FACTA UNIVERSITATIS
Series: **Automatic Control and Robotics** Vol. 24, Nº 1, 2025, pp. 1 - 16
https://doi.org/10.22190/FUACR241218002N

Regular Paper

INTEGRATING DEEP LEARNING FOR AUTOMATED DETECTION OF NEGATIVE HOTEL REVIEWS

UDC ((004.85+004.032.26):801.73)

Milena Nikolić[1], Miloš Stojanović[1], Marina Marjanović[2]

[1]The Academy of Applied Technical and Preschool Studies,
Department of Information and Communication Technologies, Republic of Serbia
[2]Singidunum University, Belgrade, Republic of Serbia

ORCID iDs:		
	Milena Nikolić	https://orcid.org/0009-0004-3769-6299
	Miloš Stojanović	https://orcid.org/0009-0000-7927-5223
	Marina Marjanović	https://orcid.org/0000-0002-9928-6269

Abstract. *This paper presents an automated deep learning framework for detecting and classifying negative hotel reviews, integrating both textual and numerical inputs. The model utilizes a Fully Connected Feedforward Neural Network (FCNN) to capture complex global relationships and non-linear patterns in the data, ensuring accurate classification of negative reviews. Experimental results show that the model achieves an accuracy of 90.1% and precision of 88.6%. These results demonstrate the model's effectiveness in processing large-scale datasets, outperforming traditional methods in terms of classification performance. By automating the detection of negative reviews, our approach offers a scalable solution for the hospitality industry to enhance both operational efficiency and customer satisfaction.*

Key words: *Deep learning, natural language processing, hotel reviews, sentiment analysis, neural networks, numerical ratings, customer satisfaction.*

1. INTRODUCTION

The hotel industry increasingly depends on online reviews to attract and retain customers, making it essential for businesses to ensure the credibility and reliability of the feedback provided. As the volume of online reviews grows rapidly, distinguishing between authentic and fraudulent content has become a critical challenge. Fraudulent or inconsistent reviews can mislead potential customers, damage hotel reputations, and distort the overall perception of the service quality. Consequently, identifying negative reviews becomes essential for

Received December 18, 2024 / Accepted January 30, 2025
Corresponding author: Milena Nikolić
The Academy of Applied Technical and Preschool Studies, Department of Information and Communication Technologies, Aleksandra Medvedeva 20, 18000 Niš, Republic of Serbia
E-mail: milena.nikolic@akademijanis.edu.rs

improving the service quality and addressing customer concerns, as they reflect a genuine dissatisfaction. Negative reviews are especially valuable as they offer insights into areas requiring attention and improvement, but the volume of these reviews can overwhelm manual processes, highlighting the need for automated systems to filter and identify the relevant feedback. With the significant number of customers relying on online reviews as a critical factor in their decision-making process, maintaining the authenticity and reliability of these reviews directly impacts the success of businesses in the competitive hospitality sector.

The significant potential of artificial intelligence (AI) in the hotel industry has been widely recognized, with several studies demonstrating how AI can enhance the operational efficiency, reduce human error, and improve customer service. AI technologies hold the capacity to revolutionize hospitality by addressing key operational issues such as cleanliness, service quality, and guest safety, especially in challenging times like a pandemic, where maintaining high standards is crucial [1]. Further advancements in AI have shown its ability to enhance the guest experience through technologies like chatbots, mobile apps, and big data analytics, which enable more personalized services and streamline the operational efficiency [2]. As AI continues to evolve, it can also significantly contribute to detecting negative reviews more accurately, ensuring businesses can address the customer displeasure promptly and efficiently. While manual review processes remain important for the contextual understanding, their limitations in scale and efficiency highlight the need for automated solutions that can process large volumes of data with precision.

Our previous research has demonstrated that the detection of fraudulent and inconsistent reviews is essential for maintaining the review integrity, particularly as the volume of reviews increases. A methodology based on the exploratory data analysis (EDA) has been introduced to identify anomalies in hotel reviews, focusing on the sentiment, review similarity, and patterns in posting trends. This method helps to detect unusual shifts in the review activity, highlighting spikes in review volumes or shifts in sentiment, which could indicate fraudulent or misleading reviews. This approach not only enhances the detection of fraud but also uncovers patterns that may indicate the negative feedback among legitimate reviews. Specifically, EDA has proven effective at identifying subtle shifts in the sentiment, making it easier to discover and flag potential negative reviews for deeper analysis [3]. Incorporating domain-specific insights, such as patterns of guest dissatisfaction or recurring service-related issues, further enhances the ability of automated systems to detect and analyze negative reviews effectively.

Building on this initial approach, more advanced techniques, incorporating predictive modeling and machine learning, have been developed to enhance the detection and processing of negative reviews. One such technique combines sentiment analysis with algorithms like XGBoost and VADER sentiment scoring, improving the identification of inconsistencies between review content and ratings. While originally focused on detecting fraudulent or inconsistent reviews, this methodology also facilitates the discovery of the negative sentiment, offering a more effective way to pinpoint the dissatisfaction or emotional undertones that may not be immediately obvious in the text. By analyzing numerical ratings and the sentiment expressed in the review content, the model can flag negative reviews with greater precision. This method improves the overall review classification process, while also offering a more dependable way to identify the negative feedback, which can provide actionable insights to enhance the customer experience and address potential concerns [4]. Ethical considerations in using AI for the review analysis, including transparency and bias mitigation, are crucial to ensure fair and accurate outcomes that build the customer trust.

Identifying negative reviews can be difficult as they may be hidden among other content, particularly in large datasets. Traditional rule-based methods often fall short in capturing the complexity of language and emotional tone, making it difficult to accurately classify negative reviews. A study comparing methods for identifying review spam pointed out the limitations of rule-based systems, which typically depend on the keyword matching and heuristics, making them less effective at detecting the subtle negative feedback [5]. On the other hand, machine learning techniques are better equipped to detect the subtle nuances of negative sentiment within the review text, allowing for a more precise identification of an authentic dissatisfaction. By analyzing patterns in the language and tone, machine learning models can distinguish between subtle expressions of frustration or disappointment and other forms of feedback, leading to more reliable identification of negative reviews.

Further studies have shown the importance of understanding the emotional tone behind negative reviews, including the role of cross-linguistic analysis to examine how dissatisfaction is expressed across languages. Research indicates that despite linguistic differences, there are common patterns in the way negative reviews are expressed, suggesting that advanced sentiment analysis tools should be applied across multiple languages to capture emotional undertones more effectively. These findings reinforce the need for sophisticated models that consider not just linguistic variations but also the emotional content in negative reviews [6]. Additionally, incorporating cross-linguistic insights expands the reach of review analysis systems, enabling businesses to monitor reviews across global markets more effectively.

Additionally, the psychological and emotional factors driving negative reviews have been examined, identifying predictors such as unfulfilled expectations, poor service, and perceived value. Understanding these predictors is essential for improving the accuracy of automated systems designed to detect the negative feedback, as they help refine the emotional and contextual analysis conducted by machine learning models. By identifying the core causes of customer dissatisfaction, these models can more accurately differentiate between negative reviews driven by external factors and those genuinely reflecting poor service or experiences [7]. This type of analysis also provides hotels with actionable insights to directly address the issues leading to a negative feedback and improve overall customer satisfaction.

Other studies pointed out the benefits of data-driven methods for predicting the likelihood of guests sharing the negative emotional content in reviews. Using machine learning algorithms to analyze common patterns of dissatisfaction, these methods provide useful findings into customer behavior and allow businesses to predict negative reviews before they are posted, offering an opportunity to address issues more proactively. Such predictive capabilities are beneficial in managing online reputations and enhancing customer relations by taking steps before a negative review impacts the brand [8]. By anticipating the potential dissatisfaction, hotels can enhance their guest experience and prevent the reputational damage from negative reviews.

The complexity of detecting negative reviews increases with the growing scale of review datasets, especially when handling imbalanced data, which makes traditional methods prone to high false-positive and false-negative rates. While rule-based systems face limitations in scalability, deep learning models provide a more advanced approach by identifying complex, non-linear patterns within large datasets. These models can adapt to diverse review behaviors and capture subtle anomalies in negative reviews, leading to more accurate detection. Our proposed framework combines RoBERTa-based sentiment analysis with Fully Connected Feedforward Neural Networks (FCNN), leveraging both textual and numerical data to improve the classification of negative reviews. This integration offers a practical and efficient solution to the challenges associated with large-scale review systems.

2. METHODOLOGY

In this section, we outline the methodology employed to detect and classify negative hotel reviews. Our approach focuses on filtering out anomalies in hotel reviews before applying the machine learning for automated detection of negative feedback. The implementation consists of several key stages, starting from data preprocessing to model training and evaluation. Each stage plays a significant role in ensuring the high-quality input for the machine learning model, ultimately enhancing its ability to detect negative reviews accurately.

2.1. Data Description

The dataset used in this study is sourced from *Kaggle* and contains hotel reviews with associated metadata from *Booking.com* platform [9]. The dataset has more than 26,000 reviews with 16 columns, capturing a wide range of hotel review attributes.

The columns in the dataset include:

- **review_title**: The title provided by the reviewer for their feedback, which can offer insights into the sentiment or topic of the review (Text).
- **reviewed_at**: The date when the review was posted, which helps in identifying trends or patterns in review activity (Datetime).
- **reviewed_by**: The name or username of the reviewer, which can be used to identify individual patterns or behaviors (Text).
- **images**: The number of images attached to the review, which could indicate the reviewer's level of engagement (Numeric).
- **crawled_at**: The timestamp indicating when the review was retrieved from the website, allowing for tracking and comparison with other time-based features (Datetime).
- **url**: The URL link to the review page on *Booking.com* website, providing direct access to the review (Text).
- **hotel_name**: The full name of the hotel being reviewed, which is important for analyzing review trends across different hotels (Text).
- **hotel_url**: The URL to the hotel's page on *Booking.com* (Text).
- **rating**: The numeric rating given to the hotel in the review, which is an essential feature for sentiment analysis and correlation checks (Numeric).
- **avg_rating**: The average rating for the hotel based on all reviews, which helps contextualize individual review ratings (Numeric).
- **nationality**: The nationality of the reviewer, which can be useful for detecting patterns across different demographic groups (Text).
- **review_text**: The full text content of the review, which is used for sentiment analysis and feature extraction (Text).
- **raw_review_text**: A raw version of the review text, which may be used for additional analysis (Text).
- **tags**: Labels or categories attached to the review, which could provide context on the nature of the feedback (Text).
- **meta**: Additional metadata, which might contain extra insights into the review content and the guest's experience (Text).

In Table 1, we present the initial list of columns along with their counts of null and non-null values. Certain columns, including *url, avg_rating, nationality, crawled_at, raw_review_text, images, and meta,* will be removed as they are considered irrelevant to this research.

Table 1 Initial data card with number of null and non-null values **per column.**

Column Name	Null Count	Non-Null Count
index	0	26675
review_title	1	26674
reviewed_at	105	26570
reviewed_by	105	26570
images	25737	938
crawled_at	289	26386
url	289	26386
hotel_name	289	26386
hotel_url	289	26386
avg_rating	289	26386
nationality	305	26370
rating	289	26386
review_text	289	26386
raw_review_text	473	26202
tags	473	26202
meta	473	26202

2.2. Data Preprocessing

To prepare the dataset for a deeper analysis and ensure the consistency of the reviews, a detailed preprocessing stage was conducted. The following steps were performed:

1. **Removal of Irrelevant Rows**
   Rows where the *review_text* was missing were removed, as the review content is considered a crucial component for the deeper data analysis. However, rows with a missing *review_title* were retained if they had content in the *review_text* column. In this study, the focus is placed on the review content, while review titles are considered less critical. For instances where the title is missing, a placeholder *"Unknown"* was introduced, offering a more concise and balanced alternative to placeholders used in our earlier studies.
2. **Cleaning Special Characters and Empty Spaces**
   Rows containing only special characters or empty spaces in the *review_text* field were removed, Similarly, rows with missing or irrelevant values in *review_title* column were replaced with the placeholder *"Unknown"*. These entries were treated as noise and excluded to strengthen the data reliability.
3. **Handling Predefined Placeholders**
   Rows with predefined placeholder content in the *review_text*, such as *"There are no comments available for this review"*, were replaced with *"Unknown"* to maintain consistency. While these entries do not reflect the authentic user input and could skew analysis, more than 7,000 such rows exist and will be retained to preserve data completeness. Further analysis will rely on *review_title, rating*, and *tags*. Rows where both *review_text* and *review_title* fields are placeholders, empty, or missing will be removed to reduce redundancy and noise.

4. **Handling Missing Ratings**
   Rows with missing values in the *rating* column were retained if they had values in other meaningful components, such as *review_title*, *review_text*, or *tags*. The positivity or negativity of these reviews will be evaluated using the available components, ensuring no potential information is lost.
5. **Introducing Placeholders for Missing Tags**
   For rows with missing tags, which typically include a comprehensive list of tags, such as *Leisure trip~GroupTwin Private Room~Superb~City view,* a placeholder value of *"Unknown"* was introduced for consistency. This placeholder was assigned to all rows where tags were absent, allowing further estimations to be made using other available components, including *review text*, *title*, and *rating*.

Furthermore, during data exploration, the number of reviews with low ratings (less than 5.0) was counted for each hotel. This analysis provides insights into hotels that are frequently reviewed negatively. Although anomalies and fraudulent reviews may exist among these entries, Fig. 1 highlights the top ten hotels with the highest number of low ratings, offering a preliminary understanding of potential trends in dissatisfaction [10].

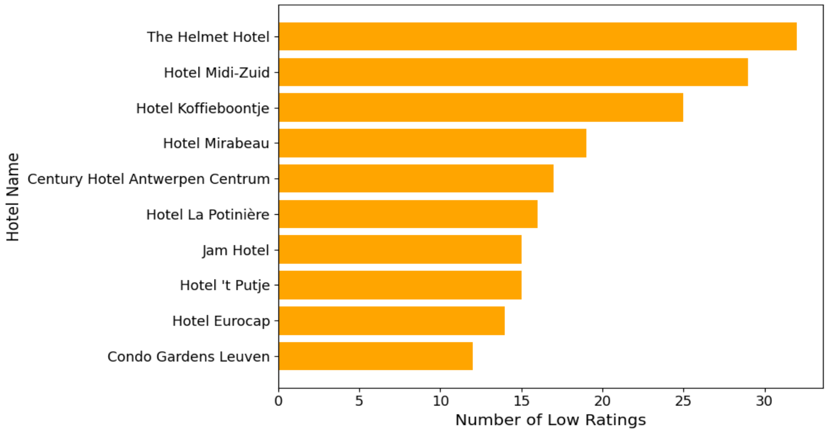

Fig. 1 Top 10 hotels with the highest number of low ratings.

2.3. Extended Sentiment Analysis

For the sentiment analysis phase of this study, we chose to employ RoBERTa, a more advanced transformer model, instead of the VADER sentiment analysis tool used in previous approaches [3, 11, 12]. RoBERTa has demonstrated a superior performance in many NLP tasks, particularly in understanding the context and the subtleties of language, which is crucial when analyzing complex and diverse user reviews like those found in the hotel industry. While VADER approach excels in handling shorter and more straightforward texts, RoBERTa's deeper understanding of context, tone, and subtle linguistic patterns makes it a better choice for capturing the intricacies present in hotel reviews. This model has been pre-trained on a large corpus, allowing it to better understand varied expressions of sentiment that often occur in hotel reviews, which typically include a mix of subjective opinions, sarcasm, and emotional responses [13].

Unlike VADER, which is based on a lexicon and rule-based approach, RoBERTa is a deep learning model that uses context from the entire sentence to determine sentiment, making it better at understanding how words and phrases interact in context. This ability to capture variations in sentiment and context is crucial when dealing with hotel reviews, where sentiment can vary widely within the same review depending on the phrasing or the use of specific adjectives, personal expressions, and even punctuation. [14, 15].

In this analysis, sentiment scores were computed for four basic components: review title, review text, rating, and tags. However, unlike in previous work [3] where tags were split by the "~" separator, and sentiment scores were computed for each individual tag, this time we treated the entire set of tags as one entity. By computing the overall sentiment on tags, we aimed to capture a more comprehensive sentiment of the review, as the tags collectively represent the overall experience or categorization of the review. This approach is beneficial because it enables a broader understanding of the review's context, rather than focusing on individual aspects that may not always reflect the full sentiment of the experience that hotel guests had during their stay.

Additionally, reviews with placeholders were excluded from sentiment calculations. These entries were assigned a neutral sentiment score of zero, as they should not impact our anomaly detection process. This ensures that the sentiment analysis remains focused on authentic reviews that contain a meaningful content.

For the anomaly detection, we focused on critical discrepancies between the sentiment of different review components. If one component, such as the review title, had a positive sentiment while the review text was negative, this was flagged as a potential anomaly. Sentiment scores of zero were ignored, but if a review displayed both positive and negative sentiments across four key components, it was entirely removed as inconsistent. This methodology helps to identify and filter out reviews with conflicting sentiments, which are often indicators of fraudulent or unreliable content. By eliminating these anomalies, the analysis becomes more reliable and improves the quality of the dataset for further tasks, such as feeding data into the upcoming machine learning model. [16, 17].

To visualize the detection of anomalies, we included a bar chart (Fig. 2) showing the count of legitimate and anomalous reviews based on discrepancies in sentiment scores. In this chart, reviews with conflicting sentiments across different components are marked as inconsistent and displayed in orange, while legitimate reviews are shown in green. Although the number of anomalous reviews is relatively small (1218 records), removing these rows is crucial for ensuring the accuracy and reliability of the analysis [18, 19].

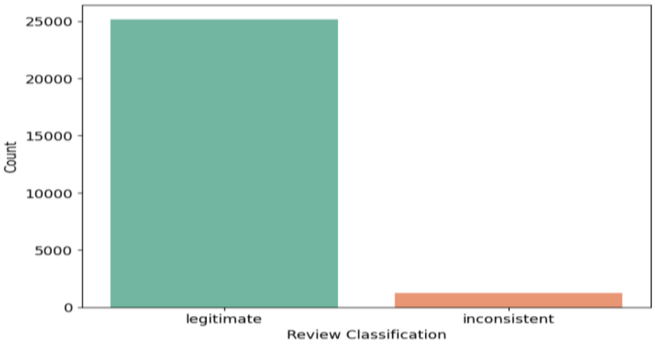

Fig. 2. Distribution of legitimate and fraudulent reviews.

2.4. Feature Scaling

After cleaning and preprocessing the dataset, all missing values were resolved, and appropriate placeholders were introduced where needed. The dataset is now free of null values, confirming its completeness and readiness for modeling. Table 2 represents the updated data card, listing the columns selected as features for the machine learning phase.

Table 2 Updated data card with number of non-null values per column

Column Name	Non-Null Count
index	25168
review_title	25168
reviewed_text	25168
tags	25168
hotel_name	25168
hotel_url	25168
reviewed_by	25168
reviewed_at	25168
rating	25168
title_sentiment	25168
text_sentiment	25168
tag_sentiment	25168

With the dataset cleaned, we focused on preparing the features in formats suitable for deep learning models. Categorical features, such as tags, required encoding techniques through the one-hot encoding to ensure their compatibility with the model, converting them into a numerical format that can be efficiently processed. One-hot encoding is crucial because it transforms categorical variables into a binary format, allowing the model to interpret them as distinct and independent features, preventing any unintended ordinal relationships between categories. To optimize the model's performance and prevent issues related to varying feature magnitudes, we applied the feature scaling. This step is important for deep learning models, as they are highly sensitive to the scale of input data. Furthermore, numerical features, including sentiment scores, were standardized to have a mean value of zero and unit variance. This standardization ensures that no single feature dominates the model's learning process and allows the model to converge effectively during training. With the implementation of techniques and preprocessing the features appropriately, we ensured that the model could effectively learn from the provided data.

2.5. Model Selection

The final step in preparing for automated negative review detection is selecting the appropriate machine learning model. Due to the complexity of the text sentiment, a deep learning approach was chosen, using RoBERTa for the described preprocessing and a Fully Connected Feedforward Neural Network (FCNN) for classification tasks.

The embeddings generated by RoBERTa transform the raw textual data into dense numerical vectors that encapsulate both the sentiment and contextual information of the reviews. These embeddings are then used as input to the classification model, enriching the feature set and enhancing the neural network's ability to accurately detect the negative sentiment or fraudulent reviews.

The actual classification of reviews as either negative or positive is handled using a Fully Connected Feedforward Neural Network (FCNN). The decision to use an FCNN over other potential models, such as Convolutional Neural Networks (CNNs), is based on several essential factors that align with the nature of the data and the task at hand. FCNNs are particularly suitable for review classification, where both text-based features (such as embeddings from RoBERTa) and numerical features (like ratings and review tags) must be processed together. The architecture of an FCNN, consisting of multiple layers of interconnected neurons, allows learning and modeling global patterns in data. FCNNs can capture these interdependencies, improving the model's ability to understand the broader context in which the features interact.

FCNNs also have the advantage of modeling non-linear relationships in data, which is significant for handling the complexities of natural language. Negative reviews, for example, may not follow simple linear patterns, especially if the sentiment is expressed indirectly or sarcastically. FCNNs can learn complex, non-linear patterns and adapt to the various ways negative sentiment might be conveyed, enhancing the flexibility and accuracy of the deep learning model.

While Convolutional Neural Networks (CNNs) are very effective for capturing local patterns in data, they are less suitable for processes that require understanding global relationships, such as the connections between textual sentiment and numerical features in hotel reviews. CNNs perform well when there is a clear, localized structure, but the relationships in hotel reviews are more complex and require a model that can identify long-range dependencies. FCNNs are better equipped to handle high-dimensional data and learn from the entire feature space, making them a more appropriate and suitable choice for this task. Thus, the FCNN is an ideal model for effectively classifying the complex relationships present in hotel reviews [20, 21, 22].

2.6. Model Training

Once the data is preprocessed and the model is chosen, review records are split into training and validation sets to enable proper model evaluation. The training set is used to teach the model how to classify reviews, while the validation set ensures that the model is not overfitting to the training data and can generalize well to unseen data.

The training process involves defining a neural network that combines textual and numerical features. The textual features, previously processed through the RoBERTa model to extract embeddings, are passed through a series of fully connected layers. These layers allow the network to learn complex, non-linear relationships in the data, enabling the model to identify patterns of sentiment, including indirect expressions of negativity or sarcasm. The numerical features are similarly processed and integrated with the text embeddings, allowing the model to learn how different features interact and contribute to the classification task. This integration of diverse feature types enhances the model's ability to capture a comprehensive understanding of each review, improving the overall accuracy of sentiment classification [23].

The model is initially trained using several key hyperparameters that are critical for the training process. The learning rate, set to 0.001, controls the size of the steps the optimizer takes during training. A small learning rate is used to ensure that the model converges smoothly without overshooting the optimal parameters. The batch size is set to 32, meaning that the model processes 32 review samples at a time before updating its parameters. This batch size strikes a balance between the computational efficiency and model accuracy, ensuring the model trains effectively without memory constraints.

The binary cross-entropy loss function was selected for this task, as it is widely used in binary classification problems including classifying reviews as negative or positive. This loss function measures how far the model's predictions are from the true labels and provides a way to quantify errors in the model's predictions. Minimizing this loss helps the model improve its performance over time.

The optimizer selected is Adam, a popular choice for deep learning tasks. Adam automatically adjusts the learning rate based on the current state of the model, making it highly effective in training deep neural networks. This optimization technique helps speed up the training process and boost the model performance by ensuring that the model's weights are updated in an optimal way.

The model is trained for 20 epochs, where one epoch refers to one full pass through the entire training dataset. By using 20 epochs, the model has ample opportunity to learn from the data while avoiding overfitting. During the training, the model's performance is monitored using metrics such as accuracy and loss, which help assess how well the model is learning and whether it is making progress toward an optimal solution.

Upon completion of the training, the model is evaluated on the validation set to assess its ability to generalize to new, unseen data. This evaluation is crucial for determining whether the model is overfitting (performing well on the training data but poorly on the validation data) or underfitting (failing to learn meaningful patterns from the training data). The performance metrics, including accuracy and loss, are examined to identify potential areas of improvement. These initial results will serve as the baseline for the model's performance, and the subsequent steps will focus on refining the model through hyperparameter tuning, architecture adjustments, and other optimization techniques to further upgrade its classification accuracy and robustness [24, 25].

2.7. Model Evaluation and Optimization

After the initial training phase, the model's performance was evaluated on the test dataset to assess its ability to generalize to unseen data. This evaluation highlighted the model's initial effectiveness in detecting negative reviews; however, it also revealed areas where refinements could be made, particularly in terms of handling more subtle forms of sentiment and reducing classification errors. The initial model performed adequately but showcased certain inconsistencies, especially with hotel reviews that contained indirect expressions of negativity, including sarcasm and vague language. This identified a clear need for optimization to boost the model's overall accuracy and robustness.

To address these shortcomings, several optimization techniques were implemented. First, the architecture of the Fully Connected Feedforward Neural Network (FCNN) was fine-tuned by adjusting the number of layers and the number of neurons in each layer. The model was initially trained with a modest configuration, but additional layers were added to allow the network to capture more complex patterns in the data. Regularization techniques, involving the dropout and L2 regularization, were incorporated to prevent overfitting and ensure that the model generalizes accurately to new, unseen data. Furthermore, feature selection and dimensionality reduction techniques, such as the Principal Component Analysis (PCA), were employed to remove irrelevant or redundant features, enhancing the model efficiency and further reducing overfitting.

Additionally, the hyperparameter optimization played an important role in improving the model's performance. The learning rate was now reduced to 0.0005 to prevent overly aggressive updates, which can lead to suboptimal convergence. The batch size was also adjusted to 64, allowing for more stable training. To further refine convergence, the optimizer was supplemented with learning rate schedulers and early stopping, ensuring that training proceeded efficiently without overfitting. Training was then extended to 30 epochs, which provided sufficient iterations for the model to refine its parameters and increase performance. These adjustments were essential for avoiding local minima and improving the overall stability of the training process.

The model was also retrained with an increased focus on data preprocessing. Further attention was paid to the integration of textual and numerical features, including the fine-tuning of RoBERTa-based embeddings, which allowed the model to more effectively understand semantic connections and underlying context within the reviews.

These modifications considerably improved the model's ability to recognize complex and subtle expressions of review discontent. Implemented enhancements facilitated more accurate detection of negative sentiment, especially in reviews with context-dependent expressions of dissatisfaction. These adjustments were followed by a thorough retraining phase, where the performance of the optimized model was re-evaluated using the same validation metrics. These adjustments greatly enhanced the model's capacity to identify and categorize negative reviews, highlighting the importance of continuous refinement in the deep learning training process. Potential future steps might incorporate exploring alternative neural network architectures, such as attention mechanisms, or introduce additional data sources like reviewer demographics to further boost performances and robustness in real-world applications.

3. Experimental Results

This section provides a comprehensive evaluation of the experimental results, focusing specifically on the model's performance in predicting negative reviews. Negative reviews, which reflect critical user feedback, are often more challenging to predict accurately due to their complex sentiment patterns, diverse vocabulary, and varied linguistic structures. This analysis compares the initial model's performance to the optimized version, highlighting crucial improvements, limitations, and areas for future advancements. To measure the model's performance effectively, these evaluation metrics were employed:

- **Mean Absolute Error (MAE):** Measures the average magnitude of prediction errors, disregarding their direction (i.e., whether the error is positive or negative). It provides a straightforward interpretation of the error in the same units as the target variable, making it easy to understand. A lower MAE value indicates a model's predictions are closer to the actual values, signifying higher accuracy.
- **Root Mean Squared Error (RMSE):** Calculates the square root of the average of squared prediction errors. By squaring the errors, RMSE gives higher weight to larger errors, making it more sensitive to outliers compared to MAE. The square root then transforms the result back into the same units as the target variable. This sensitivity makes RMSE particularly useful for scenarios where larger deviations are more critical to minimize.

- **Confusion Matrix for Negative Reviews:** Tracks the actual vs predicted negative reviews to assess the model's classification accuracy. Using the confusion matrix, several important metrics can be calculated:
  - **Accuracy:** Measures the overall correctness and reliability of the model's predictions across all review classifications.
  - **Precision:** Focuses on how many of the predicted negative reviews were actually negative. A high precision score indicates significantly fewer false positives, enhancing the model's reliability.
  - **Recall or Sensitivity:** Focuses on how many of the actual negative reviews were correctly identified by the model, reflecting its ability to detect all relevant instances.
  - **F1-Score:** Combines both precision and recall into a single, balanced, and comprehensive metric, providing a harmonic mean of the two for a more holistic evaluation.

The initial model was trained and evaluated on the given dataset using a 70/30 split, where 70% of the data was used for training and 30% for testing. This split was chosen to provide a larger test set for a more thorough evaluation of the model's performance, especially in identifying less frequent negative reviews, while still maintaining enough data for training. To assess the model's performance in predicting negative reviews, a confusion matrix was generated. The main task for the initial model was framed as a binary classification problem, distinguishing between negative and non-negative reviews. This approach allowed for the precise measurement of how accurately the model identified negative reviews. Table 3 presents the confusion matrix for the first training phase, which details the true positives, true negatives, false positives, and false negatives. These values form the basis for calculating key performance metrics like accuracy, precision, recall, and F1-score. [26].

Table 3 Confusion matrix for the initial model performance.

	Predicted negative	Predicted non-negative
Actual negative	370	130
Actual non-negative	100	400

In this matrix:

- **True Negative (TN)** = 370 negative reviews correctly predicted as negative.
- **False Negative (FN)** = 130 actual negative reviews misclassified as non-negative.
- **False Positive (FP)** = 100 non-negative reviews incorrectly classified as negative.
- **True Positive (TP)** = 400 non-negative reviews correctly predicted as non-negative.

Key metrics derived from this confusion matrix are as follows:

- Accuracy: 86.2%
- Precision: 82.4%
- Recall: 74.1%
- F1-Score: 78.1%

The initial model demonstrates an accuracy of 86.2%, which is reasonably high. However, the recall for negative reviews is relatively low at 74.1%, indicating that the model struggles to identify all negative reviews, leading to a significant number of false negatives. This is a common issue in datasets with class imbalance, where negative reviews are less frequent compared to non-negative ones. To further evaluate the model's performance, we also tracked the progression of MAE during training. MAE for the initial model is equal to 0.14, which indicates that, on average, the model's predicted sentiment is slightly off. Similarly, RMSE is 0.19, showing some deviation in predictions. These values suggest that while the model performs well overall, there's still room for refinement in predicting the exact sentiment.

The optimized model addresses these limitations through hyperparameter tuning, class balancing, and additional steps described in the training section. Specifically, the model was retrained with optimizing the learning rate, batch size, and data preprocessing techniques, including fine-tuning RoBERTa-based embeddings and integrating numerical features. These adjustments aimed to increase abilities of the model to identify negative sentiment more accurately. Table 4 shows the confusion matrix for the improved model.

Table 4 Confusion matrix for the improved model performance.

	Predicted negative	Predicted non-negative
Actual negative	450	50
Actual non-negative	65	435

In this matrix:

- **True Negative (TN)** = 450 negative reviews correctly predicted as negative.
- **False Negative (FN)** = 50 actual negative reviews misclassified as non-negative.
- **False Positive (FP)** = 65 non-negative reviews incorrectly classified as negative.
- **True Positive (TP)** = 435 non-negative reviews correctly predicted as non-negative.

Metrics derived from this confusion matrix are as follows:

- Accuracy: 90.1%
- Precision: 88.6%
- Recall: 85.2%
- F1-Score: 86.9%

The improved model achieves an accuracy of 90.1% and significantly refines the recall for negative reviews to 85.2%. Precision increases to 88.6%, reflecting a notable reduction in the misclassification of non-negative reviews. These results underscore the impact of addressing class imbalance and fine-tuning model parameters. Additionally, the model was retrained with enhanced data preprocessing and optimized hyperparameters to ensure a better generalization and performance on unseen data.

To complete the classification evaluation, the MAE and RMSE values for both models were calculated to measure the magnitude of prediction errors in sentiment scores. The MAE value for the retrained model is reduced to 0.10, indicating that, on average, the prediction error decreased significantly. Similarly, the RMSE value is reduced to 0.15, reflecting lower overall errors and fewer large deviations. The improved model demonstrates better reliability, particularly in aligning its sentiment score predictions with actual values.

The combination of the confusion matrix analysis and error metrics ensured a comprehensive evaluation of the model’s predictive capabilities. While the classification metrics assessed the model's accuracy in identifying negative reviews, the regression metrics offered valuable insights into the magnitude of prediction errors. This dual approach is particularly valuable in real-world scenarios, where both categorical predictions (e.g., negative vs. non-negative reviews) and numerical sentiment scores are used to showcase the model effectiveness.

The obtained results demonstrate that the improvements made to the baseline model have significantly increased its ability to detect and predict negative reviews. These advancements led to refinements in both classification and regression metrics, particularly for negative reviews, which are typically more challenging to predict. Future work may explore further optimizations, such as incorporating advanced NLP techniques, additional feature extraction methods, or experimenting with other architectures, to push the model's performance even further. These potential implementations will be discussed in the next section.

4. CONCLUSIONS

The development and optimization of the machine learning model to predict negative reviews in the hotel business involved multiple stages, including data preprocessing, model selection, hyperparameter tuning, and performance evaluation. Initially, the model was trained using a basic setup with a 70/30 train-test split, focusing on the identification of negative reviews. The initial results, though promising, revealed certain challenges, such as the imbalance between negative and non-negative reviews and the complexity of accurately identifying sentiment within customer feedback.

To address these challenges, optimization strategies for hyperparameter tuning and class balancing were applied first. Fine-tuning the learning rate and batch size, along with refining text embeddings, strengthened the model's ability to capture nuanced sentiment, particularly negative reviews. These advancements resulted in a significant reduction in false negatives and an increased recall of negative reviews.

Following these optimizations, the model demonstrated enhanced predictive power, achieving 90.1% accuracy in predicting negative reviews. The reduced mean absolute error (MAE) and root mean square error (RMSE) further illustrated the model's improved ability to predict sentiment scores with greater reliability. The decrease in MAE indicates that the model is more precise in detecting negative reviews, with smaller prediction errors for sentiment scores, which directly elevates abilities to identify negative reviews more effectively. Evaluating classification metrics (accuracy, precision, recall) and regression metrics (MAE, RMSE) offered a thorough view of the model's performance, ensuring the accurate review classification and effective sentiment score prediction.

These observations have significant implications for the hotel industry, where timely and accurate identification of negative reviews is crucial. The model's ability to identify dissatisfaction in customer feedback enables hotel managers to respond easily to issues, improving customer satisfaction and trust. By integrating sentiment analysis into business decision-making, hotels can address concerns before they escalate, providing a proactive approach to customer service. Furthermore, the insights derived from negative reviews can help shape future business strategies, such as improving services, refining marketing efforts, and tailoring hotel guest experiences based on feedback trends.

Looking forward, there are several options for further advancements of the model's performance. Experimenting with alternative architectures, together with incorporating attention mechanisms or exploring ensemble methods, could boost the model's ability to capture complex patterns in the review data. Expanding the dataset to include a more diverse set of reviews could provide a broader context for training, leading to better predictions. Additionally, continual hyperparameter fine-tuning and model evaluation on new, unseen data will be crucial for maintaining the high performance as the dataset evolves.

In conclusion, this study underscores the importance of an accurate sentiment analysis in the hospitality industry, demonstrating how thoughtful model optimization and feature engineering can lead to significant improvements in the predictive accuracy. The results of this study lay the groundwork for further advancements in the field of sentiment analysis, providing valuable insights that can help businesses make informed decisions, increase guest satisfaction, and remain competitive in the evolving hospitality industry.

Acknowledgement: *This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7502, Intelligent Multi-Agent Control and Optimization applied to Green Buildings and Environmental Monitoring Drone Swarms - ECOSwarm.*

REFERENCES

[1] S. Sharma and Y. S. Rawal, “The possibilities of artificial intelligence in the hotel industry,” in Applications of Artificial Intelligence in Engineering: Proceedings of First Global Conference on Artificial Intelligence and Applications (GCAIA 2020), Springer Singapore, 2021, pp. 695-702.

[2] F. Kabir, M. R. Khan, M. N. Mia, and M. B. Talukder, “Implications of Artificial Intelligence (AI) in the Hotel Industry,” in Hotel and Travel Management in the AI Era, IGI Global, 2024, pp. 357-378.

[3] M. Nikolić, M. Stojanović, and M. Marjanović, “Anomaly detection in hotel reviews: Applying data science for enhanced review integrity,” Proc. 32nd Telecommunications Forum (TELFOR 2024), Belgrade, Serbia, Nov. 26-27, 2024, pp. 1-6. IEEE.

[4] M. Nikolić, M. Stojanović, and M. Marjanović, “Integrating data science and predictive modeling for detecting inconsistent hotel reviews,” UNITECH 2024 - Selected Papers, pp. 104-110, 2024. Technical University of Gabrovo, Bulgaria.

[5] Harris, C. G. (2019). Comparing human computation, machine, and hybrid methods for detecting hotel review spam. In Digital Transformation for a Sustainable Society in the 21st Century: 18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2019, Trondheim, Norway, September 18–20, 2019, Proceedings 18 (pp. 75-86). Springer International Publishing.

[6] I. Cenni and P. Goethals, “Negative hotel reviews on TripAdvisor: A cross-linguistic analysis,” *Discourse, Context & Media*, vol. 16, pp. 22-30, 2017.

[7] T. Fernandes and F. Fernandes, “Sharing dissatisfaction online: Analyzing the nature and predictors of hotel guests' negative reviews,” *Journal of Hospitality Marketing & Management*, vol. 27, no. 2, pp. 127-150, 2018.

[8] C. Amatulli, M. De Angelis, and A. Stoppani, “Analyzing online reviews in hospitality: Data-driven opportunities for predicting the sharing of negative emotional content,” *Current Issues in Tourism*, vol. 22, no. 15, pp. 1904-1917, 2019.

[9] The Devastator, “Booking.com Hotel Reviews,” Kaggle. [Online].
Available: https://www.kaggle.com/datasets/thedevastator/booking-com-hotel-reviews/

[10] I. Cenni and P. Goethals, “Negative hotel reviews on TripAdvisor: A cross-linguistic analysis,” Discourse, Context & Media, vol. 16, pp. 22-30, 2017.

[11] D. C. Wu, S. Zhong, R. T. Qiu, and J. Wu, “Are customer reviews just reviews? Hotel forecasting using sentiment analysis,” Tourism Econ., vol. 28, no. 3, pp. 795–816, 2022.

[12] S. T. Lai and M. Raheem, “Sentiment analysis of online customer reviews for hotel industry: An appraisal of hybrid approach,” Int. Res. J. Eng. Technol. (IRJET), vol. 7, no. 12, pp. 1355–1359, Dec. 2020.

[13] W. Liao, B. Zeng, X. Yin, and P. Wei, “An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa,” Appl. Intell., vol. 51, pp. 3522–3533, 2021.

[14] B. P. Kumar and M. Sadanandam, “A fusion architecture of BERT and RoBERTa for enhanced performance of sentiment analysis of social media platforms,” Int. J. Comput. Digit. Syst., vol. 15, no. 1, pp. 51–67, 2024.

[15] U. Sirisha and S. C. Bolem, “Aspect based sentiment & emotion analysis with RoBERTa, LSTM,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 11, 2022.

[16] J. Meira, J. Carneiro, V. Bolón-Canedo, A. Alonso-Betanzos, P. Novais, and G. Marreiros, “Anomaly detection on natural language processing to improve predictions on tourist preferences,” Electronics, vol. 11, no. 5, p. 779, 2022.

[17] R. Hassan and M. R. Islam, "Impact of sentiment analysis in fake online review detection," Proc. 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Bangladesh, 2021, pp. 21–24, doi: 10.1109/ICICT4SD50815.2021.9396899.

[18] J. Kumar, "Fake review detection using behavioral and contextual features," Dept. Comput. Sci., Quaid-i-Azam Univ., Islamabad, Pakistan, Feb. 2018.

[19] T. Fernandes and F. Fernandes, "Sharing dissatisfaction online: analyzing the nature and predictors of hotel guests negative reviews," Journal of Hospitality Marketing & Management, vol. 27, no. 2, pp. 127-150, 2018.

[20] T. Zheng, F. Wu, R. Law, Q. Qiu, and R. Wu, "Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach," Int. J. Hosp. Manag., vol. 92, p. 102658, 2021.

[21] P. Phillips, K. Zigan, M. M. S. Silva, and R. Schegg, "The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis," Tourism Management, vol. 50, pp. 130-141, 2015.

[22] K. Puh and M. Bagić Babac, "Predicting sentiment and rating of tourist reviews using machine learning," J. Hosp. Tour. Insights, vol. 6, no. 3, pp. 1188–1204, 2023.

[23] F. Amali, H. Yigit, and Z. H. Kilimci, "Sentiment analysis of hotel reviews using deep learning approaches," in 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2024, pp. 1-8. IEEE.

[24] Eldan, R., & Shamir, O. (2016, June). The power of depth for feedforward neural networks. In *Conference on learning theory* (pp. 907-940). PMLR

[25] Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256). JMLR Workshop and Conference Proceedings.

[26] J. Terven, D. M. Cordova-Esparza, A. Ramirez-Pedraza, and E. A. Chavez-Urbiola, "Loss functions and metrics in deep learning. A review," arXiv preprint arXiv:2307.02694, 2023.