Tourist Reviews Sentiment Classification using Deep Learning Techniques: A Case Study in Saudi Arabia

Banan A.Alharbi, Mohammad A. Mezher, Abdullah M. Barakeh
Dept. of Computer Science, Fahad Bin Sultan University, Tabuk, Saudi Arabia

***Abstract*—Now-a-days, social media sites and travel blogs have become one of the most vital expression sources. Tourists express everything related to their experiences, reviews, and opinions about the place they visited. Moreover, the sentiment classification of tourist reviews on social media sites plays an increasingly important role in tourism growth and development. Accordingly, these reviews are valuable for both new tourists and officials to understand their needs and improve their services based on the assessment of tourists. The tourism industry anywhere also relies heavily on the opinions of former tourists. However, most tourists write their reviews in their local dialect, making sentiment classification more difficult because there are no specific rules to control the writing system. Moreover, there is a gap between Modern Standard Arabic (MSA) and local dialects. one of the most prominent issues in sentiment analysis is that the local dialect lexicon has not seen significant development. Although a few lexicons are available to the public, they are sparse and small. Thus, this paper aims to build a model capable of accurate sentiment classification in the Saudi dialect for Arabic in tourist place reviews using deep learning techniques. Machine learning techniques help classifying these reviews into (positive, negative, and neutral). In this paper, three machine learning algorithms were used, Support -Vector Machine (SVM), Long short-term memory (LSTM), and Recurrent Neural Network (RNN). These algorithms are classified using Google Map data set for tourist places in Saudi Arabia. Performance classification of these algorithms is done using various performance measures such as accuracy, precision, recall and F-score. The results show that the SVM algorithm outperforms the deep learning techniques. The result of SVM was 98%, outperforming the LSTM, and RNN had the same performance of 96%.**

Keywords*—Sentiment classification; Saudi dialect; support - vector machine; recurrent neural network; long short-term memory

I. INTRODUCTION

Due to the astounding and quick expansion of social networking sites, an increasing number of individuals are sharing their experiences and opinions on various issues, including travel, hotels, physical items, movie reviews, and health. One cannot deny that social media sites play a role in people's daily and social lives. Additionally, they assist tourists in selecting the appropriate destination via their comments on tourist destinations' social media sites, as mentioned by [1].

Moreover, social media sites have gained a significant attraction on the web. These sites have evolved into an indispensable resource for travelers whose decisions are influenced by the reviews and opinions of other travelers. Human emotions and emotional cognition influence purchasing decisions, travel, and other variables. Online reviews can help researchers and business owners understand tourists' needs and preferences correctly. It was further noted that the tourism industry's primary reliance on the opinions and perceptions of former travelers is universal. They emphasized that [2] the views expressed by tourists in the comments play a role in influencing the choices of other tourists for their travel destinations.

Moreover, travelers frequently desire to know the attractions for which a city they wish to visit is renowned. They research social media sites for recommendations, opinions, and reviews to visit tourist destinations. Given the plethora of information and evaluations, it is difficult for travelers to obtain reliable judgments and select the best hotels, restaurants, and attractions. Many people share their experiences and views spontaneously and more credibly about the tourist places they visit without a financial return. Therefore, it will be challenging for the reader to locate relevant websites when researching, rewriting, and summarizing the facts and viewpoints vital to them.

Consequently, the significance of sentiment classification reduces the time and effort required to extract relevant information for travellers. The sentiment classification process studies people's emotional state and their assessments of a particular topic or their attitudes towards a specific event, and sentiment classification is used in tourism applications, products, shopping and other areas of life. Consumers place greater credence on online reviews, personal recommendations, and comments and thus are more likely to provide product reviews after purchasing. The classification process aims to determine the polarity of the text and determine whether a person feels optimistic about a particular product, negative or neutral. Classifying reviews (positive or negative) is the goal of sentiment classification, and like the use of text, analysis has proven cost-effective. The classification is mainly based on an explained supervised learning approach [3].

As a result, sentiment classification is one of the most active and prosperous research areas in Natural Language Processing (NLP). Many researchers and those interested in this field use deep learning for sentiment classification. Data technologies reduce sentiment classification errors to ensure the highest accuracy on social media [4]. This study aims to develop a model capable of accurate sentiment classification in the Saudi dialect of Arabic by analyzing tourist location reviews using deep learning techniques.

A. Sentiment Classification

Sentiment classification refers to analyzing textual data on social media and extracting people's sentiments about different topics. The sentiment classification also provides a more understanding of people's views on a particular topic. We can observe the sentiment classification in action when determining consumer satisfaction with a specific product based on their input. In addition, we may learn about people's preferences for top attractions through their recommendations on social networking sites and by assessing the quality of hotels through reviews. These reviews can also classify opinions as positive, negative or neutral. Many English, Chinese and other languages have made remarkable progress and are high-performance using sentiment classification [5]. Likewise, as the volume of unstructured data generated by social media grows, the demand for sentiment classification has skyrocketed. There are three different levels of sentiment classification: document level, sentence level, and feature level.

- Document-level: The document is classified by general feeling, either positive, negative or neutral, and the paper is supposed to contain an opinion on one topic.
- Sentence level: At the sentence level, the classification determines whether the opinion in this sentence is positive, negative or neutral.
- Feature level: Feature level makes a more accurate analysis. Instead of looking at the structure of language (documents, paragraphs, sentences, or phrases), it directly looks at the same opinion.

It is also called aspect-based sentiment analysis. This level examines multi-opinionated sentences to determine the attitude towards entities or their aspects by examining the opinion and the entity's purpose or one of its features.

B. Sentiment Classification Techniques

It has developed classification methods in machine learning. There are roughly three classification strategies for sentiment: machine learning, lexicon-based, and hybrid.

- Machine Learning-based approach: Two proposed sets of classification sentiment problems are traditional and deep learning models [6]. Traditional models apply to classic machine learning algorithms: Naïve Bayes (NB) classifier, Maximum Entropy (ME) or SVM. Algorithm inputs also use language features such as N-grams, bi-grams, bag-of-word, parts of speech, or adjectives and adverbs. Consequently, choosing features in traditional models can affect the accuracy of the results. In contrast, sentiment analysis employing deep learning models such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and RNN can get better results than conventional methods. Accordingly, we can address classification problems at the document level, Sentence level, or Aspect or Feature level.
- Lexicon-based approach: this technique was the first to use sentiment analysis [6]. It consists of a set of sentiment terms that are translated and pre-known. It is also divided into two approaches: dictionary-based and corpus-based. In the first form of sentiment classification, a dictionary of terms is used to locate the first sentiment words, using a synonym and antonym dictionary such as WordNet SentiWordNet. In the second type, the sentiment classification corpus-based does not depend on the dictionary of pre-defined sentiment terminology but on statistical or semantic analysis for the polarity sentiment. The corpus-based uses a technique such as neighbours (k-NN), Conditional Random Field (CRF).
- A hybrid approach is prevalent and confuses both approaches, and sentiment dictionaries play a significant role in most methods [7]. The more effectively the classifier is trained, the easier and better is the future predictions. We can divide text classification methods into supervised and unsupervised learning using machine learning. Conversely, unsupervised learning is used when it is challenging to find classified training documents.

C. Deep Learning

Deep learning is a branch of machine learning as it depends on neural networks that resemble the human brain. Deep learning refers to the deep neural network introduced in 2006 by G.E. Hinton. Accordingly, deep learning can recognize speech, analyse images, and process natural language. Deep learning networks also provide supervised and unsupervised groups. Consequently, many companies have been interested in machine learning in handling big data, and models have been created to analyse big and complex data faster and more accurately [8]. However, there are many networks in deep learning, such as CNN and Deep Belief Network, and there are many benefits for neural networks in vector representation, sentence classification, sentence modelling, and text creation.

D. Models

1) Long short-term memory model: LSTM is a developer version of RNN. They were developed to overcome the problems of RNN in the vanishing and exploding gradient. They were introduced by [9]. Consequently, LSTM has achieved great success in various NLP tasks. Accordingly, [10] designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs of three basic gates: the input gate, the output gate o_t, and the forget gate f_t.

- The input gate controls the amount of information transferred from the input to the current cell.
- Forget gate *ft*. This gate decides what information should be thrown away or kept. Data from the previous hidden state and information from the current input are passed through the sigmoid function. Values come out between 0 and 1—the closer to 0 means to forget, and the closer to 1 means to keep.
- The output gate o_t controls the amount of information transferred to the output.

2) Recurrent neural network model: RNNs are artificial neural networks that process and use sequential data. In a

conventional neural network, each input and output is treated as independent of the other, which might result in poor performance. Hence, the idea of a repetitive neural network predicts the following word in the sentence based on the previous word. RNNs have the concept of 'memory' that helps them store the states or information of prior inputs to generate the following output of the sequence. In this way, the network can use history to understand the sequential nature of the data. According to this, RNNs have shown great success in many NLP tasks. Mainly concentrating on sequential data applications, such as text classification and speech recognition. On the other hand, RNNs have also shown notable success in image processing.

3) Support vector machine model: SVM algorithm is one of the supervised machine learning algorithms used for classification and regression, developed by Vapnik. SVM is used to solve various classification problems, such as the recognition of fonts, images, text classification, and other fields [11]. The concepts used in SVMs, like hyperplane, optimal margin, Kernels, etc. There are two types of SVM: linear classification and non-linear classification. Liner classification: divides the two groups of data linearly. SVM picks the hyperplane with the most significant margin to achieve maximum separation between the two classes. The margin equals the sum of the shortest distances between the separating hyperplane and the closest data point of both categories. Non-linear classification in SVM has been extended using Kernels. Kernels are mathematical functions that transform the data from a given space (known as Input Space) to a new high dimensional space (known as Feature Space) where data can be separated with a linear surface (called hyperplane).

II. Related Work

Social media sites generate vast reviews about travel, products, movies, etc. In addition, social media sites provide a suitable space for people to share their experiences more freely. Today, people are increasingly making their views and experiences available online. Therefore, millions of web pages will appear when you search for any topic online. It is challenging and time-consuming to gain an overall understanding of these reviews. A few reviews may influence an individual's perspective. Accordingly, sentiment classification aims to solve such problems by automatically classifying people's views into negative, positive, and neutral opinions. This makes it an important field of research for consumers, companies and Governments. However, resources are not always available for all areas or languages. While there has been a great deal of study on sentiment classification in English, very few scholars have examined sentiment classification in Arabic, despite the growing number of Arabic speakers.

A. Tourist Place Reviews Sentiment Classification

With the radical change in social lifestyle and people's general decisions, sentiment classification has become the subject of long-term research. A study analyzing sentiment classification revealed that numerous studies use movie or restaurant reviews as a proxy for sentiment categorisation. [12] stated that traditional sentiment classification methods often require significant human efforts. With the growing volume of web reviews, people have a big problem increasing information. The author in [1] has focused on analyzing and organising unstructured data from social media sites. In light of this, unstructured data from social media has increased the demand for sentiment analysis. The author in [13] explained that sentiment analysis could be at different levels, such as document level, sentence level, and aspect-based. The results of experimental studies [14] also showed that one of the tasks of sentiment classification at the document level is to consider the importance of sentences.

Moreover, as explained by [15], many studies use classification algorithms such as NB, ME, and, particularly, SVM to classify the polarity of reviews. Accordingly, [16] compared two approaches to supervised machine learning methods regarding sentiment classification of reviews between SVM and NB, which was highly accurate. The results indicated that NB outperformed SVM, as the training data set consists of many reviews. The approach of NB has reached a high degree of accuracy compared to SVM.

In addition, [17] compared five machine learning algorithms, K-Nearest Neighbors (KNN), Decision Tree, Artificial Neural Networks (ANNs), Naïve Bayes and SVM are used for the classification of sentiments. These algorithms are analysed on Twitter's dataset, and the performance of these methods has been compared based on different performance metrics such as accuracy, precision, recall and F measure. The results showed that SVM outperforms others and has a higher predictive ability than other algorithms. So, all the metrics, Accuracy, Precision, Recall and F-score, are high in the case of SVM.

The author in [18] compared the accuracy of LSTM and CNN by looking at the different designs of each. The training sample was taken from Booking and TripAdvisor. Thus, the results showed that LSTM neural networks outperform CNN in performance. LSTM produces higher accuracy. They also show that combining convolutional layers with recurrent LSTM layers does not benefit.

The author in [19] suggested a hybrid model using LSTM and a profound CNN model named hybrid CNN-LSTM model. Suggested results showed hybrid CNN-LSTM model outperforms traditional deep learning and machine learn ing techniques in precision, recall, f-measure, and accuracy. The training sample was IMDB movie and Amazon movie reviews.

The author in [20] has widely explained the importance of online reviews. The authors in [21], [22] have also dealt with the impact of online comments on tourist decision-making. The author in [23] shows the importance of recognizing the impact of online reviews on tourist behaviour. The intensity and effects of internet reviews on users' purchasing behaviour and decision-making have been demonstrated [24], [25]. Furthermore, [26] presented several tourists' satisfaction, while some studies employed internet reviews. Hotel visitors'

satisfaction [27] and feelings about a film [28]. Others also stated that traveller reviews are more up-to-date, entertaining, and trustworthy than those published by travel service providers [29].

B. Deep Learning Sentiment Classification

In recent years, deep learning has become an effective way to solve sentiment classification problems. Consequently, several studies have recommended deep learning-based models for sentiment classification that outperform conventional machine learning models [30]. Many deep learning models include DNN, CNN, and deep Restricted Boltzmann Machine (RBM).

One study suggested CNN for sentiment classification. The sample of training datasets was divided into three groups (movie review data, customer review data, and Stanford Sentiment Treebank data). Experimental results showed that CNN layers contributed to better performance of relatively long texts [31].

Another compared traditional models such as Logistic Regression, SVM, and deep learning models such as CNN and the simple LSTM. The results showed that the larger the data set, the better the deep learning models were than traditional models, as the LSTM results were 95.55٪ better accurate. The training sample was Amazon reviews [32].

C. Reviews Arabic Language Sentiment Classification

Research has been done on sentiment classification in English. Although Arabic is one of the most widely used languages on social media, few studies have focused on sentiment classification in Arabic. According to statistics Internet World Stats, Arabic is the fourth most commonly used language online after English, Chinese and Spanish [33]. Although Arabic-speaking users have increased online in recent years, sentiment analysis research in Arabic has not yet evolved.

The author in [34] applied the sentiment analysis of Arabic tweets from Twitter. They used unsupervised methods. They examined preprocessing methods of text and similarity to compile algorithms and discussed how they affected the results. Moreover, they used a k-means algorithm. Therefore, the brevity of the language in this context presents a challenge to the likelihood of ambiguity. In this investigation, root-based derivation yielded the best degree of accuracy, as indicated by the findings.

The author in [35] compared two supervised learning algorithms, SVM, and RNN, in terms of ability to face the challenges of Aspect-Based sentiment analysis. The training sample was Arabic hotel reviews. The results showed that the SVM approach outperformed the RNN approach in all tasks in terms of accuracy, but RNN was faster at implementation.

The author in [36] analyse the collected twitter posts in different Arabic dialects and compare the various algorithms. The measurement of performance of different algorithms is evaluated in terms of recall, precision, f-measure, and accuracy. Experimental results showed that unigram gives a higher accuracy of 99.96%, with Passive-Aggressive (PA) or Ridge Regression (RR).

The author in [37] uses two models of deep learning, CNN and LSTM. The training sample was the Arabic Sentiment Tweets Dataset (ASTD) dataset. The best results for an LSTM model. Sayed et al., 2020 used nine supervised machine learning algorithms: namely, Gradient Boosting, Logistic Regression, Ridge Classifier, SVM, Decision Tree, Random Forest, KNN, Multi-layer Perceptron (MLP) and NB. The training data sample on evaluations written in Arabic was manually prepared through hotel evaluations from Booking.com. The Experimental results showed that the Ridge Classifier (RC) appears to have the best performance in accuracy, recall, precision, training time and F1 score.

The author in [38] applied K- means and Hierarchical unsupervised learning approaches. The training sample was sentiment analysis from the tourism website TripAdvisor for Arabic reviews of Saudi hotels. Used clustering, features and preprocessing strategies to find the best models. The results showed that k-means clustering achieved the best accuracy. The author in [39] suggested combining linguistic and statistical features and sentiment classification using a tweets dataset in Arabic. They used three classifiers: SVM, KNN and ME. The results showed that SVM achieved the highest accuracy in the classification.

D. Saudi Dialect for Arabic Review Sentiment Classification

We have been increasingly interested in analyzing Arabic texts on social media sites over the past few years. However, most social media users write their comments in the local dialect of their countries rather than MSA. In addition, some work has been done to build MSA sentiment lexicons. There has been a limitation in building a dialectal Arabic lexicon, especially for the Saudi dialect. This is mainly due to the limitation in the existing natural language processing tools and the resources available for Arabic, which is developed to deal with MSA only.

The author in [5] applied deep learning to sentiment classification Saudi dialect data on Twitter. Deep learning techniques were used to compare LSTM and Bidirectional Long Short-Term Memory (Bi-LSTM) and algorithm SVM. The data sample was from 32,063 tweets. The results showed that deep learning techniques outperform the algorithm SVM. The experimental result of Bi-LSTM was 94% exceeding the LSTM's 91%, while the SVM had the lowest performance of 86.4%.

The author in [40] presented a new sentiment dictionary in the Saudi dialect called the Saudi Dialect Sentiment lexicon (SauDiSenti). The SauDiSenti comprises 4,431words, a hybrid between MSA and Saudi dialects. They compared SauDiSenti performance to that of AraSenTi. The first experience, which used only positive and negative tweets, showed the first experiment AraSenTi outperformed SauDiSenti in precision, recall, and F measure. The second experiment evaluated SauDiSenti and AraSenTi using positive, negative, and neutral tweets. The results showed that SauDiSenti outperformed AraSenTi for two values because AraSenTi identified most neutral tweets as either positive or negative. Despite the small size of SauDiSenti, they also added promising results in sentiment analysis of the Saudi dialect tweets.

The author in [41] they introduced a system to analyse the tweets of Saudi users on Twitter about Saudi universities. They used two different models suggesting sentiment analysis on Twitter. The first model depends on the use of SVM. The second model is based on using different classifier models. The results showed that SVM outperformed all other classification models.

The author in [42] compared two strategies. The first applied a translation that transforms from dialect to MSA. The second involved the designing of sentiment analysis on the resulting MSA text. Use tweets in the Saudi dialect. Use seven classifiers Logistic Regression, Passive Aggressive, SVM, Perceptron, Multinomial NB, SGD and KNN. The results showed that they proved that applying sentiment analysis techniques yields better results on MSA data than on dialect data.

In addition, there are studies on sentiment classification reviewed in other dialects such as Jordanian, Algerian, Sudanese and Egyptian. The author in [43] studied the effect of five methods of testing features on the performance of the SVM classifier. Sentiment classification was carried out on dialectical Jordanian reviews using an SVM classifier. The feature selection methods are Information Gain (IG), correlation, SVM, Gini Index (GI), and Chi-Square. He integrated some test methods to explore their ability to improve and choose the feature. The results showed that the best performance of the SVM and correlation feature selection methods was produced by combined with the uni-gram model.

The author in [44] compared the results of accurate deep learning models with classic models CNN, LSTM, SVM, and NB. They used two datasets: posts and comments collected from Algerian Facebook pages, and the second was the corpus of Algerian labelled tweets. The results showed that deep learning models are accurate compared to the classical approaches.

The author in [45] used machine learning methods for sentiment classification for text in the Sudanese dialect on Facebook. The Sudanese colloquial has no grammatical or morphological rules, as they have been made clear. Moreover, they used two different classifiers were applied SVM and NB. The results showed that SVM with lemmatisation libraries improved sentiment classification accuracy, as SVM achieved a measurement accuracy of 68.6%, while NB achieved 63.1%.

The author in [46] provided a sentiment analysis system MSA and Egyptian dialect. A different dataset (tweets, product reviews, TV program comments and Hotel reservations) was used. They demonstrated that expanding the polarity lexicon automatically affects sentiment classification. They also showed that exploitation idioms and saying lexicon with a high coverage polarity lexicon have the most significant impact on classification accuracy. Experimental results showed that the SVM classifier indicates high-resolution performance levels.

III. Methodology

A. Sentiment Classification

Many social media sites have gained tremendous popularity in the tourism and hospitality industry, such as TripAdvisor, Citysearch, Virtual Tour, Booking, and Foursquare. Consequently, these sites provided space and opportunity for people to express their opinions freely [47]. These reviews written by people across social media sites are a unique opportunity for sentiment analysis. According to this, sentiment classification aims to classify reviews into three (positive, negative, neutral) or more (positive, very positive, very negative, negative, neutral). However, without human intervention, it is difficult for a machine to discern the polarity and strength of a feeling.

B. Pesudocode of Model Sentimental Classification

Sentiment classification for tourism reviews has been accomplished by developing an algorithm called Sentiment_classification. Firstly, text Preprocessing on reviews data &lexicon, cleaning, tokenising, removing stopword, stemming text. Secondly, reviews that have been preprocessed will determine the sentiment polarity using the Saudi Sentiment Lexicon. The polarity values will determine every word in the reviews, and then the value is calculated to know the reviews' sentiment (negative, neutral or positive). We no longer need to manually label sentiments to each review (because there are too many reviews). The dataset that has been processed will be used for sentiment classification with deep learning models.

Algorithm.1: Sentimental _Classification

```
Start
        Load Raw Data
        Load Saudi Lexicon
        Text Preprocessing:
        Raw Data &Lexicon → cleaning text, text tokenisation, removing
    stop word, Stemming, normalisation
        # Determine Sentiment Polarity of Reviews Dataset
            Score=0
            for word in text:
                if (word in lexicon_positive)
                    score = score + lexicon_positive[word]
            for word in text:
                if (word in lexicon_negative):
                    score = score + lexicon_negative[word]
        # Calculation of Polarity reviews
            if (score >0)
                polarity = 'positive'
            elif (score <0)
                polarity = ' negative'
            else
                polarity= 'neutral'
                    return score, polarity
        Modelling (Reviews Dataset):
# Split the Data (with composition Training Data 80%, Testing Data 20%)
                Classifier ←Training Data
                Predict Model← Testing Data
            # Model is evaluated using its confusion matrix of test data
                Confusion matrix →Accuracy, Precision, Recall, F1-score
End
```

C. Proposed Modification

The flowchart in Fig. 1 shows the methodology followed in this research, which begins with extracting data for Arabic reviews in the Saudi dialect from Google Maps, including 14 cities in Saudi Arabia and 55 different places. After that

processing and cleaning are done for the extracted reviews. The processing process for the lexicon. The polarity of reviews is calculated. Hence, the dataset is ready for classification. Splitting the dataset is used for 80% of training data and 20% of test data. Finally, predict the model and results.

D. Dataset

Dataset is the first step of sentiment classification. The paper aims to classify Arabic reviews of tourist places in the Saudi dialect. According to this, Google Map web scraping was used to collect the dataset containing 22433 reviews and save it in a fixed format CSV [48]. The dataset included 14 cities in Saudi Arabia, thus covering most dialects in Saudi Arabia (Hijazi, Najdi, Northern, Eastern, and Southern). Given the dataset, it turned out that most of the comments were written in Najdi, Hijazi dialect and MSA. Tourist places in every city included museums, archaeological sites, Islamic places, exhibitions, parks, and recreational places. Reviews were automatically classified as positive, negative, and neutral.

E. Text Preprocessing

The text preprocessing process plays a vital role by applying the steps:

- Cleaning Text: Removing non-Arabic letters, numbers, punctuation marks, replacing newlines into space, remove characters space from both left and right text.
- Text Tokenisation: Divides text into words, for example (يستحق الزيارة منتزه جميل) after the tokenisation process ('جميل', 'منتزه', 'الزيارة', 'يستحق').
- Removing Stop-word: non-semantic words are called stop words, such as (ampersand pronouns, prepositions, and nouns). In addition, it does not affect the meaning and overall feeling of the text.
- Normalisation: The scientific of replacing some letters with other letters standardises the character forms used in writing. For example, the letters (إ،أ،آ،ا) are replaced by (ا), (ي) is replaced by (ى), (ه) is replaced by (ة), (ء) are replaced by (ؤ،ئ) and (ك) is replaced by (گ). Replace some characters that appear more than once with a two-character such as (ممتاااااز) converted to (ممتاز).
- Stemming: Stemming of words is one of the essential steps in text preprocessing; for example: removing suffixes and prefixes from words.
- Remove Prefixes: ال، وال، بال، كال، فال، لل
- Remove Suffixes:ي ،ه ، ية ، ين ، ون ، ات، ان، ها

F. Polarity of Reviews

This study used a Lexicon-based method to identify polarity, and evaluations are classified as positive, negative, and neutral, Table I. Although more polarity categories might more precisely define the polarity of the evaluation, the model has been limited to the three categories of positive, negative, and neutral. According to this, the polarity of reviews is determined based on the majority of opinion words. If the number of positive words in the review is more, it will be classified as positive.

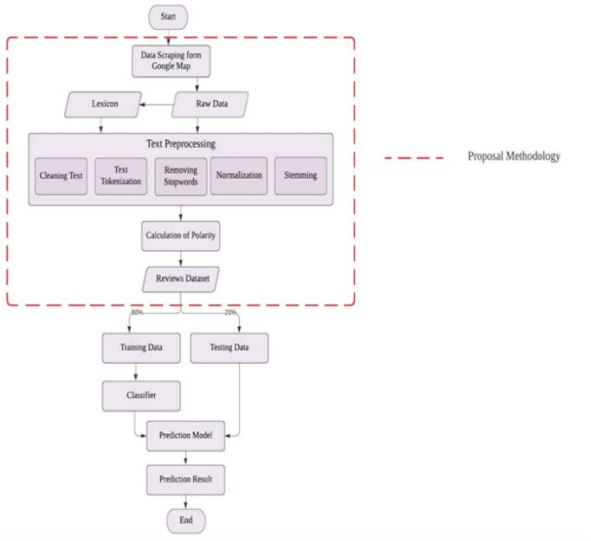

Fig. 1. Main Methodology Flowchart.

Conversely, if the number of negative words is high, the polarity of the review will be rated negatively. If the number of positive and negative words is equal, the review is neutral. In addition, if the opinion word is preceded by negation "مش،غير،لا", then the polarity of that review is reversed. Finally, the polarity of sensations is computed using the remaining poles and gathered in order to forecast the polarity of positive, negative, and neutral emotions. The Fig. 2 shows how the model calculates the scores polarities of sentiments to sentiment score calculation in reviews. Each positive word is greater than 0, and every negative word is smaller than 0. If neutral, it is equal to 0.

TABLE I. SAMPLE OF POSITIVE, NEGATIVE, NEUTRAL REVIEW

حديقة هادئة وجدا خلابة	**Positive**
لا يستحق الزيارة مبالغ في قيمة التذكرة	**Negative**
بحيرة جميله تحتاج إلى صيانة أفضل	**Neutral**

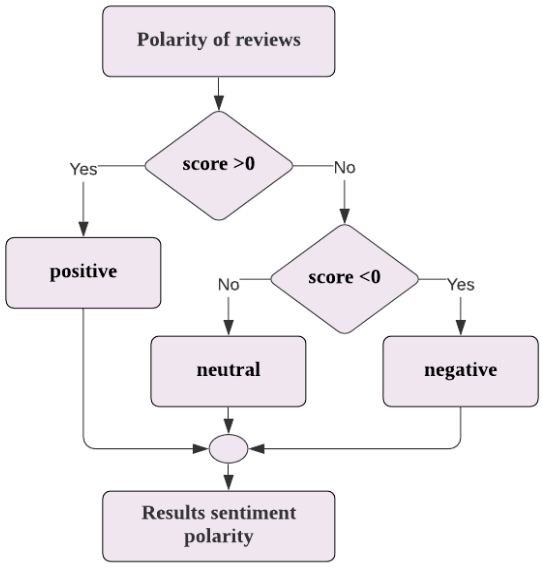

Fig. 2. Polarity of Reviews.

IV. Results and Discussion

The dataset consists of 68.8% positive, 14.6% negative, 16.6% neutral. Fig. 3 shows the number of reviews for each class.

In the Fig. 3 above, positive reviews are far more than negative and neutral reviews in the dataset. When one category is greater than the other, the data is skewed. Therefore, classification accuracy is not an appropriate statistic to optimise when dealing with imbalanced categories. Thus, accuracy may be artificially relatively high. Majority categories are favoured, while minority categories are not recognized. They result in a biased model for a particular category. Most machine learning algorithms work best when the number of samples in each category is almost equal. There are different ways to balance data: random oversampling and SMOTE. To balance the data, random oversampling was utilised to replicate the minority class samples. It also does not result in any data loss.

Evaluating the model after applying all stages is one of the most critical steps. In addition to this, there are many metrics to assess the quality of deep learning models. This research used two methods to evaluate the confusion matrix, K-Fold Cross-Validation. Indeed, the confusion matrix is one of the most informative performance measures a multi-class learning system can rely on [49]. The most commonly used measure for sentiment classifications is accuracy, precision, recall, and F1-score. They are briefly described below. There it is indicated as True Positives (TP), False Positives (FP), True Negative (TN) and False Negative (FN).

Accuracy is one of the most frequently used metrics. The ratio is expected reviews relative to the total number of reviews correctly. Calculation of the accuracy is made according to the equation.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

The ratio of properly expected positive reviews to the total expected positive reviews. The precision measure is calculated using an equation.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

It tells what fraction of all positive samples was correctly predicted as positive by the classifier. This value is calculated according to the equation.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

It combines precision and recall into a single measure. Mathematically it is the harmonic mean of precision and recall.

$$\text{F1 Score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

The results shown in Table II proved that some machine learning algorithms outperform each other in terms of accuracy, precision, recall, and F1-score, their ability to classify, discriminate and recognize reviews to be classified into (positive-negative-neutral). It was noted that the SVM model was more accurate than other classification models, with 98% accuracy.

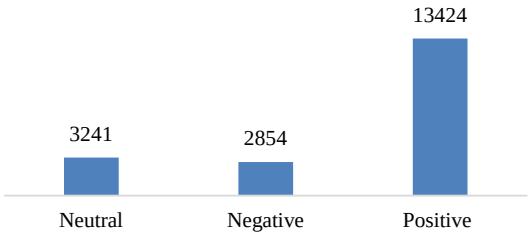

Fig. 3. Number of Reviews for Each Class.

TABLE II. Classification Methods

Classification Method	Accuracy	Precision	Recall	F1-score
SVM	98%	98.16	98.08	98.09
LSTM	96%	95.63	95.58	95.58
RNN	96%	95.99	95.96	95.95

The Fig. 4 below shows the results of Table II of the models used and compares the best results in terms of accuracy, precision, recall and F1-score. It also shows the superiority of the SVM model over the models used.

K-Fold Cross-Validation is one of the most common ways to validate ML models and bolster and corroborate our findings. The available dataset is partitioned into subsets of approximately equal size. The "K" refers to the number of producing several subsets. The first subset serves as a validation set, and the remaining four subsets serve as a training set.

Table III represents the accuracy obtained by implementing the Stratified K-Fold Cross Validation Technique on the dataset used: 99% accuracy for the SVM model, 94.7% for the LSTM model, and 93.9% for the RNN model. The accuracy obtained by each algorithm is almost similar to our findings from this confusion matrix.

Fig. 4. A Comparison of Various Performance Measurements.

TABLE III. Illustration of K-Fold Cross-Validation

Model	1-Fold	2-Fold	3-Fold	4-Fold	5-Fold
SVM	97.5%	98.51%	99.6%	99.7%	99.7 %
LSTM	94.2%	95.3%	95.2%	94.5%	94.5%
RNN	90.0%	91.0%	90.5%	99.4%	99.2%

One of the functions of machine learning is the prediction and classification of data, for which we employ one of the many machine learning models. Are tuning the parameters of models so that their behaviour can be adjusted for a given problem. According to this, a parameter can be described as a configuration variable intrinsic to the model. Multiple experiments are performed to obtain the best parameters. Table IV shows the best parametric settings for models used in the paper.

The Fig. 5 shows the training and testing accuracy score of the classifiers used in this paper.

Fig. 5 shows that the highest training accuracy score was 99% for the SVM model, and the training accuracy score for the LSTM and RNN models was 98%. Finally, the highest testing accuracy is 99% of the SVM model. Accordingly, the SVM model outperformed the rest of the classification models, as the training and testing data ratio was the highest among the classifiers; consequently, the SVM model was the best model performing among the models used.

Accordingly, almost all prior research has used machine learning algorithms to classify the Arabic text into positive, neutral or negative. Since this paper focuses on Arabic reviews in the Saudi dialect, Table V below shows the results obtained by other researchers.

TABLE IV. PARAMETRIC SETTINGS FOR THE USED MODELS

Model	Parametric settings
SVM Model	kernel='rbf'
	degree=3
	gamma=0.5
	C=1.0
	class_weight='balanced'
LSTM Model	dropout_rate=0.002
	embed_dim = 32
	hidden_unit = 160
	optimizers = RMSprop
	epochs=50
	batch_size=1500
RNN Model	dropout_rate=0.002
	embed_dim = 32
	hidden_unit = 160
	optimizers = RMSprop
	epochs=50
	batch_size=1500

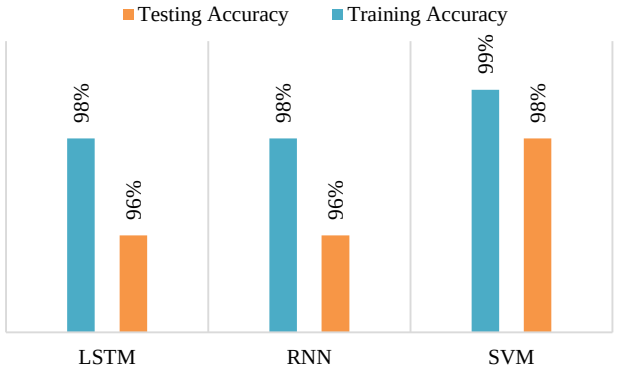

Fig. 5. A Comparison of the Training and Testing Accuracy Scores.

TABLE V. A BENCHMARK OF STUDIES USING SAUDI DIALECT DATASETS

Paper Name	Model Name	Accuracy
Sentiment Analysis of Saudi Dialect Using Deep Learning Techniques [5].	Bi-LSTM	94%
Arabic tweets sentiment analysis – a hybrid scheme [50].	SVM	84%
SDCT: Multi-Dialects Corpus Classification for Saudi Tweets [51].	SGC and LSVM	91.5%
Effect of Saudi dialect preprocessing on Arabic sentiment analysis [52].	KNN	73.3%
Sentiment Analysis of Twitter Data for Saudi Universities [41].	SVM	93.5%
Tourist Reviews Sentiment Classification Using Deep learning Technigues: A Case Study in Saudi Arabia	SVM	98%

Table V shows that all previous studies dealt with the problem of classifying reviews in the Saudi dialect. It also shows that the highest accuracy was 94 for the model Bi-LSTM [5]. Previous studies used the same models used in this paper with varying degrees of accuracy, and that is naturally due to how they processed the data. In this paper, we compared the three models after processing the dataset to see which model gives the highest accuracy results in addressing the problem of sentiment classification in the Saudi dialect. We noticed that SVM gave the best results in different places such as accuracy, precision, recall and F1-score. In addition, although the dataset used in this paper was very poor compared to previous research, the study achieved good results. Most prior research has shown that Twitter, as well as Amazon and Tripadvisor, is the most often utilized dataset source in sentiment categorization.

V. CONCLUSION

There is continued growth in tourists' search for online tourist services. This has led to an increase in the volume of online reviews, but manual reviews are time-consuming and useless. Reading a few reviews provides an incomplete understanding. Accordingly, this paper aimed to build a model capable of accurately sentiment classification in the Saudi dialect for Arabic review in the tourist place reviews using deep learning techniques. A Saudi dialect lexicon was created. The created lexicon was used to classify web scraping reviews from Google Map.

The classification of reviews was limited to (positive, negative, and neutral). This article has also provided: a summary of how-to sentiment categorization. The second section of the study is a literature survey on machine learning approaches. Thirdly, the study introduced three classification algorithms: SVM, LSTM, and RNN. The proposed algorithms' performance was measured and compared through performance metrics such as accuracy, precision, recall, and F1-score. The results showed that SVM achieved the highest classification accuracy of 98%, outperforming RNN and LSTM. All the metrics: Accuracy, Precision, Recall and F1-score high in case of SVM.

Future study will entail locating the optimal method to categorise the Saudi tone with greater precision and gathering

additional tourism ratings to expand the dataset. In addition, more classifications are needed than positive, negative, or neutral as they do not provide sufficient information about the reaction given by people and classify phrases with different emotions such as sadness, neutral, worry, love, fun, hate, happiness, and anger. This would help extract the person's natural feelings and hope to expand our lexicon by adding more slang terms for the Saudi dialect.

REFERENCES

[1] Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A.U. (2017). Sentiment Analysis Using Deep Learning Techniques: A Review. International Journal of Advanced Computer Science and Applications, 8.

[2] Asghar, M.Z., Kundi, F.M., Ahmad, S., Khan, A., & Saddozai, F.K. (2018). T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. Expert Syst. J. Knowl. Eng., 35.

[3] Humphreys, A., & Wang, R.J. (2018). Automated Text Analysis for Consumer Research. Journal of Consumer Research, 44, 1274-1306.

[4] Singh, J., Singh, G., & Singh, R. (2016). A review of sentiment analysis techniques for opinionated web text. CSI Transactions on ICT, 4, 241-247.

[5] Alahmary, R.M., Al-Dossari, H., & Emam, A.Z. (2019). Sentiment Analysis of Saudi Dialect Using Deep Learning Techniques. 2019 International Conference on Electronics, Information, and Communication (ICEIC), 1-6.

[6] Medhat, W., Hassan, A.H., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5, 1093-1113.

[7] Pandey, A.C., Rajpoot, D.S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. Inf. Process. Manag., 53, 764-779.

[8] Chen, W., Xu, Z., Zheng, X., Yu, Q., & Luo, Y. (2020). Research on Sentiment Classification of Online Travel Review Text. Applied Sciences, 10, 5275.

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9, 1735-1780.

[10] Sak, H., Senior, A.W., & Beaufays, F. (2014). Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. ArXiv, abs/1402.1128.

[11] Mustafa Abdullah, D., & Mohsin Abdulazeez, A. (2021). Machine Learning Applications based on SVM Classification A Review.

[12] Guan, Z., Chen, L., Zhao, W., Zheng, Y., Tan, S., & Cai, D. (2016). Weakly-Supervised Deep Learning for Customer Review Sentiment Classification. IJCAI.

[13] Rezapour, M. (2020). Sentiment classification of skewed shoppers' reviews using machine learning techniques, examining the textual features.

[14] Choi, G., Oh, S., & Kim, H. (2020). Improving Document-Level Sentiment Classification Using Importance of Sentences. Entropy, 22.

[15] Catal, C., & Nangir, M. (2017). A sentiment classification model based on multiple classifiers. Appl. Soft Comput., 50, 135-141.

[16] [16] Wawre, S.V., & Deshmukh, S.N. (2016). Sentiment Classification using Machine Learning Techniques.

[17] Godara, N., & Kumar, S. (2019). Opinion Mining using Machine Learning Techniques. International Journal of Engineering and Advanced Technology, 9(2).

[18] Martin, C. A., Torres, J. M., Aguilar, R. M. & Diaz, S. (2018). Using Deep Learning to Predict Sentiments: Case Study in Tourism. Complexity, 2018. DOI: https://doi.org/10.1155/2018/7408431.

[19] Rehman, A.U., Malik, A.K., Raza, B., & Ali, W. (2019). A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. Multimedia Tools and Applications, 1-17.

[20] Litvin, S.W., Goldsmith, R.E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. Tourism Management, 29, 458-468.

[21] Camilleri, A.R. (2017). The Presentation Format of Review Score Information Influences Consumer Preferences through the Attribution of Outlier Reviews. ERN: Marketing (Including Charities & the Private Sector) (Sub-Topic).

[22] Yang, Y., Park, S., & Hu, X. (2018). Electronic word of mouth and hotel performance: A meta-analysis. Tourism Management.

[23] Ernst, D., & Dolnicar, S. (2018). How to Avoid Random Market Segmentation Solutions. Journal of Travel Research, 57, 69 - 82.

[24] Park, D., Kim, S., & Han, I. (2007). The Effects of Consumer Knowledge on Message Processing of Electronic Word of Mouth via Online Consumer Reviews. ECIS.

[25] Gursoy, D. (2019). A critical review of determinants of information search behavior and utilization of online reviews in decision making process (invited paper for 'luminaries' special issue of International Journal of Hospitality Management). International Journal of Hospitality Management.

[26] Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58, 51-65.

[27] Crotts, J.C., Mason, P., & Davis, B. (2009). Measuring Guest Satisfaction and Competitive Position in the Hospitality and Tourism Industry. Journal of Travel Research, 48, 139 - 151.

[28] Mohamed Ali, N., El Hamid, M.M., & Youssif, A.A. (2019). SENTIMENT ANALYSIS FOR MOVIES REVIEWS DATASET USING DEEP LEARNING MODELS. International Journal of Data Mining & Knowledge Management Process.

[29] Gretzel, U., Kyung. Y, & Melanie, P. (2007). "Online Travel Reviews Study: Role and Impact of Online Travel Reviews." Laboratory for Intelligent Systems in Tourism, Texas A&M University.

[30] Abdi, A., Shamsuddin, S.M., Hasan, S., & Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. Inf. Process. Manag., 56, 1245-1259.

[31] Kim, H., & Jeong, Y. (2019). Sentiment Classification Using Convolutional Neural Networks. Applied Sciences.

[32] Katić, T., & Milićević, N. (2018, September). Comparing sentiment analysis and document representation methods of amazon reviews. In 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 000283-000286). IEEE.

[33] Oueslati, O., Cambria, E., Hajhmida, M.B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. ArXiv, abs/2005.12240.

[34] Abuaiadah, D., Rajendran, D., & Jarrar, M. (2017). Clustering Arabic Tweets for Sentiment Analysis. 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), 449-456.

[35] Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B.B. (2018). Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. J. Comput. Sci., 27, 386-393.

[36] Gamal, D., Alfonse, M., El-Horbaty, E.M., & Salem, A.M. (2019). Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features. Procedia Computer Science.

[37] Heikal, M., Torki, M., & El-Makky, N.M. (2018). Sentiment Analysis of Arabic Tweets using Deep Learning. ACLING.

[38] Alosaimi, S., Alharthi, M.A., Alghamdi, K.K., Alsubait, T., & Alqurashi, T. (2020). Sentiment Analysis of Arabic Reviews for Saudi Hotels Using Unsupervised Machine Learning. Journal of Computer Science, 16, 1258-1267.

[39] AL-Jumaili, A. S. (2020). A Hybrid Method of Linguistic and Statistical Features for Arabic Sentiment Analysis. Baghdad Science Journal, 17(1 (Suppl.)), 0385-0385.

[40] Al-Thubaity, A.O., Alqahtani, Q., & Aljandal, A. (2018). Sentiment lexicon for sentiment analysis of Saudi dialect tweets. ACLING.

[41] Alruily, M., & Shahin, O.R. (2020). Sentiment Analysis of Twitter Data for Saudi Universities. International Journal of Machine Learning and Computing, 10, 18-24.

[42] Rizkallah, S., Atiya, A.F., Mahgoub, H.E., & Heragy, M. (2018). Dialect Versus MSA Sentiment Analysis. AMLTA.

[43] Al-Harbi, O. (2019). A Comparative Study of Feature Selection Methods for Dialectal Arabic Sentiment Classification Using Support Vector Machine. ArXiv, abs/1902.06242.

[44] Moudjari, L., & Akli-Astouati, K. (2020). An Experimental Study on Sentiment Classification of Algerian Dialect Texts. KES.

[45] Heamida, I., S, Ahmed, E., S., Mohamed, M., N., Salih, A., A. (2020). Applying Sentiment Analysis on Arabic comments in Sudanese Dialect. International Journal of Computer Science Trends and Technology (IJCST) – Volume 8 Issue 3, May-Jun 2020.

[46] Ibrahim, H.S., Abdou, S., & Gheith, M.H. (2015). Sentiment Analysis For Modern Standard Arabic And Colloquial. ArXiv, abs/1505.03105.

[47] Zhang, Y., Sun, M., Ren, Y., & Shen, J. (2020). Sentiment Analysis of Sina Weibo Users Under the Impact of Super Typhoon Lekima Using Natural Language Processing Tools: A Multi-Tags Case Study. Procedia Computer Science, 174, 478-490.

[48] https://github.com/Banan6/Dataset-Tourist-places-reviews.

[49] Koço, S., & Capponi, C. (2013). On multi-class classification through the minimization of the confusion matrix norm. ACML.

[50] Aldayel, H. K., & Azmi, A. M. (2016). Arabic tweets sentiment analysis–a hybrid scheme. Journal of Information Science, 42(6), 782-797.

[51] Bayazed, A., Torabah, O., AlSulami, R., Alahmadi, D., Babour, A., & Saeedi, K. (2020). SDCT: Multi-Dialects Corpus Classification for Saudi Tweets. methodology, 11(11).

[52] Al-Harbi, W. A., & Emam, A. (2015). Effect of Saudi dialect preprocessing on Arabic sentiment analysis. International Journal of Advanced Computer Technology, 4(6), 91-99.