Available online at: https://journal.poltekpar-nhi.ac.id/index.php/jk/index
Jurnal Kepariwisataan: Destinasi, Hospitalitas dan Perjalanan
Volume 8 Nomor 2, 2024:178-196
DOI: 10.34013/jk.v8i2.1619

Analyzing Instagram Engagement to Forecast Domestic Tourist Trips in Lake Toba and North Sumatra: A Dual Approach with Conventional Statistics and Machine Learning Techniques

Erna Nurmawati*[1], Teguh Sugiyarto[2], Navika Artiari[3], Adelina Rahmawati[4]
Politeknik Statistika STIS[1,3,4]
Badan Pusat Statistik[2]

Abstract

The tourism industry is well known as one booster for economic development. The advance of the tourism industry will lead to the improvement of other economic sectors. Therefore, the Indonesian government is taking steps to ensure the development of its tourism industry by launching 10 super-priority destinations (DSP). Despite numerous efforts and interventions, evidence suggests that the demand for the tourism industry in certain DSPs remains unsatisfied. This also holds true for Lake Toba in North Sumatra. Therefore, it is important to understand how to promote the destination site effectively and increase the number of domestic visitors. This study is aimed at assessing the impact of digital marketing through Instagram to determine the number of domestic tourist trips. The engagement rate (ER) on Instagram posts represents the impact of digital marketing. The result reveals that the topic 'cultural tourism and its activities that develop the economy' has the highest average ER, reaching 692.48. Further analysis reveals that the LSTM model, with independent variables TPK, GTI, and ER on the topic of 'ticket information and vacation packages', is the most effective model for predicting the number of domestic tourist trips to North Sumatra. This analysis emphasizes the crucial role of digital marketing to shape the demand for the tourism industry. The conclusion is based on the significant influence of the Google Trends Index (GTI) and ER on Instagram posts, which serve as a gauge for domestic visitor numbers. The related stakeholders must consider this aspect to sustain its business.

Keywords: Instagram, Domestic Tourist, Machine Learning, Forecast

A. INTRODUCTION

Indonesia, through Presidential Degree No. 52, 2023, has identified 10 priority destinations as part of its effort to promote tourism beyond the famous Bali (Bappenas, 2019). People often refer to these destinations as the "10 New Bali's." One of the 10 New Balis is Lake Toba in North Sumatra. The Lake Toba region has potential for a priority destination as it has a combination of outstanding natural attractions and distinctive cultural heritage, such as Batak traditions and the beauty of its natural scenery (Kennedy et al., 2022). The development of Tourism Priority Destinations (DPP) not only aims to enhance destination sites but also fosters economic growth and tourism infrastructure in other areas that are captivated by diverse ethnic cultures and appealing nature (Kanwal et al., 2020; Mamirkulova et al., 2020). Tourism infrastructure development and proper promotion can provide significant economic benefits to local communities (Kanwal et al., 2020). These include the creation of new jobs, small and medium enterprise opportunities, and increasing local revenue as well as the per capita income of residents (Wanhill, 2000). Tourism can also strengthen local cultural identity and maintain cultural heritage that is important to the community (McKercher & Du Cros, 2002).

The number of tourist trips to tourism sites serves as a prominent indicator of the success of tourism development. However, attracting many visitors requires excellent planning and effective interventions. The efforts aim to increase the demand for visiting destinations. The demand for visiting the

* Corresponding author: Erna Nurmawati; Email: erna.nurmawati@stis.ac.id

Received: August 7, 2024; Revised: October 17, 2024; Accepted: October 21, 2024

destination will directly correlate with its attractiveness. Therefore, stakeholders in the tourism industry should conduct intensive and strategic promotional activities to increase the attractiveness of tourism destinations at the national and international levels. These include digital marketing (Kaur, 2017), participation in international tourism exhibitions, collaboration with travel agencies, as well as building an attractive and safe destination image for tourists (Camilleri, 2018).

The demographic characteristics of domestic tourist travelers closely influence efforts to increase domestic tourist trips. Based on data from 2023, around 60% of domestic travelers are less than 35 years old (Badan Pusat Statistik, 2024). The characteristics of the current lifestyle, such as access to technology and high activity on social media, are identical to those of the young population. Technological advances have made the younger generation more dependent on the internet and websites to support their activities. Online information also influences the young generation's choices, including travel plans. Over the last decade, researchers have identified a trend of digital marketing usage. Even in 2016, it was shown that 77% of information has been obtained from cyberspace; at the stage of determining the choice of influence, it reached 65%, while execution such as booking and payment is still relatively low at around 34% (Pitanatri & Pitana, 2016). It is believed that the trend will continue and increase, so the use of digital marketing is an important strategy to increase domestic tourist trips.

The series data reveals an increasing trend in the number of domestic tourist trips to North Sumatra Province from 2020-2024 (Badan Pusat Statistik, 2024). The number of domestic tourist trips to North Sumatra was around 14 million in 2020 and almost doubled in 2023. However, compared to the conditions in 2019, the number of domestic tourist trips to North Sumatra in 2023 remains relatively low, not yet reaching the normal pre-pandemic levels. Furthermore, the data indicates that domestic trips to North Sumatra account for 6.64% of all domestic trips in Indonesia. Regrettably, in 2023, this figure will drop to a mere 3.27%. The data clearly shows that from 2020 to 2023, tourist trips to North Sumatra grew more slowly than those to other provinces, as the province's contribution to total trips in Indonesia decreased compared to the data from 2019. This comes as a surprise, given that Lake Toba, located in North Sumatra, is part of the DPP. Therefore, the government's efforts under the 2020-2024 Medium Term Development Plan (RPJM), which aimed to increase development in 10 DPPs, have not fully achieved their objectives. This situation highlights the need for the government to enhance its interventions and focus on boosting domestic trips to North Sumatra for the best possible outcome. A better strategy is needed to increase the number of tourist trips in DPP, especially Lake Toba in North Sumatra. By focusing on priority tourism destinations, such as the Lake Toba area in North Sumatra, Indonesia can achieve its development goals in the tourism industry. This strategy utilizes the available resources, natural beauty, and cultural richness to support equal development across regions in Indonesia. This condition will bring significant economic, social, and cultural benefits to local communities and ensure Indonesia's position as a world-leading tourist destination.

Sustainability is essential to ensure that the development process has a positive impact in the future. Therefore, in the process of fostering DPP, it is important to pay attention to challenges such as environmental conservation (Baloch et al., 2023), waste management (Diaz-Farina et al., 2020), and regulating the number of tourists to prevent overtourism that can damage the environment and local culture (Abbasian et al., 2020). Future generations should enjoy the natural beauty and cultural distinctiveness by focusing on sustainable management (Liburd et al., 2022). Collaboration between the government, private sector, and local communities is essential in building sustainable tourism infrastructure (Roxas et al., 2020). Investments from the private sector can help in the development of modern and international standard tourism facilities, while the government is responsible for regulation and supervision to ensure development is in accordance with the principles of sustainability.

This study aims to determine the effectiveness of digital marketing in promoting the sustainability of the tourism industry in DSP Lake Toba, North Sumatra. The engagement rate on the Instagram accounts of tourism industry stakeholders serves as a measure of the impact of digital marketing. The study will also use the variable of searching for information about tourism in North Sumatra on the search engine Google or Google Trend Index (GTI) in predicting the number of domestic tourists. It is expected that the results of the study provide an input to increase the number of tourists in North Sumatra through an effective digital marketing strategy.

B. LITERATURE REVIEW

Tourism stakeholders play a crucial role in the advancement and promotion of tourist attractions. Three primary groups categorize the various participants in the tourism industry: the government, the private sector, and local communities (Pitana & Gyatri, 2005). Each group possesses a distinct function and impact within the tourism sector. The government is accountable for the oversight of regulations, the advancement of infrastructure, and the extensive promotion of tourism. The private sector comprises enterprises involved in the tourism industry, such as travel agents, hotels, restaurants, and tourist attractions (Simanjorang et al., 2020). Their responsibility lies in delivering services and creating experiences that attract tourists. The private sector is also crucial in marketing the tourist destination and the creation of innovative tourism offerings. Private sector investments and innovations have the potential to enhance the appeal of tourism sites and significantly impact tourists' decisions to visit a certain region. Local communities are the primary stakeholders and the most impacted by the tourism industry (Rahim, 2012). Local culture, customs, and services have the potential to enhance the tourism experience. Local communities play a vital role in ensuring the sustainability of tourism by influencing tourists' views and satisfaction. Engaging local communities in tourism initiatives, such as organizing cultural festivals or conducting village tours, can boost tourist visits and directly contribute to the economic well-being of these communities. It is imperative for all stakeholders within the tourist industry to maintain the accuracy of tourism demand forecasts as a means to mitigate the risks of improper planning and enhance corporate decision-making (Al-Jassim et al., 2022). Additionally, policymakers aim to obtain accurate forecasts to establish pricing policies and execute a sound business strategy. Studies focusing on the forecast of tourism demand primarily rely on input variables predicted to exhibit a robust correlation with tourism demand. The studies are possible to do since previously several scholars have focused on the quantitative approach to identify correlations between various observations in tourism data (Law et al., 2019).

According to UNSD (2008), a tourist is a person who travels outside of their normal environment for at least 24 hours but not more than a year, primarily for the purpose of working for the resident entity they are visiting. This definition allows us to divide tourists into three categories. First, domestic tourists (wisnus). This category includes tourists who travel to different regions within their own country. Outbound tourists are those who travel from one country to another. Additionally, inbound tourists, also known as wisman, are those who travel from other countries to visit a country. Domestic tourists (wisnus) are the backbone of national tourism, especially during the COVID-19 pandemic (Kemenparekraf, 2022). The COVID-19 pandemic has had a significant impact on the global tourism industry, including in Indonesia. According to research from Gössling et al. (2020), the COVID-19 pandemic has caused a drastic decrease in global mobility and negatively impacted the tourism sector. A significant decline in the number of tourist trips during the pandemic is evident, along with travel restrictions, destination closures, and public health concerns. Additionally, COVID-19 has a direct and significant impact on the number of tourist arrivals and TPK (Yulianto et al., 2022). Thus, it is important to include the variable of the presence or absence of COVID-19 pandemi in the prediction model of the number of tourist trips.

Several models are available for predicting the number of tourists. Among them are the autoregressive integrated moving average (ARIMA) and autoregressive integrated moving average with exogenous variable (ARIMAX) regression models. Riestiansyah et al. (2022) and Li et al. (2020) have studied these regression models. The ARIMAX model has better accuracy and produces a graph that is closer to the actual value (Riestiansyah et al., 2022). In addition, there is a SARIMA model that can overcome seasonal patterns in seasonal data (Alwi & Nurfadilah, 2021). Previous studies have utilized deep learning models such as the Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM) (Hsieh, 2021).

Google Trends measures the search volume of specific keywords within the Google search engine. The Google Trends Index (GTI) then normalizes the search value and indexes it on a scale of 0 to 100. Google Trends data is available in real time and can be deployed to determine the search preferences of topics and places at any given time. GTI can analyze travel trends, identify destination popularity, and predict tourist visitation patterns in the tourism context (Rödel, 2017). Researchers in the tourism field have utilized GTI as both primary and supplementary data (Bangwayo-Skeete et al., 2015; Li et al., 2021).

Inflation is the general increase in the prices of goods and services in an economy over a period of time. Inflation has a significant impact on consumer purchasing power, investment decisions, and overall economic stability (Astiyah, 2009). Inflation affects tourism in several ways. First, inflation can reduce the purchasing power of travelers. If the prices of goods and services in tourist destinations increase, tourists may reduce the number of trips or the duration of their stay. Second, inflation can increase operating costs for the tourism industry, including hotels, restaurants, and attractions. This can affect the prices offered to tourists and ultimately affect tourist demand (Song et al., 2008). Travelers may choose more affordable destinations or reduce their travel frequency. Thus, inflation can be one of the important variables for understanding and predicting the travel patterns of foreign tourists. A high hotel occupation room rate (TPK) indicates that the destination is popular and able to attract a lot of tourists, while a low TPK could indicate a problem in attracting tourists or overcapacity in accommodation (Fletcher et al., 2017). The higher the number of tourists coming to a destination, the higher the room occupancy rate there. Conversely, if the number of visits decreases, the TPK also tends to decrease (Butler, 1998).

Some studies have mentioned that the digital reviews and recommendations from consumers are useful for their business (Ritz et al., 2019). In the internet era, tourism stakeholders can utilize social media platforms like Instagram to promote their regions and boost the number of domestic tourist visits. Instagram plays a role as an online photo album that others can access, and users can also utilize it as a promotional channel for their businesses (Fatanti & Suyadnya, 2015). Instagram posts primarily feature images and videos, often incorporating additional information. The user can provide this information through hashtags, the location of the picture or video, and a caption or writing that explains the purpose of the post (Fiallos et al., 2018). However, it's not always appropriate to describe the uploaded image (Giannoulakis & Tsapatsoulis, 2016). In the context of tourism, captions can include destination descriptions, recommendations, travel stories, and useful tips for potential travelers. Captions can play a role in building a destination's image. In addition, effective captions can build a strong narrative, trigger emotions, and encourage user interaction, all of which are important in destination marketing. Thus, analyzing captions on Instagram posts can help in understanding what elements are most effective in attracting and influencing audiences and how they can be used to improve tourism marketing strategies.

Topic modeling is one of the techniques in natural language processing (NLP) used to discover hidden thematic structures in large text collections (Jelodar et al., 2019). This technique makes it possible to identify and extract topics from text without the need to manually read the entire document. A document is a collection of words. We refer to the collection of documents as a corpus, and we will analyze them collectively to uncover the hidden topics within. Tourism stakeholders' Instagram captions can undergo topic modeling. We can identify the main topic or discussion of each Instagram post through topic modeling. This process can provide insight into the content topics that tourists are most interested in. To assess the engagement rate on Instagram, we need a formula that considers the number of likes and comments, not the frequency of Instagram posts (Azmi & Budi, 2018). The engagement rate shows the effectiveness of posted information reaching the community. Therefore, we can use the engagement rate of posts uploaded per topic by each stakeholder as additional information to predict the number of domestic tourist trips.

This research seeks to evaluate the content and effectiveness of tourism Instagram account uploads, drawing on the previously described background. The value of upload effectiveness in the form of engagement rate (ER) is made based on the type of stakeholder account and topic modeling in the caption. The ER values will be used with the TPK, inflation, GTI, and COVID-19 dummy variables to support historical data in predicting the number of domestic tourist trips in five provinces with DSP. We have built prediction models, ARIMAX/SARIMAX and LSTM, by combining the variables used.

C. RESEARCH METHOD

Data Collection

This research uses secondary data obtained from several sources. Data on the number of tourist trips, TPK, and inflation were collected from the BPS website. To understand the trends and patterns in the number of foreign tourist trips to the five provinces designated as Super Priority Tourism Destinations (DPSP), this study employs TPK analysis. We collected the GTI based on three search keywords derived from Rödel (2017) research and adapted them to meet the needs of this study. The search keywords we utilized were "wisata (followed by) province name," "province name," and "hotel followed by province name." These search keywords were filtered by location, specifically "Indonesia," category 'Travel," and the search type “Web Search." We collected GTI data from Google's Trends Explore tool. We collect GTI based on three research-related search keywords (Rödel, 2017), tailored to meet the specific needs of this study. The list of keywords used in this research is “wisata (followed by) provinsi," "provinsi,” and “hotel (followed by) provinsi." The search keywords are filtered by location, namely "Indonesia," categories, namely "Travel," and search type in the form of “Web Search." While COVID data is dummy data with a value of 0 or 1. A value of 0 means there is no covid pandemic, while 1 indicates a covid pandemic. Data formation is based on the pandemic situation.

This study will deploy some variables related to the tourism industry and can improve the accuracy of the model according to previous research. We can add variables such as Hotel Room Occupancy Rate (TPK) (Suwanto, 2020), inflation (Riestiansyah et al., 2022), Google Trends index based on travel-related search keyword combinations (Rödel, 2017), and COVID-19 dummy data (Hsieh, 2021). Tourism-related Instagram data was collected by scraping using the Instaloader library in Python on posts from tourism stakeholder Instagram accounts. There are three types of tourism stakeholders: the government, the local community, and the business/private sector (Murphy & Murphy, 2004). The provincial tourism office's Instagram account represents the government. Local communities that act as hosts in creating a conducive tourism environment (Rahim, 2012) are represented by the DSP tourist attraction manager account. The accounts of service providers, tourism promotion agencies, and accommodation service providers represent business/private actors. The Instagram accounts of the tourism office and the DSP tourist attraction manager were selected based on the account officiality criteria. We selected the two business/private accounts based on the duration of account creation and the highest average ER. The list of tourism stakeholder Instagram accounts in North Sumatra Province used in this study can be seen in Table 1. Data collection from all Instagram posts, including attributes 'Post URL', 'Date', 'Time', 'Likes', 'Comments', and 'Caption'. The data is then saved in Excel format.

Table 1. List of Tourism Stakeholders in North Sumatra

Tourism Stakeholders	Tourism Stakeholder Instagram Account	Account Name
Government	Provincial tourism office	@disbudparekrafsumut
Local community	Tourist attraction manager	@otorita.danautoba
Business/private actors	Tourism promotion and service providers	@explorewisatasumut
Business/private actors	Accommodation service providers	@tamansimalem

Data Analysis

The flow of the overall research method carried out in this study can be seen in Figure 1.

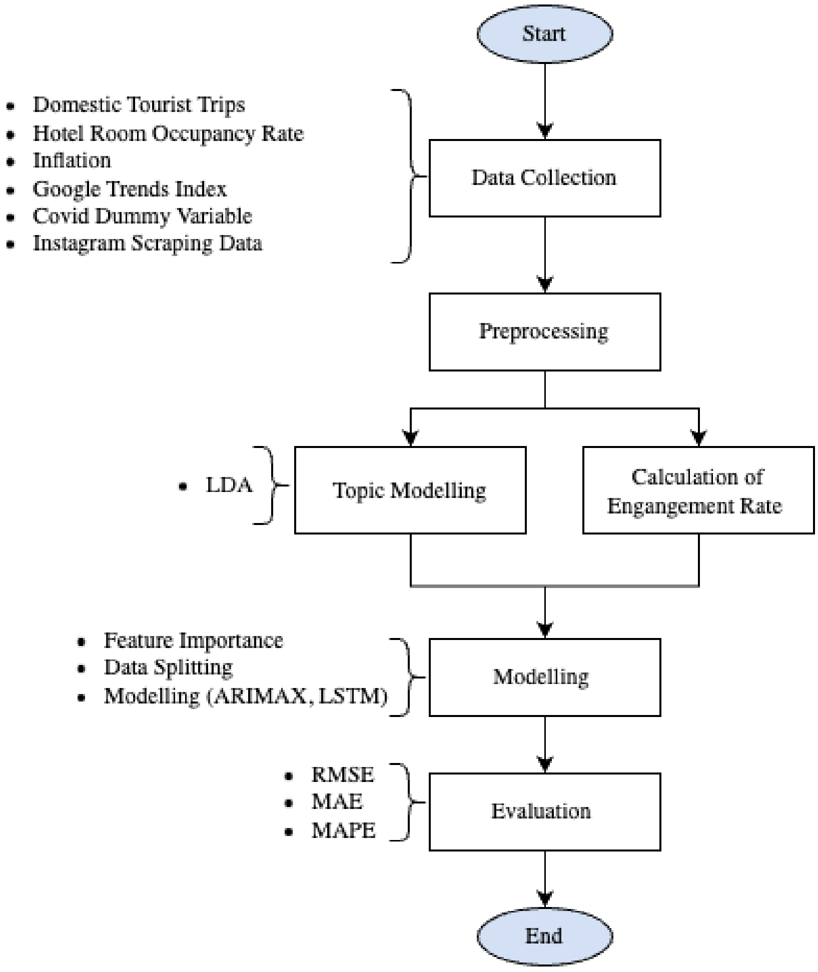

Figure 1. Flow of research methods

Data Preprocessing

The data that has been collected needs to go through a preprocessing stage before being used in modeling to prepare the data into structured data and can be processed in the next stage (Foster et al., 2016). We checked the collected Instagram post data, adjusted the research period, and deleted duplicate data lines. We carry out the preprocessing stage on captions to prepare the caption data for topic modeling. Furthermore, the preprocessing stage is carried out on caption data from Instagram posts with the stages in Figure 2.

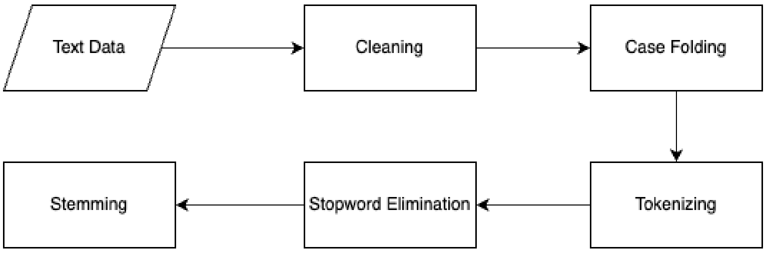

Figure 2. Chart of preprocessing stage on Caption

Topic Modeling

We combined the preprocessed caption data by province. The topic modeling stage will use the data to identify the topics present in the Instagram posts of tourism accounts across the five DSP provinces. We perform topic modeling using the Latent Dirichlet Allocation (LDA) method, utilizing the Gensimm library in Python.

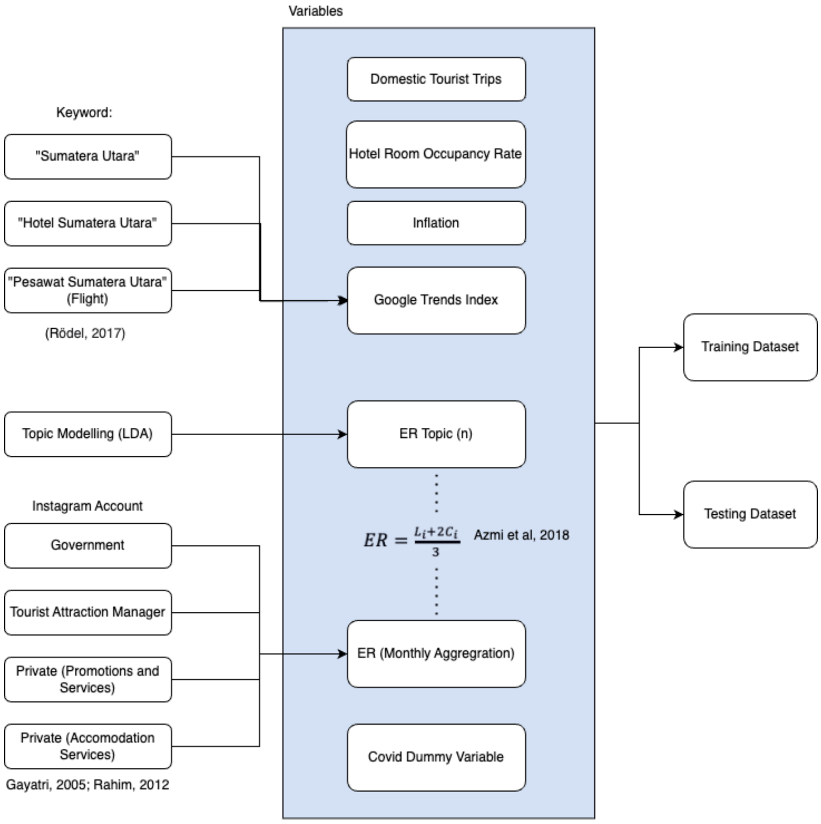

Figure 3. Chart of research variable formation

Engangement Rate Calculation

The next step, following the modeling of the topic in the caption, involves calculating the Effective Rate (ER). The ER value serves as a key metric for assessing the effectiveness of information dissemination through content. We calculate the engagement rate on each Instagram post based on the number of likes and comments using the previous equation (1). Upon completing the topic modeling and ER calculation stages, we will formulate variables for the modeling process. The account ER variable is the average ER of all posts from stakeholder accounts in each DSP province. The study generates a specific number of topic ER variables by averaging the ER posts for each DSP province's topics. Thus, the flowchart of variable formation to data preparation for modeling in this study can be seen in Figure 3.

Prediction Model Analysis

a. Feature Importance and variable combinations

We use the feature-important technique to identify the variables that have the greatest influence on the number of tourist trips in each DSP province. Feature importance is done by applying the random forest method using the RandomForestRegressor library from sklearn in Python.

b. Data Splitting

We first divide the data into training and testing datasets before modeling the number of tourist trips. Training data and testing data are divided in a ratio of 85% for training data and the remaining 15% for testing data.

c. Data Modeling

We use time series methods, specifically ARIMAX/SARIMAX, and machine learning models, specifically LSTM, for modeling. We perform ARIMAX/SARIMAX modeling based on the best ARIMA/SARIMA model's order. The general equation of the ARIMAX model can be written as follows:

$$\mathrm{y_t} = \beta_0 + \beta_1 \mathrm{x}_{1,\mathrm{t}} + \cdots + \beta_\mathrm{k} \mathrm{x}_{\mathrm{k},\mathrm{t}} + \varepsilon_\mathrm{t} \qquad (1)$$

Where y is a linear function of k predictor variables, x-1, t., ..., x-k, t., β is the coefficient vector of exogenous variables, and ε-**t**. is an uncorrelated error, commonly called white noise. Then, by adding the seasonal component to the model, the SARIMAX model is written in the following equation:

$$\left(1 - \phi_1 \mathrm{B} - \phi_2 \mathrm{B}^2 - \cdots - \phi_\mathrm{p} \mathrm{B}^\mathrm{p}\right)(1 - \Phi_1 \mathrm{B}^\mathrm{m} - \Phi_2 \mathrm{B}^{2\mathrm{m}} - \cdots - \Phi_\mathrm{P} \mathrm{B}^{\mathrm{Pm}})(1-\mathrm{B})^\mathrm{d}(1-\mathrm{B}^\mathrm{m})^\mathrm{D} \mathrm{y_t} = \\ \left(1 + \theta_1 \mathrm{B} + \theta_2 \mathrm{B}^2 + \cdots + \theta_\mathrm{q} \mathrm{B}^\mathrm{q}\right)\left(1 + \Theta_1 \mathrm{B}^\mathrm{m} + \Theta_\mathrm{s} \mathrm{B}^{2\mathrm{m}} + \cdots + \Theta_\mathrm{Q} \mathrm{B}^{\mathrm{Qm}}\right)\varepsilon_\mathrm{t} + \beta \mathrm{X_t} \qquad (2)$$

Where B is the backshift operator (lag operator), ϕ is the AR coefficient, Φ is the AR seasonal coefficient, θ is the MA coefficient, Θ is the MA seasonal coefficient, d is the number of non-seasonal differencing transformations, and D is the number of non-seasonal differencing transformations, and X is the value of exogenous variables. The formation of the ARIMA/SARIMA model was carried out with the help of library auto.arima() in Python. Meanwhile, we select the best model based on the smallest AIC value. Furthermore, modeling is performed using LSTM with the architecture as shown in Figure 4.

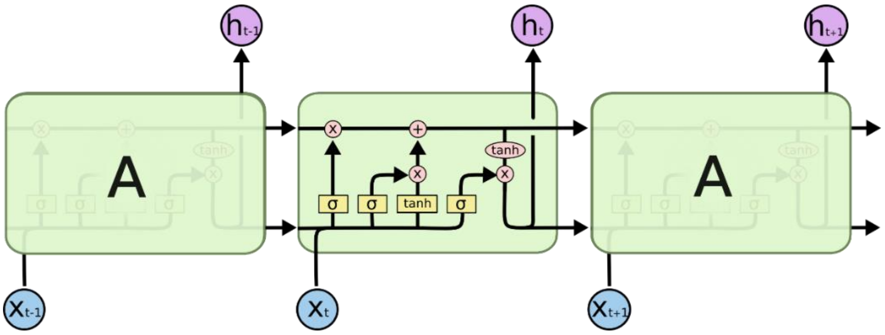

Figure 4. A long short-term memory architecture (Khumaidi & Nirmala, 2022)

We construct the LSTM method by applying hyperparameter tuning with the grid search method (Larochelle et al., 2007). The hyperparameter size combination used in this study can be seen in Table 2.

Table 2. Hyperparameter LSTM

Hyperparameter	Size
Batch size	32
Dropout rate	0,05; 0,15; 0,25
Epoch	50, 100, 150
Learning rate	0,0001; 0,001; 0,01
LSTM units	128, 192, 256
Optimizer	adam, sigmoid

In addition, there is a hyperparameter n_past, which is a lag in the LSTM model set before modeling. The higher the value of n_past, the more complex the model formed, but the time required is also longer. This study determines the value of n_past by taking into account the limited data length and the comparability between the LSTM and ARIMAX/SARIMAX models. Thus, n_past used is the number of lags used in the ARIMAX/SARIMAX model.

Model Evaluation

This research applies three accuracy matrices to the model, namely MAPE, MAE, and RMSE. The calculation of the three accuracy matrices is done by importing functions from the library sklearn.metrics.

D. RESULTS AND DISCUSSIONS

Topic Modeling

We conducted topic modeling for all uploads from the four stakeholder accounts for each DSP province. We conducted topic modeling experiments by varying the number of topics from 3 to 10 and the number of passes from 1 to 20. Modeling was selected based on the highest coherence score value. Each province's topic modeling results yield a distinct number of topic categories and discussions. Table 3 displays the topic modeling results for North Sumatra, demonstrating a coherence score of 0.716 across 3 topics and 5 passes. Based on the constituent keywords, Topic 0 is a topic with a focus on offering tickets and vacation packages at promo prices. Topic 1 discusses how cultural tourism and related activities can develop the local economy. While keywords such as 'reservation', 'hotline', 'info', ‘park’, ‘beautifull’, etc. on Topic 2 show the information and contact services for booking accommodation and the natural beauty facilities offered.

Table 3. North Sumatra Topic Modeling Results

Topic Type	Keywords	Discussion
0	ticket, info, promo, open, island, people, repeat, healthy, price, package	Tickets and vacation packages at promo prices
1	tourism, travel, culture, Indonesia, village, website, enterprising, area, economy, flower	Cultural tourism and its activities can develop the economy
2	reservation, hotline, info, office, stay, night, room, park, wonderful, nature	Accommodation reservation offer and contact details

Engangement Rate Calculation

Engagement Rate (ER) calculation is done based on equation (3). The ER value of each account upload is calculated and then averaged by month to finally obtain the average monthly account ER. The results of the ER calculation in North Sumatra province can be seen in Table 4.

$$ER_i = \frac{L_i + 2C_i}{3} \qquad (3)$$

ER_i = Engagement rate of the i-th post
L_i = Likes of the i-th post
C_i = Comments of the i-th post

Table 4. North Sumatra Engagement Rater Score

Number of Posts	Number of Likes	Number of Comments	Average ER
6,725	8,356,532	207,533	422,211

While the details of the average ER of each tourism stakeholder account in North Sumatra can be observed in Table 5.

Table 5. Average ER by Stakeholder type

Stakeholder Type	Average ER
DSP Tourist Attraction Manager	76,359
Tourism Office	42,290
Tourism Promotion and Service Providers	942,274
Accommodation providers	57,076

Descriptive Analysis

The modeling used a dataset consisting of 59 rows, based on the stages carried out. The independent variables that were utilized in the modeling to predict the number of tourist trips to North Sumatra include TPK, inflation, GTI, COVID, ER account, ER_topic0, ER_topic1, and ER_topic2. Table 6 presents a descriptive statistical analysis of the data used in the modeling of the number of foreign trips to North Sumatra. The tourist variable has an average of 2,157,482.22 with a standard deviation of 1,128,406.67. The TPK variable has an average of 41.43 with a standard deviation of 9.16. In North Sumatra, the inflation rate averages 0.23, with a standard deviation of 0.58.

Table 6. Descriptive Analysis

Variable	Mean	Standard Deviation	Min.	Median	Max.
Wisnus	2.157.482,22	1.128.406,67	442.07	1.847.654	5.774.370
TPK	41,43	9,16	11,93	43,24	54,44
Inflasi	0,23	0,58	-1,81	0,22	1,63
GTI	19,13	5,55	9,75	18,30	36,88
Covid	0,68	0,47	0,00	1,00	1,00
ER akun	466,99	218,17	144,43	463,23	973,92
ER_topic0	369,94	331,45	72,59	270,23	2216,00
ER_topic1	204,22	140,28	36,72	165,25	684,40
ER_topic2	692,48	349,39	182,45	721,62	1656,03

The independent variables obtained from big data also have diverse descriptive statistics. The average GTI of North Sumatra is 19.13 with a standard deviation of 5.55. ER stakeholder accounts in North Sumatra have an average of 466.99 with a standard deviation of 218.17. ER topic 0, which contains discussions about ticket information and vacation packages at promo prices, has an average of 369.94 and a standard deviation of 331.45. ER Topic 1 is a topic with discussions about cultural tourism and its activities that can develop the economy. It has an average of 204.22 and a standard deviation of 140.28. The average for ER topic 2, which discusses accommodation reservation information with special discounts, is 692.48, with a standard deviation of 349.39.

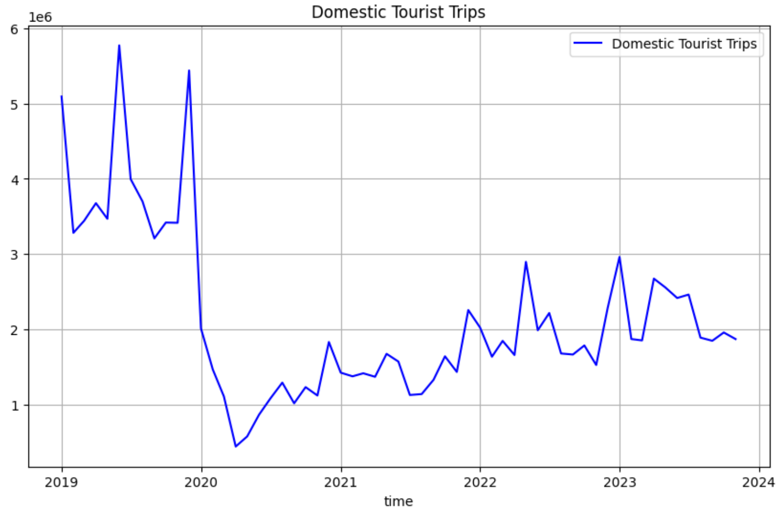

Figure 5. Number of domestic tourist trips to North Sumatra from January 2019 - November 2023

Figure 5's graph indicates a fluctuating pattern in the number of foreign tourist trips to North Sumatra. The number of foreign tourist trips to North Sumatra peaked in June 2019 and experienced the most drastic decline in January 2020. The decline continued until April 2020 and tended not to be able to return to its highest number until the end of 2023. The decline almost coincided with the COVID-19 pandemic in Indonesia, which was officially determined for the first time in March 2020.

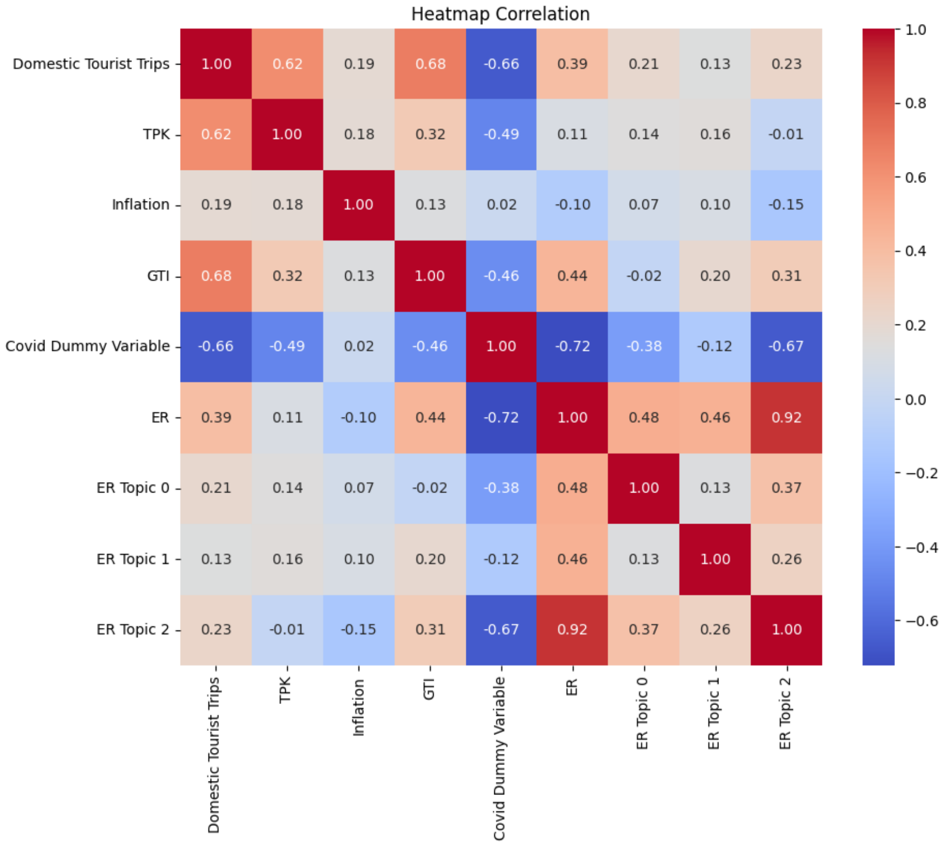

Figure 6. Correlation plot between the number of domestic tourist trips to North Sumatra and the independent variables used

The three variables that produce the highest positive correlation value with the number of tourist trips are GTI, TPK, and ER, which are 0.68, 0.62, and 0.39, respectively. Meanwhile, the dummy variable COVID-19 is a variable that has a negative and quite high correlation with the number of tourist trips.

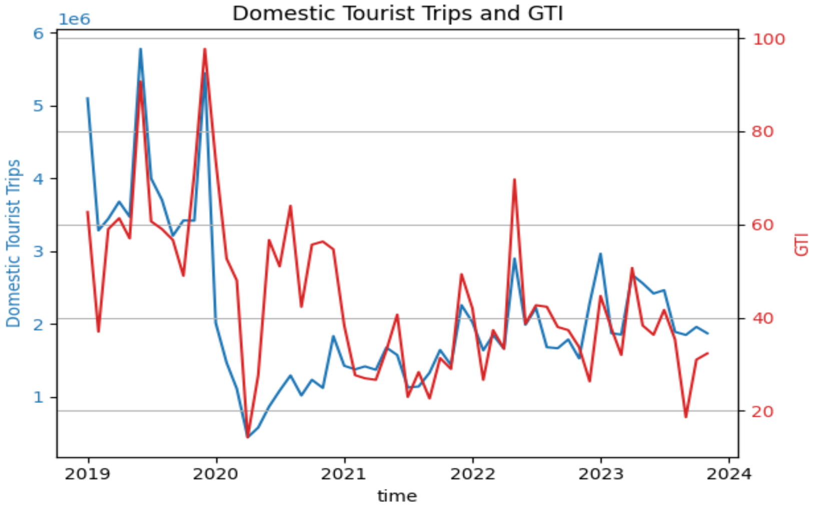

Figure 7. Graph of the number of tourist trips and GTI searches from January 2019-November 2023

We obtain GTI using the three keywords 'Wisata Sumatera Utara', 'Sumatera Utara', and 'Hotel Sumatera Utara'. The graph in Figure 7 illustrates a similar pattern between the number of foreign trips to North Sumatra and the GTI derived from search keywords. This is in accordance with the previous Pearson correlation calculation of the number of foreign trips and GTI of North Sumatra which is 0.68.

Prediction Modeling of Domestic Tourists Trips

We calculate the importance scores of the independent variables from the data in advance before entering the modeling phase. Figure 8 illustrates the results of this calculation.

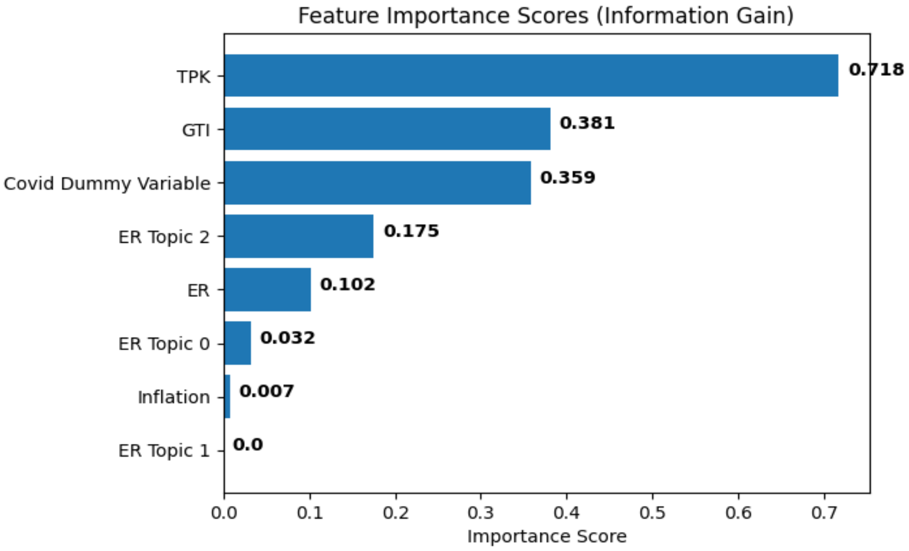

Figure 8. Feature Importance Result Graph

TPK has the highest feature importance score of 0.72; followed by GTI, covid, and ER_topic 2 with a feature importance score of 0.38; 0.36; and 0;18, respectively. So that the independent variables TPK, GTI, covid, and ER_topic2 will be used in modeling the number of tourist trips to North Sumatra based on a combination of variable feature importance score results.

Table 7. North Sumatra SARIMAX Prediction Model

Model	MAPE(%)	MAE	RMSE
SARIMAX(2,0,0)(0,0,1)(12) (wisnus + TPK + GTI + covid + ER_topic2)	36.446	758,244.49	829,922.09
SARIMAX(2,0,0)(0,0,1)(12) (wisnus + TPK + GTI + covid)	39.115	810,986.45	892,278.07
SARIMAX(2,0,0)(0,0,1)(12) (wisnus + TPK + GTI + ER_topic2)	11.59	253,103.69	293,195.99
SARIMAX(2,0,0)(0,0,1)(12) (wisnus + TPK + covid + ER_topic2)	59.98	1,208,878.42	1,470,230.82
SARIMAX(2,0,0)(0,0,1)(12) (wisnus + GTI + covid + ER_topic2)	26.42	563,540.32	652,971.24
SARIMAX(2,0,0)(0,0,1)(12) (wisnus + TPK + Inflasi + GTI + covid + ER akun + ER_topic0 + ER_topic1 + ER_topic2)	56.65	1,165,563.76	1,503,109.70

The SARIMA (2, 0, 0) (0, 0, 1) (12) model is a prediction model for the number of tourist trips to North Sumatra that has the smallest AIC value. Therefore, we build the SARIMAX model using a combination of variables in the same order. Table 7 contains an evaluation of each SARIMAX (2, 0, 0) (0, 0, 1) (12) model built using a combination of variables. The SARIMAX(2,0,0)(0,0,1)(12) model, which incorporates independent variables such as TPK, inflation, GTI, and ER in Topic 2 on 'accommodation reservation information', has an evaluation value of 11.59%, a MAE of 253,103.69, and an RMSE of 293,195.99. The three evaluation values are the lowest among the five SARIMAX models. Thus, the SARIMAX(2,0,0)(0,0,1)(12) model with independent variables of TPK, inflation, GTI, and ER topic 2 with the discussion of 'accommodation reservation information' is the best SARIMAX model in predicting the number of tourist trips to North Sumatra and falls into the good category.

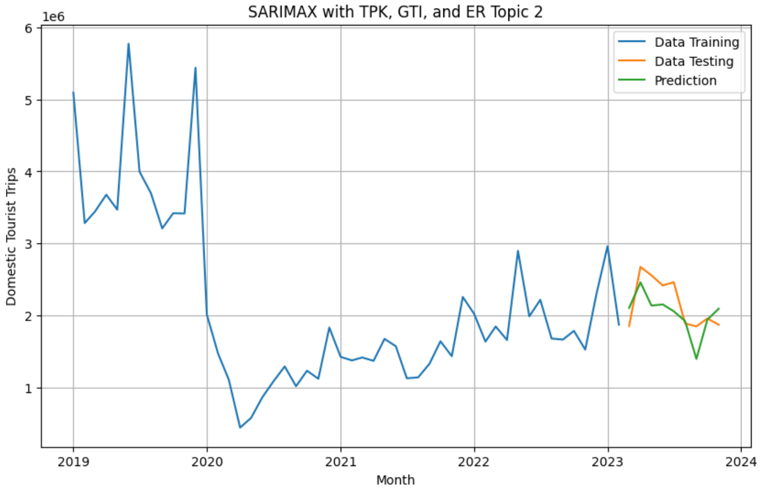

Figure 9. Graph of the best SARIMAX Prediction Results

Figure 9 is a graph of the best prediction model for the number of tourist trips to North Sumatra using SARIMAX(2,0,0)(0,0,1)(12). The graph of the prediction data has followed the pattern of the test data and has been quite close together at some points. Hyperparameter tunning on the LSTM model with a combination of variables in order can be observed in table 8. Based on these results, LSTM modeling will then be carried out using these hyperparameters with n_past= 2. The evaluation of the LSTM model built based on the best combination of variables and hyperparameters is shown in Table 9.

Table 8. Best hyperparameters of LSTM

Variable Combination	Hyperparameter	Size
(wisnus + TPK + GTI + covid + ER_topic2)	Batch size	32
	Dropout rate	0,05
	Epoch	150
	Learning rate	0,01
	LSTM units	192
	Optimizer	adam
(wisnus + TPK + GTI + covid)	Batch size	32
	Dropout rate	0,15
	Epoch	150

Variable Combination	Hyperparameter	Size
(wisnus + TPK + GTI + ER_topic2)	Learning rate	0,01
	LSTM units	128
	Optimizer	adam
	Batch size	32
	Dropout rate	0,05
	Epoch	100
(wisnus + TPK+ covid + ER_topic2)	Learning rate	0,0001
	LSTM units	192
	Optimizer	adam
	Batch size	32
	Dropout rate	0,15
	Epoch	100
(wisnus + GTI + covid + ER_topic2)	Learning rate	0,001
	LSTM units	128,
	Optimizer	adam
	Batch size	32
	Dropout rate	0,25
	Epoch	150
all variabel	Learning rate	0,0001
	LSTM units	256
	Optimizer	adam
	Batch size	32
	Dropout rate	0,05
	Epoch	100
	Learning rate	0,0001
	LSTM units	192
	Optimizer	adam

Table 9: Evaluation results of the LSTM model

Model	MAPE	MAE	RMSE
LSTM (wisnus+TPK+GTI+covid+ER_topic2)	63,29	1.261.577,66	1.557.983,18
LSTM (wisnus+TPK+GTI+covid)	60,59	1.203.554,30	1.458.137,17
LSTM (wisnus+TPK+I GT+ER_topic2)	8,24	183.440,95	215.890,00
LSTM (wisnus+TPK+covid+ER_topic2)	37,39	783.917,64	829.485,48
LSTM (wisnus+GTI+ covid+ER_topic2)	21,09	451573,27	489.974,87
LSTM (wisnus + TPK + Inflasi + GTI + covid + ER akun + ER_topic0 + ER_topic1 + ER_topic2)	11,97	256.595,84	275.880,69

Table 9 is an evaluation table of each LSTM model with the best combination of variables and hyperparameters from the grid search. The LSTM model with independent variables TPK, GTI, and ER Topic 2, namely with the topic 'accommodation reservation information', has an evaluation value of MAPE of 8.24%, MAE of 183,440.95, and RMSE of 215,890.00. These three evaluation values are the lowest among the five LSTM models. So that the LSTM model with the independent variables TPK, GTI, and ER topic 2 with the discussion of'reservation and accommodation information' is the best LSTM model in predicting the number of tourist trips to North Sumatra and based on the resulting MAPE value is included in the highly accurate category.

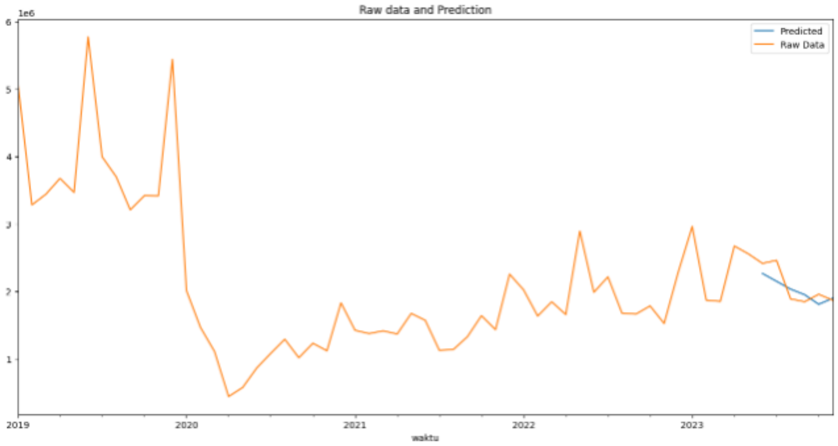

Figure 10. Graph of North Sumatra LSTM prediction results

Figure 10 is a graph of the prediction results of the number of tourist trips to North Sumatra using the best LSTM model. The graph of the prediction data is enough to follow the test data pattern and coincide with the test data.

E. CONCLUSION

The model has a MAPE evaluation of 8.24%; MAE of 183,440.95; and RMSE of 215,890.00. Topic modeling using LDA on the captions of North Sumatra stakeholder Instagram posts resulted in 3 topics, namely about 'ticket information and vacation packages', 'cultural tourism and its activities to develop the economy', and 'accommodation reservation offers'. The average monthly ER value of North Sumatra reached 422,211. In North Sumatra, the topic 'cultural tourism and its activities that develop the economy' has the highest average ER, reaching 692.48. This finding indicates that posting about destination attractions and tourist activities gets the highest attention from the potential traveler.

Further analysis finds that the best model for predicting the number of domestic tourist trips to North Sumatra is LSTM with independent variables of TPK, IGT, and ER on the topic of ‘accommodation reservation offer and contact details’. The model has a MAPE evaluation of 8.24%; MAE of 183,440.95; and RMSE of 215,890.00. Posting on Instagram about cultural tourism and its activities does not appear as an important variable to determine the number of domestic visitors. This result cannot be concluded that the stakeholders in the tourism industry should not pay attention to this aspect to promote the destination. This analysis suggests that promoting the destination must still concern cultural tourism and its activities based on the ER, which has the highest score. This condition will influence the potential tourist in deciding to visit the destination.

Then, after the decision is made, the action is finding further information on ticket promotions and so on. Therefore, the topic of 'tickets and vacation packages at promo prices' emerges as a significant variable to determine the number of visitors. This analysis emphasizes the crucial role of digital marketing to shape the demand for the tourism industry. This conclusion comes from the significant impact of GTI and ER on Instagram postings to determine domestic visitors. The related stakeholders must consider this aspect to sustain its business.

This study has several limitations but opens up opportunities for further exploration. Future research may explore various avenues to expand upon the findings of this study, such as incorporating additional exogenous variables that have a significant impact on tourist visits, including reviews on platforms like Google Maps, TripAdvisor, and Traveloka. In the topic modeling phase, a more efficient method may be selected. Furthermore, additional analysis of the results from topic modeling can be conducted to enrich the research.

REFERENCES

Abbasian, S., Onn, G., & Arnautovic, D. (2020). Overtourism in Dubrovnik in the eyes of local tourism employees: A qualitative study. *Cogent Social Sciences*, *6*(1), 1775944. https://doi.org/10.1080/23311886.2020.1775944

Al-Jassim, R. S., Jetly, K., Abushakra, A., & Al Mansori, S. (2022). A review of the methods and techniques used in tourism demand forecasting. *EAI Endorsed Transactions on Creative Technologies*, *9*(4), e1. https://doi.org/10.4108/eetct.v9i31.2986

Alwi, W., & Nurfadilah, K. (2021). Penerapan Metode SARIMA untuk Peramalan Jumlah Pengunjung Wisata Taman Nasional Bantimurung Bulusaraung Maros. *JOMTA Journal of Mathematics: Theory and Applications*, *3*(1), 1–7. https://doi.org/https://doi.org/10.31605/jomta.v3i1.1221

Astiyah, S. (2009). *Inflasi* (1st ed.). Pusat Pendidikan dan Studi Kebanksentralan. http://www.bi.go.id

Azmi, A. F., & Budi, I. (2018). Exploring practices and engagement of Instagram by Indonesia Government Ministries. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 18–21. https://doi.org/10.1109/ICITEED.2018.8534799

Badan Pusat Statistik. (2024). *Domestics Tourism Statistics 2023*.

Baloch, Q. B., Shah, S. N., Iqbal, N., Sheeraz, M., Asadullah, M., Mahar, S., & Khan, A. U. (2023). Impact of tourism development upon environmental sustainability: a suggested framework for sustainable ecotourism. *Environmental Science and Pollution Research*, *30*(3), 5917–5930. https://doi.org/10.1007/s11356-022-22496-w

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46,* 454-464. https://doi.org/10.1016/j.tourman.2014.07.014

Bappenas. (2019). *Presidential degree no 52 year 2023: the government work plan 2024*. Edisi 11, 12(November), 1–68. https://peraturan.bpk.go.id/Details/234926/perpu-no-2-tahun-2022%0Awww.djpk.depkeu.go.id

Butler, R. (1998). Seasonality in tourism: Issues and implications. *The Tourist Review*, *53*(3), 18–24. https://doi.org/10.1108/eb058278

Camilleri, M. A. (2018). The Tourism Industry: An Overview. In: *Travel Marketing, Tourism Economics and the Airline Product. Tourism, Hospitality & Event Management.* Springer, Cham. https://doi.org/10.1007/978-3-319-49849-2_1

Diaz-Farina, E., Díaz-Hernández, J. J., & Padrón-Fumero, N. (2020). The contribution of tourism to municipal solid waste generation: A mixed demand-supply approach on the island of Tenerife. *Waste Management*, *102*, 587–597. https://doi.org/10.1016/j.wasman.2019.11.023

Fatanti, M. N., & Suyadnya, I. W. (2015). Beyond User Gaze: How Instagram Creates Tourism Destination Brand?. *Procedia - Social and Behavioral Sciences*, *211*, 1089–1095. https://doi.org/10.1016/j.sbspro.2015.11.145

Fiallos, A., Jimenes, K., Fiallos, C., & Figueroa, S. (2018). Detecting topics and locations on Instagram photos. In *2018 International Conference on eDemocracy & eGovernment (ICEDEG)* (pp. 246-250). IEEE. https://doi.org/10.1109/ICEDEG.2018.8372314

Fletcher, J., Fyall, A., Gilbert, D., & Wanhill, S. (2017). Tourism: Principles and practice. Pearson UK.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2016). *Big data and social science: A practical guide to methods and tools*. Chapman and Hall/CRC.

Giannoulakis, S., & Tsapatsoulis, N. (2016). Evaluating the descriptive power of Instagram hashtags. *Journal of Innovation in Digital Ecosystems*, *3*(2), 114–129. https://doi.org/10.1016/j.jides.2016.10.001

Gössling, S., Scott, D., & Hall, C. M. (2020). Pandemics, tourism and global change: a rapid assessment of COVID-19. *Journal of Sustainable Tourism, 29*(1), 1-20. https://doi.org/10.1080/09669582.2020.1758708

Hsieh, S. C. (2021). Tourism demand forecasting based on an lstm network and its variants. *Algorithms*, *14*(8). 243. https://doi.org/10.3390/a14080243

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Kanwal, S., Rasheed, M. I., Pitafi, A. H., Pitafi, A., & Ren, M. (2020). Road and transport infrastructure development and community support for tourism: The role of perceived benefits, and community satisfaction. *Tourism Management*, *77*, 104014. https://doi.org/10.1016/j.tourman.2019.104014

Kaur, G. (2017). The importance of digital marketing in the tourism industry. *International Journal of Research-Granthaalayah*, *5*(6), 72–77. https://doi.org/10.5281/zenodo.815854

Kemenparekraf. (2022). *Siaran Pers: Menparekraf: Kekayaan SDA Modal Pengembangan Pariwisata Berkualitas Berkelanjutan Mendukung Kebangkitan Ekonomi*. Kemenparekraf. https://www.kemenparekraf.go.id/berita/siaran-pers-menparekraf-kekayaan-sda-modal-pengembangan-pariwisata-berkualitas-berkelanjutan-mendukung-kebangkitan-ekonomi

Kennedy, P. S. J., Tobing, S. J. L., & Toruan, R. L. (2022). Marketing strategy with marketing mix for Lake Toba tourism destination. *Journal of Sustainable Tourism and Entrepreneurship (JoSTE)*, *3*(3), 157–174. https://doi.org/10.35912/joste.v3i3.1515

Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research*, 75, 410-423. https://doi.org/10.1016/j.annals.2019.01.014

Li, H., Hu, M., & Li, G. (2020). Forecasting tourism demand with multisource big data. *Annals of Tourism Research, 83*, 102912. https://doi.org/10.1016/j.annals.2020.102912

Li, X., Law, R., Xie, G., & Wang, S. (2021). Review of tourism forecasting research with internet data. *Tourism Management, 83*, 104245. https://doi.org/10.1016/j.tourman.2020.104245

Liburd, J., Duedahl, E., & Heape, C. (2022). Co-designing tourism for sustainable development. *Journal of Sustainable Tourism*, *30*(10), 2298–2317.

Mamirkulova, G., Mi, J., Abbas, J., Mahmood, S., Mubeen, R., & Ziapour, A. (2020). New Silk Road infrastructure opportunities in developing tourism environment for residents better quality of life. *Global Ecology and Conservation*, *24*, e01194. https://doi.org/10.1016/j.gecco.2020.e01194

McKercher, B., & Du Cros, H. (2002). *Cultural tourism: The partnership between tourism and cultural heritage management*. Routledge.

Murphy, P., & Murphy, A. (2004). Strategic Management for Tourism Communities: Bridging the Gaps. In *Strategic Management for Tourism Communities*. https://doi.org/10.21832/9781873150856

Pitana, I. G., & Gyatri, P. G. (2005). *Sosiologi Pariwisata: Kajian Sosiologi terhadap Struktur, Sistem, dan Dampak-Dampak Pariwisata.* Yogyakarta: Andi Offset.

Pitanatri, P. D., & Pitana, I. (2016). Digital Marketing in Tourism: The More Global, The More Personal. In *International TourismConference Promoting Cultural & Heritage Tourism. Udayana University*

Rahim, F. (2012). *Pedoman Kelompok Sadar Wisata* (1st ed.). Kemenparekraf.

Riestiansyah, F., Damayanti, D., Reswara, M., & Susetyoko, R. (2022). Perbandingan metode ARIMA dan ARIMAX dalam Memprediksi Jumlah Wisatawan Nusantara di Pulau Bali. *Jurnal Infomedia: Teknik Informatika, Multimedia & Jaringan*, *7*(2), 58–62.

Rödel, E. L. (2017). *Forecasting tourism demand in Amsterdam with Google Trends: A research into the forecasting potential of Google Trends for tourism demand in Amsterdam* (Master's thesis, University of Twente).

Roxas, F. M. Y., Rivera, J. P. R., & Gutierrez, E. L. M. (2020). Mapping stakeholders' roles in governing sustainable tourism destinations. *Journal of Hospitality and Tourism Management*, *45*, 387–398. https://doi.org/10.1016/j.jhtm.2020.09.005

Simanjorang, F., Hakim, L., & Sunarti. (2020). Peran Stakeholder Dalam Pembangunan Pariwisata Di Pulau Samosir. *Jurnal Profit*, *14*(1), 42–52. https://doi.org/10.21776/ub.profit.2020.014.01.5

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. *Tourism Management, 29*(2), 203-220. https://doi.org/10.1016/j.tourman.2007.07.016

Suwanto, S. (2020). Hubungan Jumlah Kunjungan Wisatawan Mancanegara dengan Rata-Rata Tingkat Penghunian Kamar Hotel Provinsi DKI Jakarta Tahun 2012-2018. *Jurnal Kepariwisataan Indonesia: Jurnal Penelitian Dan Pengembangan Kepariwisataan Indonesia, 14*(1), 9–20. https://doi.org/10.47608/jki.v14i12020.9-20

UNSD. (2008). Role of the International Recommendations for Tourism Statistics 2008. In *International Recommendations for Tourism Statistics 2008*. (pp 5-13). https://doi.org/10.18356/05265168-en

Wanhill, S. (2000). Small and medium tourism enterprises. *Annals of Tourism Research*, *27*(1), 132–147. https://doi.org/10.1016/S0160-7383(99)00072-9

Yulianto, A., Putri, E. D. H., & Wardani, D. M. (2022). Dampak Pandemi COVID-19 Terhadap Jumlah Kunjungan Wisatawan dan Tingkat Hunian Kamar Hotel D.I Yogyakarta. *Jurnal Pariwisata, 9*(1), 53–63. https://doi.org/10.31294/par.v9i1.12331