approximately 96% accuracy and F1-scores with TF-IDF and BOW. The XGBoost model also produced high performance with different representations on the same topic. These results demonstrate that Topic 2 contains distinct and differentiated patterns in terms of sentiment orientation.

The results obtained for Topic 1 also indicate strong and balanced performance. In particular, the Linear SVC and LR models achieved over 90% accuracy with the TF-IDF representation. The XGBoost model reached high accuracy values using the Word2Vec. Under Topic 0, the models produced stable but relatively more limited performance values.

When Figure 3 and Table 6 are considered together, it is observed that the correlation between customer and employee comments emerges with varying intensities depending on the topic. While a strong and consistent emotional overlap is observed, particularly in Topic 2, the correlation in other topics appears to be contextually sensitive. These results demonstrate that the topic-based approach offers a more detailed and interpretable framework for identifying correlations.

The results confirm that the correlation between customer and employee perceptions varies depending on the topic. While high classification performance and significant sentiment overlap were observed in some topic headings, sentiment distributions exhibited a more diversified structure in others. These findings suggest that the correlation exhibits a context-sensitive and multidimensional structure rather than a uniform pattern.

5 Conclusion and Recommendations

This study proposes a novel multi-stage sentiment correlation detection framework that integrates semi-supervised preprocessing, multi-source sentiment classification, and LDA-based topic modeling into a unified analytical pipeline. Beyond its application to hotel reviews, the proposed framework contributes to the emerging field of AI-driven service intelligence by demonstrating how heterogeneous user-generated text from multiple stakeholder groups can be jointly analyzed to reveal perceptual alignment and divergence in a shared sentiment space. The proposed approach has been shown to offer high accuracy and consistency in both single and combined datasets, using different text representations and classifier models.

Customer reviews, containing direct and explicit expressions of emotion, achieved high success with frequency-based representations and linear classifiers. Employee reviews, on the other hand, have a more contextual and critical language structure, exhibiting consistent and reliable performance with TF-IDF-based linear models. The combined analysis of customer and employee reviews revealed a high degree of overlap in perceptions between the two groups, effectively representing them within a single sentiment space. Topic-based correlation detection added a more detailed and interpretable dimension to sentiment analysis by showing how perceptual overlap varies according to specific themes. This method allows for the examination of customer and employee perceptions not only at a general level but also on a topic-specific basis.

The proposed methods offer a decision support mechanism for both customers and hotel employees through automated sentiment summaries and topic-based analyses. Within this framework, sentiment trends in service areas can be monitored, and improvement processes can be strategically guided.

Limitations of the study include the use of only positive and negative sentiment classes, the focus of the datasets on specific languages and sectors, and the exclusion of context-sensitive models based on deep learning. Future studies are recommended to incorporate neutral sentiment classes, adopt large language models such as BERT, RoBERTa, and GPT-based architectures, and extend the correlation detection framework to multilingual and cross-domain settings. The integration of temporal dynamics and real-time review streams also represents a promising emerging direction for AI-driven service intelligence systems. In conclusion, this study provides a methodological contribution to the academic literature and proposes an analytical framework that offers concrete, interpretable, and strategic insights for users and hotel businesses in practice.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This study is based exclusively on publicly available online hotel reviews. No direct interaction with human participants, collection of personal data, or use of private identifying information was involved. Ethical review and approval were therefore waived for this study. All data were used in accordance with the terms of service of the respective platforms from which they were obtained.

References

[1] Ameur, A., Hamdi, S., & Ben Yahia, S. (2023). Sentiment analysis for hotel reviews: A systematic literature review. *ACM Computing Surveys, 56*(2), 1-38. [CrossRef]

[2] Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer science review, 41*, 100413. [CrossRef]

[3] Mehraliyev, F., Chan, I. C. C., & Kirilenko, A. P. (2022). Sentiment analysis in hospitality and tourism: a thematic and methodological review. *International Journal of Contemporary Hospitality Management, 34*(1), 46-77. [CrossRef]

[4] Liu, B. (2012). *Sentiment analysis and opinion mining*. Springer. [CrossRef]

[5] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval, 2*(1-2), 1-135. [CrossRef]

[6] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems, 28*(2), 15-21. [CrossRef]

[7] Yurdasever, E. (2024). İşletmelerde çalışanlara yönelik duygu analizinin uygulanması: Potansiyel faydalar ve zorluklar. *Academic Review of Economics and Administrative Sciences, 17*(3), 462-488. [CrossRef]

[8] Oğul, B. B., & Ercan, G. (2016, May). Sentiment classification on Turkish hotel reviews. In *2016 24th Signal Processing and Communication Application Conference* (*SIU*) (pp. 497-500). IEEE. [CrossRef]

[9] 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. [CrossRef]

[10] Gunawan, R., & Hendry, H. (2025). Analisis sentimen ulasan tamu terhadap layanan hotel menggunakan pendekatan machine learning. *IT-Explore: Jurnal Penerapan Teknologi Informasi dan Komunikasi, 4*(3), 295-306. [CrossRef]

[11] Muhammad, P. F., Kusumaningrum, R., & Wibowo, A. (2021). Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews. *Procedia Computer Science, 179*, 728-735. [CrossRef]

[12] Ahmetoğlu, H., & Daş, R. (2020). Türkçe otel yorumlarıyla eğitilen kelime vektörü modellerinin duygu analizi ile incelenmesi. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 24*(2), 455-463. [CrossRef]

[13] Erdoğan, D., Kayakuş, M., Çelik Çaylak, P., Ekşili, N., Moiceanu, G., Kabas, O., & Ichimov, M. A. M. (2025). Developing a deep learning-based sentiment analysis system of hotel customer reviews for sustainable tourism. *Sustainability, 17*(13), 5756. [CrossRef]

[14] Jayanto, R., Kusumaningrum, R., & Wibowo, A. (2022). Aspect-based sentiment analysis for hotel reviews using an improved model of long short-term memory. *International Journal of Advances in Intelligent Informatics, 8*(3), 391. [CrossRef]

[15] Kusumaningrum, R., Nisa, I. Z., Jayanto, R., Nawangsari, R. P., & Wibowo, A. (2023). Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews. *Heliyon, 9*(6). [CrossRef]

[16] Wen, Y., Liang, Y., & Zhu, X. (2023). Sentiment analysis of hotel online reviews using the BERT model and ERNIE model—Data from China. *Plos one, 18*(3), e0275382. [CrossRef]

[17] Puh, K., & Bagić Babac, M. (2023). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of hospitality and tourism insights, 6*(3), 1188-1204. [CrossRef]

[18] Özdemir, A. O., Giritli, E. B., & Can, Y. S. (2024, October). Sentiment analysis for hotel reviews in turkish by using llms. In *2024 9th International Conference on Computer Science and Engineering* (*UBMK*) (pp. 210-214). IEEE. [CrossRef]

[19] Adıgüzel, F., Elsherbiny, M., Aguiar Quintana, J. T., & González-Martel, C. (2021). Topic modelling application for luxury hotel reviews. In *25th PPAD Marketing Congress*.

Pazarlama ve Pazarlama Araştırmaları Derneği. https://ntu-repository.worktribe.com/output/1763496

[20] Bilge, A. C. (2024). Online Otel Yorumlarının Duygu Analizi ile İncelenmesi: Konya Beş Yıldızlı Oteller Örneği. *Selçuk Üniversitesi Akşehir Meslek Yüksekokulu Sosyal Bilimler Dergisi*, (17), 84-93. https://izlik.org/JA55PA79RM

[21] Ali, T., Omar, B., & Soulaimane, K. (2022). Analyzing tourism reviews using an LDA topic-based sentiment analysis approach. *MethodsX, 9*, 101894. [CrossRef]

[22] Nadeem, A., Missen, M. M. S., Al Reshan, M. S., Memon, M. A., Asiri, Y., Nizamani, M. A., ... & Shaikh, A. (2025). Resolving ambiguity in natural language for enhancement of aspect-based sentiment analysis of hotel reviews. *PeerJ Computer Science, 11*, e2635. [CrossRef]

[23] Gaye, B., Zhang, D., & Wulamu, A. (2021). Sentiment classification for employees reviews using regression vector-stochastic gradient descent classifier (RV-SGDC). *PeerJ Computer Science, 7*, e712. [CrossRef]

[24] Das, M., & Alphonse, P. J. A. (2023). A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset. *arXiv preprint arXiv:2308.04037*. [CrossRef]

[25] Soumya George, K., & Joseph, S. (2014). Text classification by augmenting bag of words (BOW) representation with co-occurrence feature. *IOSR Journal of Computer Engineering, 16*(1), 34–38. [CrossRef]

[26] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. [CrossRef]

[27] Karcioğlu, A. A., & Aydin, T. (2019, April). Sentiment analysis of Turkish and english twitter feeds using Word2Vec model. In *2019 27th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE. [CrossRef]

[28] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297. [CrossRef]

[29] Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137-142). Berlin, Heidelberg: Springer Berlin Heidelberg. [CrossRef]

[30] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408. [CrossRef]

[31] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). *Applied logistic regression* (Vol. 2, pp. 116-128). New York: wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387

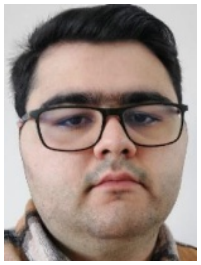
[32] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning, 63*(1), 3-42. [CrossRef]

[33] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). [CrossRef]

[34] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022. https://jmlr.org/papers/v3/blei03a.html

[35] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*. [CrossRef]

Abdullah Ammar Karcioglu received the B.Eng. degree in Computer Engineering from Ege University, Türkiye, in 2015. He received the M.Eng. degree in computer engineering from Atatürk University in 2018 and his Ph.D degree in computer engineering from Ege University in 2022. His research interests include AI, NLP, LLMs, and related areas. (Email: ammar.karcioglu@atauni.edu.tr)



Berat Bakan received the B.Eng. degree in Software Engineering from Fırat University, Türkiye, in 2023. He is currently a master's student in the department of software engineering at Atatürk University. His research interests include AI, NLP, image processing, and related areas. (Email: berat.bakan.2523@gmail.com)

Esra Katırcıoğlu received her bachelor's degree in Department of Translation and Interpreting from Hacettepe University, Türkiye in 2005. She received the master's degree in the department of tourism management from Suleyman Demirel University in 2017 and her Ph.D degree in the department of tourism management from Afyon Kocatepe University in 2022. Her research interests include tourism-hotel management, labor economics, industrial relations, and related areas. (Email: esra.katircioglu@omu.edu.tr)