A Sentiment Analysis of the Tourist Reviews of the Attractions in Saudi Arabia Using Deep Learning Models

Raneem Alharbi
Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Makkah Region, Saudi Arabia
raneem.alharbi@outlook.sa (corresponding author)

Areej Alshutayri
Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Makkah Region, Saudi Arabia
aoalshutayri@uj.edu.sa

Shahd Alahdal
Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Makkah Region, Saudi Arabia
saalahdal@uj.edu.sa

Received: 15 June 2025 | Revised: 18 September 2025, 9 November 2025, 14 December 2025, and 17 December 2025 | Accepted: 18 December 2025

 DOI: https://doi.org/10.48084/etasr.12750

ABSTRACT

This paper aims to enhance the standards of tourism and tourist experiences in Saudi Arabia, thereby contributing to the achievement of Saudi Vision 2030 goals. One of these goals is to stimulate economic growth by increasing tourism-related commerce, as Saudi Arabia is becoming an increasingly popular tourist destination. This study employs sentiment analysis of tourist experiences in places recognized as popular attractions by the Saudi Ministry of Tourism. Arabic language tourist reviews of Boulevard Riyadh City, Al-Ula Old Town, the Al-Balad district in Jeddah, the Heritage Village in Dammam, and the Al-Hada cable car from Google Maps are collected, and the sentiment of the reviews is classified as positive, negative, or neutral. The textual representation techniques employed in the model word embedding methods, such as AraVec, ArBERT, Qarib, and MARBERT. Moreover, various models, including baseline models, Deep Learning (DL) models, and transformer-based models, were implemented to predict sentiment. Various metrics, including accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC), were used to evaluate the models' performance. The results demonstrated that QARiB achieved the best performance.

Keywords-sentiment analysis; CNN; deep learning; transformer-based classification

I. INTRODUCTION

Technology in all its forms has developed, making it easier for people to share their thoughts and experiences on social media platforms. Consequently, an enormous amount of data has been collected, with approximately 64.2 ZB of data as of 2021 [1]. These data are highly valuable to organizations, businesses, communities, and countries. As people use text, video, and images to openly share their experiences [2], customer feedback has become highly important in industries such as tourism. Tourists choose their destinations based on their own experiences, personal preferences, and other personal variables, such as the purpose of travel, health and safety concerns, and time constraints. As reviews enable them to choose the best places to visit, the government is aiming to enhance its services and identify the most popular tourist destinations [3, 4].

According to [5], Arabic is one of the main languages used on social media platforms. Based on statistics, Arabic is the fourth most utilized language on social media [6], following Chinese, Spanish, and English (Internet World Statistics). Although there are various difficulties, such as content regulation, language attrition, and the influence of new communication methods, it has been indicated that Arabic is widely used on social media platforms, with different dialects

being utilized based on age and culture [7, 8]. However, the increasing use of informal Arabic and the tendency to combine Arabic with foreign phrases have added to the decline of the Arabic language, a result of social media use [8].

The Arabic language is complex compared to other languages, as there are different dialects in different regions and countries. In addition, the roots of Arabic words change the context of sentences, and punctuation impacts the meaning of words [9]. The complexity of Arabic dialects is a significant challenge for automatic translation using artificial intelligence [7, 8]. Two basic concepts characterize Arabic morphology: word formation (derivational morphology) and word interaction with syntax (inflectional morphology) [10]. With approximately 300,000 part-of-speech tags in Modern Standard Arabic (MSA) and an average of 12 morphological analyses per word, Arabic is morphologically rich but also very ambiguous [11]. Arabic contains many irregular forms and intricate morpho-syntactic agreement requirements, while more than half of its plurals are irregular [10]. Arabic training datasets are widely available in audio, text, and image formats and are essential for Machine Learning (ML) and Natural Language Processing (NLP) in Arabic [12].

Sentiment analysis, also known as opinion mining, is an important research tool that uses NLP, text analysis, and computational linguistics to extract subjective information from text data. It makes textual data valuable and meaningful by analyzing them to determine people's impressions based on their writing. It offers numerous benefits to businesses, including improved business intelligence, customer satisfaction analysis, and real-time problem detection [13]. By applying the same criteria to all data, businesses can gain better insights and address customer issues in real time, especially on social media platforms where customers can express their discontent. Overall, sentiment analysis provides valuable insights for product improvement and market competitiveness. Sentiment classification enables authorities, such as governments and businesses, to understand people's opinions regarding their experiences, including feedback and recommendations. The classification procedure examines the written text and the polarity of the sentiments (positive, negative, or neutral) to determine the writer's impression.

There has been a growing interest among researchers in the tourism field focused on Saudi Arabia to examine and analyze tourists' experiences and opinions. Hotel reviews, social media, and tweets are data sources utilized in these studies. Different studies have collected data from platforms, such as Twitter, Google Maps, Booking.com, Tripadvisor.com, Instagram, and Snapchat, and implemented baseline and DL models to conduct sentiment analysis. Sentiment analysis provides valuable knowledge into various sectors by understanding tourists' experiences and evaluating products and services. All studies agree that sentiment analysis is essential for understanding public opinion and developing effective strategies. Therefore, sentiment analysis is essential for businesses, governments, and other organizations to make informed decisions.

Authors in [14] employed DL models for sentiment classification using a dataset collected from Google Maps reviews of tourist attractions across 14 cities in Saudi Arabia. However, their study remains limited to gathering a dataset of Saudi dialects. They used the Lexicon-based method to classify reviews into positive, negative, and neutral. Moreover, their study involved the use of three classifiers: Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Recurrent Neural Networks (RNNs). The results show that the SVM classifier achieved an accuracy of 98%, while the other classifier achieved a similar accuracy of 96%.

Authors in [15] used Google Maps reviews along with the SVM model to investigate the effect of celebrities on people's decisions to choose restaurants in Riyadh. The workflow of this study was as follows: First, the authors searched for famous restaurants by making a questionnaire, revealing that most of them were promoted by Snapchat celebrities. Then, they selected the 30 most frequent restaurants, and people's reviews of these restaurants were collected from Google Maps. A lexicon-based technique was then used to detect phrases related to food opinions in the Saudi dialect, with positive or negative polarity. The results of the sentiment analysis were compared with a previously conducted questionnaire, and it was found that most celebrity restaurants have adverse effects on customers.

In contrast, authors in [16] collected Twitter tweets and filtered them based on keywords related to Saudi Arabia's tourism. A sentiment analysis was performed on this dataset using ML and big data algorithms. The AraSenti corpus was utilized as labeled data to train ML models for sentiment analysis. This dataset facilitated the development and evaluation of models by providing pre-annotated examples that capture sentiment expressions in Arabic text. Various ML models were employed, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Naïve Bayes (NB), with LR achieving the highest performance at 86%. To analyze the most popular tourist cities in Saudi Arabia, the authors searched for Saudi city names (50 words were used) and then extracted the most frequent ones in a dataset using the "normalize location" local function. Their findings suggest that Riyadh, Alula, Hail, Taif, and Tabuk were the most popular cities in Saudi Arabia.

Authors in [17] investigated tourist reviews to explore the experience of the first tourist cruise trip in Saudi Arabia. The dataset was collected from social media platforms, including Instagram, Snapchat, and Twitter, suggesting that Instagram had the highest interaction rate. The research employed sentiment analysis utilizing various ML models, including Multilayer Perceptron (MLP), NB, RF, SVM, and voting. For the Snapchat samples, the RF model achieved an accuracy rate of 100%, representing overfitting, a situation in which the model succeeds in training data but struggles to generalize to new and untested data. Moreover, DL models were also used in [18], presenting an Adaptive Particle Grey Wolf Optimizer with Deep Learning Sentiment Analysis (APGWO-DLSA) method, which uses a Deep Belief Network (DBN) model and an APGWO algorithm for sentiment classification. This method achieved an accuracy of 94.77% and 85.31% by utilizing APGWO to tune the hyperparameters.

Authors in [19] analyzed Arabic hotel reviews using unsupervised ML models, employing a feature-based sentiment

analysis method for online Arabic reviews. The study extracted features and classified reviews as neutral, negative, or positive by integrating ML models with NLP. Arabic reviews for Saudi hotels were collected from the TripAdvisor website and then subjected to Term Frequency-Inverse Document Frequency (TF-IDF) to extract key features. Unsupervised learning methods, including K-means and hierarchical algorithms with two distance metrics (cosine and Euclidean), were employed, while it was found that preprocessing steps were required to reduce ambiguity in Arabic. In the preprocessed dataset, K-means clustering achieved the highest accuracy, with a purity of 75%.

II. METHODOLOGY

The dataset used in this study was collected from Google Maps, followed by data annotation and preprocessing. The proposed methodology for Arabic sentiment analysis is shown in Figure 1.

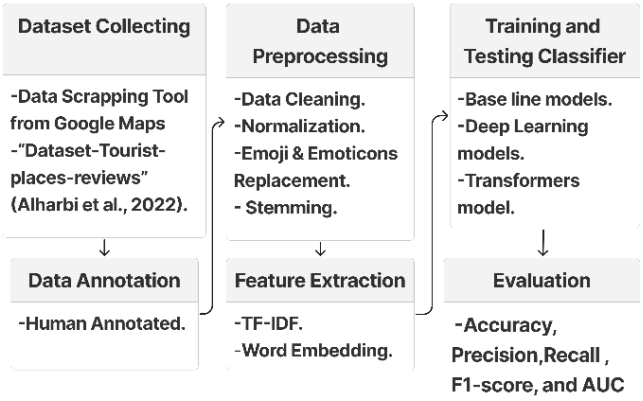

Fig. 1. Proposed methodology for Arabic sentiment analysis.

A. Dataset

The present study uses the Arabic tourist reviews dataset [20] collected in [14] from Google Maps. The former is based on five major Saudi tourist attractions visited by both local and international tourists. These attractions are Boulevard Riyadh City, Al-Ula Old Town, the Al-Balad district in Jeddah, the Heritage Village in Dammam, and the Al-Hada cable car. An initial set of 3,772 raw reviews was manually extracted from these locations. The collected user reviews reflect real experiences and contain both MSA and multiple regional dialects.

During the initial inspection of the dataset, an imbalance between sentiment classes was observed. To reduce this imbalance and improve model learning, additional reviews were collected using the Instant Data Scraper tool [21]. These additional records were merged with the original dataset, resulting in a new dataset [22] containing a total of 3,895 reviews. At this stage, the dataset contained four main attributes: the raw extracted review text (div_text), the city name, the place name, and, later, the sentiment label. Table I displays representative samples from the collected dataset before annotation.

TABLE I. SAMPLE OF A DATASET USED FOR SENTIMENT ANALYSIS

div_text	City	Place
ممتع	الطايف	ٍلفريك الهدا
الأسعار منخفضه وانصح كل من أراد التسوق	الدمام	القرية الشعبية
حلو ويغلب عليه الكافيهات مكان رايق وهادي	الرياض	بوليفارد الرياض
لا بأس	العلا	بلدة العلا القديمة
ذكرى جميلة	الدمام	القرية الشعبية

B. Data Annotation

The dataset was manually annotated by three volunteer native Arabic speakers who were assigned to the annotation task and provided with clear labeling guidelines, as described in [23]. The reviews were categorized into three sentiment classes: positive, negative, and neutral. In cases of disagreement, the final sentiment label was determined by majority voting (two out of three annotators). Table II presents sample reviews and their final sentiment labels after annotation, where A_1 represents annotator 1, A_2 represents 2, A_3 represents 3, and the final label after comparison. Table III presents several examples of review annotations and categorizes tourist reviewers' opinions. The labels are "1" for positive, "0" for negative, and "2" for neutral. After labelling, the dataset had 1510 positive reviews, 1269 negative reviews, and 1115 neutral reviews.

TABLE II. INTER-ANNOTATOR AGREEMENT AND FINAL SENTIMENT LABELS

Sentence in Arabic	Translated into English	A_1	A_2	A_3	Final label
" المرافق الخاصة بالتلفريك نظيفة."	"The cable car facilities are clean."	2	1	1	1
" للأسف المكان مهجور ومهمل"	"Unfortunately, the place is abandoned and neglected."	0	1	0	0
" قريةٍ تكلم عن الماضي وعن حياه الناس سابقا"	"A village that tells the story of the past and how people used to live."	2	2	0	2

TABLE III. SAMPLE ANNOTATED REVIEWS FOR EACH SENTIMENT CLASS

Class	Sentence in Arabic	Translated into English
Positive	" المرافق الخاصة بالتلفريك نظيفة."	"The cable car facilities are clean."
Negative	" أفشل مجمع بالعالم الضيق والزحمة"	"The worst mall in the world cramped and crowded."
Neutral	" قريةٍ تكلم عن الماضي وعن حياه الناس سابقا"	"A village that tells the story of the past and how people used to live."

C. Dataset Preprocessing

The data were cleaned to remove influences that may lead to reading them incorrectly. This data cleaning process includes many steps, as presented in Figure 2. The former involves deleting problematic content, such as stop words. For example, prepositions, conjunctions, pronouns, and other functional words, such as "في" (in), "و" (and), "كما" (as), "الذي" (which), and "هذا" (this), non-Arabic words, numbers, punctuation, URLs, small sentences, extra spaces, hashtags, mentions, and diacritics. Normalization was used to standardize the word characters by replacing characters with others (e.g., replacing

(أ,آ,إ) with (ا), ((ي) with (ى)), replacing duplicated characters with two characters, and replacing emojis and emoticons with their corresponding semantic representations. Stemming was used to obtain the base or root word by removing prefixes and suffixes, e.g., the word "الأماكن", which means "places", will be "مكان", "place". Also, duplicated reviews were checked, and 186 duplicated reviews were deleted. The reviews after preprocessing are presented on column "text," as depicted in Table IV. After preprocessing, the final dataset contained 3,249 reviews, of which 1,196 were positive, 1,028 were negative, and 1,025 were neutral.

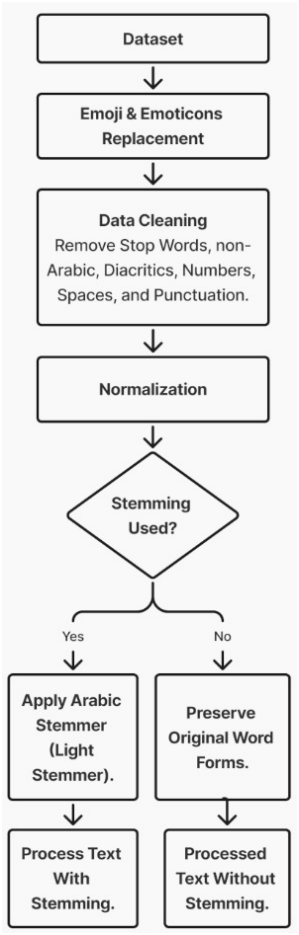

Fig. 2. Data preprocessing flowchart.

TABLE IV. OVERVIEW OF THE DATASET AFTER PREPROCESSING

div_text	Label	City	Place	Text
□دفع دخولية 70ريال للشخص للدخول وقبل الطلب على شان في متحف المدينة	0	الدمام	القرية الشعبية	□دفع دخولية للشخص الدخول وقبل الطلب شان متحف المدينة
عبق الماضي ووجد□ه في هذا المكان	2	الدمام	القرية الشعبية	عبق الماضي وجد□ه المكان
الحين□حت الصيانه□أمنى أن يتم إصلاحه في اقرب وقت	2	الدمام	القرية الشعبية	الحين الصيانه□أمنى يتم إصلاحه اقرب وقت

D. Feature Extraction

The present study adopts two different text representation strategies, depending on the model type. First, the ML classifiers rely on TF–IDF features to represent the preprocessed reviews [24]. In contrast, DL and transformer-based models exploit word embeddings to capture semantic information at different levels. The models were applied in a zero-fine-tuning setting, where the pre-trained parameters remained unchanged throughout the experiment.

1) Machine Learning

In TF–IDF, TF represents the number of times a word is used in a text, and IDF measures the rarity of words within a set of texts [24]. This method represents text numerically to show the relative importance of each word. It is designed to reduce the weight of words that do not have meaning (such as "the" or "and"). The TfidfVectorizer is a tool in the scikit-learn library that automatically splits texts into words, removes extraneous tokens (such as punctuation), and converts texts to numerical form when passed to the model.

2) Deep Learning Models

The following DL models and word embedding techniques were employed for Arabic sentiment analysis:

- AraVec is an open-source, pre-trained Arabic word embedding model [25], which uses a Continuous Bag-of-Words model, a neural network-based model with 300-dimensional word embeddings. Its training dataset includes text from Arabic Wikipedia, tweets, and World Wide Web pages. AraVec has almost 3,300,000,000 tokens, normalization, and non-Arabic content filtering.
- The MARBERT model [26] has been pre-trained on a masked language model trained on 1 billion tweets, containing text in both MSA and various Arabic dialects. The text equates to 128 gigabytes, with 15.6 billion tokens and 160 million trainable parameters in the corpora.
- The ArBERT model is proposed in [26], pre-trained on text corpora from news, books, and Wikipedia. It has 6.2 billion tokens, totaling about 61 gigabytes, and consists of over 163 million trainable parameters, 12 self-attention heads, 768 hidden units, and 12 layers of transformer blocks.
- QARiB model, introduced in [27], was trained on text from different Arabic Dialects, including informal Tweets and MSA, as well as news and movie/TV subtitles. Qarib was pre-trained on a corpus of 14 billion tokens, roughly 127 gigabytes of text.

III. RESULTS AND DISCUSSION

All baseline, DL, and transformer-based models were trained and evaluated using a unified experimental protocol to ensure fair comparison. The dataset was split into training, validation, and testing subsets. The validation set was used exclusively for selecting the best model during training, while the test set was used only for final performance reporting.

A. Baseline Model

The baseline models employed include SVM [28], LR [29], RF [30], and DT [31]. These models were deployed with default parameters from the scikit-learn package without hyperparameter optimization. TF-IDF feature extraction was also applied to all baseline models. Table V summarizes the hyperparameters used for the baseline model.

TABLE V. SUMMARY OF HYPERPARAMETERS FOR THE BASELINE MODEL

Model	Hyperparameters
LR	*C* = 1.0, solver = 'lbfgs', multi_class ='auto', max_iter = 1000
SVM	kernel = 'rbf', C=1.0
DT	criterion = 'gini', max_depth = None, min_samples_split = 2
RF	n_estimators = 100, max_depth = None, criterion = 'gini'

The baseline model's result comparison is presented in Tables VI and VII. It is observed that SVM with stemming achieved 84% accuracy, an F1-score of 84%, and an AUC of 0.94, outperforming all models. Furthermore, as shown in Figure 3, the LR and SVM produced better reliability and strong results, demonstrating the strength of basic linear models. When compared with LR and SVM, the RF displayed less efficiency at classification but demonstrated stable performance. Due to its tendency to overfit training data, the DT model performed poorly. ML models' overall performance dropped without stemming from the data, as illustrated in Tables VI and VII and Figure 3. They struggled to recognize similar patterns. With stemming, the models understood words more clearly, improving the results.

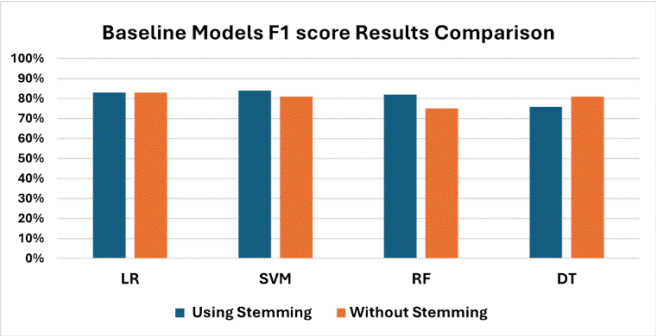

Fig. 3. Performance comparison of baseline models.

TABLE VI. PERFORMANCE OF BASELINE MODELS ON THE DATASET WITH STEMMING

Model	Accuracy	Precision	Recall	F1- score	AUC
LR	83%	83%	83%	83%	0.95
SVM	84%	84%	85%	84%	0.94
DT	76%	76%	76%	84%	0.82
RF	82%	82%	83%	82%	0.92

TABLE VII. PERFORMANCE COMPARISON OF BASELINE MODELS ON THE DATASET WITHOUT STEMMING

Model	Accuracy	Precision	Recall	F1- score	AUC
LR	83%	83%	83%	83%	0.94
SVM	81%	81%	81%	81%	0.93
DT	74%	75%	75%	75%	0.81
RF	80%	81%	81%	81%	0.90

B. Deep Learning Models

The study further employed DL models, such as Bidirectional LSTM (BiLSTM) [32], Convolutional Neural Networks (CNNs) [33], and RNNs [34]. All DL models were trained using the same hyperparameters under controlled conditions. The Adam optimizer was trained with CrossEntropyLoss. The AraVec model used static Word2Vec embeddings and a 0.001 learning rate over 20 epochs. However, contextual embeddings, like ArBERT, MARBERT, and QARiB, were frozen during training to maximize their pre-learned linguistic representations. These models were trained across 40 epochs at a 0.001 learning rate. A consistent batch size of 8 was used in all experiments, and after each run, the model achieving the highest validation accuracy was saved. Table VIII presents the complete configuration settings for the DL models.

However, there were significant differences in performance across different embedding models when comparing the outcomes of the DL models using stemming. As shown in Tables IX and X, ArBERT and MARBERT provided relatively stable representations, but MARBERT generally outperformed ArBERT for all models. With the QARiB and MARBERT models, CNN utilized its capacity to collect fixed characteristics to achieve maximum accuracy, resulting in the best performance, with 86% accuracy, 85% F1-score, and an AUC of 0.95. The RNN model yielded poor results with AraVec (36% accuracy, 29% F1-score, AUC 0.61) but performed well with QARiB and Marber, as portrayed in Figure 4.

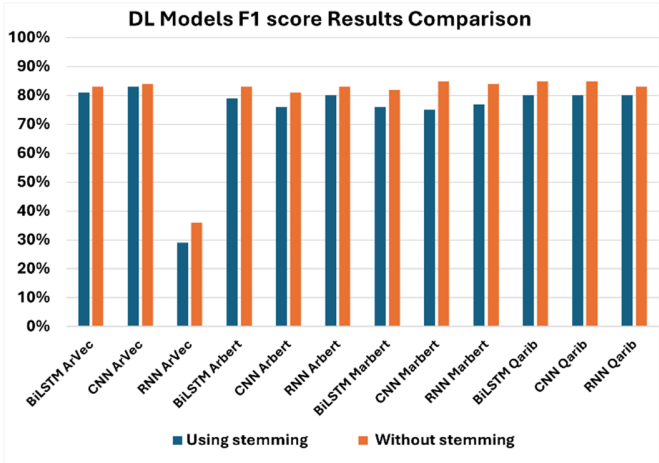

Fig. 4. Performance comparison of DL models.

In contrast, the BiLSTM model had excellent results with QARiB embedding. For embedding models, the QARiB model, which was pre-trained on various datasets and dynamically interpreted contextual cues within brief words, provided additional flexibility in performance. Conversely, the AraVec embedding model produced a wide range of results, indicating unstable performance.

Some DL models, especially CNNs and BiLSTM, improved performance when data stemming was not used, as

shown in Figure 4 and Tables IX and X. For example, BiLSTM accuracy increased from 86% to 89% with ArBERT embedding, and its F1-score increased from 80% to 85%. Also, the AUC of the models with MARBERT increased from 0.91 to 0.95. However, RNN models did not improve much in either case, and AraVec results were also poor. This suggests that some DL models benefit from preserving words in their original form, as they depend on the whole word sequence and meaning. Stemming is unnecessary for them, as with baseline. In summary, stemming affected DL models, especially the CNN model. Furthermore, QARiB's contextual awareness allowed it to perform well for all DL models with and without stemming. Overall, the results highlight the impact of stemming in achieving practical sentiment analysis.

TABLE VIII. HYPERPARAMETERS OF DL MODELS

Embedding	**AraVec**	**ArBERT**	**MARBERT**	**QARiB**
Model	CNN / RNN / BiLSTM	CNN / RNN / BiLSTM	CNN / RNN / BiLSTM	CNN / RNN / BiLSTM
Loss function	Cross entropy	Cross entropy	Cross entropy	Cross entropy
Optimizer	Adam	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001	0.001
Batch size	8	8	8	8
Epochs	20	40	40	40
Encoder setting	–	Frozen	Frozen	Frozen
Pooling	300	768	768	768
Classifier head	Conv / Seq / BiLSTM	Conv / Seq / BiLSTM	Conv / Seq / BiLSTM	Conv / Seq / BiLSTM

TABLE IX. PERFORMANCE COMPARISON OF DL MODELS ON THE DATASET WITH STEMMING

Embedding	**Model**	**Accuracy**	**Precision**	**Recall**	**F1- score**	**AUC**
AraVec	BiLSTM	81%	81%	81%	81%	0.93
	CNN	83%	83%	83%	83%	0.94
	RNN	36%	49%	38%	29%	0.61
ArBERT	BiLSTM	78%	78%	77%	79%	0.92
	CNN	77%	78%	77%	76%	0.91
	RNN	80%	80%	80%	80%	0.92
MARBER-T	BiLSTM	76%	78%	77%	76%	0.91
	CNN	76%	75%	75%	75%	0.91
	RNN	77%	77%	77%	77%	0.91
QARiB	BiLSTM	80%	80%	80%	80%	0.93
	CNN	81%	80%	80%	80%	0.93
	RNN	80%	80%	80%	80%	0.93

TABLE X. PERFORMANCE COMPARISON OF DL MODELS ON DATASET WITHOUT STEMMING

Embedding	**Model**	**Accuracy**	**Precision**	**Recall**	**F1- score**	**AUC**
AraVec	BiLSTM	83%	83%	83%	83%	0.93
	CNN	84%	84%	84%	84%	0.94
	RNN	36%	53%	43%	36%	0.62
ArBERT	BiLSTM	83%	83%	83%	83%	0.94
	CNN	82%	82%	81%	81%	0.94
	RNN	84%	83%	84%	83%	0.94
MARBER-T	BiLSTM	83%	83%	83%	82%	0.95
	CNN	85%	85%	85%	85%	0.95
	RNN	84%	84%	84%	84%	0.95
QARiB	BiLSTM	85%	85%	85%	85%	0.92
	CNN	86%	85%	85%	85%	0.95
	RNN	83%	83%	83%	83%	0.94

C. Transformer-Based Classification

Previous studies have used transformer-based approaches for sentiment analysis applications [35]. However, the present study uses the AutoModelForSequenceClassification fine-tuned model from the Hugging Face library. This approach updates all encoder layers during training rather than keeping them frozen, allowing the models to adapt the characteristics of the sentiment dataset effectively. Each BERT-based model employs the [CLS] token to represent the entire input sequence, followed by a fully connected layer for classification. In addition, all models were trained for five epochs with 16 batches, 2×10^{-5} learning rate, 0.01 weight decay, and 500 warmup steps. Table XI presents the transformer-based model hyperparameters used in the study.

TABLE XI. FINE-TUNING PARAMETERS AND CONFIGURATION FOR THE TRANSFORMER MODELS

Parameter	**Value**
Fine-tuning epochs	5
Batch size	16
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup steps	500
Optimizer	AdamW

After each period, the evaluation and saving procedures were applied using the AdamW optimizer. The best model was automatically selected based on the maximum validation accuracy.

The results of transformer models are presented in Tables XII and XIII. QARiB results are more robust and more accurate, with 88% accuracy and 87% F1-score. While AraBERT and ArBERT were trained in classical Arabic and MARBERT in informal Arabic, QARiB benefited from being pre-trained in both, making it ideal for tourist reviews. QARiB and MARBERT utilize only MLM since they handle brief Tweet data that do not require deep context understanding. As shown in Figure 5, the Transformers models have maintained and improved their performance despite using stemming.

TABLE XII. PERFORMANCE COMPARISON OF TRANSFORMER CLASSIFICATION MODELS ON DATASET WITH STEMMING

Model	Accuracy	Precision	Recall	F1- score	AUC
AraBERT	87%	87%	86%	86%	0.96
MARBERT	88%	88%	87%	87%	0.96
ArBERT	85%	85%	85%	85%	0.95
QARiB	86%	86%	86%	86%	0.96

TABLE XIII. PERFORMANCE COMPARISON OF TRANSFORMER CLASSIFICATION MODELS ON DATASET WITHOUT STEMMING

Model	Accuracy	Precision	Recall	F1- score	AUC
AraBERT	88%	88%	88%	88%	0.96
MARBERT	88%	88%	87%	87%	0.95
ArBERT	88%	88%	88%	88%	0.95
QARiB	88%	88%	88%	88%	0.96

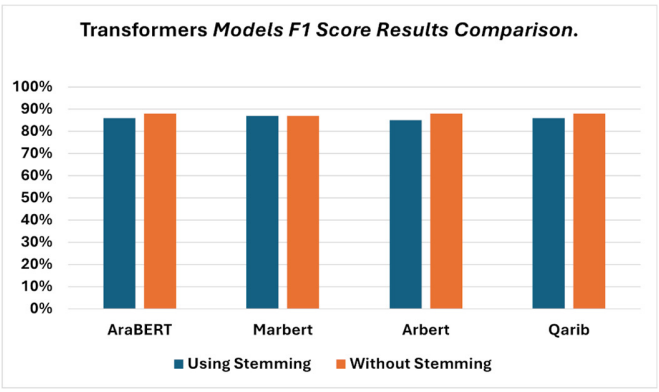

Fig. 5. Performance comparison of transformer-based classification models.

D. Error Analysis

To understand how the model's errors are distributed across different categories, a calibrated confusion matrix, as depicted in Figure 6, was analyzed. QARiB attained an accuracy of 92.3% for positive reviews, 90.1 for negative reviews, and 78.4% for neutral reviews. Despite notable misclassifications between the neutral and positive classes, the results indicate the model's robust capacity for accurately identifying review sentiment.

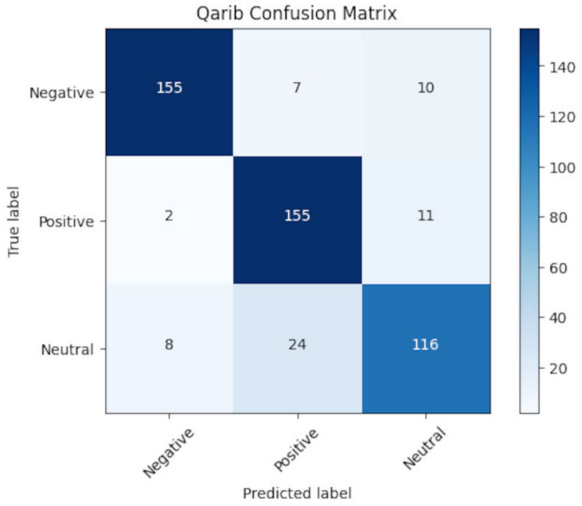

Fig. 6. Confusion matrix of the QARiB model.

E. Models Comparison

Transformer-based models are known for their contextual understanding. As displayed in Figure 5 and Tables XII and XIII, QARiB is a more suitable model for evaluating reviews as it has contextual and sequential processing capabilities, which improve speed and the overall comprehension of the content. This method enables QARiB to detect sentiment in reviews, resulting in a more accurate and reliable analysis than previous models. QARiB is an effective sentiment analysis tool in this regard.

Figures 7 and 8 illustrate a comparison of all models using different evaluation metrics. The accuracy of transformers, as observed in Figure 7, was higher than that of baseline and DL models. When using stemming, the ArBERT model achieved 85% accuracy, and QARiB and MARBERT's results were stable.

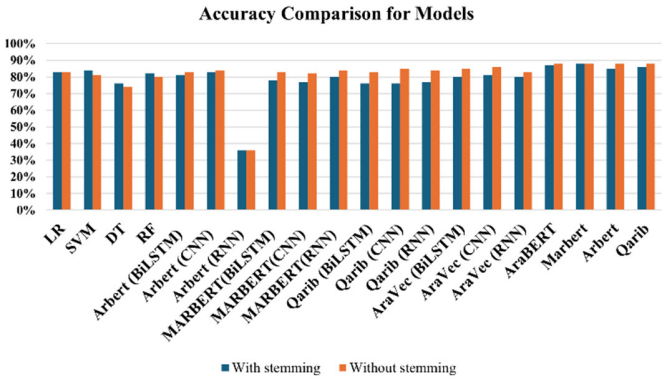

Fig. 7. Accuracy performance comparison of all evaluated models.

In terms of AUC, AraBERT and QARiB achieved the highest value of 0.96, as shown in Figure 8. DL model performance showed good results with MARBERT and QARiB. However, the AUC was unstable in ML, with LR achieving a value of 0.95. Overall, the transformer-based models exhibited better performance.

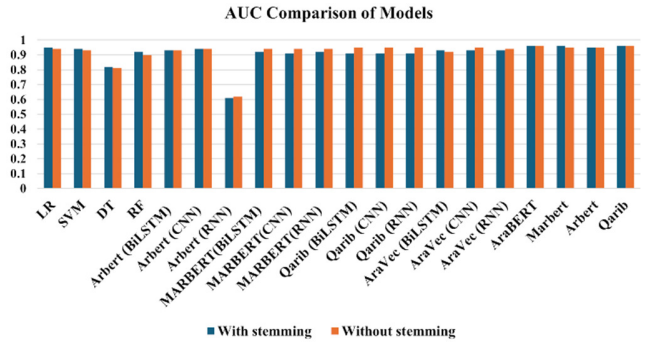

Fig. 8. AUC performance comparison of all evaluated models.

Stemming has a higher negative effect on baseline models compared to DL models; the transformer-based models have demonstrated improved performance without stemming. Stemming also affects models that operate entirely on words' context, such as transformers. The Approximate Randomization test was used to check if the results were statistically significant. Table XIV presents statistical results for different models—showing F1-scores, p-value, significance (Yes/No), and identification. The results indicate statistically significant differences ($p < 0.005$), confirming that QARiB performs better compared to other models and achieves the highest F1-score.

The Bootstrap technique was used to calculate confidence intervals for the main performance measures. In many comparisons, the differences were statistically significant (p-value = 0.0), which confirms the stability and reliability of the QARiB model's superior performance. For the strongest transformer-based corresponding model, the significance was not always achieved, but QARiB results remained comparable or higher.

TABLE XIV. COMPARATIVE F1-SCORE PERFORMANCE AND STATISTICAL SIGNIFICANCE OF THE QARiB MODEL VERSUS ML, DL, AND TRANSFORMER-BASED MODELS

Metric	Model 1	Model 2	F1-Score (Model 1)	F1-Score (Model 2)	p-value	Sig.	Best
F1-Score	AraBERT	QARiB	0.88	0.89	0.136	No	QARiB
F1-Score	MARBERT	QARiB	0.87	0.89	0.647	No	QARiB
F1-Score	ArBERT	QARiB	0.88	0.89	0.412	No	QARiB
F1-Score	Others*	QARiB	<0.85	0.89	<0.005	Yes	QARiB

*Others: All others tested DL and ML models except for AraBERT, MARBERT, and ArBERT.

Sig.: Statistical significance at α=0.05.

Best: Model with the highest F1-Score in each comparison

.

The results indicate that the transformer-based models are suitable for sentiment classification tasks in tourist reviews. Stemming demonstrated outstanding results and better model performance, while QARiB achieved robust and trustworthy results.

Compared to the present study, authors in [14, 15, 17] report very high accuracies (95–100%) using ML and DL models [18]; however, these results are often due to training and evaluation on the same datasets, making them susceptible to overfitting. In the present study, the most reliable performance was achieved by QARiB and MARBERT, which, after task-specific fine-tuning, attained strong accuracy without signs of overfitting. This performance is attributed to the superior generalization of QARIB and MARBERT to their pre-training on Arabic, including colloquial varieties, which closely match the linguistic characteristics of the dataset used in this study.

IV. CONCLUSION AND FUTURE WORK

Evaluating clients' feedback and opinions is crucial for any business, especially in the tourism sector. Thus, this study collected a dataset of tourist reviews in Arabic extracted from Google Maps of famous Saudi tourist attractions: the Boulevard Riyadh City, Al-Ula Old Town, the Al-Balad district in Jeddah, the Heritage Village in Dammam, and the Al-Hada cable car.

Various models were implemented and compared, including Deep Learning (DL) models such as Convolutional Neural Network (CNN), Recurrent Neural Networks (RNNs), and Bidirectional-Long Short-Term Memory (BiLSTM), with different word embedding techniques and baseline classifiers such as Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). The study also employed transformer-based models, including AraBERT, ArBERT, MARBERT, and QARiB, to classify evaluations of popular tourist destinations into three classes: neutral, negative, and positive. The findings suggest that the QARiB model performs the best across various measures, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC).

For future work, adding more reviews to the dataset is expected to enhance the performance of DL models. Additionally, models' performance across different cities or tourism domains should be explored to test models' generalization capabilities.

REFERENCES

[1] J. Gantz and D. Reinsel, *Extracting Value from Chaos*. Boston MA, USA: International Data Corporation, 2011.

[2] M. Iansiti, "The Value of Data and its Impact on Competition," *SSRN Electronic Journal*, 2021, https://doi.org/10.2139/ssrn.3890387.

[3] Q. T. Ain *et al.*, "Sentiment Analysis using Deep Learning Techniques: A Review," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017, https://doi.org/10.14569/IJACSA.2017.080657.

[4] S. Hlee, H. Lee, C. Koo, and N. Chung, "Will the Relevance of Review Language and Destination Attractions be Helpful? A Data-Driven Approach," *Journal of Vacation Marketing*, vol. 27, no. 1, pp. 61–81, Jan. 2021, https://doi.org/10.1177/1356766720950356.

[5] F. Salem, *Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World: Potential, Limits and Concerns*, 7th ed. Dubai, UAE: Mohammed Bin Rashid School of Government, 2017.

[6] O. Oueslati, E. Cambria, M. B. Haj-Hmida, and H. Ounelli, "A Review of Sentiment Analysis Research in Arabic Language," *Future Generation Computer Systems*, vol. 112, pp. 408–430, Nov. 2020, https://doi.org/10.1016/j.future.2020.05.034.

[7] H. S. Rashid, "The Arabic Language in Social Medias' Era," *Utopía y Praxis Latinoamericana*, vol. 25, no. 1, pp. 356–363, 2020.

[8] R. Al-Jarf, "Effect of Social Media on Arabic Language Attrition," in *Sylvester Confab Book of Readings*, Abuja, Nigeria: Cissus World Press Books, 2019, pp. 38–48.

[9] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," *ACM Transactions on Asian Language Information Processing*, vol. 8, no. 4, pp. 1–22, Dec. 2009, https://doi.org/10.1145/1644879.1644881.

[10] I. Al-Huri, "Arabic Language: Historic and Sociolinguistic Characteristics," *English Literature and Language Review*, vol. 1, no. 4, pp. 28–36, 2016, https://doi.org/10.13140/RG.2.2.16163.66089/1.

[11] N. Habash, "Arabic Natural Language Processing," in *Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Dubai, UAE, 2022, pp. 9–10, https://doi.org/10.18653/v1/2022.emnlp-tutorials.2.

[12] E. Boujou, H. Chataoui, A. E. Mekki, S. Benjelloun, I. Chairi, and I. Berrada, "An Open Access NLP Dataset for Arabic Dialects: Data Collection, Labeling, and Model Construction." arXiv, Feb. 07, 2021, https://doi.org/10.48550/arXiv.2102.11000.

[13] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A Survey on Sentiment Analysis Methods, Applications, and Challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, https://doi.org/10.1007/s10462-022-10144-1.

[14] B. A. Alharbi, M. A. Mezher, and A. M. Barakeh, "Tourist Reviews Sentiment Classification using Deep Learning Techniques: A Case Study in Saudi Arabia," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 717–726, 2022, https://doi.org/10.14569/IJACSA.2022.0130685.

[15] R. AlQadi, H. Al-Nojaidi, L. Alabdulkareem, M. Alrazgan, N. Alghamdi, and M. M. Kamruzzaman, "How Social Media Influencers Affect Consumers' Restaurant Selection: Statistical and Sentiment Analysis," in *2nd International Conference on Computer and Information Sciences*, Sakaka, Saudi Arabia, Oct. 2020, pp. 1–6, https://doi.org/10.1109/ICCIS49240.2020.9257636.

[16] W. A. Alasmari, H. A. Abdelhafez, and H. A. Abdelhafez, "Twitter Sentiment Analysis for Reviewing Tourist Destinations in Saudi Arabia using Apache Spark and Machine Learning Algorithms," *Journal of Computer Science*, vol. 18, no. 3, pp. 215–226, Mar. 2022, https://doi.org/10.3844/jcssp.2022.215.226.

[17] B. Al Sari *et al.*, "Sentiment Analysis for Cruises in Saudi Arabia on Social Media Platforms Using Machine Learning Algorithms," *Journal of Big Data*, vol. 9, no. 1, Dec. 2022, Art. no. 21, https://doi.org/10.1186/s40537-022-00568-5.

[18] D. Elangovan and V. Subedha, "Adaptive Particle Grey Wolf Optimizer with Deep Learning-Based Sentiment Analysis on Online Product Reviews," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10989–10993, Jun. 2023, https://doi.org/10.48084/etasr.5787.

[19] S. Alosaimi, M. Alharthi, K. Alghamdi, T. Alsubait, and T. Alqurashi, "Sentiment Analysis of Arabic Reviews for Saudi Hotels Using Unsupervised Machine Learning," *Journal of Computer Science*, vol. 16, no. 9, pp. 1258–1267, Sept. 2020, https://doi.org/10.3844/jcssp.2020.1258.1267.

[20] Banan, "Dataset-Tourist-Places-Reviews." Github, June 2022, [Online]. Available: https://github.com/Banan6/Dataset-Tourist-places-reviews.

[21] P. Jonaitis, "Instant Data Scraper." Web Robots, Vilnius, Lithuania, Jan. 2025, [Online]. Available: https://chromewebstore.google.com/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah.

[22] Raneem, "Saudi Tourism Reviews Sentiment Dataset." Github, Nov. 2025, [Online]. Available: https://github.com/Raneem224/Tourist_Reviews_in-Saudi_Arabia.

[23] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," *Procedia Computer Science*, vol. 117, pp. 63–72, 2017, https://doi.org/10.1016/j.procs.2017.10.094.

[24] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, Jan. 1988, https://doi.org/10.1016/0306-4573(88)90021-0.

[25] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," *Procedia Computer Science*, vol. 117, pp. 256–265, 2017, https://doi.org/10.1016/j.procs.2017.10.117.

[26] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT and MARBERT: Deep Bidirectional Transformers for Arabic." arXiv, June 07, 2021, https://doi.org/10.48550/arXiv.2101.01785.

[27] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-Training BERT on Arabic Tweets: Practical Considerations." arXiv, Feb. 21, 2021, https://doi.org/10.48550/arXiv.2102.10684.

[28] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sept. 1995, https://doi.org/10.1023/A:1022627411411.

[29] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 1st ed. Hoboken, NJ, USA: Wiley, 2013.

[30] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, https://doi.org/10.1023/A:1010933404324.

[31] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision Trees: From Efficient Prediction to Responsible AI," *Frontiers in Artificial Intelligence*, vol. 6, July 2023, Art. no. 1124553, https://doi.org/10.3389/frai.2023.1124553.

[32] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, https://doi.org/10.1109/78.650093.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, https://doi.org/10.1038/nature14539.

[34] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1724–1734, https://doi.org/10.3115/v1/D14-1179.

[35] S. Tzimiris, S. Nikiforos, M. N. Nikiforos, D. Mouratidis, and K. L. Kermanidis, "A Comparative Evaluation of Transformer-Based Language Models for Topic-Based Sentiment Analysis," *Electronics*, vol. 14, no. 15, July 2025, Art. no. 2957, https://doi.org/10.3390/electronics14152957.