



Appraising Customer Experience from Hotel Reviews using Text-Based Transformers

Fiqi AMALI ¹ , Halil YİĞİT ^{2,*} , Zeynep Hilal KİLİMCİ ³

¹ Department of Information Systems Engineering, Kocaeli University, Kocaeli, Türkiye, **ORCID:** 0009-0003-2871-0311

² Department of Information Systems Engineering, Kocaeli University, Kocaeli, Türkiye, **ORCID:** 0000-0003-0932-6966

³ Department of Information Systems Engineering, Kocaeli University, Kocaeli, Türkiye, **ORCID:** 0000-0003-1497-305X

Abstract

Article Info

Research paper

Received : February 11, 2026

Accepted : March 31, 2026

Keywords

Customer experience appraisal
Sentiment Analysis
Transformer architectures
ELECTRA
Data augmentation

In the rapidly evolving landscape of the hospitality industry, comprehending and enhancing customer experience through digital feedback has become a strategic imperative. This study introduces a high-precision framework powered by advanced transformer-based architectures to evaluate customer sentiments in hotel reviews. A significant challenge in this domain, class imbalance, was addressed through an innovative data augmentation strategy utilizing ChatGPT to generate semantically consistent paraphrased samples for minority classes. We conducted a comprehensive comparative analysis of five state-of-the-art models: BERT, DistilBERT, RoBERTa, ELECTRA, and GPT-2. The methodology was validated using a specialized dataset, the Bali Hotel Review Corpus (BHRC), alongside four globally recognized benchmarks (IMDb, SST-1, SST-2, and MR). Experimental results demonstrate that the proposed framework achieves exceptional accuracy levels, with the ELECTRA model outperforming others by reaching a peak accuracy of 99.96% under balanced conditions. This performance is attributed to ELECTRA's efficient discriminative pre-training objective, which proves superior to standard masked or generative approaches for domain-specific sequence classification. Our findings provide a robust and resource-efficient benchmark for real-time customer experience appraisal, offering hotel managers actionable insights to drive service quality in the modern digital tourism economy.

1. Introduction

Today's online platforms enable continuous, large-scale interaction between customers and the products or services they purchase. In this setting, user-generated reviews play a central role: they shape prospective customers' preferences and simultaneously influence perceptions of a brand's reputation and trustworthiness¹. Within the hospitality industry, guests routinely share experiences and assign ratings via web- and mobile-based online travel agencies (OTAs) such as Booking.com, TripAdvisor, and Expedia². Feedback is typically expressed through free-form text and/or numerical scores reflecting satisfaction. These reviews often cover multiple service dimensions, including service quality, cleanliness, room comfort, location, and available facilities, thereby providing potential customers with a detailed view of

strengths and weaknesses and affecting both hotel selection and booking intention³. Beyond their influence on consumers, reviews constitute an important operational resource: compared with traditional advertising signals, experiential narratives can offer more direct evidence for identifying areas requiring improvement in service delivery⁴.

Sentiment analysis (SA), a core task in natural language processing (NLP), aims to infer affective polarity or emotion from opinionated text⁵. In recent years, SA has become a practical tool for summarizing and interpreting customer attitudes across products and services, and it is particularly relevant to hospitality contexts where customer satisfaction is tightly linked to competitiveness and revenue^{6,7}. Prior studies have approached hotel-review sentiment classification using conventional machine learning methods—such as Naïve Bayes, support vector machines (SVM), k-nearest neighbors (KNN), logistic regression, decision trees, and shallow neural networks⁸⁻¹⁴—as well as deep learning architectures including

* Corresponding Author: halilyigit@kocaeli.edu.tr



convolutional neural networks (CNN), long short-term memory (LSTM), bidirectional LSTM (BiLSTM), and recurrent neural networks (RNN)¹⁵⁻¹⁸. While these approaches have shown promising results, they can be limited in modeling nuanced contextual cues and long-range dependencies that frequently occur in opinion-rich reviews.

Transformer-based models address these limitations through attention mechanisms that explicitly model token-to-token interactions, enabling richer contextual representations and improved handling of subtle sentiment cues^{19,20}. Motivated by these advances, this study investigates the effectiveness of five transformer-based models—BERT, DistilBERT, RoBERTa, ELECTRA, and GPT-2—for binary sentiment classification (positive vs. negative) of hotel reviews. We construct a domain-specific dataset by collecting 5000 customer reviews from TripAdvisor and Hotels.com for hotels in the Seminyak region of Bali, Indonesia, and use it to evaluate model performance under realistic hospitality-domain conditions.

This work makes the following contributions:

- i. We provide a systematic comparative evaluation of multiple pre-trained transformer models (BERT, DistilBERT, RoBERTa, ELECTRA, and GPT-2) for hotel-review sentiment classification.
- ii. To mitigate class imbalance, we employ ChatGPT-based paraphrasing for data augmentation and examine its effect on classification performance, particularly for the minority (negative) class.
- iii. We introduce a domain-specific review dataset collected from TripAdvisor and Hotels.com to support transformer-based customer experience appraisal in hospitality settings.
- iv. We validate the generalizability of the proposed framework using widely adopted benchmark datasets, including IMDb, SST-1, SST-2, and MR, and report comparative results against recent literature.

Although the field has recently shifted toward large-scale generative models, high-precision sequence classification in constrained domains often benefits from specialized transformer encoders. Models such as ELECTRA and RoBERTa learn strong bidirectional representations—via objectives such as replaced token detection and dynamic masking—which can be advantageous for discriminative tasks. Moreover, compared with large language model (LLM) deployments, fine-tuning encoder-centric transformers typically offers a more reproducible and computationally efficient solution for large-volume, real-time customer feedback analysis. Accordingly, this study establishes a practical benchmark for hospitality sentiment analysis, emphasizing performance under domain constraints while accounting for computational considerations.

The remainder of this paper is organized as follows. Section 2 reviews related studies employing machine learning and deep learning methods for hotel-review sentiment classification. Section 3 summarizes the transformer-based models considered in this work. Section 4 describes the datasets used in the experiments. Section 5 details the proposed methodology. Section 6 presents experimental results and performance analysis. Finally, Section 7 concludes the paper and outlines future directions.

2. Related Works

This section reviews sentiment analysis studies focusing on hotel and tourism-related customer reviews.

In [10], a framework was proposed for categorizing and predicting customer sentiment using reviews collected from an international hotel. Several machine learning algorithms, including SVM, ANN, Naïve Bayes, Decision Tree, C4.5, and KNN, were evaluated. Among them, the Decision Tree achieved the best performance and was further utilized to identify key factors influencing positive customer experiences. The study also highlighted the impact of feature dimensionality on classification performance.

The study in [11] optimized a Naïve Bayes classifier for hotel review sentiment analysis through a three-stage pipeline comprising text pre-processing, N-gram feature extraction, and classification. Experiments conducted on the OpinRank Hotels Review dataset showed that unigram features outperformed bigram and trigram representations, achieving an average accuracy of 81.30%.

In [12], a multidimensional sentiment analysis framework was introduced for Arabic hotel reviews collected from Booking.com, covering both standard and dialectal forms. Nine machine learning classifiers were evaluated under different data representation and pre-processing scenarios. The Ridge Classifier demonstrated superior performance across accuracy, precision, recall, F1-score, and training time, while pre-processing techniques such as stop-word removal and stemming further improved results.

The performance of the proposed CNN-based approach for the sentiment classification task was reported in [13]. Based on experimental evaluations conducted on the AWARE dataset, the model achieved accuracy rates of 80% in the productivity domain, 70% in the gaming domain, and 86% in the social networking domain. These results demonstrate notable performance improvements over the existing baseline methods, with gains of 16.43% in productivity, 3.72% in gaming, and 23.35% in social networking.

In [14], a dynamic framework was proposed for measuring and evaluating hotel customer satisfaction through sentiment analysis of online reviews. The framework integrated review preprocessing, topic extraction, sentiment orientation analysis, and dynamic importance-performance competitor analysis to generate managerial insights. Its feasibility was demonstrated using online reviews of two five-star hotels collected from Ctrip.com, with the results compared against manual annotations and findings reported in the relevant literature.

Deep learning approaches were explored in [15], where an LSTM-based sentiment classification model incorporating word embeddings was evaluated on a large hotel review dataset. The proposed model outperformed several state-of-the-art methods, achieving an accuracy of 97% and demonstrating the effectiveness of embedding-enhanced recurrent architectures for sentiment classification.

In [16], an RNN-based model named DSWE-GRNN was introduced for multi-class review rating prediction using domain-specific word embeddings and GRU units. Without relying on reviewer metadata, the model effectively captured latent contextual information and benefited from downsampling for data balancing. Experimental results on IMDb and hotel review datasets showed performance comparable to or exceeding common baseline methods.

The work in [17] investigated machine learning and deep learning models for joint sentiment and rating prediction using TripAdvisor reviews. While traditional classifiers employed TF-IDF representations, deep learning models utilized GloVe embeddings. The BiLSTM architecture emerged as the most effective model, outperforming others in both accuracy and efficiency.

Hybrid feature-based sentiment classification methods were examined in [18], comparing machine learning, ensemble, and deep learning approaches for detecting praise and complaints in hotel reviews. Results indicated that ensemble classifiers and deep neural networks achieved superior performance across accuracy, precision, recall, and F1-score metrics.

Finally, [19] introduced a hotel recommendation system based on sentiment analysis and aspect-based review categorization. The proposed framework combined BERT-based sentiment modeling handcrafted textual features, fuzzy logic, and random forest classification. The system achieved a Macro F1-score of 84% and a test accuracy of 92.36%, demonstrating the effectiveness of integrating sentiment analysis into recommendation systems for the hospitality domain.

Building on the above literature, we next introduce transformer-based architectures, which have become the dominant paradigm in NLP due to their attention-driven

ability to model global context and improve sequence classification performance.

3. Models

This section gives a summary of transformer models evaluated in this study: Bidirectional Encoder Representations from Transformers (BERT), Distilled BERT (DistilBERT), A Robustly Optimized BERT Pretraining Approach (RoBERTa), Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), and Generative Pre-Training-2 (GPT-2).

3.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a transformer-based encoder designed to learn deep bidirectional representations by conditioning on both left and right context. Introduced by Google in 2018, BERT relies on self-attention to capture contextual dependencies in sequential text and serves as a general-purpose backbone for downstream NLP tasks [20]. BERT is typically trained in two stages: (i) pretraining on large-scale unlabeled corpora to learn language regularities and (ii) task-specific fine-tuning using labeled data. This paradigm has enabled strong performance across a wide range of applications, including sentiment analysis and question answering²¹.

Despite its effectiveness, BERT's large parameter count can entail substantial computational and memory overhead during training and inference, which may limit deployment in resource-constrained settings²².

3.2. Distilled BERT (DistilBERT)

DistilBERT is a compressed variant of BERT developed via knowledge distillation, where a smaller "student" model is trained to approximate the behavior of a larger "teacher" model during pretraining. This strategy preserves much of BERT's language understanding capacity while reducing model size and improving inference speed, making it better suited to latency- and resource-sensitive applications²³.

3.3. RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) builds on BERT by refining the pretraining recipe and training setup. Compared to the original BERT configuration, RoBERTa removes the next sentence prediction objective, adopts improved hyperparameter settings, and applies dynamic masking, where the masked

tokens vary across epochs. These adjustments, together with training on larger corpora, yield stronger contextual representations and improved generalization in downstream tasks²⁴.

3.4. ELECTRA

ELECTRA introduces a sample-efficient pretraining objective based on replaced token detection. Instead of masking tokens as in BERT, ELECTRA replaces some tokens with plausible alternatives and trains a discriminator to identify which tokens were replaced. Because the model learns from all tokens in the sequence, this objective improves training efficiency and often provides competitive performance with fewer parameters, which is advantageous under limited computational budgets²⁵.

3.5. GPT-2

GPT-2 is an autoregressive transformer language model pretrained to predict the next token given prior context. This causal language modeling objective enables GPT-2 to capture long-range dependencies and generate coherent text, after which the model can be adapted to various downstream tasks via fine-tuning²⁶. Relative to earlier GPT variants, GPT-2 scales model capacity substantially (125M–1.5B parameters) and incorporates architectural refinements, contributing to improved general-purpose performance across diverse NLP settings²⁷.

The fundamental architectural distinctions and the specific training objectives of the aforementioned transformer models are summarized in Table 1. This comparison highlights the transition from standard masked language modeling to more sample-efficient discriminative approaches, which significantly influences their classification performance in the hospitality domain.

4. Methodology

4.1. Data Collection

The experimental data for this study is a dataset of hotel reviews for W Seminyak Hotel in Indonesia obtained from two of the most popular online travel platforms, namely TripAdvisor and Hotels.com. The extensive and abundant content generated by users on both platforms significantly increases the chance of gathering comprehensive and sufficient data for analysis. The W Seminyak Hotel has been chosen for this study for two reasons. Firstly, W Seminyak Hotel is a luxury hotel chain primarily targeting the younger demographic and owned

by Marriott International. Younger individuals tend to express their complaints more prominently online. Secondly, Google Trends is used to compare different hotel brands such as Hilton, Marriott, Hyatt, Sheraton, Holiday Inn, etc., revealing that "W Seminyak" is the most frequently searched keyword among all hotel brands in Bali.

Table 1. Summary of Evaluated Transformer Architectures

Model	Architecture Type	Primary Training Objective	Technical Driver for Performance
BERT	Bidirectional Encoder	Masked Language Modeling (MLM)	Standard bidirectional context; limited by 15% token masking efficiency.
RoBERTa	Optimized Bidirectional Encoder	Dynamic MLM & No NSP	Enhanced robustness through larger batches and dynamic masking strategies.
ELECTRA	Discriminative Encoder	Replaced Token Detection (RTD)	Superior sample efficiency; learns from 100% of tokens in every sequence.
GPT-2	Unidirectional Decoder	Causal Language Modeling	High capacity for long-range dependencies, but less optimized for classification than encoders.
DistilBERT	Distilled Encoder	Knowledge Distillation	Maintains ~97% of BERT's performance with 40% fewer parameters.

Figure 1 illustrates the multi-stage process employed to construct the Bali Hotel Review Corpus (BHRC). This methodology results in two distinct datasets: the original BHRC dataset and a balanced version derived from it. While the original BHRC retains an imbalanced class distribution, the balanced BHRC dataset is obtained through targeted adjustments designed to achieve a more uniform class representation.

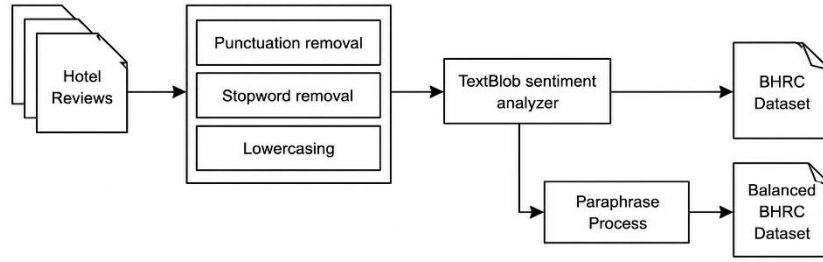


Figure 1. Pre-processing stage for BHRC dataset.

The raw dataset consists of 4518 hotel reviews. The retrieval of customer review data is done by crawling via WebHarvy. Before the analysis of the proposed system, the following data pre-processing steps are ensured to eliminate inconsistencies by cleaning the reviews in the collected dataset:

- i. If a review is not written in English, then it is discarded.
- ii. Neutral comments, if any, are not included.
- iii. Punctuation marks, numbers, and emojis are removed from the reviews.

The Natural Language Toolkit (NLTK) was employed to pre-process and clean the collected textual data. The pre-processing pipeline included stop-word removal to eliminate frequently occurring but semantically insignificant terms such as “is,” “the,” “and” and “a.” In addition, all text was converted to lowercase to ensure consistency, and punctuation marks were removed to reduce noise in the input data. Examples illustrating the application of these pre-processing steps to hotel reviews in the dataset are presented in Table 2.

Table 2. Example of data pre-processing on hotel reviews.

Reviewed Text	Pre-processed Text
W Bali Seminyak is absolutely a stunning property to rejuvenate in a great location. It's worth the money you spend here. With the Marriott Bonvoy Platinum elite status I was able to enjoy additional benefits which was really nice. Everything from the minute I checked and to check out it was extraordinary.	“w”,“bali”,“seminyak”, “absolutely”, “stunning”,“property”, “rejuvenate”, “great”, “location”, “worth”,“money”,“spend”,“here”, “marriott”, “bonvoy”, “platinum”,“elite”,“status”,“able”, “enjoy”,“additional” “benefits”, “really”, “nice”, “everything”, “minute”, “checked”, “check” “out”, “extraordinary”.
The W Hotel was the best part of my Bali trip. The service at the W Hotel is fantastic and the whole vibe in general. I would recommend W Hotel to anyone going to Bali. There is no down side of my stay at the W Hotel.	“w”,“hotel”,“best”,“part”,“bali”,“trip”,“service”,“w”,“hotel”,“fantastic”,“whole” “vibe”,“general”,“would”,“recommend”,“w”,“hotel”, “anyone”,“going”,“bali”,“no”,“down”,“side”,“my”,“stay”,“w”,“hotel”
W seminyak was the 5th hotel we visited in Bali. Unfortunately, from the moment we arrived at the hotel we didn't feel much welcome after a few jokes by the receptionist that may have had the opposite intention but were expressed poorly.	“w”,“seminyak”,“5”,“hotel”,“visited”,“bali”,“unfortunately”,“moment”,“arrived”,“hotel”,“did”,“not”,“feel”,“much”,“welcome”,“after”,“jokes”,“receptionist”,“may”,“have”,“had”,“opposite”,“intention”,“expressed”,“poorly”
I been W bali couple hours ago. Seriously, u're not 5 stars grade. Bad services ever. Please take care more ur guest. I'm so disappointed with this situation. I saw in trip advisor u're the best. But poor u are. U are The bad ever hotel. Low grade ever!!	“w”,“bali”,“couple”,“hours”,“ago”,“seriously”,“not”,“5”,“stars”,“grade”,“bad” “services”,“ever”,“please”,“take”,“care”,“more”,“guest”, “so”,“disappointed” “situation”,“saw”,“trip”,“advisor”,“best”,“poor”, “bad”,“ever”,“hotel”,“low”,“grade” “ever”

The polarity scores generated by the TextBlob library were used to assign sentiment labels to the hotel reviews. TextBlob calculates a rule-based polarity score ranging from −1 to 1. Reviews with scores below zero were labeled as negative, whereas those with scores above zero were labeled as positive. Although this rule-based method enabled efficient labeling of the large-scale BHRC dataset, it may not capture the nuances reflected in human annotations and may introduce systematic bias into the transformer training process. Nevertheless, this limitation was considered an acceptable trade-off for resource-

efficient labeling. Table 3 provides representative post-pre-processing labeling examples. Using this procedure, the Bali Hotel Review Corpus (BHRC) comprises 4094 positive and 424 negative reviews, indicating a substantial class imbalance. To address this issue, ChatGPT-based data augmentation was performed before the train-test split to align the negative class with the positive class size. This sequence was strategically chosen to ensure that the minority class achieved sufficient statistical representativeness in the evaluation phase, as a pre-split augmentation would have otherwise resulted in an

excessively sparse test set. To mitigate potential data leakage, we utilized high-variance prompting to ensure that the generated paraphrases introduced significant structural and lexical diversity from the original samples. Additionally, a subset of the generated samples was manually verified to preserve sentiment polarity. Table 4 summarizes the sentiment distribution before and after the class-balancing procedure.

Table 4. The distribution of sentiments in the dataset

Dataset	Positive	Negative	Total
BHRC	4094	424	4518
Balanced BHRC	4094	4094	8188

Table 3. Sentiment labels on the dataset after pre-processing.

Review	Sentiment
w bali seminyak absolutely stunning property rejuvenate great location worth money spend marriott bonvoy platinum elite status enjoy additional benefits really nice everything minute checked check out extraordinary	Positive
w hotel best part bali trip service w hotel fantastic whole vibe general would recommend w hotel anyone going bali no down side my stay w hotel	Positive
w seminyak 5th hotel visited bali unfortunately moment arrived hotel did not feel much welcome after jokes receptionist may have had opposite intention expressed poorly	Negative
w bali couple hours ago seriously not 5 stars grade bad services ever please take care more guest so disappointed situation saw trip advisor best but poor bad ever hotel low grade ever	Negative

Additionally, experiments are conducted on four commonly used benchmark datasets: SST-1, SST-2, IMDb, and Movie Reviews (MR). These datasets are chosen for their widespread adoption in the field, encompassing a variety of linguistic and sentiment analysis tasks. Each dataset presents unique features and complexities that enable a comprehensive evaluation of the proposed methodologies. The rationale for selecting these specific datasets lies in their representativeness and adherence to established criteria in the field, ensuring a robust evaluation of experimental results against established standards. A detailed description of the datasets is provided in Table 5.

4.2. Proposed Framework

In this study, the proposed framework is designed to classify hotel reviews as positive or negative employing transformer models. Moreover, the performances of the models are compared on a wide scale using different versions of publicly available datasets in the literature.

The general flowchart of the proposed framework is given in Figure 2. It is important to note that the pre-processing stage is only applied to our custom dataset to

compare the proposed framework with literature studies. The pre-processing processes depicted in Figure 1 are applied only to balanced and unbalanced BHRC datasets. The processed datasets are then fed into the BERT, DistilBERT, RoBERTa, and GPT-2 transformer models. Finally, the sentiment classification performance of the models is assessed using accuracy, precision, recall, F1-score, and AUC evaluation metrics.

Table 5. Characteristics of the datasets where c: Number of classes, A_1 : Average sentence length, M_1 : Maximum sentence length, V_{Train} : Training set, V_{Test} : Test set, V_{Valid} : Validation set

Dataset	c	A_1	M_1	V_{Train}	V_{Test}	V_{Valid}
BHRC	2	76	256	2710	904	904
Balanced BHRC	2	76	256	5151	1717	1637
MR ¹	2	20	51	8655	1064	2166
SST-1 ²	5	18	56	8544	2210	1101
SST-2 ²	2	19	56	6920	1821	872
IMDb ³	2	21	252	30000	10000	10000

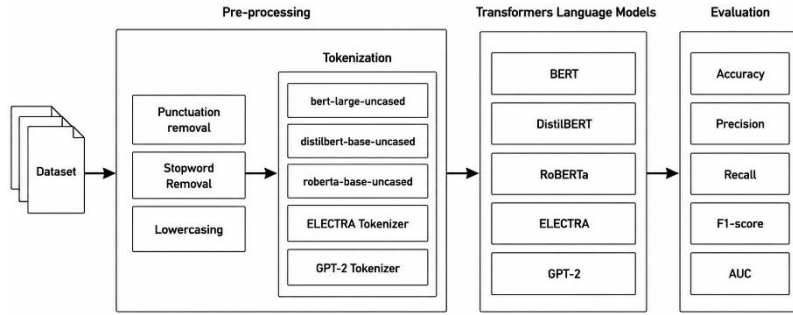


Figure 2. The general flowchart of the proposed framework

During the fine-tuning of the pre-trained transformer models, most hyperparameters were retained as specified in their original publications, with only the batch size and maximum token length being adjusted. These specific adjustments were informed by preliminary experiments and the memory limits of the Tesla T4 GPU to achieve a balance between computational efficiency and stable model performance. The learning rate is set to $2e-5$, and the batch size is set to 32. A maximum sequence length of 256 tokens was selected based on the distribution of review lengths in the BHRC dataset, allowing most reviews to be processed without truncation. The models were trained for 10 epochs, which preliminary experiments indicated was sufficient to achieve convergence while limiting the risk of overfitting. All models are implemented using the Hugging Face transformer library. The same computing environment (Tesla T4 GPU) is utilized for the training process. To achieve optimal performance in practice, the dataset is divided into three subsets for training, validation, and testing, with proportions of 60%, 20%, and 20%, respectively. Detailed hyperparameter settings for all models are provided in Table 6.

Table 6. Hyperparameter settings for all models.

Hyperparameters	Value
Learning rate	$2e-5$
Token Length	256
Batch Size	32
Epochs	10
Optimizer	Adam

5. Results

In this section, experiment results are demonstrated on both balanced and imbalanced datasets, presenting the performance of different transformer models. In addition, state-of-the-art (SOTA) results are also provided to show the effectiveness of the proposed framework.

Table 7. Comparison of Evaluation Metrics for Imbalanced BHRC Dataset

Model	Class	Precision	Recall	F1-score
BERT	N	51.67	57.88	54.97
	P	98.03	98.35	98.82
DistilBERT	N	67.63	80.17	73.84
	P	99.49	98.55	99.57
RoBERTa	N	67.02	76.08	71.30
	P	99.42	98.72	99.46
ELECTRA	N	75.58	71.54	73.25
	P	99.72	99.79	99.86
GPT-2	N	60.35	48.56	54.09
	P	98.75	99.72	98.44

N: Negative, P: Positive

Table 7 summarizes the performance of five models—BERT, DistilBERT, RoBERTa, ELECTRA, and GPT-2—on the BHRC dataset. The evaluation considers precision, recall, and F1-score, providing detailed insights into the models' ability to perform sentiment classification on this dataset. A notable pattern emerges in the results for the negative class, for which all three metrics indicate suboptimal performance. This underperformance may be partly attributed to the substantial class imbalance in the original BHRC dataset, in which negative reviews are considerably underrepresented relative to positive reviews. The limited number and diversity of negative samples may reduce the models' ability to learn representative features for this class, resulting in lower precision, recall, and F1-score values.

The results outlined in Table 7 underscore the pivotal influence of a balanced dataset on precision, recall, and F1-score performance metrics across diverse models. Significantly, this comparative analysis explicitly illustrates a substantial enhancement in the performance metrics when models undergo evaluation on a balanced dataset. Each model exhibits markedly improved scores in precision, recall, and F1-score when both training and testing occur on a dataset characterized by a balanced distribution between classes. In Table 8, it is clearly observed that ELECTRA exhibits superior performance compared to the other transformer models, both for negative and positive predictions.

Table 8. Comparison of Evaluation Metrics for Balanced BHRC Dataset

Model	Class	Precision	Recall	F1-score
BERT	N	97.61	98.14	97.24
	P	98.28	96.33	97.36
DistilBERT	N	99.58	100	99.49
	P	100	99.62	99.30
RoBERTa	N	99.82	100	99.46
	P	100	99.73	99.51
ELECTRA	N	99.85	99.96	99.91
	P	99.94	99.89	99.82
GPT-2	N	98.71	99.56	99.67
	P	99.65	98.87	99.75

N: Negative, P: Positive

To validate the performance of the proposed

framework, experiments are carried out on widely used datasets in the literature namely MR, SST-1, SST-2, and IMDB. Table 9 enumerates the classification accuracies achieved by the models across these diverse datasets. When the results of pre-trained models on the imbalanced and balanced BHRC datasets are examined, ELECTRA performs better than other models with an accuracy of 98.21% and 99.96%, respectively. ELECTRA achieves an accuracy gain of 2.07% over the poorly performing BERT. BERT exhibits comparatively inferior performance when juxtaposed with other models, with its notable exception being the SST-2 dataset. In this specific dataset, BERT attains a higher accuracy score of 98.31% compared to the GPT-2 model.

Table 9. The accuracy results of the proposed models across various datasets

Model	BHRC	Balanced BHRC	MR	SST-1	SST-2	IMDb
BERT	96.14	98.62	74.13	40.31	98.31	77.91
DistilBERT	97.98	99.70	84.75	53.17	98.42	85.24
RoBERTa	97.53	99.88	88.56	50.41	98.33	83.66
ELECTRA	98.21	99.96	89.27	57.46	99.07	87.46
GPT-2	98.03	99.85	83.40	50.24	89.15	82.37

Table 10 contains the comparison of the proposed transformer models with state-of-the-art studies which are concentrated on different deep learning methods. It is clearly seen from the experiments that relatively lower accuracy values are obtained with BERT in MR and SST-1 datasets. In the SST-2 dataset, all models exhibit significant performance compared to basic methods and models with joint architecture. In the IMDb dataset, the

models give remarkable results. ELECTRA achieves superior accuracy values in all studies except the IMDb dataset. To sum up, ELECTRA model outperforms literature studies by achieving 89.27% accuracy for MR dataset, 57.46% accuracy for SST-1 dataset, 99.07% accuracy for SST-2 dataset, and 87.46% accuracy for IMDb dataset.

Table 10. Comparison of the proposed models with deep learning methods mentioned in the literature

Model	Year	MR	SST-1	SST-2	IMDb
DCNN ²⁸	2014	-	48.5	86.8	-
CNN - non-static ²⁹	2014	81.5	48.0	87.2	-
CNN - multi-channel ²⁹	2014	81.1	47.4	88.1	-
LSTM ³⁰	2015	-	46.4	84.9	-
Bidirectional LSTM ³⁰	2015	-	49.1	87.5	-
Constituency Tree-LSTM ³⁰	2015	-	51.0	88.0	-
CNN-GRU-word2vec ³¹	2016	82.28	50.68	89.95	-
Non-static GloVe+word2vec CNN ³²	2016	81.24	47.08	85.49	-
ConvLSTM ³³	2017	-	47.5	88.3	-
BiLSTM-CRF ³⁴	2018	82.3	48.5	88.3	-
CNN + LSTM ³⁵	2018	-	48.8	89.2	93.2
AGCNN-SELU-3-channel ³⁶	2018	81.3	49.4	87.6	-
AGCNN-NLReLU-3-channel ³⁶	2018	81.9	49.4	87.4	-
Memristive LSTM ³⁷	2019	-	-	-	84.30

Table 10. (Cont.) Comparison of the proposed models with deep learning methods mentioned in the literature

Model	Year	MR	SST-1	SST-2	IMDb
ATTPooling RNN-pre ³⁸	2020	83.64	51.14	89.62	-
Word2Vec + CNN ³⁹	2020	-	-	-	84.30
TF-IDF + CNN ⁴⁰	2020	-	-	-	80.62
Word Embedding + DNN ⁴⁰	2020	-	-	-	80.25
BERT	2026	74.13	40.31	98.31	77.91
DistilBERT	2026	84.75	53.17	98.42	85.24
RoBERTa	2026	88.56	50.41	98.33	83.66
ELECTRA	2026	89.27	57.46	99.07	87.46
GPT-2	2026	83.40	50.24	89.15	82.37

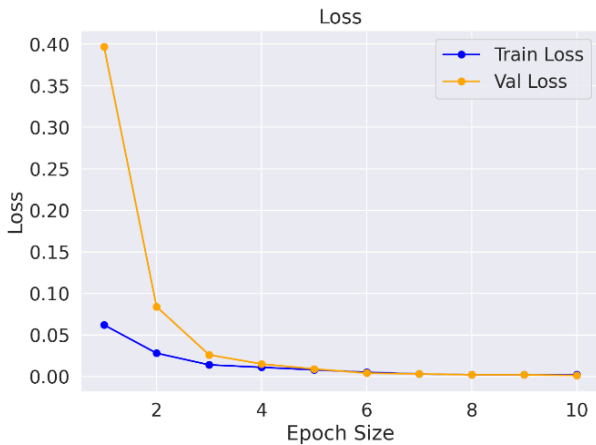


Figure 3. The loss graphic of ELECTRA model for balanced BHRC dataset

In Fig. 3, the training and validation loss curves for the balanced BHRC dataset are presented across epochs. Both curves exhibit similar trends, with loss rates decreasing as the epoch size increases. In the graph, it is observed that the curves of the training and validation datasets move in parallel, and after the 4th epoch, the graphs overlap. As clearly depicted in the graph, there is no evidence of overfitting issues during the training and validation phases.

The Receiver Operating Characteristic (ROC) curve functions as a metric for assessing the efficacy of a classification model. The degree of steepness exhibited by the curve is indicative of the model's proficiency in classification and its error rate. A more pronounced ROC curve generally denotes a well-performing model. In such instances, the model is likely to manifest high accuracy, a low error rate, and adept classification capabilities, suggesting that its predictions offer a more distinct separation between the true classes. Conversely, a less steep ROC curve may suggest diminished classification performance or an increased incidence of misclassifications in the confusion matrix. The steepness of the ROC curve can be contingent upon diverse factors, including the model's employed features, the training

dataset, the model's complexity, and the classification algorithm applied. The optimization or modification of these factors has the potential to impact the steepness of the ROC curve.

In summary, a sharper ROC curve signifies that the model exhibits superior classification performance and imparts a clearer differentiation between its predictions and the true classes. Rather than concentrating exclusively on prediction rates, investigators have utilized the Area under the Curve (AUC) as a metric, as depicted in Figure 4. The ROC curve was used to evaluate the models' discriminative performance across different classification thresholds. By illustrating the relationship between the true positive rate and the false positive rate, the ROC curve provides a threshold-independent assessment of each model's ability to distinguish between positive and negative classes. However, the ROC curve and the corresponding area under the curve (AUC) should be interpreted together with other performance metrics, such as precision, recall, and F1-score, particularly in the presence of class imbalance. Among the evaluated models, ELECTRA achieved the highest AUC value of 0.99, as shown in Figure 4.

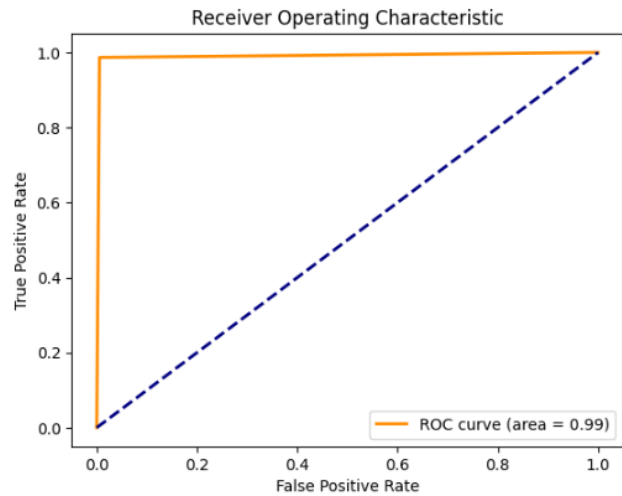


Figure 4. The ROC curve of ELECTRA model for balanced BHRC dataset

6. Conclusion

This study systematically evaluated the performance of five state-of-the-art transformer-based language models—BERT, DistilBERT, RoBERTa, ELECTRA, and GPT-2—for customer experience analysis in the hospitality domain. Experiments were conducted using the proposed Bali Hotel Review Corpus (BHRC), compiled from TripAdvisor and Hotels.com reviews, in conjunction with four widely used benchmark datasets, enabling a robust and comparative assessment of contemporary NLP architecture across both domain-specific and general sentiment classification settings.

The experimental results support several important conclusions. First, transformer-based models demonstrated strong effectiveness in capturing fine-grained sentiment patterns in hotel reviews, achieving accuracy levels above 98% on balanced datasets. Second, the application of a targeted data augmentation strategy based on ChatGPT-generated paraphrasing substantially mitigated class imbalance, leading to notable improvements in F1-scores for the minority (negative) class. These findings highlight the critical role of data-centric strategies in enhancing model robustness, even when employing advanced transformer architectures. Finally, the introduction of the BHRC dataset represents a meaningful contribution to the literature, offering a valuable resource for future research on domain-adapted sentiment analysis in tourism and hospitality contexts.

Overall, the results indicate that transformer-based frameworks provide reliable and scalable solutions for extracting actionable insights from large volumes of customer-generated textual feedback. As such, these models hold practical relevance for hospitality managers and analysts seeking to support evidence-based decision-making and continuous service quality improvement.

7. Discussion

The ELECTRA model demonstrated the highest performance in this study, reaching an accuracy of 99.96% on the balanced BHRC dataset. This result can be primarily attributed to ELECTRA's Replaced Token Detection (RTD) objective, which differs fundamentally from the Masked Language Modeling (MLM) strategy employed by BERT and its variants. By framing pretraining as a discriminative task, ELECTRA leverages information from all input tokens rather than a masked subset, resulting in improved sample efficiency and stronger contextual sensitivity. Consistent with the findings of Clark et al.⁴¹, these results suggest that, despite the growing prominence of large generative language models,

discriminative transformer encoders remain highly effective and computationally efficient for domain-specific classification problems.

The comparative analysis with conventional deep learning architectures, including CNN and LSTM models, further confirms the advantages of transformer-based approaches. Across benchmark datasets such as SST-2 and IMDb, the proposed framework consistently outperformed traditional models, indicating that attention mechanisms are better suited for capturing long-range semantic dependencies in opinion-rich and linguistically complex texts. From an application perspective, the strong performance of compact transformer models is particularly relevant for the hospitality sector, where scalability and real-time inference over large volumes of user-generated content are critical operational requirements.

An additional contribution of this study is the empirical demonstration of the impact of data balancing on model performance. The increase in BERT accuracy from 96.14% on the imbalanced BHRC dataset to 98.62% on the balanced version highlights that even advanced transformer architectures remain sensitive to data distribution and class imbalance. This finding reinforces the importance of careful dataset construction and pre-processing in mitigating bias and improving generalization.

Despite the high accuracy levels achieved, several limitations should be acknowledged. In particular, data augmentation techniques may introduce unintended semantic artifacts that affect model reliability, an issue that warrants further systematic investigation. Moreover, future work may explore the integration of multimodal information, such as numerical ratings or visual content, to enable a more comprehensive representation of customer experience. In line with Ameer et al.⁴², the results indicate that future research in tourism sentiment analysis will benefit from combining efficient transformer architectures with robust, data-centric enhancement strategies.

Declaration of Ethical Standards

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Fu JR, Lu IW, Chen JHF, Farn CK. Investigating consumers' online social shopping intention: an information processing perspective. *International Journal of Information Management*. 2020;54:102189. Available from: <https://doi.org/10.1016/j.ijinfomgt.2020.102189>
- [2] Tang L, Wang X, Kim E. Predicting conversion rates in online hotel bookings with customer reviews. *Journal of Theoretical and Applied Electronic Commerce Research*. 2022;17(4):1264-1278. Available from: <https://doi.org/10.3390/jtaer17040064>
- [3] Güler YC, Huseynov F. The impact of online consumer reviews on online hotel booking intention. *Journal of Business Research-Turk*. 2021;13(3):2634-2652. Available from: <https://doi.org/10.20491/isarder.2021.1282>
- [4] Touni R, Kim WG, Haldorai K, Rady A. Customer engagement and hotel booking intention: the mediating and moderating roles of customer-perceived value and brand reputation. *International Journal of Hospitality Management*. 2022;104:103246. Available from: <https://doi.org/10.1016/j.ijhm.2022.103246>
- [5] Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*. 2022;55(7):5731-5780. Available from: <https://doi.org/10.1007/s10462-022-10144-1>
- [6] Aakash A, Gupta Aggarwal A. Assessment of hotel performance and guest satisfaction through eWOM: big data for better insights. *International Journal of Hospitality & Tourism Administration*. 2022;23(2):317-346. Available from: <https://doi.org/10.1080/15256480.2020.1746218>
- [7] Sann R, Lai PC. Understanding homophily of service failure within the hotel guest cycle: applying NLP-aspect-based sentiment analysis to the hospitality industry. *International Journal of Hospitality Management*. 2020;91:102678. Available from: <https://doi.org/10.1016/j.ijhm.2020.102678>
- [8] Obiedat R, Qaddoura R, Al-Zoubi AM, Al-Qaisi L, Harfoushi O, Alrefai M, Faris H. Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution. *IEEE Access*. 2022;10:22260-22273. Available from: <https://doi.org/10.1109/ACCESS.2022.3149482>
- [9] Anis S, Saad S, Aref M. Sentiment analysis of hotel reviews using machine learning techniques. In: Hassanien AE, Slowik A, Snášel V, El-Deeb H, Tolba FM, editors. *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020*. Cham: Springer; 2021. p. 227-234. Available from: https://doi.org/10.1007/978-3-030-58669-0_21
- [10] Noori B. Classification of customer reviews using machine learning algorithms. *Applied Artificial Intelligence*. 2021;35(8):567-588. Available from: <https://doi.org/10.1080/08839514.2021.1922843>
- [11] Tiffani IE. Optimization of naïve Bayes classifier by implemented unigram, bigram, trigram for sentiment analysis of hotel review. *Journal of Soft Computing Exploration*. 2020;1(1):1-7. Available from: <https://doi.org/10.52465/joscex.v1i1.4>
- [12] Sayed AA, Elgeldawi E, Zaki AM, Galal AR. Sentiment analysis for Arabic reviews using machine learning classification algorithms. 2020 *International Conference on Innovative Trends in Communication and Computer Engineering*; 2020 Feb 8-9; Aswan, Egypt. p. 56-63. Available from: <https://doi.org/10.1109/ITCE48509.2020.9047822>
- [13] Gunathilaka S, De Silva N. Aspect-based sentiment analysis on mobile application reviews. 22nd *International Conference on Advances in ICT for Emerging Regions (ICTer)*; 2022 Nov 2-3; Colombo, Sri Lanka. p. 183-188. Available from: <https://doi.org/10.1109/ICTer58063.2022.10024070>
- [14] Zhou G, Liao C. Dynamic measurement and evaluation of hotel customer satisfaction through sentiment analysis on online reviews. *Journal of Organizational and End User Computing*. 2021;33(6):1-27. Available from: <https://doi.org/10.4018/JOEUC.20211101.0a8>
- [15] Ishaq A, Umer M, Mushtaq MF, Medaglia C, Siddiqui HUR, Mehmood A, Choi GS. Extensive hotel reviews classification using long short-term memory. *Journal of Ambient Intelligence and Humanized Computing*. 2021;12(10):9375-9385.

- Available from: <https://doi.org/10.1007/s12652-020-02654-z>
- [16] Hassan J, Shoaib U. Multi-class review rating classification using deep recurrent neural network. *Neural Processing Letters*. 2020;51(2):1031-1048. Available from: <https://doi.org/10.1007/s11063-019-10125-6>
- [17] Puh K, Bagić Babac M. Predicting sentiment and rating of tourist reviews using machine learning. *Journal of Hospitality and Tourism Insights*. 2023;6(3):1188-1204. Available from: <https://doi.org/10.1108/JHTI-02-2022-0078>
- [18] Khedkar, S., Shinde, S. Deep learning and ensemble approach for praise or complaint classification. *Procedia Computer Science* 2020, 167, 449-458.
- [19] Ray B, Garain A, Sarkar R. An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing*. 2021;98:106935. Available from: <https://doi.org/10.1016/j.asoc.2020.106935>
- [20] Kotei E, Thirunavukarasu R. A systematic review of transformer-based pre-trained language models through self-supervised learning. *Information*. 2023;14(3):187. Available from: <https://doi.org/10.3390/info14030187>
- [21] Acheampong FA, Nunoo-Mensah H, Chen W. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*. 2021;54(8):5789-5829. Available from: <https://doi.org/10.1007/s10462-021-09958-2>
- [22] Greco CM, Simeri A, Tagarelli A, Zumpano E. Transformer-based language models for mental health issues: a survey. *Pattern Recognition Letters*. 2023;167:204-211. Available from: <https://doi.org/10.1016/j.patrec.2023.02.016>
- [23] Vakili YZ, Fallah A, Sajedi H. Distilled BERT model in natural language processing. 14th International Conference on Computer and Knowledge Engineering (ICCKE); 2024 Nov 19-20; Mashhad, Iran. p. 243-250. Available from: <https://doi.org/10.1109/ICCKE65377.2024.10874673>
- [24] Areshey AM, Mathkour H. Exploring transformer models for sentiment classification: a comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet. *Expert Systems*. 2024;41(11):e13701. Available from: <https://doi.org/10.1111/exsy.13701>
- [25] Dong C, Liu L, Cheng H, Shang J, Gao J, Liu X. Fast-ELECTRA for efficient pre-training. 12th International Conference on Learning Representations (ICLR); 2024 May 7-11; Vienna, Austria.
- [26] Yenduri G, Ramalingam M, Selvi GC, Supriya Y, Srivastava G, Maddikunta PKR, Raj DG, Jhaveri RH, Prabadevi B, Wang W, Vasilakos AV, Gadekallu TR. GPT (Generative Pre-Trained Transformer): a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*. 2024;12:54608-54649. Available from: <https://doi.org/10.1109/ACCESS.2024.3389497>
- [27] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Published 2019. Accessed 2026 Jul 17
- [28] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. 52nd Annual Meeting of the Association for Computational Linguistics; 2014 Jun 22-27; Baltimore, Maryland, USA. p. 655-665. Available from: <https://doi.org/10.3115/v1/P14-1062>
- [29] Kim Y. Convolutional neural networks for sentence classification. 2014 Conference on Empirical Methods in Natural Language Processing; 2014 Oct 25-29; Doha, Qatar. p. 1746-1751. Available from: <https://doi.org/10.3115/v1/D14-1181>
- [30] Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015 Jul 26-31; Beijing, China. p. 1556-1566. Available from: <https://doi.org/10.3115/v1/P15-1150>
- [31] Wang X, Jiang W, Luo Z. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. *Proceedings of COLING 2016, the 26th International Conference on*

- Computational Linguistics; 2016 Dec 11-16; Osaka, Japan. p. 2428-2437. Available from: <https://aclanthology.org/C16-1229/>
- [32] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. 8th International Joint Conference on Natural Language Processing; 2017 Nov 27-Dec 1; Taipei, Taiwan. p. 253-263. Available from: <https://aclanthology.org/I17-1026/>
- [33] Hassan A, Mahmood A. Deep learning approach for sentiment analysis of short texts. 3rd International Conference on Control, Automation and Robotics (ICCAR); 2017 Apr 24-26; Nagoya, Japan. p. 705-710. Available from: <https://doi.org/10.1109/ICCAR.2017.7942788>
- [34] Chen T, Xu R, He Y, Wang X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications. 2017;72:221-230. Available from: <https://doi.org/10.1016/j.eswa.2016.10.065>
- [35] Hassan A, Mahmood A. Convolutional recurrent deep learning model for sentence classification. IEEE Access. 2018;6:13949-13957. Available from: <https://doi.org/10.1109/ACCESS.2018.2814818>
- [36] Liu Y, Ji L, Huang R, Ming T, Gao C, Zhang J. An attention-gated convolutional neural network for sentence classification. Intelligent Data Analysis. 2019;23(5):1091-1107. Available from: <https://doi.org/10.3233/IDA-184311>
- [37] Wen S, Wei H, Yang Y, Guo Z, Zeng Z, Huang T, Chen Y. Memristive LSTM network for sentiment analysis. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2021;51(3):1794-1804. Available from: <https://doi.org/10.1109/TSMC.2019.2906098>
- [38] Usama M, Ahmad B, Song E, Hossain MS, Alrashoud M, Muhammad G. Attention-based sentiment analysis using convolutional and recurrent neural network. Future Generation Computer Systems. 2020;113:571-578. Available from: <https://doi.org/10.1016/j.future.2020.07.022>
- [39] Tammina S, Annareddy S. Sentiment analysis on customer reviews using convolutional neural network. International Conference on Computer Communication and Informatics (ICCCI); 2020 Jan 22-24; Coimbatore, India. p. 1-6. Available from: <https://doi.org/10.1109/ICCCI48352.2020.9104086>
- [40] Dang NC, Moreno-García MN, De la Prieta F. Sentiment analysis based on deep learning: a comparative study. Electronics. 2020;9(3):483. Available from: <https://doi.org/10.3390/electronics9030483>
- [41] Cortiz D. Exploring transformer models for emotion recognition: a comparison of BERT, DistilBERT, RoBERTa, XLNet and ELECTRA. In: Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System (CCRIS '22); 2022 Aug 26-28; Virtual Event, China. New York (NY): Association for Computing Machinery; 2022. p. 1-7. Available from: <https://doi.org/10.1145/3562007.3562051>
- [42] Ameer A, Hamdi S, Ben Yahia S. Sentiment analysis for hotel reviews: a systematic literature review. ACM Computing Surveys. 2023;56(2):Article 51, 1-38. Available from: <https://doi.org/10.1145/3605152>