Post Pandemic Tourism: Sentiment Analysis using Support Vector Machine Based on TikTok Data

Norlina Mohd Sabri*, Siti Nur Athira Muhamad Subki, Ummu Fatihah Mohd Bahrin, Mazidah Puteh

College of Computing, Informatics and Mathematics

Universiti Teknologi MARA Cawangan Terengganu, Kampus Kuala Terengganu, Malaysia

Abstract—The tourism industry is one of the hard hit businesses during the Covid-19 pandemic and has been struggling for backup ever since. However, nowadays the industry has started to bloom again with the lifting of all of the restrictions of Covid-19. This research aims to analyze the sentiments of the tourists using the Support Vector Machine (SVM) algorithm to know their views on the tourist spots after the pandemic. The scope of the research covers the state of Terengganu which is popularly known for its islands and unique culture on the east coast of Malaysia. TikTok data has been used as the source of data as social media currently has become one of the top mediums for reviewing, selling and promoting products and services. The objective of the research is to explore the SVM algorithm in the sentiment classification of tourist spots in Terengganu. This research is expected to help the Tourism Terengganu to improve their tourist spots and their services. The phases of the research include collecting data from TikTok, data pre-processing, data labelling, feature extraction, model creation using SVM, graphical user interface development and performance evaluation. The evaluation results showed that the performance of the SVM classifier model was good and reliable, with 90.68% accuracy. The future work would be collecting more data from TikTok regularly to further improve the accuracy of the algorithm.

Keywords—*Post pandemic tourism; support vector machine; sentiment classification; TikTok data*

I. INTRODUCTION

The Covid-19 pandemic has once brought tourism to stop and become one of the hardest hit businesses in the global economy [1]. However, this industry is starting to bloom again with the lifting of all the Covid-19 restrictions and also the less severe effects of virus infections. In the early days of the opening of tourism, people have been educated about the new normal procedures such as wearing masks, social distancing and sanitizing hands [2]. Since then, people have not been afraid to go anywhere in the world for a holiday getaway. Tourism is now slowly recovering and is back again as one of the country's sources of economic growth. The outbreak of the Covid-19 virus is not the biggest concern nowadays since people can get treatment if infected [3].

Terengganu is one of the states in Malaysia, located on the east coast of the Malaysian peninsular. It is a coastal state, facing the South China Sea with diverse tourist attractions such as natural tourism, cultural tourism and marine tourism. Terengganu is renowned for its islands, stunning beaches, and abundant marine life. Since the opening of the tourism businesses, social media has grown to be a very powerful information source for tourists to share their experiences. Social media such as Facebook, TikTok, Twitter and Instagram allow tourists to describe their own experience with the hotel, restaurants, and other tourist attractions. These shared sentiments have a big impact on local tourism.

Sentiment analysis (SA), also known as opinion mining is the computation of people's opinions, judgments and emotions through entities, events and attributes held by the users [4]. Sentiment analysis or opinion mining uses natural language processing and text analytics to locate and extract subjective information from source materials. To determine whether a statement indicates a positive or negative opinion towards the subject, sentiment analysis is frequently used to extract sentiments, opinions, and subjectivity from texts [5].

The ability of locals and visitors to express their sentiments is essential for the development of tourism. The expressed opinions and emotions in the reviews from social media could be extracted and analysed for the improvement of the business. Summary findings from sentiment analysis will aid tourists in selecting their tour destination and itinerary [6]. Tourists may do an information search to select the right location, which can be difficult due to the abundance of options and information on the Internet [7]. Sentiment analysis could be used to gather feedback for any tourist spots, thus helping people to choose the right vacations for them.

Based on these motivations, this research has proposed the sentiment analysis for Terengganu tourist spots after the pandemic using TikTok data. TikTok has been chosen as it has become the top medium for today's online business and socialization. TikTok is one of the social media platforms that have the power to spread news and information. Moreover, many travel and tourism companies nowadays use TikTok to promote their tourist attractions or activities. Reviews from TikTok could be classified into positive or negative sentiments. This sentiment analysis is based on the machine learning approach and Support Vector Machine (SVM) has been chosen as the classifier. SVM has proven to be able to produce good performance in sentiment classification problems [8] [9] [10] [11]. The objective of the research is to explore the capability of SVM in the classification of Terengganu's tourist spot reviews after the Covid-19 pandemic using TikTok data. The analysis results are expected to help the Tourism Agency and also the tourists to know the current conditions of the tourist spots in Terengganu, especially after the Covid-19 pandemic. This paper is arranged into five main sections which are the Introduction in Section I, Brief Literature Review in Section II, Material and Method in Section III, Result and Discussion in

Section IV and finally paper is concluded in Section V. The Brief Literature Review section provides explanation on SVM, its advantage, limitation and the previous works that have implemented the algorithm.

II. BRIEF LITERATURE REVIEW

A. Support Vector Machine (SVM)

Support Vector Machine is a global classification algorithm under the supervised learning method. SVM uses the hyperplane, which separates new inputs and produces the output [12]. The basic concept of SVM is to identify the optimum hyperplane that divides two different classes which are positive and negative classes. A separator in a d-dimensional space known as a hyperplane has d-1 dimensions. The data points closest to the hyper-plane, known as support vectors, have an impact on the hyper-position planes and orientation. The margin, or the distance between the support vectors and the hyperplane, must be maximal for the hyperplane that has been chosen. The hyper-plane can be altered by even a slight interference in the location of these support vectors. There are different types of kernel functions in SVM. The kernel's job is to accept input data and transform it into the required shape. The kernels are Linear, Polynomial and Radial Basis Function (RBF). Each of the kernels has its formula based on their concepts [13].

The advantage of SVM is that the binary classifier SVM is very efficient and has the benefit of being able to categorize with a minimal amount of information [14]. However, SVM also has its limitations such as being time-consuming when used with big amounts of data. Also, this method must be modified to classify data into more than two classes since it was designed to classify with only two classes [15]. In this research, SVM has been selected due to its suitability to process an ample amount of data and only two classes are needed in the sentiment classification.

B. Implementation of SVM Algorithm in Various Problems

SVM has been implemented in various classification problems and the results have proven to be promising. Reference [16] uses SVM, K-Nearest Neighbour (KNN) and Naive Bayes algorithm in the classification of wheat grain. This study aimed to determine the most discriminatory features and a suitable classifier that may classify the given wheat sample into classes 'fresh' or 'rotten'. Since wheat is the body's principal source of energy in the form of protein, it needs to be stored with a good storage management system. The results showed that the SVM classifier outperformed other classifiers by achieving an accuracy of 93%. Another implementation of SVM was the analysis of lung cancer classification using multiple feature extraction with SVM and KNN [17]. The lung cancer was a result from the unchecked cell proliferation in a lung area. This project aimed to classify the lung CT images as normal or damaged. The results showed that SVM has achieved the highest accuracy of 96.42%.

Reference [18] has conducted a research on movie recommendation and sentiment analysis using Naïve Bayes and SVM. This project aimed to perform sentiment analysis on the movie's reviews and to deal with the vast volume of data. The overall accuracies have shown that SVM has achieved 98.63% whereas the accuracy of Naïve Bayes was 97.33%. Reference [19] has conducted a research on sentiment analysis based on the reviews of the smartphone. There are so many smartphone products on the market which provide high-efficiency features and customers were more likely to write reviews about them. The results of this research showed that SVM has produced 79.5% accuracy. Reference in [20] has conducted a research to classify the online class student feedback in new semester. An analytical approach is required to learn the feelings of the students as they begin the new semester with online learning. In the research, the SVM method has obtained a good accuracy of 84%. Based on the previous problems, SVM has generated good performance in sentiment classification problems. Based on the algorithm's capability, it is worth exploring the algorithm in this classification problem.

III. MATERIAL AND METHOD

A. Experimental Data

In this research, the TikTok application has been utilized to gather information for identifying tourist attractions in Terengganu. There are many tourist attractions in Terengganu such as the islands of Redang, Perhentian, Kapas, Tenggol and also on the mainland which are Pasar Payang, Taman Tamadun Islam, Batu Buruk Beach and the Terengganu State Musem. The data was collected from TikTok by using Chrome Developer Console and JavaScript has been used to automate data extraction from the TikTok website. The search keys were "Amazing Terengganu", "Tourism Terengganu", "Terengganu best places" and "Terengganu aesthetic". All those keys were searched one by one, and the data scrapped was saved into the .csv file format. The data were scrapped from March to July 2023. A total of 1311 rows of reviews had been scrapped from TikTok and the data contained 11 attributes which are comment number (id), nickname, user@, user URL, comment, time, likes, profile picture URL, 2nd level comment, the user replied to and number of replies. Since this project was only focused on analysing the comments made by the user, all other attributes were discarded during the pre-processing stage.

B. Data Pre-processing

The data pre-processing is an essential part of natural language processing when it comes to text classification [21]. The unstructured data from TikTok were processed in the next stage via the data pre-processing techniques. The steps of the TikTok data pre-processing are removing duplicated data, lowercase conversion, removing TikTok mention and punctuation, tokenization and stop words removal, POS-tags labelling and lemmatization.

The first step is to remove the duplicated data from the raw dataset. The number of the raw data was 1311, which has been reduced to 1305 when all duplicated data were deleted. For lowercase conversion, the lower() method has been used to convert all of the comments to lowercase. After that, the hashtag symbol (#), user handles (@) and non-letter characters were then removed from the remark by replacing them with a blank string. These TikTok mentions and punctuation was eliminated to avoid interfering later with the main process. The next step is the tokenization and the stop words removal, which

is necessary to improve the readability and transformability during feature extraction. Tokenization divides all text sentences into smaller pieces called tokens. Following tokenization, stopwords removal was applied to the clean dataset, removing stopwords from the NLTK package. The lambda function is used to eliminate the stopwords. Then, the POS tagging established the word class based on the word's placement in the sentences, indicating whether the word was a noun, adjective, verb, and so on, to allow for future lemmatization use. The POS tags of a word are necessary for correctly obtaining the word's lemma. Finally, the lemmatization step was applied to the dataset to produce meaningful root words. Lemmatization was chosen over stemming because it generated better results by analysing the word's portion and constructing actual dictionary words.

Fig. 1 shows the outcome of the data pre-processing steps, which demonstrates that most of the data have been adequately cleaned.

Fig. 1.   Data pre-processing results.

C. Data Labelling

After data cleaning, the data must be labelled to train the classifier model. Textblob is used in this project to label the text polarity to determine the text sentiment. Positive sentiment and negative sentiment are the two sentiment classes that have been assigned to the user review data. The positive class classification was based on comments that have supportive, agreeable, and positive-sounding terms. The negative class was based on user complaints or unhappiness with the situation. The dataset has been cleaned up of neutral remarks and comments that have nothing to do with the application's opinion [21].

The polarity score ranges from [-1.0,1.0] where a negative statement receives a score of -1 and a positive statement receives a score of 1. The polarity value for this project was between 1 and -1, where 1 stands for "Positive" and -1 for "Negative". Table I shows the data labelling in this research. Fig. 2 shows the result of text labelling. After removing the neutral sentiment texts, the total number of data was reduced from 1304 to 1178. This numerical labelling was done to facilitate further processing that requires numerical labels instead of textual labels.

TABLE I.   DATA LABELLING

Polarity	Data Labelling	Sentiment Class
More than 0	1	Positive Review
Equal or less than 0	-1	Negative Review

	lemmatized	label	polarity
0	['fun', 'long', 'beach', 'fireworks', 'show', ...	positive	0.18
1	['fun', 'perhentian', 'island', 'long', 'beach...	negative	-0.01
2	['yesterday', 'pay', 'rain', 'continuously', '...	negative	-0.20
3	['reserve', 'table', 'may', 'slightly', 'expen...	negative	-0.50
4	['fun', 'beach']	positive	0.30
...	...	...	...
1174	['sultan', 'zainal', 'abidin', 'museum', ':', ...	negative	-0.04
1175	['saw', 'garbage', 'beach', 'make', 'difficult...	negative	-0.05
1176	['crowd', 'large', 'difficult', 'find', 'place...	negative	-0.14
1177	['garbage', 'major', 'turnoff', 'suggest', 'ga...	negative	-0.32
1178	['price', 'high', 'suggests', 'price', 'expens...	negative	-0.01

Fig. 2.   Sample of labelled dataset.

D. Feature Extraction

Feature extraction helps identify characteristic sentences in tourist attractions, review data and turn them into features. The ability to produce performance in machine learning is determined by the features employed, hence the feature extraction stage was regarded as being of utmost importance [22]. The technique used for feature extraction is the TF-IDF (Term Frequency-Inverse Document Frequency) score. The TF-IDF technique is an unsupervised feature extraction method that functions at the level of a language's words or lexicon. The review text's word items are converted into numerical data using TF-IDF, making it simple for the following machine learning approach to build the classification model [23]. The term frequency (TF), inverse document frequency (IDF), and the TF-IDF score are the three calculations used to assess the significance of each word inside the data. Eq. (1) to Eq. (3) represent the three equations respectively. The TF-IDF values for each word within the dataset are collected in the corpus, creating a vocabulary of unique words. Words with high TF-IDF values are considered more important and distinctive. These words are often indicative of the specific content or theme of a document. Identifying these key features helps in understanding the focus of each document. Fig. 3 shows the sample of words and their corresponding TF-IDF values.

- Term Frequency (TF)

Using TF, one may determine how many terms are contained in a document.

$$Term\ Frequency,\ TF = \frac{Frequency\ of\ word\ (W)\ appearing\ in\ a\ document\ (D)}{Total\ number\ of\ words\ (W)\ in\ document\ (D)} \quad (1)$$

- Inverse Document Frequency (IDF)

The Inverse Document Format Frequency (IDF) gives the text's uncommon words priority.

$$IDF = log_e \frac{\text{Total number of documents (D)}}{\text{Total number of document containing word (W)}} \qquad (2)$$

- TF-IDF Score

The TF-IDF score for each word is calculated. Higher-scoring words are thought to be more important, whereas lower-scoring words are thought to be less relevant.

$$TF\text{-}IDF\ (W) = TF(W) \times IDF\ (W) \qquad (3)$$

```
bay: 0.0
bazaar: 0.0
bbq: 0.0
bbqs: 0.0
beach: 0.15795974963882325
beachgoers: 0.0
beautiful: 0.0
beautifulattractive: 0.0
beautifully: 0.0
beautifulthat: 0.0
beautifulwe: 0.0
beauty: 0.0
become: 0.0
becomes: 0.0
bed: 0.0
beer: 0.0
beginner: 0.0
behavior: 0.0
```

Fig. 3. System architecture of SVM classifier model.

E. System Architecture

The phases of the research include collecting data from TikTok, data pre-processing, data labelling, feature extraction, model creation using the Support Vector Machine (SVM) method, graphical user interface development and performance evaluation. Fig. 4 depicts the system architecture for the SVM classifier model. The initial phase involves data collection, where data was collected from TikTok using the Chrome Developer Console. Data pre-processing requires several phases, which include duplicate data removal, case conversion, punctuation removal, word expansion, hashtag removal, short word removal, tokenization, stop word removal, POS tag labelling, and lemmatization. The next phase is the data labelling with Text Blob, where the entire dataset will be labelled as "Positive Review" or "Negative Review." Positive reviews are labelled with the value 1, while negative reviews are labelled with the value -1. In the feature extraction phase, the dataset will be vectorized using TF-IDF technique. This is to convert the text data into numerical form so that it can be processed by the algorithm. The dataset will then be separated into two datasets, training and testing using the hold-out method. The training data is used to train the SVM classifier, while the test data is used to evaluate the performance of the algorithm. The values of accuracy, recall, precision, F1-Score and AUC are measured during the performance evaluation. The classifier model is then integrated with the graphical user interface to be used by the end user. The graphical user interface enables the system to collect user input, process the text, classify the text using the model and display the output to the user.
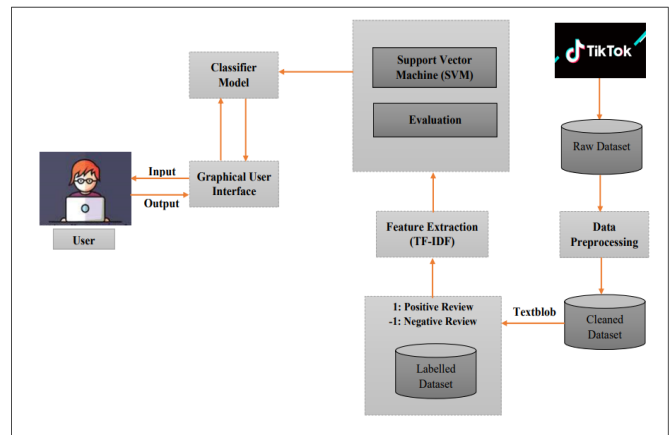
Fig. 4. System architecture of SVM classifier model.

F. Performance Evaluation

After building the classification model, it is important to undertake a performance evaluation to determine how accurately the suggested model produces the classification. To help the model perform better, several performance metrics could be applied. These sections explain the confusion matrix, which includes the F1 scores, accuracy, precision, recall, and ROC curve.

1) Confusion matrix: Accuracy, precision, recall, and F1 score values are computed using the confusion matrix. How accurately the model can classify objects is described by its accuracy. Precision is the proportion of correctly made positive forecasts to all correctly made positive predictions. The next step is to determine how many of the positive groups that were accurately predicted are supported by the data. The capacity to determine the number of data that would count positively for a certain attribute is known as recall. The harmonic mean between recall and precision is the definition of the F1 score. The F1 score is a statistical metric used to evaluate the average performance, depending on precision and recall [24]. A confusion matrix presents a matrix-style summary of the data set's entries based on the two standards of actual value and predicted value. The matrix's columns reflect the expected values, whereas the rows of the matrix represent the true values [25].

Confusion matrix is a technique for evaluating how well classifiers perform at each label level. It can compute the F1-measure, recall rate, precision, and accuracy rate [26]. True Positive (TP) represents the number of positive tourist attraction reviews that are correctly predicted while False Positive (FP) represents the number of positive tourist attraction reviews that are predicted as negative by the classifier. True Negative (TN) is the number of negative reviews correctly predicted and False Negative (FN) is the number of negative reviews predicted as positive by the classifier. Eq. (4) to Eq. (7) represents the Accuracy Rate, Precision, Recall Rate and F1-Measure.

- Accuracy Rate

The accuracy ratio is the number of correct samples divided by the total number of samples. The Eq. (4) below can be used to calculate the classifier's accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

- Precision

Precision shows the percentage of predictions for this kind of result that were accurate. The accuracy of the model prediction increases with increasing value. The Eq. (5) below can be used to calculate the precision.

$$Precision = \frac{TP}{TP+FP} \qquad (5)$$

- Recall Rate

Recall is calculated as the ratio of the number of positive predictions to the number of positive class values in the test data. The accuracy of the classifiers is measured by recall, and it can be calculated with Eq. (6).

$$Recall = \frac{TP}{TP+FN} \qquad (6)$$

- F1-Measure

F1-Measure effectively communicates the harmony of memory and precision. The harmonic mean of recall and precision is known as the balanced F1-score. F1-Measure obtained from the Eq. (4) below:

$$F1\text{-}measure = \frac{2*Precision*Recall}{Precision+Recall} \qquad (7)$$

2) Receiver Operating Characteristic (ROC): One of the model measurement metrics is the receiver operating curve or ROC. The AUC stands for the area under the ROC curve. The performance of the model is improved by a higher AUC. It utilizes positive and negative numbers, demonstrating its capacity for classification [27]. The corresponding coordinates for the threshold's highest value are (0, 0), and for the threshold's smallest value are (1, 1) [25].

IV. RESULT AND DISCUSSION

In this research, the first analysis conducted was the exploratory data analysis on the collected TikTok data. The second analysis was on the performance of the Support Vector Machine Classifier. The second analysis covers the accuracy, Confusion Matrix and ROC results.

A. Exploratory Data Analysis

The labelling of the data has resulted in more positive comments than negative ones. Fig. 5 shows the number of positive comments (773) is larger than the number of negative comments (407). Some of the positive comments include beautiful beaches, nice scenery and good food in Terengganu. The most mentioned tourist spots were Pulau Redang, Pulau Perhentian, Pantai Batu Buruk, Pasar Payang, Taman Tamadun Islam, Masjid Kristal, Kuala Ibai and Muzium Negeri. There were also comments on the improvement of some of the tourist spots after the Covid-19 movement control order. The negative comments were mostly about the cleanliness of certain areas of

the tourist spots. At certain tourist spots such as beaches, some irresponsible tourists were being negligent on environmental cleanliness by throwing rubbish everywhere and dirtying the places. Overall, the tourists were happy to have their holidays at Terengganu. They seemed not bothered about the Covid-19 virus which still existed in this post-pandemic. This might be because people were tired and could not care less after being in the movement control order for almost two years and fortunately the virus had also evolved to be more benign.

After processing the numeric information in the dataset, word clouds were generated for two categories: positive and negative. Word clouds are commonly used to visualize and analyse qualitative data. In this case, the comment text from the "stopword_removed" column is utilized to create the word clouds. The purpose of these word clouds is to gain insights into the main topics being discussed. Fig. 6 shows the positive word cloud, highlighting the most prominent words mentioned in TikTok comments about tourist attractions in Terengganu. The keywords include "beautiful," "Terengganu," "place," "best," "clean," "beach," "good," "view," and "island." These words suggest that users frequently mentioned these positive aspects of the tourist attractions in Terengganu in their comments.
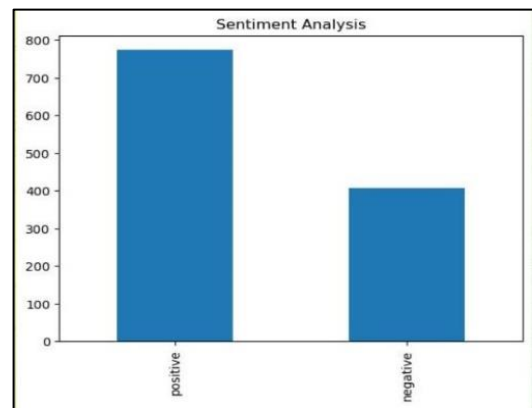
Fig. 5.   Number of positive and negative comments.

On the other hand, Fig. 7 displays the negative word cloud. The words "dirty," "Terengganu",”place”, "beach," and "difficult" are considered negative, indicating that people were dissatisfied with the environmental conditions at some of the tourist attractions in Terengganu. Perhaps proper management of trash and also warning or fine should be imposed on tourists who throw rubbish at improper places.

Fig. 6.   Positive word cloud.

Fig. 7. Negative word cloud.

B. Support Vector Machine Classifier Performance Evaluation

This section provides the evaluation result for the performance of the Support Vector Machine (SVM) model that has been developed from scratch. The evaluation covers SVM accuracy, confusion matrix and the ROC curve.

1) SVM accuracy: In this research, the holdout method was used, and the dataset was split into the training and testing sets based on three different percentage splits: 80:20, 70:30 and 60:40. The accuracy results for each split are shown in Table II. Among the splits, the 80:20 split achieved the highest accuracy of 90.68%. Based on this result, the 80:20 split was selected for further model development since it produced the highest accuracy. Fig. 8 shows the comparison of accuracies among the data splits. In this research, the accuracy result was getting better with more data used for training. The accuracy result of 90.68% has shown that the SVM model is good and acceptable as the sentiment classifier. This is also on par with other SVM performances, as the results of SVM in other research have also produced more than 90% accuracy [16] [18] [17]. In future, the accuracy is expected to be improved if more data are scrapped and anayzed.

2) Comparison between the similar works: This section presents the accuracy of algorithms that have been implemented in similar works, which are tourism-related sentiment analysis. Table III shows the accuracy results for each of the research, using RNN, CNN, LTSM and the proposed SVM models. Based on the table, the proposed SVM model has achieved an accuracy of 90.68% which is higher than the reported RNN (80%) and LSTM (84%) accuracies. Based on this comparison, the accuracy result of the SVM model has proven to be good and able to achieve higher accuracy in its problem compared to certain deep learning models.

3) Confusion matrix: Fig. 9 shows the confusion matrix plot to illustrate the value of True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) for negative (0) and positive (1) labels. Confusion matrix is

the calculation of the correct and wrong predictions, which gives an insight into the error made by the classifier model and the type of error being made. Based on Fig. 9, 73 is the value of TP, which indicates the number of positive tourist attraction reviews correctly predicted, while the FP is 13, which was the positive tourist attraction reviews predicted as negative by the classifier. The value of FN is 9, which is the number of negative reviews predicted as positive by the classifier. Lastly, the TN value is which indicates 141 the number of negative reviews correctly predicted by the prototype. From the results, it can be observed that the model has made 214 correct predictions out of 236 predictions made. From the confusion matrix plot, it could be seen that the model has succeeded in predicting 90.68% of data in this sentiment classification problem.

TABLE II. DATA SPLIT RESULTS

Data Split	Training Data	Testing Data	Accuracy (%)
60:40	708	472	88.35%
70:30	826	354	89.27%
80:20	944	236	90.68%

Fig. 8. Accuracy comparison among the data splits.

TABLE III. COMPARISON OF ACCURACY BETWEEN SIMILAR WORKS

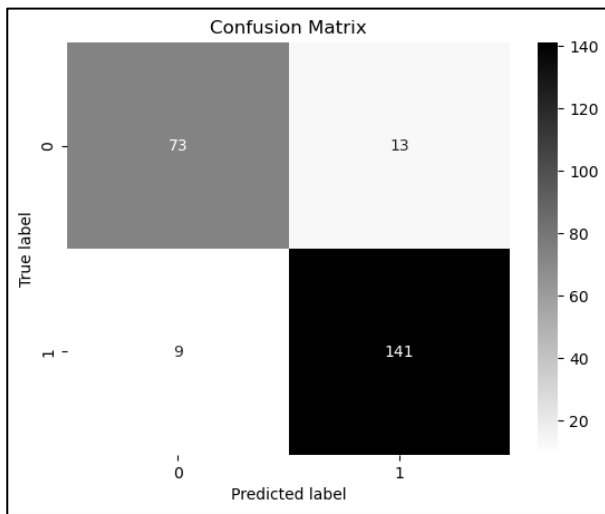| Title	Algorithm	Accuracy	Reference
Sentiment Analysis in Reviews About Beaches in Bali on Tripadvisor Using Recurrent Neural Network (RNN)	Recurrent Neural Network (RNN)	80%	[28]
Convolutional Neural Networks for Indonesian Aspect-Based Sentiment Analysis Tourism Review	Convolutional Neural Network (CNN)	95.22%	[29]
LSTM-based Deep Learning Architecture of Tourist Review in Tripadvisor	Long Short Term Memory (LSTM) method	84%	[30]
Proposed SVM model	Support Vector Machine (SVM)	90.68%	-

Fig. 9.   Confusion matrix plot.

Fig. 10 shows the classification report of the accuracy, precision, recall and F1-score for the Support Vector Machine (SVM) classifier model. The average precision obtained was 0.92, which indicates how many instances the model correctly predicted out of all the instances. For recall, the average value is 0.94, indicating the instances that the model correctly predicted in the particular class. Finally, the average F1-score is 0.93, which represents the weighted average of precision and recall. The F1-score value has shown the good and acceptable performance of the model in correctly identifying both positive and negative instances.


```
Classification Report:
              precision    recall  f1-score   support

         -1       0.89      0.85      0.87        86
          1       0.92      0.94      0.93       150

   accuracy                          0.91       236
  macro avg       0.90      0.89      0.90       236
weighted avg      0.91      0.91      0.91       236
```
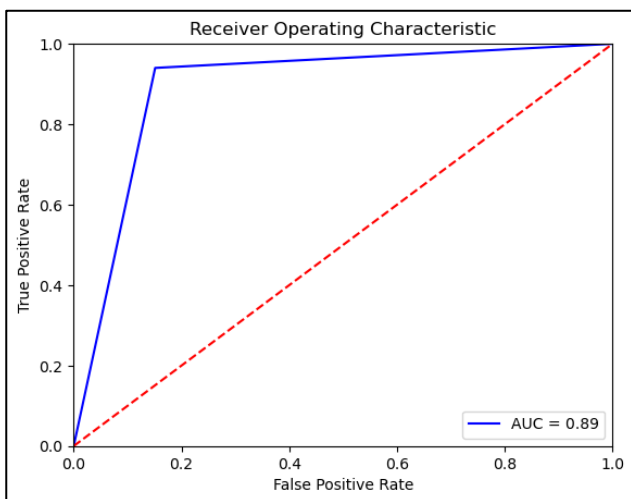
Fig. 10. Classification report.

Fig. 11. ROC curve.

4) Receiver Operating Characteristic (ROC) Curve: Fig. 11 displays the ROC curve for the Support Vector Machine (SVM) classifier model. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate represents the proportion of positive observations correctly identified as positive out of all positive observations (TP/ (TP + FN)). The ROC curve's position closer to the upper left corner indicates a more effective classification of data into categories by the model. To quantify the area covered by the curve, the AUC (area under the curve) value was used. For the SVM model, the AUC calculated was 0.89, which indicates a good performance in correctly classifying data points. The AUC value could be improved more with the improvement of the true positive rate.

V.   CONCLUSION

This research has met its objective in exploring the capability of SVM in the sentiment classification of the tourist spots in Terengganu after the Covid-19 pandemic era. The SVM model has produced good and reliable performance in this sentiment classification problem with an accuracy of 90.68%. In this research, the SVM model has successfully classified the tourists' sentiments and it was found that the reviews from TikTok were mostly positive about the tourist spots in Terengganu in the post pandemic. People were not afraid anymore of the Covid-19 virus and were mostly positive about coming to Terengganu for getaways. The effects of the Covid-19 virus have gradually become mild and nowadays people can do their treatments if affected. The research findings can be used by the tourism industry in Terengganu to improve the tourist spots and their services. This research also provides informed decisions about which places to visit in Terengganu, enabling tourists to choose their ideal destinations and have a memorable vacation experience. Positive feedback could significantly influence users' final decisions and make their vacation planning process easier, ensuring a smooth and well-organized trip. The recommendation for future work is to establish automated data collection techniques. This would allow the system to regularly scrap the TikTok data, enabling the model to be trained with the latest and relevant data. It is expected that this future work could further improve the classifier accuracy in the sentiment classification of the tourist spots in Terengganu. Moreover, the SVM classifier performance would also be compared with other classification algorithms such as the Naive Bayes and other deep learning algorithms.

REFERENCES

[1]   J. A. Salinas Fernández, J. M. Guaita Martínez, and J. M. Martín Martín, "An analysis of the competitiveness of the tourism industry in a context of economic recovery following the COVID19 pandemic," Technol.

Forecast. Soc. Change, vol. 174, 2022, doi: 10.1016/j.techfore.2021.121301.

[2] S. H. A. Samsudin, N. M. Sabri, N. Isa, and U. F. M. Bahrin, "Sentiment Analysis on Acceptance of New Normal in COVID-19 Pandemic using Naïve Bayes Algorithm," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 9, pp. 581–588, 2022, doi: 10.14569/IJACSA.2022.0130968.

[3] N. M. Sabri, J. N. Azman Norman, N. Isa, and U. F. Mohd Bahrin, "Sentiment Analysis on Covid-19 Outbreak Awareness Using Naïve Bayes Algorithm," in IVIT 2022 - Proceedings of 1st International Visualization, Informatics and Technology Conference, 2022, pp. 278–283, doi: 10.1109/IVIT55443.2022.10033379.

[4] C. Steven and W. Wella, "The Right Sentiment Analysis Method of Indonesian Tourism in Social Media Twitter Case Study: The City of Bali," Int. J. New Media Technol., vol. 7, no. 2, 2020, doi: https://doi.org/10.31937/ijnmt.v7i2.1732.

[5] N. H. Alharbi and J. H. Alkhateeb, "Sentiment Analysis of Arabic Tweets Related to COVID-19 Using Deep Neural Network," 2021, doi: 10.1109/ICOTEN52080.2021.9493467.

[6] A. A. Wadhe and S. S. Suratkar, "Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques," 2020, doi: 10.1109/I4Tech48345.2020.9102673.

[7] T. Alenezi and S. Hirtle, "Normalized Attraction Travel Personality Representation for Improving Travel Recommender Systems," IEEE Access, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3178439.

[8] Helen and R. Kurniawan, "Sentiment Analysis of The Tourist Destination Using Support Vector Machine Algorithm on Twitter Post," in Proceedings of 2023 International Conference on Information Management and Technology, ICIMTech 2023, 2023, pp. 316–321, doi: 10.1109/ICIMTech59029.2023.10277866.

[9] M. I. Alhari, O. N. Pratiwi, and M. Lubis, "Sentiment Analysis of The Public Perspective Electric Cars in Indonesia Using Support Vector Machine Algorithm," in 2022 International Conference of Science and Information Technology in Smart Administration, ICSINTESA 2022, 2022, pp. 155–160, doi: 10.1109/ICSINTESA56431.2022.10041604.

[10] R. N. Satrya, O. N. Pratiwi, R. Y. Farifah, and J. Abawajy, "Cryptocurrency Sentiment Analysis on the Twitter Platform Using Support Vector Machine (SVM) Algorithm," in Proceedings - International Conference Advancement in Data Science, E-Learning and Information Systems, ICADEIS 2022, 2022, pp. 2–6, doi: 10.1109/ICADEIS56544.2022.10037413.

[11] D. A. Kristiyanti, R. Aulianita, D. A. Putri, L. A. Utami, F. Agustini, and Z. I. Alfianti, "Sentiment Classification Twitter of LRT, MRT, and Transjakarta Transportation using Support Vector Machine," in 2022 International Conference of Science and Information Technology in Smart Administration, ICSINTESA 2022, 2022, pp. 143–148, doi: 10.1109/ICSINTESA56431.2022.10041651.

[12] A. Yekurke, G. Sonawane, H. Chavan, P. Thote, and A. Phapale, "Sentiment Analysis using Naïve Bayes, CNN, SVM," Int. Res. J. Eng. Technol., vol. 9, no. 4, pp. 694–698, 2022, [Online]. Available: www.irjet.net.

[13] A. F. Rochim, K. Widyaningrum, and D. Eridani, "Performance Comparison of Support Vector Machine Kernel Functions in Classifying COVID-19 Sentiment," in 2021 4th International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2021, 2021, pp. 224–228, doi: 10.1109/ISRITI54043.2021.9702845.

[14] P. D. Poornima and P. Nath Singh, "Masked Unmasked Face Recognition Using Support Vector Machine Classifier," in 2021 IEEE International Conference on Mobile Networks and Wireless Communications, ICMNWC 2021, 2021, pp. 1–4, doi: 10.1109/ICMNWC52512.2021.9688542.

[15] N. Hasanati, Q. Aini, and A. Nuri, "Implementation of Support Vector Machine with Lexicon Based for Sentiment Analysis on Twitter," in 2022 10th International Conference on Cyber and IT Service Management, CITSM 2022, 2022, pp. 1–4, doi: 10.1109/CITSM56380.2022.9935887.

[16] D. Agarwal, Sweta, and P. Bachan, "Machine learning approach for the classification of wheat grains," Smart Agric. Technol., vol. 3, p. 100136, 2023, doi: 10.1016/j.atech.2022.100136.

[17] S. S. Ashwini, M. Z. Kurain, and M. Nagaraja, "Performance Analysis of Lung Cancer Classification using Multiple Feature Extraction with SVM and KNN Classifiers," in 2021 IEEE International Conference on Mobile Networks and Wireless Communications, ICMNWC 2021, 2021, pp. 1–4, doi: 10.1109/ICMNWC52512.2021.9688404.

[18] N. Pavitha, V. Pungliya, A. Raut, R. Bhonsle, and A. Purohit, "Movie recommendation and sentiment analysis using machine learning," Glob. Transitions Proc., vol. 3, pp. 279–284, 2022, doi: 10.1016/j.gltp.2022.03.012.

[19] K. Chitra, T. M. Saravanan, S. N. Prasath, G. Robin, and N. K. S. Babu, "Sentiment Analysis on Smartphone Using Support Vector Machine," in 2022 International Conference on Computer Communication and Informatics, ICCCI 2022, 2022, vol. 3, pp. 1–6, doi: 10.1109/ICCCI54329.2022.9740882.

[20] C. Kurniawan and F. Wahyuni, "Sentiment analysis of online learning students feedback for facing new semester: A support vector machine approach," in Proceedings - 2021 7th International Conference on Education and Technology, ICET 2021, 2021, pp. 1–6, doi: 10.1109/ICET53279.2021.9575116.

[21] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, "Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm," in 2021 International Seminar on Machine Learning, Optimization, and Data Science, ISMODE 2021, 2022, pp. 163–167, doi: 10.1109/ISMODE53584.2022.9742906.

[22] F. E. Cahyanti, Adiwijaya, and S. Al Faraby, "On the Feature Extraction for Sentiment Analysis of Movie Reviews Based on SVM," 2020, doi: 10.1109/ICoICT49345.2020.9166397.

[23] J. C. Sugitomo, N. Kevin, N. Jannatri, and D. Suhartono, "Sentiment Analysis using SVM and Naïve Bayes Classifiers on Restaurant Review Dataset," in Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021, 2021, vol. 1, pp. 100–108, doi: 10.1109/ICCSAI53272.2021.9609776.

[24] A. Setiawan and V. C. Mawardi, "Android Application For Analysis Review On Google Playstore Using Support Vector Machine Method," in ICOIACT 2022 - 5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era, Proceeding, 2022, pp. 331–336, doi: 10.1109/ICOIACT55506.2022.9972122.

[25] Y. L. Zhang, J. ping Su, S. gang Cui, J. yu Zhang, and X. li Wang, "Haematococcus pluvialis cell based on support vector machine Classification research," in Proceeding - 2021 China Automation Congress, CAC 2021, 2021, pp. 7255–7258, doi: 10.1109/CAC53003.2021.9727504.

[26] Y. Huang, R. Wang, B. Huang, B. Wei, S. L. Zheng, and M. Chen, "Sentiment Classification of Crowdsourcing Participants' Reviews Text Based on LDA Topic Model," IEEE Access, vol. 9, pp. 108131–108143, 2021, doi: 10.1109/ACCESS.2021.3101565.

[27] R. Tanna and T. Sharma, "Binary Classification of Melanoma Skin Cancer using SVM and CNN," in Proceedings - 2021 1st IEEE International Conference on Artificial Intelligence and Machine Vision, AIMV 2021, 2021, pp. 2021–2024, doi: 10.1109/AIMV53313.2021.9670894.

[28] N. K. Shadrina, E. Sutoyo, and V. P. Widartha, "Sentiment Analysis in Reviews About Beaches in Bali on Tripadvisor Using Recurrent Neural Network (RNN)," in Proceedings - 2021 IEEE 7th Information Technology International Seminar, ITIS 2021, 2021, pp. 1–6, doi: 10.1109/ITIS53497.2021.9791501.

[29] R. A. N. Nayoan, A. Fathan Hidayatullah, and D. H. Fudholi, "Convolutional Neural Networks for Indonesian Aspect-Based Sentiment Analysis Tourism Review," in 2021 9th International Conference on Information and Communication Technology, ICoICT 2021, 2021, pp. 60–65, doi: 10.1109/ICoICT52021.2021.9527518.

[30] [A. Ramadhani, E. Sutoyo, and V. P. Widartha, "LSTM-based Deep Learning Architecture of Tourist Review in Tripadvisor," in 2021 6th International Conference on Informatics and Computing, ICIC 2021, 2021, pp. 1–6, doi: 10.1109/ICIC54025.2021.9632967.