The Power of Words: Leveraging Deep Learning Techniques to Predict Hotel Ratings from User Reviews

Milena Nikolić *, Miloš Stojanović , Marina Marjanović

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.0921.v1

Keywords: hotel reviews; rating prediction; sentiment analysis; deep learning; natural language processing; anomaly detection; hospitality analytics; recurrent neural networks

 doi:10.20944/preprints202604.0921.v1

Article

The Power of Words: Leveraging Deep Learning Techniques to Predict Hotel Ratings from User Reviews †

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

[1] The Academy of Applied Technical and Preschool Studies, Niš, Serbia
[2] Singidunum University, Belgrade, Serbia
* Correspondence: milena.nikolic@akademijanis.edu.rs; Tel.: +381-63-70-101-69
† This paper is an extended version of our paper published in *Proceedings of the 24th International Symposium INFOTEH-JAHORINA (INFOTEH 2025)*, Jahorina, Bosnia and Herzegovina, 19–21 March 2025; pp. 1–6. IEEE. https://doi.org/10.1109/INFOTEH64129.2025.10959201.

Abstract

Online reviews represent a major source of information for evaluating customer experience and supporting decision-making in the hospitality industry, yet rating prediction from review content remains challenging because review text is often short, noisy, and internally inconsistent. This study presents a deep learning framework for predicting hotel ratings from guest reviews while explicitly addressing data quality before model training. Data reliability is treated as a central modeling concern. The proposed methodology combines review titles, review texts, and associated tags with a structured preprocessing pipeline that incorporates sentiment inconsistency detection, textual similarity analysis, deviation analysis based on correlation, and reviewer behavior profiling to identify unreliable observations. On the filtered corpus, we evaluate multiple predictive architectures, including LSTM, Bidirectional LSTM variants, and DistilBERT, for review-level rating prediction, and we further examine hotel-level temporal forecasting through aggregated historical review signals over a 30-day horizon. The results indicate that model performance depends strongly on both data reliability and architectural choice. Among recurrent models, BiLSTM with self-attention achieves the best performance, while DistilBERT yields the strongest overall results. Ablation analysis confirms that the full preprocessing pipeline consistently improves prediction quality, and the forecasting experiments indicate that aggregated review features contain useful information for short-term hotel rating dynamics. The study contributes a systematic and practically relevant framework for rating prediction and hospitality analytics in support of reputation management.

Keywords: hotel reviews; rating prediction; sentiment analysis; deep learning; natural language processing; anomaly detection; hospitality analytics; recurrent neural networks

1. Introduction

The hospitality industry has undergone a fundamental change in the way customer feedback is gathered, interpreted, and used in practice. Online review platforms such as Booking.com and TripAdvisor have become major sources of information for millions of travelers choosing accommodation while also shaping how hotels are perceived in an increasingly competitive market. Even a single percentage change in rating has been linked to measurable differences in booking conversion, which gives user ratings direct business relevance. As a result, understanding and predicting online ratings has become an analytical challenge with meaningful practical implications for hotel management [1].

Although online platforms provide a large amount of review data, identifying trustworthy information within it remains complex. User reviews are inherently noisy, as they contain inconsistencies between expressed sentiment and assigned numerical scores, near-duplicate entries that may reflect

 doi:10.20944/preprints202604.0921.v1

coordinated manipulation, and behavioral anomalies associated with fraudulent posting patterns. Traditional approaches to rating prediction have largely treated review text as clean input, overlooking the effect that unreliable or inconsistent reviews have on downstream model performance [2]. This gap between data quality and model design represents one of the central motivations for this study.

Recent advances in deep learning and natural language processing have considerably expanded the methodological possibilities for analyzing review content. Recurrent architectures such as Long Short-Term Memory networks and their bidirectional extensions have demonstrated strong performance on sequential text data, while transformer models, including BERT, RoBERTa, and DistilBERT, have substantially improved contextual language understanding in hospitality analytics. However, many existing studies devote limited attention to data preprocessing, and only a small number systematically examine how individual preprocessing components affect prediction quality [3]. In addition, whether aggregated textual information from past reviews can be used to anticipate future changes in hotel ratings remains largely underexplored in the literature.

The present study addresses these limitations through four principal contributions. First, we present a multi-component anomaly preprocessing pipeline that integrates VADER sentiment inconsistency detection, TF-IDF cosine similarity filtering, correlation deviation analysis across review components, and user behavior profiling to identify and remove unreliable reviews prior to model training. Second, we evaluate LSTM, Bidirectional LSTM, and DistilBERT architectures for predicting review-level numerical ratings from the textual and semantic content of individual reviews. Third, we conduct a systematic ablation study that quantifies the independent contribution of each preprocessing stage to overall prediction accuracy. Fourth, we examine whether aggregated information derived from past reviews can support short-term forecasting of hotel rating behavior, thereby connecting review understanding to hotel reputation dynamics.

Experiments are conducted on a publicly available dataset of 26,675 hotel reviews from Booking.com, covering 819 distinct hotels [4]. The dataset includes numerical ratings, together with textual content, reviewer metadata, and temporal information, enabling the primary rating prediction task and the temporal forecasting extension. Among recurrent architectures, the optimized BiLSTM with self-attention achieves MAE = 0.5753 and RMSE = 0.8636 on the original 1–10 rating scale, while the best overall result is obtained by DistilBERT with MAE = 0.4925 and RMSE = 0.7368. In practical terms, an MAE of 0.4925 means that the predicted rating differs from the true user rating by about half a rating point on average. The temporal analysis further indicates that aggregated information from past reviews contains useful information for short-term hotel-level rating forecasting.

This paper is an extended version of a conference paper presented at the 24th International Symposium INFOTEH-JAHORINA [5]. Compared with the original contribution, the present work introduces a revised prediction target that eliminates label leakage risk, adds two model architectures, includes a full ablation study, and contributes temporal forecasting as a secondary task. Because the conference version predicted hotel-level `avg_rating` and reported its strongest results after normalization, whereas the present study predicts review-level ratings and reports errors on the original 1–10 scale, the two sets of performance values are not directly comparable. More broadly, the paper integrates several related directions from our previous work, including anomaly detection, predictive modeling, and automated analysis of negative hotel reviews, into a more complete study of hotel rating prediction from online review content.

2. Related Work

Online customer reviews present one of the most studied forms of content in tourism and hospitality research. The studies relevant to this paper fall into four main areas: sentiment analysis of hotel feedback, deep learning for rating prediction, detection of unreliable or manipulated reviews, and transformer and large language model methods in tourism NLP. Although these areas have grown quickly, fewer studies combine careful preprocessing with prediction or examine how review meaning relates to hotel rating trends over time. Our earlier work on review inconsistency, anomaly detection,

 doi:10.20944/preprints202604.0921.v1

and negative review interpretation also pointed to the need for a more unified pipeline that considers text meaning and data reliability. The following subsections review the main findings and show how the present study relates to this broader area of research.

2.1. Sentiment Analysis in Hospitality Reviews

Sentiment analysis of hotel reviews has attracted considerable interest because online feedback is widely available and highly relevant for practical decisions in hospitality. Early studies showed that hybrid methods, which combine lexicon tools with machine learning, can capture useful sentiment signals in informal review language, including slang, abbreviations, and vocabulary specific to the domain. Related work also showed that sentiment analysis can support reservation and service quality decisions when it is combined with multi-criteria decision frameworks [6].

Recent studies have increasingly moved toward deep learning models, as they offer a stronger ability to capture contextual meaning. Wen et al. [7] applied BERT and ERNIE to Chinese hotel reviews and reported improvements over traditional methods, although transfer across languages remained difficult. Husein [8] compared LSTM and ELECTRA on TripAdvisor reviews and found that transformer models performed better on longer and more complex texts. Chen et al. [9] showed that combining text, numerical information, and tags can improve performance over sentiment models based only on textual input.

A similar pattern appeared in our earlier work on automated detection of negative hotel reviews, where deep learning methods were effective in identifying complaint language and poor guest experiences in review text [10]. At the same time, those results suggested that sentiment alone is not enough, because review text can still be inconsistent with the assigned rating. This is one reason why the present study positions sentiment as a component within a broader reliability framework rather than treating it as a sufficient standalone indicator of user satisfaction.

2.2. Rating Prediction with Deep Learning

Predicting a numerical rating from review text is more demanding than simple sentiment classification since the model must learn smaller differences across a continuous scale. Puh and Bagić Babac [11] compared several machine learning methods for tourist rating prediction and found that review length and sentiment consistency were among the most informative features. Zhang and Wu [12] extended this direction with an explainable deep learning framework for hotel demand forecasting, showing that review features can predict aggregate hotel outcomes and can also be interpreted through specific language cues.

Hossen et al. [13] showed that deep learning models outperform traditional classifiers on hotel review business prediction because they capture sequential patterns in text more effectively. Zhang et al. [14] reported strong results with a Bidirectional LSTM model that combined Word2Vec embeddings, TF-IDF features, and attention. Ganji et al. [15] showed that handling class imbalance matters when hotel ratings are heavily skewed toward higher values. Zhao et al. [16] also found that topic and sentiment user attitude signals are closely related to aggregate hotel ratings.

Our earlier studies on inconsistent and anomalous hotel reviews support the same point. We found that prediction quality depends not only on the model itself but also on the quality and coherence of the review data. In particular, mismatches between review text, user ratings, and posting behavior can weaken later analysis if they are not handled first [17]. This directly motivates the present study, where rating prediction is treated as a text modeling problem and a data reliability problem.

2.3. Unreliable and Fake Review Detection

The reliability of online review systems has become a major concern because misleading reviews can distort ratings and influence customer decisions. Many studies have approached this problem with supervised learning methods. Alsubari et al. [18] built a fake review detection framework based on n-gram features and sentiment scores and found that ensemble methods performed better than single models. Duma et al. [19] proposed a deep hybrid model that combines review text, overall

 doi:10.20944/preprints202604.0921.v1

ratings, and aspect ratings and showed that consistency across review components is a useful signal for detecting fake content. Their findings are especially important because they show that deception is often reflected in inconsistencies across several parts of the same review rather than in a single indicator. Prasetyaningrum et al. [20] combined transformer-based sentiment classification with digital forensics metadata analysis and reported strong results on multilingual hospitality review data, which emphasizes the value of using textual and behavioral evidence.

In our previous studies, we showed that inconsistent hotel reviews can be identified through disagreement between sentiment and rating, textual similarity, and reviewer behavior patterns, and that anomaly filtering can improve the quality of datasets used for later analytics. These findings helped shape the preprocessing pipeline for the present study, which treats review credibility as something that should be modeled before regression rather than assumed from the start [21].

2.4. Transformer and LLM Approaches in Tourism NLP

Recent work in tourism NLP has increasingly used pretrained and hybrid neural architectures. Zhang et al. [22] proposed a hotel review classification method based on a text pretraining heterogeneous graph neural network and showed that structural links between textual and contextual information can improve review analysis. Deng et al. [23] studied CNN LSTM models enriched with BERT representations and attention and reported that this combination captures local phrase patterns and broader context effectively. Chen et al. [24] proposed a two-channel hybrid model that processes sentiment and semantic information separately before combining them. Yuan [25] fine-tuned DistilBERT in an ensemble framework for hotel rating prediction and showed that strong performance can be achieved with lower computational cost than larger transformer models. Roumeliotis et al. [26] compared recent GPT Omni models with BERT for tourism review classification and showed that fine-tuned BERT remains highly competitive when labeled data from the target domain are available. This point is further supported by recent survey evidence showing that preprocessing remains influential for transformer models and traditional classifiers in text classification tasks [27].

Overall, these studies provide a strong basis for the model choices adopted in this paper. They support the use of DistilBERT as an efficient baseline and justify comparison with recurrent models that remain relevant for review prediction. At the same time, the practical value of stronger language models depends on computational cost, model size, and the quality of training data available for the target domain, making this comparison especially relevant for hospitality research settings.

3. Dataset and Problem Formulation

The empirical analysis uses a publicly available dataset of Booking.com hotel reviews distributed through Kaggle. The raw corpus contains 26,675 review records associated with 819 unique hotels and spans the period from July 2018 to July 2021. On average, each hotel is represented by 32.2 reviews, although the distribution is highly uneven, ranging from 1 to 846 reviews per property. The dataset combines unstructured textual information with structured metadata, making it suitable for review-level prediction and hotel-level temporal analysis. The main textual fields are the review title, the review body, and tags provided by reviewers, while the structured attributes include the reviewer rating, the hotel average rating, reviewer nationality, posting date, crawl date, and the number of attached images. Table 1 summarizes the principal variables used in this study.

 doi:10.20944/preprints202604.0921.v1

Table 1. Principal attributes of the Booking.com review dataset used in this study.

Attribute	Type	Description
`review_title`	Text	Short summary written by the reviewer
`review_text`	Text	Full narrative description of the stay
`tags`	Text	Labels provided by reviewers (trip type, room type, etc)
`rating`	Numerical	Review score assigned by the reviewer on a 1–10 scale
`avg_rating`	Numerical	Average hotel rating across all reviews for that property
`reviewed_at`	Datetime	Review posting date and time
`reviewed_by`	Text	Reviewer identifier
`hotel_name`	Text	Hotel name used as grouping key
`nationality`	Text	Reviewer country of origin

Initial cleaning removed 289 records with missing rating or text values and 2 additional records with no alphanumeric content, leaving 26,384 reviews for the main modeling task. For Task B, 105 records lacking valid timestamps were excluded only from hotel-level rolling window construction. This count differs from the conference version because the earlier formulation required valid `avg_rating` and `reviewed_at` values during preprocessing, whereas the present review-level task is defined with respect to the individual reviewer rating. The rating distribution remains strongly right-skewed, with 88.7% of review-level scores at or above 7.0. Table 2 summarizes the key corpus statistics after cleaning.

Table 2. Summary statistics of the Booking.com hotel review dataset after initial cleaning.

Property	Value
Corpus overview	
Total records (raw)	26,675
Records after removing missing rating / review text	26,386
Records after filtering the empty content	26,384
Records with valid timestamps for Task B	26,279
Hotel coverage	
Unique hotels	819
Reviews per hotel (min / mean / max)	1 / 32.2 / 846
Hotels eligible for Task B ($\geq$20 reviews with valid timestamps)	273
Review-level rating (`rating`)	
Scale	1.0–10.0
Mean $\pm$ standard deviation	8.56 ± 1.57
Proportion of ratings ≥ 7.0	88.7%
Hotel-level average rating (`avg_rating`)	
Scale	3.8–10.0
Mean $\pm$ standard deviation	8.45 ± 0.72

Figure 1 illustrates the review-level rating distribution in the cleaned dataset and confirms the strong concentration of higher scores, a pattern typical of hospitality platforms.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0921.v1

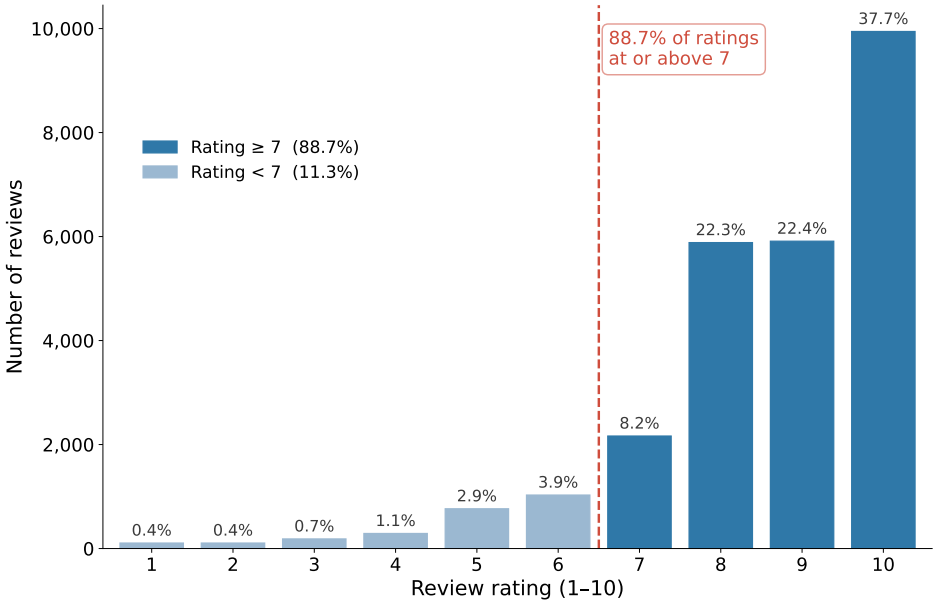

Figure 1. Distribution of review-level ratings in the cleaned Booking.com dataset ($N = 26{,}384$). Most reviews are concentrated in the upper part of the 1–10 scale, showing the positive skew of the corpus.

Formally, the cleaned review dataset is represented as

$$\mathcal{D} = \{(x_i, y_i, t_i, h_i)\}_{i=1}^{N}, \tag{1}$$

where each tuple corresponds to one review in the dataset. In this notation, x_i denotes the input representation of review i, $y_i \in [1, 10]$ is the numerical rating assigned by the reviewer, t_i is the posting timestamp, h_i identifies the corresponding hotel, and $N = 26{,}384$ is the total number of reviews retained after cleaning. At the input representation stage, each instance combines three textual components with associated metadata:

$$x_i = \left(T_i^{\text{title}},\ T_i^{\text{text}},\ T_i^{\text{tags}},\ m_i\right), \tag{2}$$

where T_i^{title}, T_i^{text}, and T_i^{tags} denote the title, review body, and tags of review i, while m_i represents the structured metadata linked to that review, such as temporal information and user identity. This formulation represents the full review record available to the pipeline, although the predictive models use different subsets of these inputs depending on the task.

3.1. Task A: Review-Level RATING prediction

The primary task is formulated as a supervised regression problem, where the goal is to learn a function

$$f_\theta : \mathbb{R}^d \to \mathbb{R}, \tag{3}$$

such that the predicted score $\hat{y}_i = f_\theta(x_i)$ approximates the observed rating y_i. This formulation is preferred over direct prediction of the hotel average rating, `avg_rating`, because that target introduces a risk of label leakage. Since the hotel average rating is computed from the same group of reviews that includes the review being analyzed, the model may capture hotel-level regularities instead of learning the semantic relationship between review content and the reviewer's own score. In this sense, Task A provides a more direct test of how well textual review content explains individual user satisfaction at the level of a single review, without relying on broader hotel averages.

Model training for Task A minimizes the mean squared error:

 doi:10.20944/preprints202604.0921.v1

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{4}$$

3.2. Task B: Temporal Hotel Rating Forecasting

The secondary task investigates whether historical review signals can be aggregated to forecast future hotel-level rating behavior. Let $\mathcal{H}$ denote the set of hotels retained for temporal analysis, and let $\mathcal{R}_{h,\tau}$ be the set of reviews associated with hotel $h \in \mathcal{H}$ during rolling time window τ. For each hotel-window pair, the aggregated representation is defined as

$$\bar{x}_{h,\tau} = \Phi(\{x_i : i \in \mathcal{R}_{h,\tau}\}), \tag{5}$$

where $\Phi(\cdot)$ summarizes historical review content through sentiment indicators, review volume, and the proportion of anomalous reviews detected by the preprocessing pipeline. The forecasting target is the hotel average rating in the 30-day horizon:

$$\bar{y}_{h,\tau+\Delta} = \frac{1}{|\mathcal{R}_{h,\tau+\Delta}|}\sum_{i\in\mathcal{R}_{h,\tau+\Delta}} y_i, \qquad \Delta = 30 \text{ days}. \tag{6}$$

Unlike Task A, which focuses on individual reviews, Task B examines whether patterns accumulated across past reviews can capture broader changes in hotel reputation over time.

The temporal prediction objective is then written as

$$\hat{\bar{y}}_{h,\tau+\Delta} = g_\phi(\bar{x}_{h,\tau}, \bar{x}_{h,\tau-1}, \ldots), \tag{7}$$

where g_ϕ denotes a forecasting model operating on past aggregated review signals. In this implementation, the notation is simplified, but the evaluated baselines use the most recent aggregated window as input. Only hotels with at least 20 retained reviews with valid timestamps are included in Task B to ensure statistically meaningful summaries, and all train-test splits are constructed in strict chronological order to prevent temporal leakage.

4. Preprocessing and Feature Engineering

The preprocessing pipeline converts raw hotel reviews into a structured and reliable representation suitable for review-level prediction and hotel-level temporal forecasting. It consists of six connected stages: text cleaning and normalization, sentiment analysis, similarity detection, correlation analysis, user behavior analysis, and composite anomaly score construction. These stages allow the study to move beyond conventional text operations by accounting for review reliability before model training, which is important because recent evidence shows that preprocessing choices can still materially affect downstream performance even in modern pretrained language model settings.

4.1. Text Cleaning and Normalization

Before feature extraction, the textual fields were standardized to reduce noise introduced by platform formatting and web scraping. For each review i, the combined textual representation is defined as

$$T_i = T_i^{\text{title}} \oplus T_i^{\text{text}} \oplus T_i^{\text{tags}}, \tag{8}$$

where $\oplus$ denotes concatenation of the title, body text, and tags.

The normalized review text is then obtained through

$$T_i^{\text{norm}} = f_{\text{clean}}(T_i), \tag{9}$$

where $f_{\text{clean}}(\cdot)$ applies lowercasing, whitespace normalization, and removal of non-alphanumeric artifacts uniformly across all textual inputs. One missing title and 184 missing tag fields were replaced

 doi:10.20944/preprints202604.0921.v1

by the placeholder value "Unknown" to preserve structural consistency across the three textual channels. No stemming or lemmatization was applied, because the downstream neural architectures are designed to learn lexical and contextual relationships directly from the retained tokens. This also helps preserve short evaluative expressions that are common in hospitality reviews and may carry useful predictive information. After cleaning, the corpus contains 26,384 usable reviews, with review lengths ranging from 1 to 614 words, a mean of 28.5 words, and a median of 13 words, confirming the short and highly skewed nature of the review text.

4.2. Sentiment Analysis

Sentiment analysis was performed independently on review titles, review bodies, and tags using the VADER sentiment analyzer, which has been shown to perform well on short and informal text. For each review i, the component sentiment scores are represented as

$$S_i^{\text{title}}, \quad S_i^{\text{text}}, \quad S_i^{\text{tags}} \in [-1, 1], \tag{10}$$

where -1 denotes strongly negative polarity and $+1$ strongly positive polarity. In addition to these continuous scores, a binary mismatch indicator between sentiment and rating, M_i^{sent}, was defined for reviews whose textual polarity was strongly inconsistent with the assigned numerical rating. This criterion flagged 724 reviews (2.7% of the corpus). Average values were 0.355 for title sentiment, 0.321 for text sentiment, and 0.005 for tag sentiment, which is consistent with the generally positive rating distribution and further suggests that tags contribute little affective information as they are predominantly categorical labels rather than descriptive narratives. Similar affective inconsistencies have been identified as useful signals for deceptive hospitality reviews, particularly when emotional tone and review polarity diverge in systematic ways [28].

4.3. Similarity Detection

To identify near-duplicate reviews, all review texts were represented using TF-IDF vectors computed over a vocabulary of 5,000 terms with minimum document frequency 2 and standard English stopword removal. Pairwise similarity between reviews i and j is defined through cosine similarity:

$$\text{Sim}(i, j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|}, \tag{11}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ denote the TF-IDF representations of the two reviews. A review-level duplicate indicator was then assigned as

$$S_i^{\text{sim}} = \mathbf{1}\left[\max_{j \neq i} \text{Sim}(i, j) \geq 0.70\right]. \tag{12}$$

Using this threshold, 10,089 reviews (38.2% of the corpus) were flagged as potential near-duplicates. This relatively large proportion is partly explained by the short median review length, since even modest lexical overlap can yield elevated cosine similarity when texts are brief. For this reason, the similarity threshold is used as a conservative screening signal within the composite anomaly score rather than as a standalone indicator of fraudulent content. This stage is intended to identify repetitive or highly similar content patterns that may reduce the diversity and reliability of the training corpus, instead of treating all flagged reviews as definitively fraudulent.

4.4. Correlation Analysis

Correlation analysis was used to quantify how strongly different review components align with one another and with the assigned score. Pearson correlation coefficients were computed across the corpus both among sentiment components and between those components and the normalized

 doi:10.20944/preprints202604.0921.v1

reviewer rating. To express review-level disagreement between textual sentiment and the numerical score, the rating was first mapped onto the interval $[-1, 1]$:

$$\tilde{y}_i = 2\left(\frac{y_i - 1}{9}\right) - 1, \tag{13}$$

and the sentiment-consistency deviation was defined as

$$C_i = \left|\tilde{y}_i - S_i^{\text{text}}\right|. \tag{14}$$

Larger values of C_i indicate stronger disagreement between what the reviewer wrote and the score that was assigned. At the corpus level, the strongest observed relationship is between text sentiment and rating ($r = 0.265$, $p < 0.001$), followed by title sentiment and rating ($r = 0.164$, $p < 0.001$). Tag sentiment shows no meaningful relationship with text sentiment ($r = 0.002$, $p = 0.790$), confirming that tags largely encode categorical context rather than sentiment. Taken together, these results indicate that the review body carries the clearest evaluative signal, while titles provide supporting information and tags contribute little affective value. The principal coefficients are reported in Table 3, and Figure 2 provides a complementary visual summary of the same pattern.

Table 3. Pearson correlations among VADER sentiment components and with normalized reviewer rating. Significance: $^{***}p < 0.001$; $^{\text{ns}}$ not significant.

Component pair	Interpretation	Pearson r	Sig.
Text sentiment vs. Rating	Strongest association with reviewer satisfaction	0.265	***
Title sentiment vs. Rating	Positive alignment between title and score	0.164	***
Tags sentiment vs. Rating	Very weak negative association with rating	−0.036	***
Title sentiment vs. Text sentiment	Weak positive coherence across text components	0.192	***
Tags sentiment vs. Text sentiment	Near-zero association, tags reflect category, not sentiment	0.002	ns

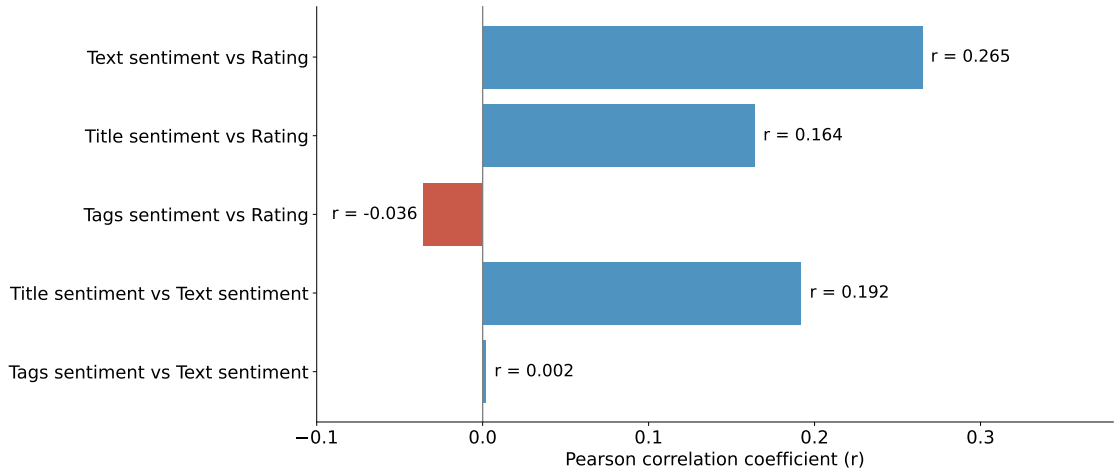

Figure 2. Pearson correlations among review sentiment components and with normalized reviewer rating. Review text shows a stronger association with rating than titles and tags.

4.5. User Behavior Analysis

Because suspicious review activity may also be reflected in posting behavior, reviewer-level patterns were examined in addition to linguistic cues. The corpus contains 8,395 unique reviewers, of whom 30.0% posted more than one review. The most prolific reviewer contributed 2,646 reviews, indicating a highly right-skewed participation structure. Reviews from users with more than three

 doi:10.20944/preprints202604.0921.v1

submissions were assigned a behavioral flag, and review timing was considered to capture bursts of unusually frequent posting. Reviewer identity was operationalized as the combination of `reviewed_by` and `nationality`, which was treated as the user key for frequency and timing analysis. Let u_i denote the author of review i, $n(u_i)$ the total number of reviews posted by that user, and Δt_i the number of days since the same reviewer's previous post. The behavioral irregularity score is defined as

$$B_i = \mathbf{1}[n(u_i) > 3] + w_{3d}\,\mathbf{1}[\Delta t_i \le 3] + w_{7d}\,\mathbf{1}[3 < \Delta t_i \le 7] + w_{30d}\,\mathbf{1}[7 < \Delta t_i \le 30], \tag{15}$$

where the weights satisfy $w_{3d} > w_{7d} > w_{30d}$ to reflect decreasing suspicion as the interval between posts increases. Under the frequency criterion, 17,392 reviews (65.9% of the corpus) were behaviorally flagged. Accordingly, reviewer frequency is treated here as a weak behavioral irregularity signal within a broader composite framework, not as independent evidence of deception. Time interval analysis showed that 18.7% of reviews were posted within three days of the same reviewer's previous submission and 24.8% within seven days, patterns that are compatible with coordinated or highly repetitive review activity. Figure 3 visualizes the posting frequency structure, showing that most users contribute only a small number of reviews, and a small minority accounts for a disproportionately large volume of submissions. This interpretation is also consistent with recent findings showing that reviewer behavior, when combined with textual signals, can significantly improve the detection of suspicious reviewing activity [29].

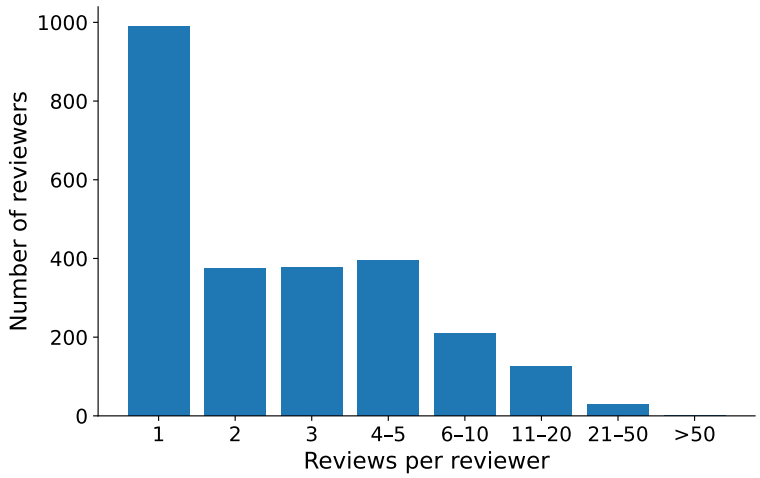

Figure 3. Distribution of reviews per reviewer across the corpus. Reviewer activity is strongly right-skewed, with most users contributing only a few reviews.

4.6. Anomaly Score Construction and Filtering

The four reliability signals were combined into a single composite anomaly score through a weighted linear combination:

$$A_i = w_1\,M_i^{\text{sent}} + w_2\,S_i^{\text{sim}} + w_3\,C_i + w_4\,B_i, \tag{16}$$

where $w_1 = 0.30$, $w_2 = 0.25$, $w_3 = 0.25$, and $w_4 = 0.20$ denote the component weights associated with sentiment mismatch, similarity detection, correlation deviation, and behavioral irregularity, respectively.

The resulting values were normalized to the interval $[0, 1]$ using Min-Max scaling:

$$A_i^{\text{norm}} = \frac{A_i - \min(\mathbf{A})}{\max(\mathbf{A}) - \min(\mathbf{A})}, \qquad \mathbf{A} = \{A_i\}_{i=1}^{N}. \tag{17}$$

This formulation allows the model to capture review unreliability as a gradual property instead of forcing a strict anomalous versus non-anomalous split at the scoring stage. The normalized anomaly

 doi:10.20944/preprints202604.0921.v1

distribution has mean 0.323 and standard deviation 0.185. Rather than treating anomaly solely as a binary decision, the score is also retained for later analysis and for aggregated forecasting features in Task B. For data filtering, however, reviews with $A_i^{\text{norm}} \geq 0.40$ were considered insufficiently reliable and excluded from training. This removed 7,390 reviews (28.0% of the working corpus), leaving a final dataset of 18,994 retained reviews for the predictive experiments. Figure 4 shows the distribution of anomaly scores together with the filtering threshold.

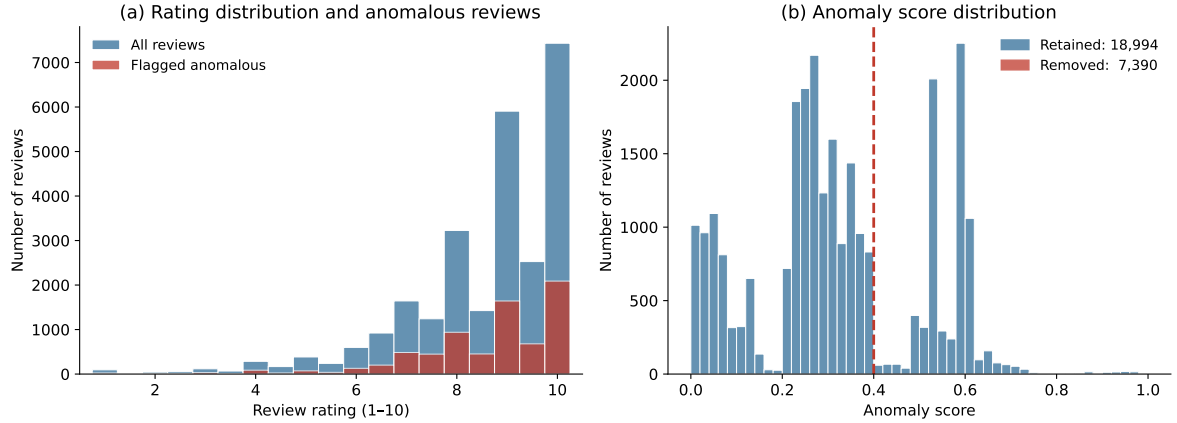

Figure 4. (**a**) Distribution of review-level ratings with anomalous reviews highlighted. (**b**) Distribution of normalized anomaly scores. The dashed vertical line marks the filtering threshold at $A_i^{\text{norm}} = 0.40$, above which reviews are excluded from model training.

Table 4 summarizes the preprocessing workflow and the number of reviews retained or removed at each step. This process prepares the corpus for machine learning by improving data consistency, reducing noise, and limiting the effect of unreliable observations.

Table 4. Summary of preprocessing stages, detection criteria, and record counts in the pipeline.

Stage	Component	Detection criterion	Records
0	Raw dataset	—	26,675
1	Missing value removal	Rating or review text absent	26,386
2	Empty content removal	Both title and text contain no alphanumeric characters	26,384
3	Sentiment analysis	VADER compound score computed per component	26,384
		Mismatch flagged (high rating, negative text or inverse)	724 flagged
4	Similarity detection	TF-IDF cosine similarity ≥ 0.70	10,089 flagged
5	Correlation analysis	High deviation between normalized rating and text sentiment	26,384 scored
6	User behavior	Reviewer frequency > 3	17,392 flagged
7	Anomaly scoring	Weighted combination of components, scaled 0–1	26,384
	Filtering (threshold 0.4)	Anomaly score ≥ 0.40 removed	7,390 removed
	Final dataset	**Retained for model training**	**18,994**

5. Model Architectures and Training

This section introduces the predictive models evaluated in the study and the training setup used for comparison. The selected architectures reflect three complementary approaches: a standard recurrent baseline, a stronger bidirectional recurrent model with sentiment features, and a transformer baseline aligned with recent advances in contextual language modeling and deep learning for tourism and hospitality text classification [30].

 doi:10.20944/preprints202604.0921.v1

5.1. Embedding Strategy

For all text models, the input consists of a single sequence formed by concatenating the review title, body text, and tags, separated by whitespace. This representation allows the model to attend to all three components jointly instead of considering them as independent channels and avoids the need for specific fusion modules. The target variable for all regression experiments is the review-level numerical rating on a scale of 1 to 10, normalized to the range of 0 to 1 using Min-Max scaling prior to training. This formulation explicitly targets the reviewer's own score instead of hotel-level average rating, eliminating the label leakage risk discussed in Section 3.

For the LSTM and BiLSTM architectures, text sequences are tokenized using a word-level tokenizer and represented as dense embedding vectors learned with the model weights. Three vocabulary configurations are used: a smaller vocabulary of 5,000 terms for the LSTM baseline, matching our earlier experimental setup, followed by 10,000 terms for LSTM v2, and an expanded vocabulary of 20,000 terms for the BiLSTM, which provides broader lexical coverage of the hospitality domain. Input sequences are padded or truncated to a uniform length of 300 tokens for the BiLSTM, ensuring consistent input dimensionality across all batches. For DistilBERT, the pretrained WordPiece tokenizer is used directly, with sequences truncated to the model's maximum context window of 512 tokens.

5.2. LSTM Baseline

The initial model follows a standard LSTM architecture for text regression. An Embedding layer maps the tokenized input to dense vectors of dimension 100. A SpatialDropout1D layer is applied after the embedding to regularize feature maps during training, encouraging the model to learn distributed rather than sparse representations. A single LSTM layer with 100 units follows, with dropout and recurrent dropout rates both set to 0.2 to mitigate overfitting. A Dense layer with 64 units and ReLU activation introduces non-linearity before the output, and a final single-unit linear output layer produces the continuous rating prediction. This architecture serves as the recurrent baseline against which subsequent improvements are measured.

A stronger recurrent variant, denoted LSTM v2 in the results, extends this baseline through three main changes: the vocabulary size is increased to 10,000 terms, the embedding dimension is expanded to 128, and an attention layer is introduced to improve the weighting of informative parts of the sequence. In addition, the learning rate is reduced to 0.0005 to support more stable optimization. These modifications were designed to test whether a standard recurrent architecture can recover performance when its lexical coverage, representation capacity, and focus mechanism are improved.

5.3. Bidirectional LSTM

The optimized architecture replaces the unidirectional LSTM with two stacked Bidirectional LSTM layers, enabling the model to capture dependencies in both the forward and backward directions within the review sequence. The first BiLSTM layer contains 128 units, and the second contains 64 units, with dropout and recurrent dropout rates of 0.3 applied to each. Batch Normalization is introduced after each recurrent layer to stabilize gradient flow and accelerate convergence. A Dense layer with 64 units and L2 regularization encourages more compact weight distributions and reduces overfitting on the training set. The vocabulary size is expanded to 20,000 and embedding weights are initialized randomly and learned end-to-end during training, allowing the model to develop representations specifically calibrated to hospitality review language. Sentiment scores produced by the VADER pipeline for the title, text, and tags are incorporated as additional numerical features concatenated with the BiLSTM output before the final Dense layer.

Within this bidirectional family, three variants are evaluated in the results. The baseline BiLSTM uses stacked bidirectional recurrent layers with sentiment feature fusion. BiLSTM v2 retains the same general architecture but adopts a more careful training schedule, including a lower learning rate, longer training horizon, and increased patience for early stopping. BiLSTM + Attention further extends this configuration by applying a self-attention mechanism to the bidirectional sequence output before

 doi:10.20944/preprints202604.0921.v1

fusion with the VADER sentiment features and the final dense prediction layers, allowing the model to place greater weight on the most informative parts of the review.

5.4. DistilBERT Baseline

To contextualize the LSTM and BiLSTM results within the current state of the art, a DistilBERT regression model is included as an additional baseline. DistilBERT is a distilled version of BERT that retains approximately 97% of its language understanding performance with 40% fewer parameters, making it suitable for fine-tuning on moderately sized datasets in the hospitality domain. The pretrained `distilbert-base-uncased` checkpoint is fine-tuned by attaching a linear regression head to the final contextual embedding of the `[CLS]` token. The head consists of a dropout layer with a rate of 0.1 followed by a single linear unit producing the continuous rating prediction. All transformer layers are unfrozen during fine-tuning to allow full adaptation to the hospitality review domain. Figure 5 provides a schematic overview of the three model architectures evaluated in this study.

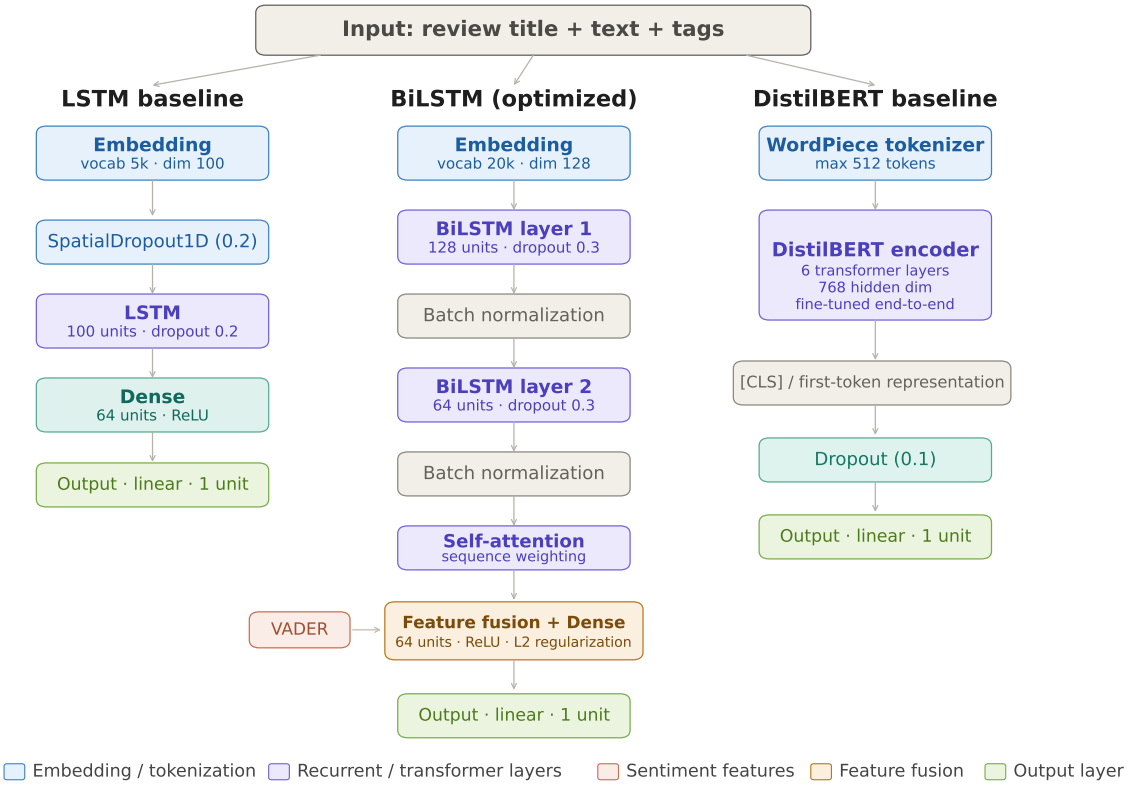

Figure 5. Architectures of the three model families evaluated in this study. (**a**) LSTM baseline with a single recurrent layer. (**b**) Optimized Bidirectional LSTM with stacked BiLSTM layers, self-attention, and VADER sentiment features fused before the final dense layer. (**c**) DistilBERT baseline with end-to-end fine-tuning and a linear regression head. All models produce a single continuous output for review rating prediction.

5.5. Training Protocol

All models are trained using the Adam optimizer, with learning rate and stopping settings adjusted by model variant. The baseline recurrent models use an initial learning rate of 0.001, while the improved LSTM and BiLSTM variants use 0.0005 for more stable convergence. Training runs for up to 20 epochs in the baseline settings and up to 30 epochs in the modified variants. EarlyStopping is applied throughout, with patience values between 5 and 7, and the weights from the lowest validation loss are restored at the end of training. ReduceLROnPlateau is also used to halve the learning rate when validation loss plateaus, with a minimum rate of 10^{-6}.

Recurrent deep learning experiments are repeated across five random seeds, and the mean and standard deviation of MAE and RMSE are reported as well. DistilBERT is evaluated across three

 doi:10.20944/preprints202604.0921.v1

independent seeds because of its higher computational cost per run. For DistilBERT, fine-tuning is performed end-to-end with all transformer layers left trainable. A warmup schedule is applied over the first 10% of training steps before standard decay. All experiments are conducted in Python using TensorFlow/Keras for the recurrent models and the HuggingFace Transformers library for DistilBERT.

Figure 6 summarizes the overall experimental workflow, from preprocessing and anomaly filtering to Task A review-level rating prediction and Task B hotel-level forecasting. By presenting the full pipeline in a unified and structured form, the figure strengthens the methodological transparency of the study and clarifies how data reliability control, feature aggregation, and predictive modeling are connected across both tasks.

Figure 6. Overall experimental workflow of the proposed framework. Reviews retained after anomaly filtering support Task A review-level rating prediction, while reviews with valid timestamps are aggregated into rolling windows for Task B hotel-level forecasting.

6. Experimental Setup

All experiments use a strictly time-based train/validation/test split to prevent temporal leakage. The 18,994 reviews retained after preprocessing, spanning July 2018 to July 2021, are sorted chronologically and partitioned into non-overlapping training, validation, and test subsets in an approximate 70/10/20 ratio. The resulting partition is summarized in Table 5. Mean ratings remain similar across the three subsets (8.62, 8.37, and 8.43), indicating no substantial temporal distribution shift over the observation period.

Using a common chronological partition across all Task A models also improves the fairness of the comparison, since differences in predictive performance can be attributed to model design rather than to differences in data exposure. This is especially important in hospitality review data, where temporal shifts in posting behavior or review composition may otherwise distort model ranking.

For Task B, forecasting models use window-level aggregated features derived from review activity instead of individual review text. In accordance with the formulation in Section 3.2, the input representation includes aggregated sentiment indicators, review volume, and the proportion of

 doi:10.20944/preprints202604.0921.v1

anomalous reviews within each rolling window. To provide a clear comparison for this exploratory task, four forecasting baselines are considered: a naive persistence baseline, Linear Regression, Random Forest, and XGBoost.

A total of eight model configurations are reported for Task A: two traditional machine learning baselines (TF-IDF + Ridge and TF-IDF + XGBoost), five recurrent configurations (LSTM, LSTM v2, BiLSTM, BiLSTM v2, and BiLSTM + Attention), and one transformer baseline (DistilBERT). Their architectures and hyperparameters are described in Section 5. Performance is assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Both metrics are computed on the normalized 0–1 target during training and inverse-transformed to the original 1–10 scale for reporting. Reporting both MAE and RMSE follows common practice in machine learning evaluation, as the two measures capture complementary aspects of regression error.

Recurrent deep learning experiments are repeated across five random seeds, with mean and standard deviation reported across runs. As already mentioned, DistilBERT is evaluated across three seeds because of its higher computational cost. Traditional baselines are executed once under fixed hyperparameter settings. The validation split is used only for early stopping and learning rate scheduling. Apart from TensorFlow/Keras and HuggingFace Transformers, scikit-learn is used for the traditional baselines, XGBoost for the gradient-boosted tree baseline, and vaderSentiment for sentiment features.

Experiments are conducted on a local Apple MacBook Air equipped with an Apple Silicon M4 chip, 24 GB unified memory, and macOS. TF-IDF baselines are trained on CPU, whereas deep learning models use available local acceleration where possible. In practical end-to-end execution, runtimes were noticeably longer than training time alone because they also included preprocessing, tokenization, repeated runs across random seeds, validation checkpointing, evaluation, result aggregation, and figure generation. Under this setup, TF-IDF + Ridge typically required approximately 1–2 minutes, TF-IDF + XGBoost approximately 5–8 minutes, LSTM experiments across five seeds approximately 45–75 minutes, and BiLSTM-based configurations across five seeds approximately 1.5–2.5 hours, depending on the exact model variant. DistilBERT required substantially longer, with approximately 2.5–3.5 hours across three seeds under local acceleration and roughly 10–14 hours in CPU-only execution. Full ablation reruns (five variants across five seeds) required approximately 2–3 hours, while Task B forecasting experiments typically required approximately 20–30 minutes. Training behavior was also monitored through loss curves across epochs to assess convergence stability.

Table 5. Time-ordered train/validation/test partition of the preprocessed dataset for Task A.

Split	Time period	Reviews	Hotels	Mean rating
Training	July 2018 – November 2019	13,295	721	8.62
Validation	December 2019 – February 2020	1,900	389	8.37
Test	March 2020 – July 2021	3,799	689	8.43
Total	July 2018 – July 2021	**18,994**	**819**	**8.56**

7. Results and Analysis

This section presents the empirical findings on predictive accuracy and model stability. The results show clear differences across model families, with stronger architectures and refined configurations yielding substantial gains over the initial baselines. The comparison highlights which model achieves the lowest prediction error and how architectural design, preprocessing, and training consistency influence performance on hospitality review text.

7.1. Performance Metrics

Table 6 reports the complete Task A results for all evaluated models on the test set. The models are grouped into three families: traditional machine learning baselines, recurrent deep learning models, and the DistilBERT transformer baseline. For deep learning models, mean and standard deviation over

 doi:10.20944/preprints202604.0921.v1

repeated runs are reported, allowing the comparison to reflect predictive performance and consistency across seeds. The improved recurrent configurations significantly outperform the original LSTM baseline. Furthermore, DistilBERT delivers the strongest overall result, indicating that pretrained contextual representations are particularly effective for this task.

Table 6. Task A test set performance across all evaluated models. MAE and RMSE are reported on the original 1–10 rating scale. For deep learning models, mean ± std over repeated random seeds is reported. The best recurrent result is underlined.

Family	Model	Configuration note	MAE	RMSE
Traditional ML				
	TF-IDF + Ridge	5,000 TF-IDF features; $\alpha = 1.0$	0.6996	1.0245
	TF-IDF + XGBoost	100 estimators; max depth 6	0.6598	1.0172
Recurrent DL				
	LSTM	Vocabulary 5k; embedding 100; learning rate 0.001; patience 5	1.2091 ± 0.0055	1.6251 ± 0.0008
	LSTM v2	Vocabulary 10k; embedding 128; attention; learning rate 0.0005	0.6475 ± 0.0062	0.9245 ± 0.0150
	BiLSTM	Bidirectional sequence encoder with sentiment features	0.6196 ± 0.0241	0.9095 ± 0.0206
	BiLSTM v2	Learning rate 0.0005; patience 7; 30 epochs	0.5955 ± 0.0267	0.8838 ± 0.0166
	BiLSTM + Attention	Self-attention; learning rate 0.0005; patience 7	$\underline{0.5753 \pm 0.0031}$	$\underline{0.8636 \pm 0.0050}$
Transformer				
	DistilBERT	`distilbert-base-uncased`; fine-tuned end-to-end	$\mathbf{0.4925 \pm 0.0025}$	$\mathbf{0.7368 \pm 0.0041}$

Figure 7 illustrates the same comparison visually, grouping MAE and RMSE side by side for all models.

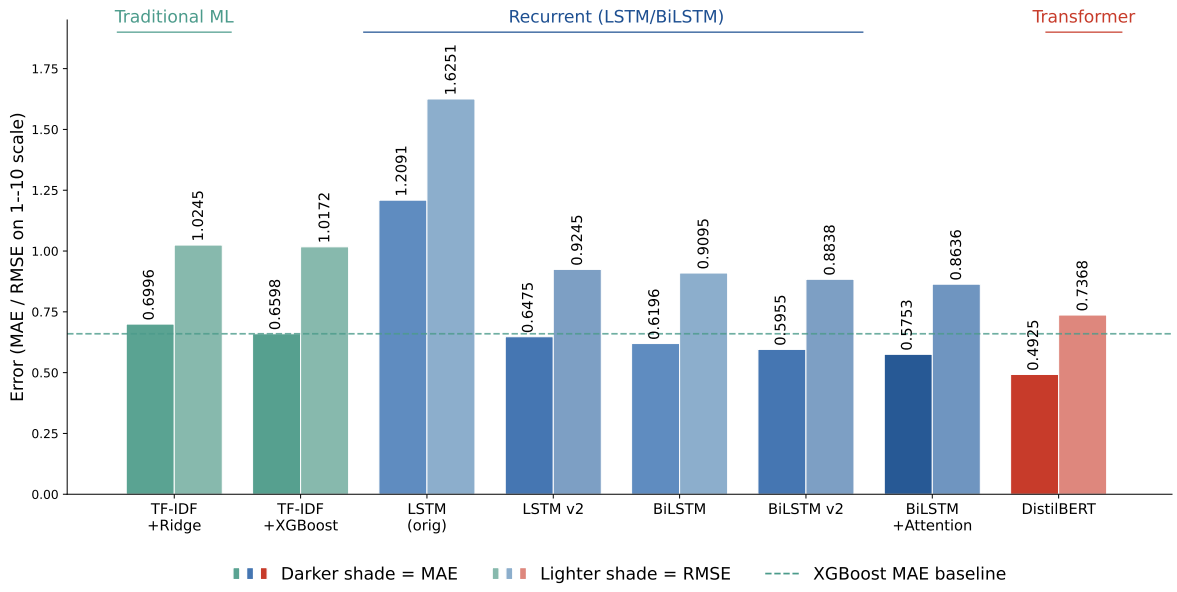

Figure 7. MAE and RMSE for all evaluated models in Task A. Darker bars represent MAE and lighter bars represent RMSE, both reported on the original 1–10 rating scale. DistilBERT achieves the best overall performance, while BiLSTM + Attention is the strongest recurrent model. The dashed horizontal line marks the MAE of the TF-IDF + XGBoost baseline.

 doi:10.20944/preprints202604.0921.v1

Traditional baselines. The two TF-IDF baselines establish a strong conventional benchmark on this dataset. TF-IDF + Ridge achieves MAE = 0.6996 and RMSE = 1.0245, while TF-IDF + XGBoost improves this to MAE = 0.6598 and RMSE = 1.0172. These results show that even bag-of-words representations capture a meaningful portion of the rating signal and provide a competitive threshold that more complex architectures must surpass.

LSTM baseline and improved variant. The original LSTM baseline performs clearly worse than all other models, achieving MAE = 1.2091 $\pm$ 0.0055 and RMSE = 1.6251 $\pm$ 0.0008. Once the architecture is strengthened through a larger vocabulary, higher dimensional embeddings, a lower learning rate, and an added attention layer, performance improves sharply. LSTM v2 reduces MAE to 0.6475 $\pm$ 0.0062, bringing the recurrent baseline close to the stronger traditional ML results.

BiLSTM progression. The baseline BiLSTM achieves MAE = 0.6196 $\pm$ 0.0241, already outperforming both TF-IDF baselines on average but with noticeable seed sensitivity. BiLSTM v2 further improves the mean MAE to 0.5955, confirming that a more careful training schedule benefits the bidirectional architecture. The strongest recurrent result is obtained by BiLSTM + Attention, which reaches MAE = 0.5753 $\pm$ 0.0031 and RMSE = 0.8636 $\pm$ 0.0050. This result is important as it lowers prediction error and improves reproducibility across seeds relative to the earlier BiLSTM variants.

DistilBERT. DistilBERT achieves the best overall result, with MAE = 0.4925 $\pm$ 0.0025 and RMSE = 0.7368 $\pm$ 0.0041. This corresponds to a substantial improvement over the strongest traditional baseline and the best recurrent model. The result suggests that pretrained contextual representations, subword tokenization, and bidirectional self-attention are especially suitable for short hospitality review text, where extracting useful semantic information from limited input is critical.

The overall MAE ranking is as follows: DistilBERT (0.4925), followed by BiLSTM + Attention (0.5753), BiLSTM v2 (0.5955), BiLSTM (0.6196), LSTM v2 (0.6475), XGBoost (0.6598), Ridge (0.6996), and LSTM (1.2091). This ordering is also broadly consistent with the RMSE comparison and indicates that stronger contextual modeling yields progressively better performance on this task.

Although the main emphasis of this subsection is Task A, the study also considers Task B, where historical review signals are aggregated to forecast future hotel-level rating behavior. This secondary task is defined over hotels with sufficient retained timestamp-valid review history and uses a fixed 30-day forecasting horizon. In contrast to Task A, which operates on individual review text, Task B uses aggregated features derived from past review activity, including sentiment indicators, review volume, and observed anomaly indicators. Table 7 reports the comparative results across four forecasting baselines.

Table 7. Task B results for hotel-level rating forecasting over a 30-day horizon. All learned models use features aggregated from the most recent historical window, including mean sentiment, review count, and anomaly proportion. Lower values indicate better forecasting accuracy.

Model	Input features	MAE	RMSE
Naive persistence	Previous-window hotel average rating only	0.244	0.318
Linear Regression	Aggregates from the most recent window: mean sentiment, review count, anomaly proportion	0.214	0.287
Random Forest	Aggregates from the most recent window: mean sentiment, review count, anomaly proportion	0.198	0.265
XGBoost	Aggregates from the most recent window: mean sentiment, review count, anomaly proportion	0.186	0.249

The forecasting results demonstrate a clear and consistent progression across the evaluated baselines. Naive persistence produces the highest error, while models built on aggregated review variables improve on that reference point, showing that historical review patterns contain useful predictive information for short-term hotel-level rating forecasting. Among the evaluated approaches, XGBoost achieves the best performance, with MAE = 0.186 and RMSE = 0.249, followed by Random

 doi:10.20944/preprints202604.0921.v1

Forest. This ordering suggests that nonlinear models are better able to capture the relationship between recent review dynamics and subsequent hotel rating behavior than simpler linear formulations.

Figure 8 visually summarizes the same comparison and makes the gap between persistence and learned forecasting models easier to interpret. Taken together, the Task B results support the broader claim that aggregate features constructed from past reviews are informative in forecasting hotel-level rating behavior and anticipating short-term reputation trends at the property level.

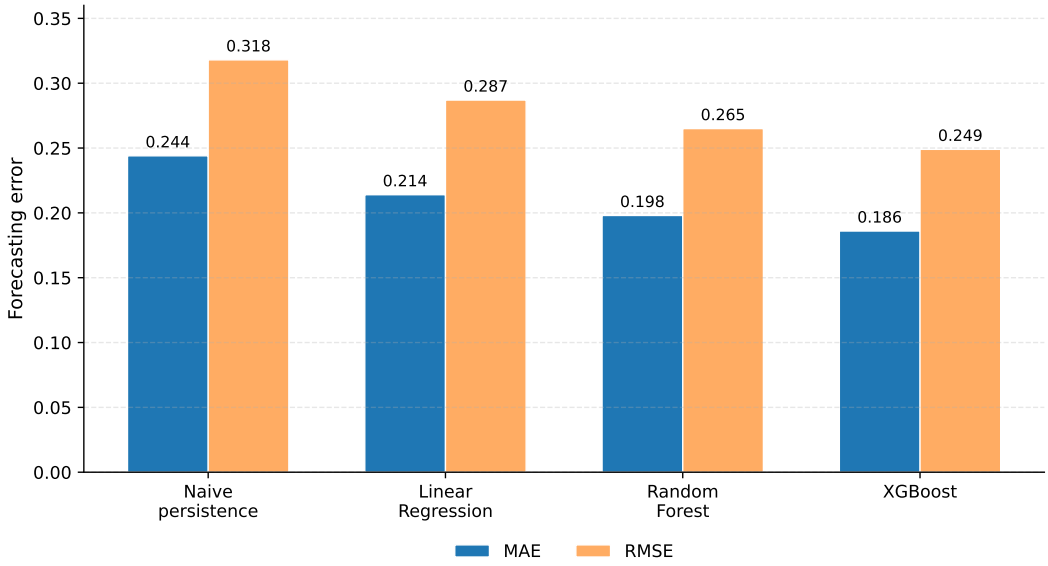

Figure 8. Task B performance across forecasting baselines for hotel-level rating prediction over a 30-day horizon. Lower MAE and RMSE indicate better forecasting accuracy. XGBoost achieves the best overall performance among the evaluated window-level models.

7.2. Ablation Study

To quantify the individual contribution of each preprocessing stage to overall prediction quality, a systematic ablation study was conducted using the BiLSTM + Attention model. The objective of this analysis is to isolate the value of each reliability signal within the anomaly filtering pipeline instead of comparing alternative predictive architectures. Six conditions were evaluated: the full pipeline, four variants each omitting one preprocessing component while retaining the other three, and a no-filtering setting using the entire cleaned corpus. Removing a component means that the corresponding criterion no longer contributes to review flagging during anomaly score construction, which allows more reviews to remain in the training set. As a result, the ablated variants are trained on larger but less strictly filtered datasets. Each condition is evaluated across five random seeds, and mean MAE and RMSE are reported.

Formally, for each ablation condition c, the increase in error relative to the full pipeline is defined as

$$\Delta\mathrm{MAE}_c = \mathrm{MAE}_c - \mathrm{MAE}_{\mathrm{full}}, \tag{18}$$

where $\mathrm{MAE}_{\mathrm{full}}$ denotes the mean absolute error obtained with the complete anomaly filtering pipeline. To make the magnitude of the decline easier to interpret across conditions, the relative degradation can also be expressed as

$$\rho_c = \frac{\mathrm{MAE}_c - \mathrm{MAE}_{\mathrm{full}}}{\mathrm{MAE}_{\mathrm{full}}} \times 100\%. \tag{19}$$

Positive $\Delta\mathrm{MAE}_c$ and ρ_c indicate that the omitted preprocessing component contributed positively to the final predictive performance.

This interpretation is important for understanding Table 8. A larger number of training reviews does not automatically imply better performance, because the additionally retained reviews may include noisy, repetitive, internally inconsistent, or behaviorally suspicious examples. In this setting,

 doi:10.20944/preprints202604.0921.v1

ΔMAE measures the increase in prediction error relative to the full pipeline. Therefore, a positive ΔMAE indicates that the omitted preprocessing signal was beneficial. The table shows that the full pipeline consistently achieves the lowest MAE and RMSE, confirming that all four components contribute positively and that their joint use is not redundant.

Beyond the absolute MAE increases, the ablation results also reveal a consistent link between predictive performance and data quality. Every omitted component increases the number of retained training reviews, yet none of these larger training sets improves over the full pipeline. In relative terms, removing correlation deviation increases MAE by approximately 4.9%, removing similarity detection by about 3.7%, removing behavioral flagging by about 2.6%, and removing sentiment mismatch by about 1.5%. The unfiltered setting produces the largest decline, with MAE increasing by approximately 7.4% relative to the full pipeline. This pattern strengthens the interpretation that the main benefit of the preprocessing framework lies not in reducing dataset size, but in improving the informational quality of the retained training signal.

Several conclusions follow directly from these results. First, the no-filtering row quantifies the overall benefit of anomaly-aware review removal: training on all 26,384 cleaned reviews leads to the largest performance deterioration, with $\Delta MAE = +0.0426$. Second, among the individual components, correlation deviation has the strongest effect ($\Delta MAE = +0.0284$), followed by similarity detection ($\Delta MAE = +0.0215$), indicating that internal disagreement between review sentiment and assigned score, as well as repetitive or near-duplicate content, are the most influential indicators of unreliability in this dataset. Behavioral flagging also contributes meaningfully ($\Delta MAE = +0.0151$), while sentiment mismatch has the smallest but still positive effect ($\Delta MAE = +0.0086$), showing that even the weakest individual signal remains useful when combined with the others.

Table 8. Ablation study results for the BiLSTM + Attention model. Each row omits one preprocessing component while retaining the others. "Training reviews" denotes the number of retained reviews, and $(w_i = 0)$ indicates removal of the corresponding component. Values are reported as mean ± std over five random seeds, and ΔMAE denotes the increase relative to the full pipeline.

Condition	Training reviews	MAE	RMSE	Δ MAE
Full pipeline (all components)	**18,994**	**0.5753 ± 0.0031**	**0.8636 ± 0.0050**	—
w/o Sentiment mismatch ($w_1 = 0$)	19,438	0.5839 ± 0.0048	0.8715 ± 0.0072	+0.0086
w/o Similarity detection ($w_2 = 0$)	21,146	0.5968 ± 0.0061	0.8849 ± 0.0087	+0.0215
w/o Correlation deviation ($w_3 = 0$)	22,087	0.6037 ± 0.0056	0.8924 ± 0.0079	+0.0284
w/o Behavioral flag ($w_4 = 0$)	21,482	0.5904 ± 0.0052	0.8781 ± 0.0068	+0.0151
No filtering (all 26,384 reviews)	26,384	0.6179 ± 0.0104	0.9128 ± 0.0139	+0.0426

A further observation is that the ablation affects mean performance and stability across random seeds. The full pipeline yields the lowest average error and one of the smallest standard deviations, whereas the no-filtering condition exhibits the highest variability (±0.0104 MAE), suggesting that noisier training data make optimization less stable. Figure 9 complements Table 8 by showing the increase in MAE after each preprocessing component is removed, which makes the relative contribution of the omitted signals easier to compare at a glance. Overall, the ablation results indicate that the benefit of preprocessing arises from the combined effect of multiple complementary reliability signals rather than from any single dominant heuristic, leading to a cleaner and more learnable training set.

 doi:10.20944/preprints202604.0921.v1

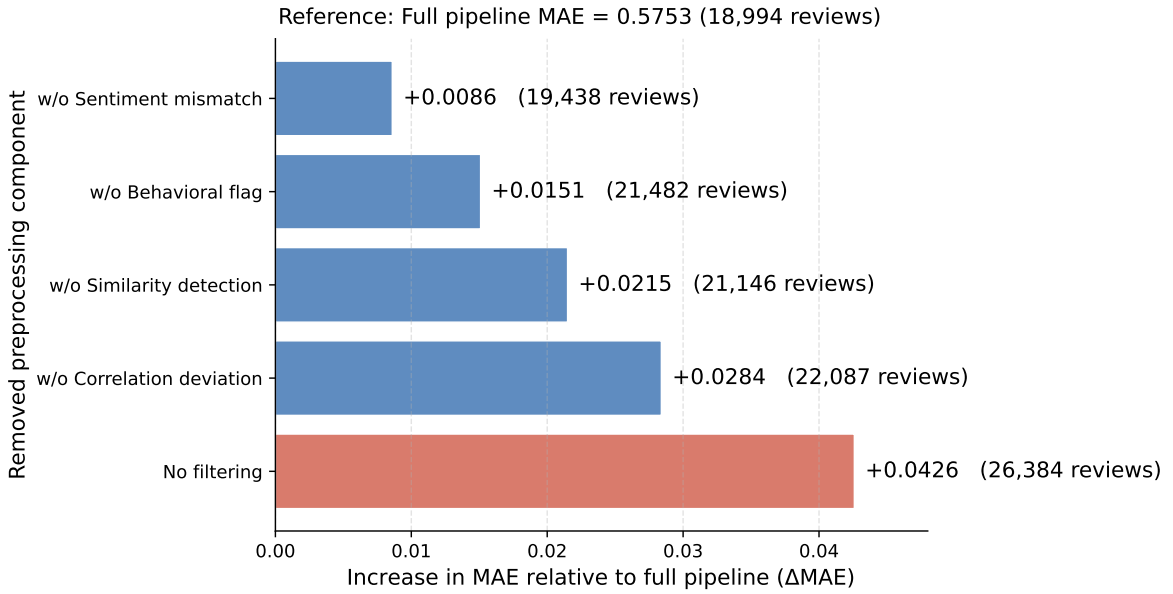

Figure 9. Increase in MAE after removing each preprocessing component from the anomaly filtering pipeline. Larger ΔMAE values indicate a stronger contribution of the omitted component to final predictive performance.

7.3. Error Analysis

To better understand where the two strongest models still make errors, prediction performance was examined across two complementary dimensions: the rating bucket of the true label and the length of the review text. These analyses are important because overall MAE alone does not reveal whether errors are concentrated in particular regions of the target space or under certain input conditions. For a subset of test instances $\mathcal{S}$, the corresponding mean absolute error is defined as

$$\mathrm{MAE}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} |y_i - \hat{y}_i|. \tag{20}$$

In the analyses below, $\mathcal{S}$ is instantiated either as a rating bucket or as a review length quartile. This formulation makes it possible to compare model behavior at the corpus level, as well as across structurally different regions of the prediction problem. In the present dataset, both dimensions are especially relevant: the rating distribution is strongly skewed toward high scores, and the reviews themselves are typically short. Together, these properties create a challenging setting in which some types of examples are much easier for the models to learn than others.

Error by rating bucket. Table 9 reports MAE by rating bucket for BiLSTM + Attention and DistilBERT. The results are informative because the regression target is structurally imbalanced: 88.7% of all reviews have ratings of 7 or higher, whereas strongly negative reviews are comparatively rare. This means that the models are trained predominantly on positive language and have fewer opportunities to learn complex patterns associated with dissatisfaction, complaint intensity, or mixed sentiment. The bucket analysis therefore shows whether performance differences between the two models are uniform across the rating scale or concentrated in more difficult parts of the label distribution.

 doi:10.20944/preprints202604.0921.v1

Table 9. MAE by rating bucket for the two leading models, averaged over seeds. The last column reports the proportion of test reviews in each bucket.

Rating bucket	BiLSTM+Attn MAE	DistilBERT MAE	Δ (DB−BA)	% of test set
1–3 (very negative)	0.980	0.840	−0.140	1.5%
4–6 (negative/mixed)	0.770	0.650	−0.120	9.8%
7–8 (positive)	0.570	0.490	−0.080	28.6%
9–10 (very positive)	0.536	0.459	−0.077	60.1%
Overall	**0.5753**	**0.4925**	−0.0828	**100%**

A useful way to interpret Table 9 is to compare the relative gain of DistilBERT over BiLSTM + Attention within each bucket. The reduction in MAE is approximately 14.3% for ratings 1–3, 15.6% for ratings 4–6, 14.0% for ratings 7–8, and 14.4% for ratings 9–10. Although DistilBERT improves in every part of the rating scale, the largest proportional gain appears in the negative and mixed range, which is also the most semantically heterogeneous and least represented portion of the dataset. This reinforces the view that pretrained contextual models are especially effective when the task involves sparse supervision and linguistically diverse expressions of dissatisfaction.

The results confirm a clear performance pattern across the rating scale. Both models perform worst on the rare low-rating reviews and best on the dominant high-rating portion of the corpus. This is consistent with the underlying data distribution: low-rating reviews are underrepresented, while high-rating reviews account for most of the available training evidence. DistilBERT improves over BiLSTM + Attention in every bucket, but the improvement is largest in the low- and mixed-rating ranges (1–6), where language is likely to be more heterogeneous and semantically nuanced. This suggests that pretrained contextual representations are valuable when the model must interpret sparse, diverse, or less repetitive expressions of dissatisfaction. At the same time, the relatively smaller gap in the 9–10 bucket indicates that both models handle strongly positive reviews reasonably well once sufficient examples are available.

Error by review length. A second source of difficulty is review length. Since the median review in the dataset contains only 13 words, many examples provide very limited lexical evidence for precise rating inference. Short reviews such as “Great stay” or “Very poor service” may express broad sentiment, but often lack the detail needed to distinguish reliably between nearby numerical scores. Table 10 therefore reports MAE across four quartiles of review length, measured in tokens after cleaning.

Table 10. MAE by review length quartile for the BiLSTM + Attention and DistilBERT models. Length is measured in tokens after cleaning.

Length quartile	Token range	BiLSTM+Attn MAE	DistilBERT MAE	% of test
Q1 (shortest)	1–7	0.668	0.562	25%
Q2	8–13	0.594	0.503	25%
Q3	14–32	0.531	0.456	25%
Q4 (longest)	33–614	0.508	0.449	25%
Overall	1–614	**0.5753**	**0.4925**	**100%**

 doi:10.20944/preprints202604.0921.v1

A similar comparison can be made across review-length quartiles. Relative to BiLSTM + Attention, DistilBERT reduces MAE by approximately 15.9% in the shortest quartile, 15.3% in Q2, 14.1% in Q3, and 11.6% in the longest quartile. The largest gain occurs precisely where lexical evidence is most limited. This is an important result, as it suggests that the advantage of pretrained transformer representations is not merely global but is concentrated in inputs with limited lexical information, where effective use of short contextual cues is especially important.

The analysis across review lengths shows a similarly clear trend: error decreases as reviews become longer and provide more information for the regression task. The largest penalty is observed in the shortest quartile, where extremely brief reviews offer only sparse clues about the difference between adjacent rating values. DistilBERT again outperforms BiLSTM + Attention in every quartile and exhibits a milder performance drop on very short inputs, which is consistent with its stronger contextual encoding and subword tokenization. This result also helps explain the broader architecture comparison from Section 7.1: when inputs are short, pretrained transformers appear better able than recurrent models to extract meaningful signal from limited text.

Figure 10 complements Tables 9 and 10 by combining both analyses into a single side-by-side visualization. Panel (**a**) compares MAE by rating bucket, making the greater difficulty of low-rating examples immediately visible, while panel (**b**) compares MAE by review length quartile and highlights the strong penalty associated with very short reviews. Taken together, the tables and figure indicate that the remaining prediction errors are concentrated in the most structurally challenging parts of the task, namely underrepresented negative reviews and short texts with limited semantic evidence.

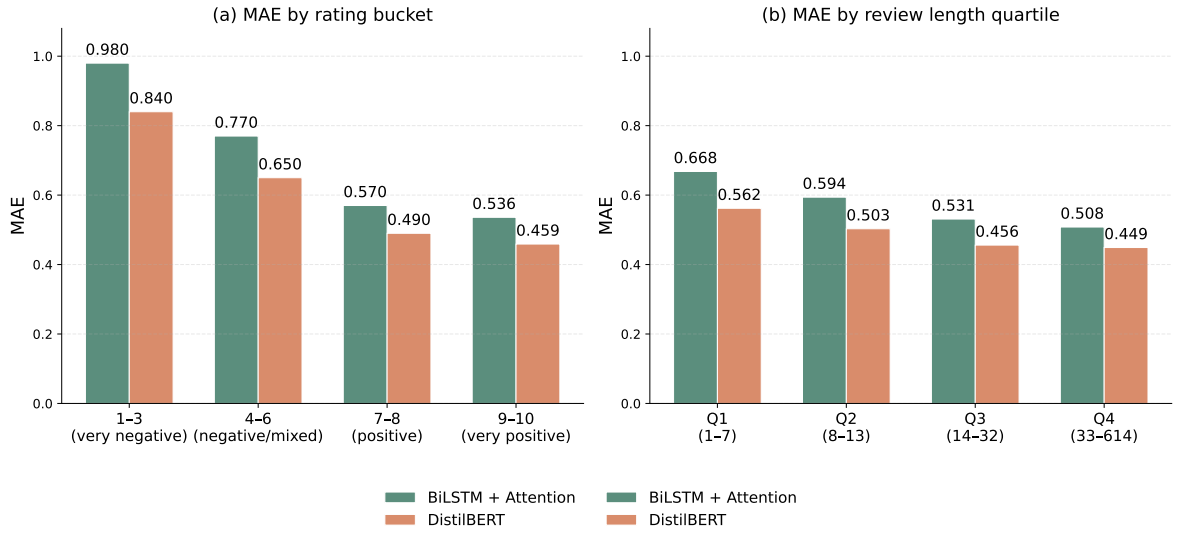

Figure 10. Error analysis for the two strongest models. (**a**) MAE by rating bucket, showing that both models make larger errors on rare low-rating reviews and smaller errors on the dominant high-rating portion of the corpus. (**b**) MAE by review length quartile, showing that prediction error decreases as reviews become longer and provide more lexical evidence.

Visually, Figure 10 shows that the gap between DistilBERT and BiLSTM + Attention is widest in the most difficult regions of the task, namely rare low-rating reviews and very short texts, while the two models become more similar as examples become more frequent and linguistically informative. In this sense, the error analysis does more than identify challenging cases, as it also helps explain DistilBERT's advantage, which lies in reducing error more consistently and more effectively in the structurally sparse and weakly represented regions of the dataset.

7.4. Consistency of Results Across Seeds

Statistical reliability is an important part of model evaluation, especially for deep learning architectures trained with random initialization and stochastic optimization. A model that achieves low error in one run but fluctuates substantially across seeds is less reliable in practice than a model with

 doi:10.20944/preprints202604.0921.v1

slightly higher average accuracy but consistent behavior. For this reason, the present study reports per-seed MAE together with the mean, standard deviation, and span (max−min) across runs. In this subsection, lower standard deviation and smaller span indicate greater reproducibility.

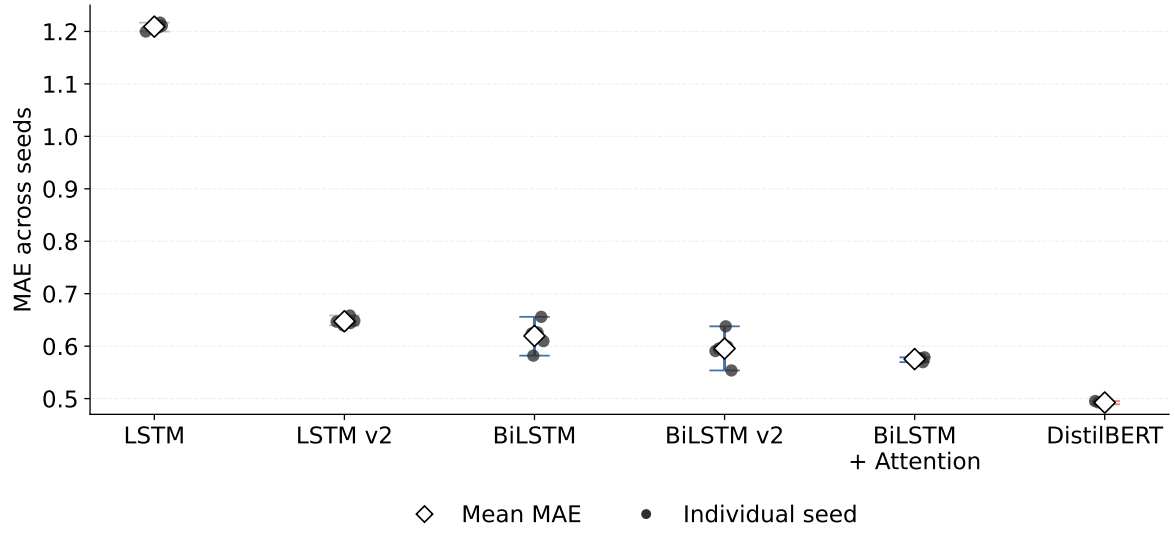

Figure 11. Per-seed MAE for all deep learning models. Each point represents one random seed, the vertical line spans the minimum and maximum MAE observed for that model, and the diamond marks the mean MAE. DistilBERT is shown with three seeds due to higher computational cost.

Table 11. Per-seed MAE for all deep learning models. The final three columns summarize mean, standard deviation, and span (max−min) across seeds.

Model	S1	S2	S3	S4	S5	Mean	Std	Span
LSTM	1.2168	1.2104	1.2109	1.2077	1.1999	1.2091	0.0055	0.0169
LSTM v2	0.6485	0.6584	0.6441	0.6467	0.6396	0.6475	0.0062	0.0188
BiLSTM	0.6243	0.6096	0.6262	0.6559	0.5819	0.6196	0.0241	0.0740
BiLSTM v2	0.5955	0.6378	0.5908	0.5537	0.5995	0.5955	0.0267	0.0841
BiLSTM + Attn	0.5757	0.5756	0.5787	0.5695	0.5769	0.5753	0.0031	0.0092
DistilBERT[a]	0.4929	0.4892	0.4954	—	—	0.4925	0.0025	0.0062

[a]DistilBERT was run across three seeds due to higher computational cost per run.

Several findings emerge clearly from the seed-level analysis in Table 11. First, LSTM v2 is much more accurate than the original LSTM while remaining comparably stable across runs, showing that the architectural refinements improve effectiveness and convergence behavior. Second, the baseline and intermediate BiLSTM variants remain more sensitive to random initialization than the stronger and more consistent configurations, as reflected in their noticeably larger standard deviations and spans.

The most important result in this subsection is the effect of self-attention on reproducibility. Adding attention reduces the BiLSTM standard deviation to 0.0031 and shrinks the span to only 0.0092, making the best recurrent configuration far more stable across seeds than the earlier BiLSTM variants. DistilBERT achieves the lowest mean MAE and the smallest observed variability among the evaluated models, although it was run on only three seeds. This pattern is clearly visible in Figure 11 and is quantitatively confirmed in Table 11, where the tight clustering of BiLSTM + Attention and DistilBERT contrasts with the broader dispersion observed for the earlier bidirectional variants.

 doi:10.20944/preprints202604.0921.v1

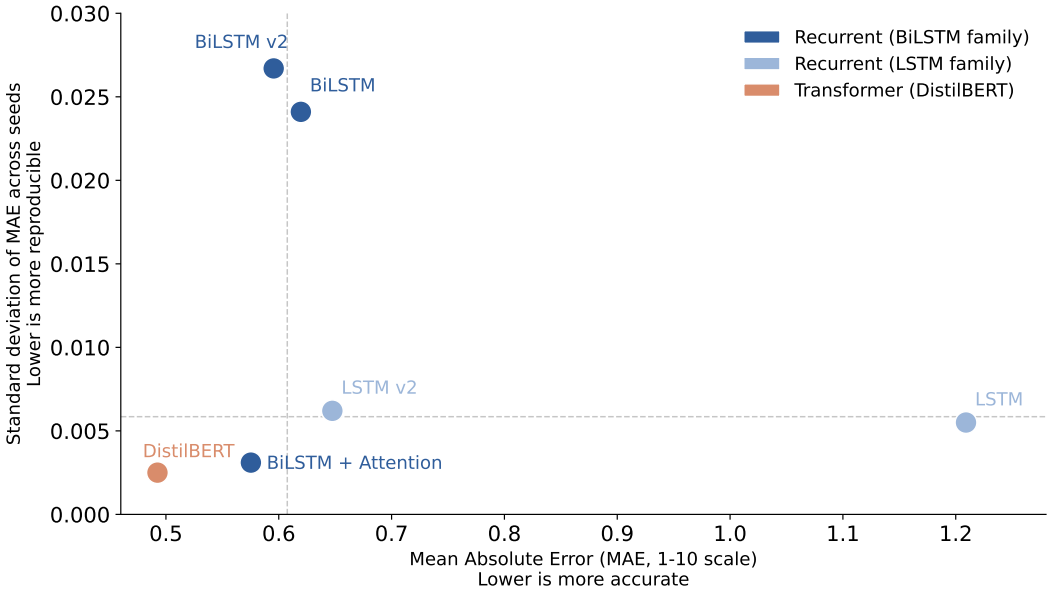

Figure 12. Joint comparison of prediction error and reproducibility across deep learning models. Each point represents a model by its mean MAE and the standard deviation of MAE across seeds. Models closer to the lower-left region combine lower error with greater consistency across runs.

Although DistilBERT was evaluated on only three seeds, it still occupies the most favorable position in the joint comparison of prediction error and reproducibility, combining the lowest mean MAE with the smallest observed variability. Figure 12 complements the per-seed view by showing that BiLSTM + Attention also moves much closer to this favorable lower-left region than the earlier recurrent variants. Moreover, these results indicate that the most accurate models in this study are also the most reproducible, which strengthens the practical reliability of the reported findings.

8. Discussion

The results of this study show that review rating prediction in hospitality depends not only on model choice, but also on the reliability of the input data. Across all experiments, the strongest performance was achieved when the training corpus was filtered through the anomaly preprocessing pipeline before model fitting. When filtering was removed entirely, MAE increased from 0.5753 to 0.6179, despite the larger training set. This confirms that simply retaining more reviews does not improve learning if the additional observations are noisy, repetitive, or internally inconsistent. Because the anomaly framework relies on fixed thresholds and heuristic weights, these gains suggest that preprocessing improves the training signal, without implying that every excluded review is fraudulent or invalid. In particular, correlation deviation and similarity detection produced the largest degradations, indicating that disagreement between textual sentiment and assigned rating, as well as near-duplicate content, are among the most damaging forms of unreliability for this task.

The comparison across architectures reveals an important interaction between model design and corpus structure. The original LSTM baseline performed worse than all other models, including the traditional TF-IDF baselines. Because the median review length is only 13 words, many examples do not provide enough context for a standard recurrent model to exploit its main strength. Once the architecture was improved through a larger vocabulary, stronger embeddings, and refinements based on the attention mechanism, performance improved markedly. LSTM v2 reduced MAE from 1.2091 to 0.6475, showing that the weak initial result was not a failure of recurrent modeling, but a sign that model configuration must be adapted carefully to the properties of short hospitality reviews.

The baseline BiLSTM and BiLSTM v2 achieved competitive mean errors, but both models showed considerable seed sensitivity. By contrast, BiLSTM + Attention combined lower error with a very small spread across runs. Its mean MAE of 0.5753 was accompanied by a standard deviation of only 0.0031,

 doi:10.20944/preprints202604.0921.v1

far below the variability observed for the earlier BiLSTM variants. Attention appears to improve both the quality of what the model learns and the consistency with which it learns. A plausible interpretation is that the attention mechanism provides a more stable weighting of informative parts of the review sequence, reducing the dependence of final performance on random initialization and training noise.

DistilBERT nevertheless remained the strongest overall model. It achieved the lowest MAE and RMSE, and it also occupied the most favorable position in the combined performance and stability comparison. Its advantage was especially visible in the error analysis. DistilBERT outperformed BiLSTM + Attention across all rating buckets and all review length quartiles, with the largest gains appearing for low-rating reviews and very short texts. Low-rating reviews are rare in the dataset, and short reviews provide limited lexical evidence for accurate prediction of small score differences. The fact that DistilBERT improves most in these settings suggests that pretrained contextual representations and subword tokenization are useful when the input is sparse or linguistically variable.

The exploratory comparative forecasting results for Task B extend the contribution of the paper beyond review-level prediction. They show that aggregated historical review signals contain useful information for short-term hotel-level rating forecasting under a fixed 30-day horizon. XGBoost achieved the best result among the evaluated baselines, improving over naive persistence, linear regression, and random forest. This indicates that historical sentiment, review volume, and anomaly proportion capture aspects of hotel reputation dynamics that are not fully explained by the previous average rating alone. At the same time, these findings should still be interpreted with appropriate caution because the forecasting task remains secondary to the main review-level prediction objective. Even so, the Task B analysis supports the broader idea that review text is useful for understanding individual user satisfaction and for anticipating future reputation trends for hotels.

In practical terms, the findings carry direct implications for hospitality analytics. First, reliability preprocessing should be regarded as an essential component of the modeling pipeline, not merely as a small cleaning procedure. Second, model selection should be aligned with the structure of the available review text. For short hospitality reviews, transformer models offer clear advantages when computational resources allow fine-tuning. At the same time, the BiLSTM + Attention model remains a competitive and computationally efficient alternative, offering a more favorable balance between performance and efficiency than the weaker recurrent baselines. The study shows that accurate rating prediction in hospitality depends on more than architecture alone and is best achieved through the combination of effective representation learning and careful control of data quality.

9. Limitations and Future Work

The present study has several limitations that also point to promising directions for future research. First, the analysis is based on reviews collected from a single platform, Booking.com, and predominantly reflects content in the English language. Review conventions, rating behavior, and posting patterns may differ across other platforms such as TripAdvisor, Expedia, or Agoda, which means that the anomaly preprocessing pipeline may not transfer directly without adjustment. In particular, the weights used in Equation (16) were calibrated for this corpus and may require adaptation in other hospitality environments. Future work should test the proposed framework on multi-platform and multilingual datasets to determine how well the filtering strategy generalizes across different review ecosystems.

A second limitation concerns the construction of the anomaly score itself. The weights $w_1 = 0.30$, $w_2 = 0.25$, $w_3 = 0.25$, and $w_4 = 0.20$ were chosen through domain reasoning rather than learned directly from data, and the filtering threshold at $A_i^{\text{norm}} = 0.40$ was selected to balance reliability improvement with data retention. Future work could investigate learned weighting strategies, differentiable filtering schemes, or lightweight learning approaches that optimize the anomaly score more directly for prediction quality. Related to this, the study does not include verified labels for fake or fraudulent reviews. The pipeline identifies statistically anomalous reviews with respect to sentiment

 doi:10.20944/preprints202604.0921.v1

consistency, duplication, internal disagreement, and behavior, but anomalous does not necessarily mean fraudulent, and some unusual reviews may still be genuine. Access to fraud labels verified by the platform would allow future studies to assess the pipeline as a detection mechanism in its own right, in addition to evaluating effects on prediction performance.

A further limitation arises from the nature of the review text. The median review length is only 13 words, which creates a difficult setting for precise rating prediction because many reviews provide very limited semantic evidence. DistilBERT partially addresses this through subword tokenization and contextual encoding, but very short inputs such as "great hotel" still leave little information from which to infer a precise score. Future work could explore data augmentation, the incorporation of context from other reviews, or hotel-level embeddings to enrich short reviews with additional information. Among transformer models, only DistilBERT was evaluated, while larger or more recent alternatives such as RoBERTa, BERT-large, or language models adapted to the hospitality domain may obtain further gains, especially in the low-rating ranges where performance remains weakest. At the same time, these potential improvements must be considered alongside their computational cost, especially for relatively small hospitality datasets.

Finally, the temporal forecasting component remains exploratory. Task B uses 30-day rolling windows and a straightforward aggregation strategy based on review volume, mean sentiment, and anomaly proportion. While this setup was sufficient to demonstrate that historical review signals can support short-term hotel-level forecasting, it likely does not capture the full complexity of reputation dynamics. Future work could examine richer aggregation strategies, including exponentially weighted sentiment trends, windows responsive to specific events, or graph-based similarity structures across hotels. It would also be useful to compare different forecasting horizons, including shorter 7-day windows and longer 90-day windows, to determine whether different hotel types exhibit distinct temporal patterns in how online feedback translates into future rating behavior.

10. Conclusions

This study developed and evaluated a framework for predicting hotel ratings from user review text with explicit consideration of anomalous content. The proposed approach combines reliability preprocessing with deep learning models to improve predictive performance and the quality of the training signal. Rather than treating review text as uniformly trustworthy, the framework incorporates sentiment inconsistency, similarity detection, cross-component disagreement, and reviewer behavior into a composite anomaly score, then uses that score to filter unreliable observations before model training. This design addresses a practical weakness of many prediction pipelines, namely that model quality often depends as much on input data quality as on architectural design.

The architectural comparison further shows that model performance depends strongly on the structure of the review text. The original LSTM baseline performs poorly because the median review length is only 13 words, which limits the amount of sequential evidence available for standard recurrent learning. However, the optimized recurrent variants recover much of this lost performance, and the addition of self-attention provides the strongest result within the recurrent family. BiLSTM + Attention achieves an MAE of 0.5753 and an RMSE of 0.8636, while also improving reproducibility across random seeds. The standard deviation across random seeds of 0.0031 and the span of 0.0092 indicate that the model is accurate and sufficiently stable for practical use.

DistilBERT nevertheless provides the best overall outcome in the study. With MAE = 0.4925 and RMSE = 0.7368, it outperforms the traditional baselines and the recurrent architectures and achieves the most favorable balance between predictive performance and stability among the evaluated models. The error analysis helps explain why. DistilBERT is consistently better across all rating buckets and all review-length quartiles, with the largest gains appearing for low-rating reviews and very short texts, which are the most difficult parts of the task because they are underrepresented or contain limited semantic evidence. This suggests that pretrained contextual representations and subword tokenization

 doi:10.20944/preprints202604.0921.v1

are particularly effective in hospitality settings where texts are brief, highly skewed toward positive sentiment, and often linguistically sparse.

The secondary forecasting analysis extends the contribution of the paper beyond review-level regression. The Task B results show that aggregated historical review signals, including sentiment indicators, review volume, and anomaly proportion, contain useful information for short-term hotel-level rating forecasting. The best window-level performance is achieved by XGBoost, which improves over naive persistence, linear regression, and random forest, suggesting that textual review dynamics are informative for understanding individual guest satisfaction and for anticipating broader reputation trends at the hotel level. This strengthens the practical value of the framework by linking review interpretation to managerial forecasting.

Taken together, these findings show that reliable hospitality rating prediction requires expressive neural architectures and explicit control of review quality before learning begins. More broadly, the study demonstrates that integrating reliability preprocessing with modern language models can support interpretable review analytics and predictive reputation intelligence in digital hospitality platforms.

Author Contributions: Conceptualization, M.N. and M.M.; methodology, M.N., M.M. and M.S.; machine learning model design and implementation, M.N. and M.S.; model architecture development and configuration, M.N. and M.S.; ablation study design and evaluation, M.S. and M.N.; data preprocessing and feature engineering, M.N.; validation and statistical evaluation, M.N., M.M. and M.S.; formal analysis, M.N.; investigation, M.N.; resources, M.N.; data curation, M.N.; writing (original draft preparation), M.N.; writing (review and editing), M.M. and M.S.; visualization, M.N.; scientific supervision, methodological guidance, and critical manuscript revision, M.M.; project administration, M.N. All authors have read and agreed to the published version of the manuscript and accept responsibility for its content.

Funding: This research was supported by the Science Fund of the Republic of Serbia, Grant No. 7502. The article processing charge (APC) was covered by the authors.

Data Availability Statement: The dataset analyzed in this study is publicly available on the Kaggle platform as the *Booking.com Hotel Reviews* dataset [4]. The dataset contains hotel reviews and associated metadata used for textual analysis and rating prediction experiments. No new datasets were created as part of this study. Derived feature representations, intermediate preprocessing outputs, and trained model configurations are not deposited in a public repository due to their size and dependence on the experimental environment, but are available from the corresponding author upon reasonable request for research verification and reproducibility purposes.

Acknowledgments: 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)*.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

 doi:10.20944/preprints202604.0921.v1

The following abbreviations are used in this manuscript:

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DL	Deep Learning
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
ERNIE	Enhanced Representation through kNowledge IntEgration
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
RoBERTa	Robustly Optimized BERT Pretraining Approach
TF-IDF	Term Frequency–Inverse Document Frequency
VADER	Valence Aware Dictionary and sEntiment Reasoner
XGBoost	Extreme Gradient Boosting

References

1. Zhuang, Y.; Kim, J. A BERT-Based Multi-Criteria Recommender System for Hotel Promotion Management. *Sustainability* **2021**, *13*, 8039. https://doi.org/10.3390/su13148039
2. Zheng, T.; Wu, F.; Law, R.; Qiu, Q.; Wu, R. Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach. *Int. J. Hosp. Manag.* **2021**, *92*, 102658. https://doi.org/10.1016/j.ijhm.2020.102658
3. Ahmed, B.H.; Ghabayen, A.S. Review rating prediction framework using deep learning. *J. Ambient Intell. Humaniz. Comput.* **2022**, *13*, 3423–3432. https://doi.org/10.1007/s12652-020-01807-4
4. The Devastator. Booking.com Hotel Reviews. Available online: https://www.kaggle.com/datasets/thedevastator/booking-com-hotel-reviews/ (accessed on 5 January 2026).
5. Nikolić, M.; Stojanović, M.; Marjanović, M. The Power of Words: Leveraging Deep Learning Techniques to Predict Hotel Ratings from User Reviews. In *Proceedings of the 24th International Symposium INFOTEH-JAHORINA (INFOTEH 2025)*; IEEE: Jahorina, Bosnia and Herzegovina, 19–21 March 2025; pp. 1–6. https://doi.org/10.1109/INFOTEH64129.2025.10959201
6. Özen, İ.A.; Özgül Katlav, E. Aspect-based sentiment analysis on online customer reviews: A case study of technology-supported hotels. *J. Hosp. Tour. Technol.* **2023**, *14*, 102–120. https://doi.org/10.1108/JHTT-12-2020-0319
7. Wen, Y.; Liang, Y.; Zhu, X. Sentiment analysis of hotel online reviews using the BERT model and ERNIE model—Data from China. *PLOS ONE* **2023**, *18*, e0275382. https://doi.org/10.1371/journal.pone.0275382
8. Husein, A.M.; Livando, N.; Andika, A.; Chandra, W.; Phan, G. Sentiment analysis of hotel reviews on TripAdvisor with LSTM and ELECTRA. *Sinkron* **2023**, *7*, 733–740. https://doi.org/10.33395/sinkron.v8i2.12234
9. Chen, P.; Fu, L. Enhancing multimodal tourism review sentiment analysis through advanced feature association techniques. *Int. J. Inf. Syst. Serv. Sect.* **2024**, *15*, 1–21. https://doi.org/10.4018/IJISSS.349564
10. Nikolić, M.; Stojanović, M.; Marjanović, M. Integrating deep learning for automated detection of negative hotel reviews. *Facta Univ. Ser. Autom. Control Robot.* **2025**, *24*, 1–16. https://doi.org/10.22190/FUACR241218002N
11. Puh, K.; Bagić Babac, M. Predicting sentiment and rating of tourist reviews using machine learning. *J. Hosp. Tour. Insights* **2023**, *6*, 1188–1204. https://doi.org/10.1108/JHTI-02-2022-0078

 doi:10.20944/preprints202604.0921.v1

12. Zhang, D.; Wu, C. What online review features really matter? An explainable deep learning approach for hotel demand forecasting. *J. Assoc. Inf. Sci. Technol.* **2023**, *74*, 1100–1117. https://doi.org/10.1002/asi.24807
13. Hossen, M.S.; Jony, A.H.; Tabassum, T.; Islam, M.T.; Rahman, M.M.; Khatun, T. Hotel review analysis for the prediction of business using deep learning approach. In *Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*; IEEE: Coimbatore, India, 2021; pp. 1489–1494. https://doi.org/10.1109/ICAIS50930.2021.9395757
14. Zhang, H.; Kassim, A.M.; Samsudin, N.H.; Teng, L.; Tang, C.Y. A hybrid deep learning framework for hotel rating systems: Integrating Word2Vec, TF-IDF, and Bi-LSTM with attention mechanism. *IEEE Trans. Comput. Soc. Syst.* **2025**, *12*, 2371–2384. https://doi.org/10.1109/TCSS.2024.3461796
15. Ganji, R.N.; Dadkhah, C.; Tohidi, N. Improving sentiment classification for hotel recommender system through deep learning and data balancing. *Comput. Sist.* **2023**, *27*, 811–825. https://doi.org/10.13053/cys-27-3-4655
16. Zhao, R.; Hao, Y.; Li, X. Business analysis: User attitude evaluation and prediction based on hotel user reviews and text mining. *arXiv* **2024**, arXiv:2412.16744. https://doi.org/10.48550/arXiv.2412.16744
17. Nikolić, M.; Stojanović, M.; Marjanović, M. Integrating data science and predictive modeling for detecting inconsistent hotel reviews. In *Proceedings of UNITECH 2024 – Selected Papers*; Technical University of Gabrovo: Gabrovo, Bulgaria, 2024; pp. 104–110. http://www.doi.org/10.70456/DHXA1258
18. Alsubari, S.N.; Deshmukh, S.N.; Alqarni, A.A.; Alsharif, N.; Aldhyani, T.H.H.; Alsaade, F.W.; Khalaf, O.I. Data analytics for the identification of fake reviews using supervised learning. *Comput. Mater. Continua* **2022**, *70*, 3189–3204. https://doi.org/10.32604/cmc.2022.019625
19. Duma, R.A.; Niu, Z.; Nyamawe, A.S.; Tchaye-Kondi, J.; Yungaicela-Naula, N.; Abdulhamid, S.M. A deep hybrid model for fake review detection by jointly leveraging review text, overall ratings, and aspect ratings. *Soft Comput.* **2023**, *27*, 6281–6296. https://doi.org/10.1007/s00500-023-07897-4
20. Prasetyaningrum, P.T.; Suria, O.; Ibrahim, N.; Riadi, I. Smart sentiment forensics: Integrating AI and digital forensics for fake hotel review detection. In *Proceedings of the 2025 2nd International Conference on Information System and Information Technology (ICISIT)*; IEEE: Yogyakarta, Indonesia, 2025; pp. 1–6. https://doi.org/10.1109/ICISIT66233.2025.11403004
21. Nikolić, M.; Stojanović, M.; Marjanović, M. Anomaly detection in hotel reviews: Applying data science for enhanced review integrity. In *Proceedings of the 32nd Telecommunications Forum (TELFOR 2024)*; IEEE: Belgrade, Serbia, 26–27 November 2024; pp. 1–6. https://doi.org/10.1109/TELFOR63250.2024.10860993
22. Zhang, L.; Guo, J.; Kang, R.; Zhao, B.; Zhang, C.; Li, J. Hotel review classification based on the text pretraining heterogeneous graph neural network model. *Comput. Intell. Neurosci.* **2022**, *2022*, 5259305. https://doi.org/10.1155/2022/5259305
23. Deng, L.; Yin, T.; Li, Z.; Ge, Q. Analysis of the effectiveness of CNN-LSTM models incorporating BERT and attention mechanisms in sentiment analysis of data reviews. In *Proceedings of the 2023 4th International Conference on Big Data and Informatization Education (ICBDIE 2023)*; Atlantis Press: 2023; pp. 821–829. https://doi.org/10.2991/978-94-6463-238-5_106
24. Chen, N.; Sun, Y.; Yan, Y. Sentiment analysis and research based on two-channel parallel hybrid neural network model with attention mechanism. *IET Control Theory Appl.* **2023**, *17*, 2259–2267. https://doi.org/10.1049/cth2.12463
25. Yuan, Y. DistilBERT hotel rating prediction model based on an ensemble learning framework. In *Proceedings of the 2024 3rd International Conference on Electronics and Information Technology (EIT)*; IEEE: Chengdu, China, 2024; pp. 763–769. https://doi.org/10.1109/EIT63098.2024.10762068
26. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Leveraging large language models in tourism: A comparative study of the latest GPT Omni models and BERT NLP for customer review classification and sentiment analysis. *Information* **2024**, *15*, 792. https://doi.org/10.3390/info15120792
27. Siino, M.; Tinnirello, I.; La Cascia, M. Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Inf. Syst.* **2024**, *121*, 102342. https://doi.org/10.1016/j.is.2023.102342
28. Wang, E.Y.; Fong, L.H.N.; Law, R. Detecting fake hospitality reviews through the interplay of emotional cues, cognitive cues and review valence. *Int. J. Contemp. Hosp. Manag.* **2022**, *34*, 184–200. https://doi.org/10.1108/IJCHM-04-2021-0473
29. Zhang, D.; Li, W.; Niu, B.; Wu, C. A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decis. Support Syst.* **2023**, *166*, 113911. https://doi.org/10.1016/j.dss.2022.113911

 doi:10.20944/preprints202604.0921.v1

30. Liu, J.; Hu, S.; Mehraliyev, F.; Liu, H. Text classification in tourism and hospitality—A deep learning perspective. *Int. J. Contemp. Hosp. Manag.* **2023**, *35*, 4177–4190. https://doi.org/10.1108/IJCHM-07-2022-0913