SBMITB
School of Business & Management

The Predictive Analytics Model of International Tourist Arrivals in Indonesia amid the COVID-19 Pandemic using Multisource Internet Data

Thesis Defense - Master of Science in Management

Dinda Thalia Andariesta

29020007

Supervisor : Meditya Wasesa, Ph.D.
Examiners : 1. Santi Novani, Ph.D.
2. Gatot Yudoko, Ph.D.

Bandung, 13 April 2022

The Outline of Presentation

01 | INTRODUCTION

- Background and Problem Statement
- Research Questions and Research Objectives
- Research Method
- Research Scope

02 | LITERATURE REVIEW

- Tourism Demand Prediction Model
- Tourism Demand Prediction using Single-source Internet Data
- Tourism Demand Prediction using Multisource Internet Data
- Tourism Demand Prediction during COVID-19
- Research Positioning

03 | METHODOLOGY

- Research Philosophical Position
- Research Framework
- Data Collection
- Data Preparation
- Model Development
- Model Evaluation, Model Use, and Reporting

04 | RESULT AND ANALYSIS

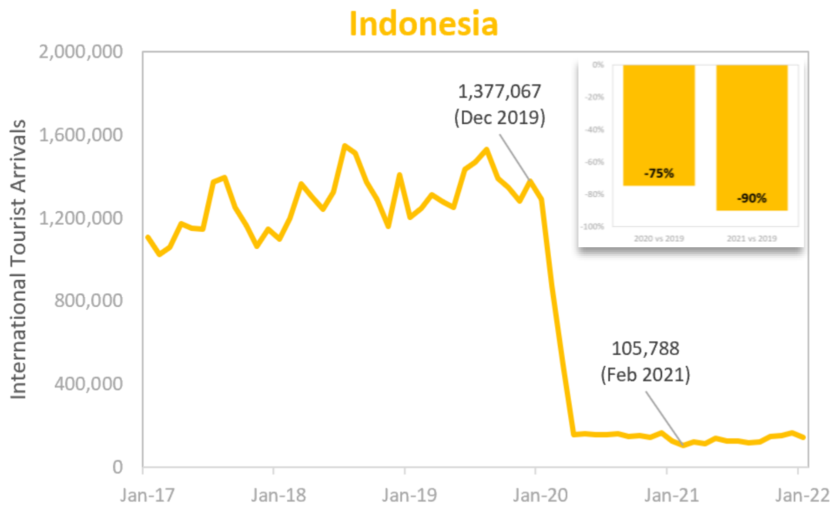
- Descriptive Analysis
- Textual Analysis
- Prediction Result and Analysis
- Forecast Result and Analysis

05 | CONCLUSION

- General Conclusion
- Research and Practical Implications
- Limitations and Further Research

SBMITB School of Business & Management

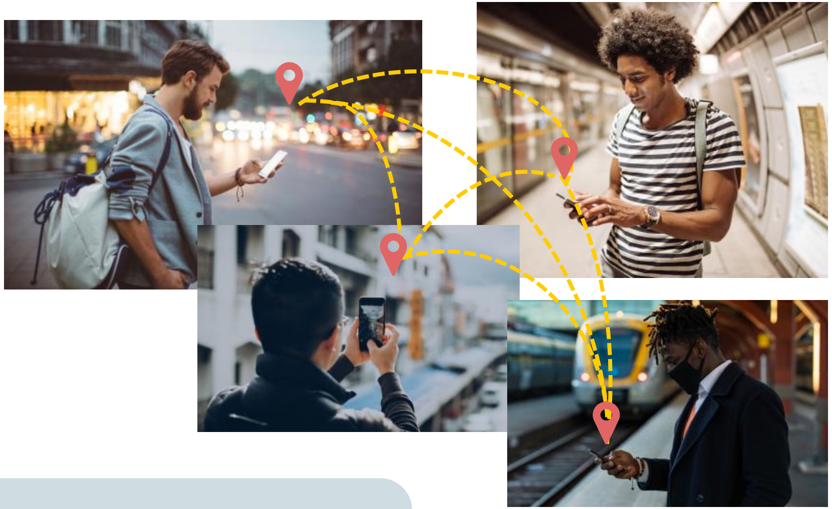
Background and Problem Statement



Source: World Tourism Organization (UNWTO)

The **devastating impact** of the **COVID-19** pandemic on global tourism has **carried on into 2022**

Understanding **the outlook of international tourist arrivals** in the recovery period become **critical** to ensure the safe restart of tourism and avoid another year of massive losses



Need to develop the prediction model

Employing Internet data to provide accurate tourism demand prediction under COVID-19

Integrating the data and verifying empirical applications

A vast amount of **Internet data** potential for the emerging **analytics domain** in the **tourism** context

SBMITB School of Business & Management

Research Questions and Research Objectives

RQ1 *"How to develop a predictive analytics model of international tourist arrivals in Indonesia amid the COVID-19 pandemic using multisource Internet data?"*

Objective: To develop a predictive analytics model of international tourist arrivals in Indonesia amid the COVID-19 pandemic using multisource Internet data

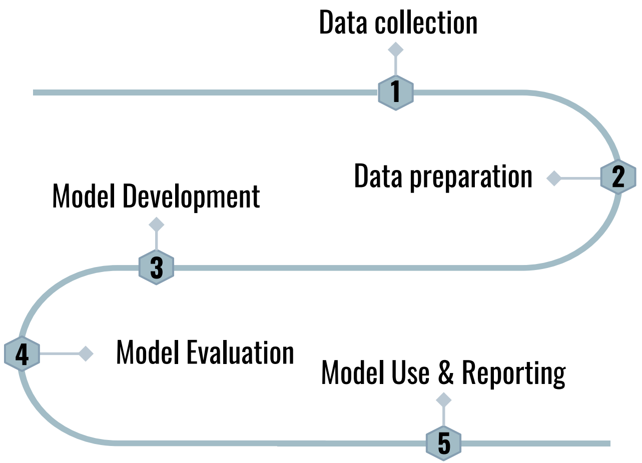
RQ2 *"Does utilizing multisource Internet data lead to a more accurate tourist arrivals prediction during the COVID-19 pandemic than single-source Internet data and historical tourist arrivals data?"*

Objective: To evaluate the performance of tourist arrivals prediction models during the COVID-19 pandemic based on multisource Internet data compared to single-source Internet data and historical tourist arrivals data

Research Method and Research Scope

Research Method

This study focusing on the development of machine learning-based **predictive analytics** model (Shmueli & Koppius, 2011)



Research Scope

01 This study focuses on the prediction of **international tourist arrivals** in **Indonesia** during **the COVID-19**

02 This study uses **social media** data from TripAdvisor and **search engine** data from Google Trends

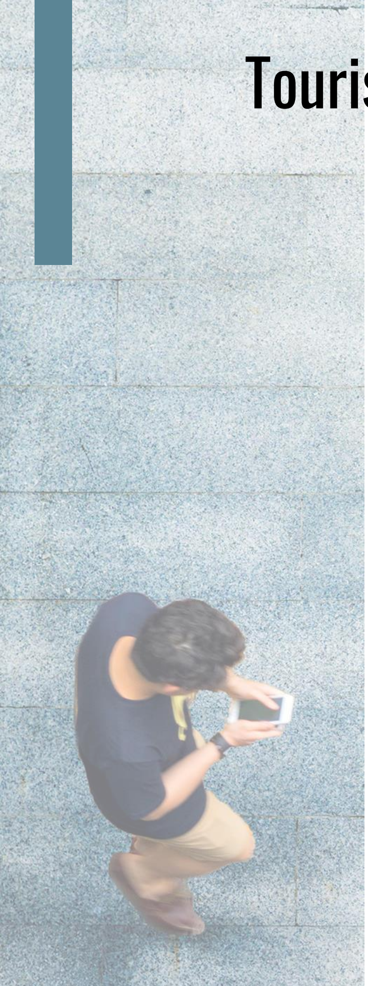
03 This study examines the proposed **multisource Internet data** using three **machine learning models**, namely artificial neural network, support vector regression, random forest

Tourism Demand Prediction Model

Quantitative Methods for Tourism Demand Prediction

	Time Series	Econometric	Artificial Intelligence
Pros	• Provide simplicity • Provide reasonably accurate predictions, especially for short forecasting horizons	• Focus on the causality of various explanatory variables • Improve accuracy in more extended time horizons	• Capable to describe nonlinear data without a prior understanding of the correlations between input and output • Strong flexibility for processing nonlinear data • Improving the forecasting performance
Cons	• Tend to overlook the influencing factors of tourism demand other than temporal factors • Need to check the assumptions	• Need to check the statistical assumptions • All variables should be stationary to avoid spurious results	• Questionable explanatory value of input variables • Poor interpretations of analytical outcomes
Example models	AR, Naïve, ARMA, ARIMA, SARIMA, BSM, ES	ARMAX, ARIMAX, SARIMAX, ADLM, VAR, TVP	ANN, SVR, RF, LSTM, CNN

Tourism Demand Prediction using Single-source Internet Data



Internet Data Sources

	Web Traffic	Search Engine	Social Media
Pros	• Provide structured time series data • Reflect websites' visits or popularity	• Provide structured time series data • Reflect tourists' attention or interests	• Provide user-generated content in various form • Reflect tourists' behavior and sentiments
Cons	• Data not publicly available • Difficulty in applying the data to broader forecasting contexts	• Noises may affect the accuracy • Need to adjust the keywords according to the contexts	• Require tools to obtain data • Require advance data preprocessing
Data sources	Google Analytics	Baidu, Google Trends	Online Forum, Social Network, Microblog, News Articles
Result	The use of single-source Internet data can improve forecasting accuracy		

SBMITB School of Business & Management

Tourism Demand Prediction using Multisource Internet Data

Study	Internet data sources	Predictor variables	Predicted variable	Prediction methods
Fronzetti Colladon et al. (2019)	• Social media • Search engine	• Online forum • Google Trends	International airport arrivals to seven major European capital cities	AR, FAAR, FABM, BM
Gunter et al. (2019)	• Social media • Search engine	• Facebook • Google Trends	Tourist arrivals to four Austrian cities	ARMA, Naïve, ETS, ADLM, MIDAS
Li et al. (2020)	• Social media • Search engine	• Online reviews • Baidu	Tourist arrivals to Mount Siguniang, China	ARIMA, Seasonal Naïve, ARIMAX, ETS, SVR, RF

- Only few studies combining volume-based and sentiment-based data as the explanatory variables
- No previous study uses multisource Internet data in the context of country and under COVID-19
- Most studies use time series and econometric models

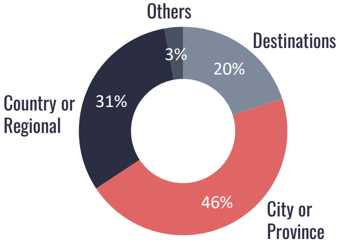
Tourism Demand Prediction during COVID-19

Study	Internet data source	Predictor variables	Predicted variable	Prediction methods
Polyzos et al. (2020)	No Internet data	Tourist arrivals in the SARS 2003 outbreak	Chinese tourist arrivals to US and Australia	LSTM
Kim et al. (2021)	• Search engine • Others	Google Trends, Dummy politics, disease, and seasonal variables	Tourism demand in South Korea	1D-CNN, Bidirectional LSTM, CNN-LSTM, MHAC
Kourentzes et al. (2021)	No Internet data	GDP, PPP, Implied PPP conversion rate, Judgmental adjustments	Tourist arrivals in 20 countries	• Quantitative (Naïve, Seasonal Naïve, ETS, Theta, ARIMA, MLP, ELM, RF, hierarchical) • Qualitative (judgmental adjustments)
Liu et al. (2021)	No Internet data	AR sub-index, SP sub-index, Judgmental adjustments	Tourist arrivals in 20 countries	• Quantitative (Seasonal Naïve, SARIMA, ETS, STL, NN, RF, SVR, hybrid models) • Qualitative (judgmental adjustments)
Qiu et al. (2021)	No Internet data	GDP, CPI, Exchange rate, Judgmental adjustments	Tourist arrivals in 20 countries	• Quantitative (Seasonal Naïve, SARIMA, ETS, STL, TBATS, ELM, MLP, ADLM, SRTVP, Multivariate ELM, Multivariate MLP) • Qualitative (judgmental adjustments)
Zhang et al. (2021)	No Internet data	GDP, CPI, Exchange rate, Dummy SARS 2003, global financial crisis 2008, and social unrest Hong Kong 2019, Judgmental adjustments	Tourism demand in Hong Kong	• Quantitative (ARDL-ECM) • Qualitative (judgmental adjustments)

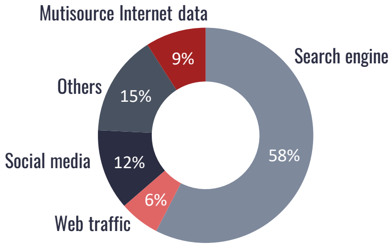
- Most studies use economic and dummy crisis variables
- Most studies use judgmental adjustments and require accuracy assessment for predictions during COVID-19
- The adjustments using Delphi surveys during the COVID-19 period should be updated promptly
- Google Trends data significantly influence tourist arrivals prediction during COVID-19 period

Research Positioning

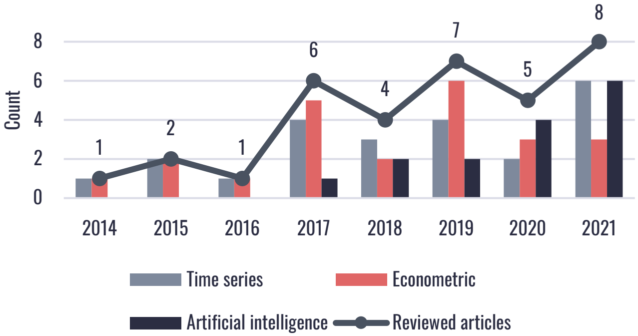
Prediction contexts



Data sources



Methods



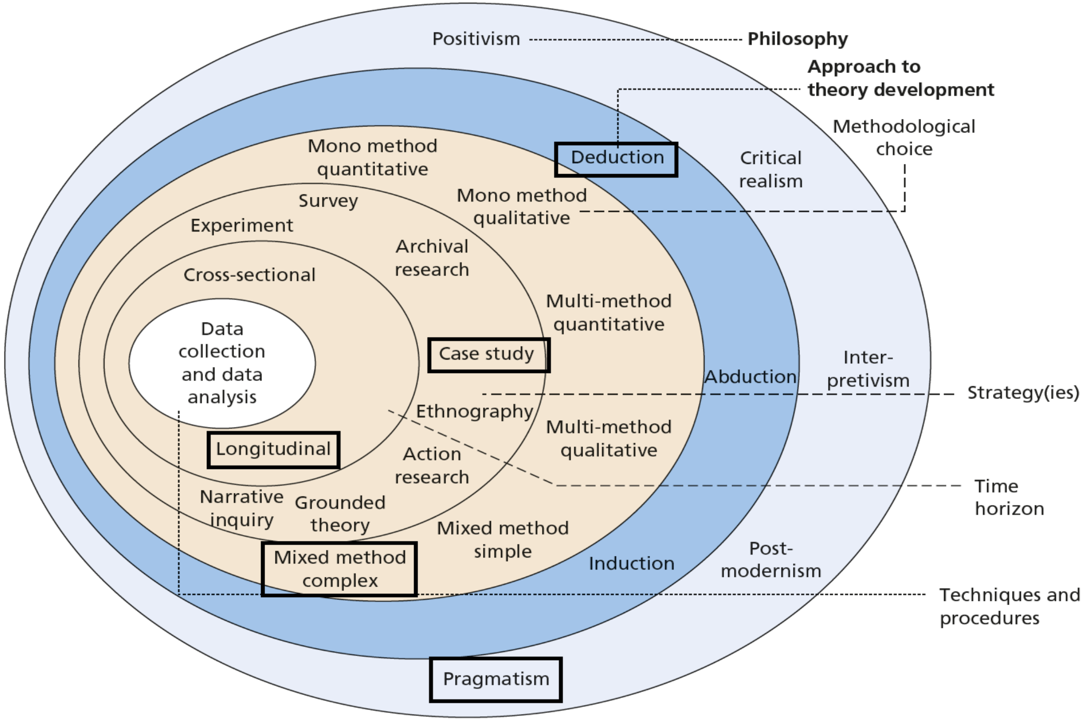
From a total of 34 reviewed articles:

- Most studies aim to predict tourist arrivals to a particular city or province
- Studies in tourism demand prediction using Internet data are **saturated with single-source Internet data**
- Time series and econometric have been well explored while **artificial intelligence-based** models become **more popular** in recent years
- Most studies of tourism demand prediction amid COVID-19 use **economic** and **crisis variables**, and combine quantitative methods with judgmental approach

This study proposes the machine learning-based model using multisource Internet data to predict international tourist arrivals in Indonesia during COVID-19

SBMITB
School of Business & Management

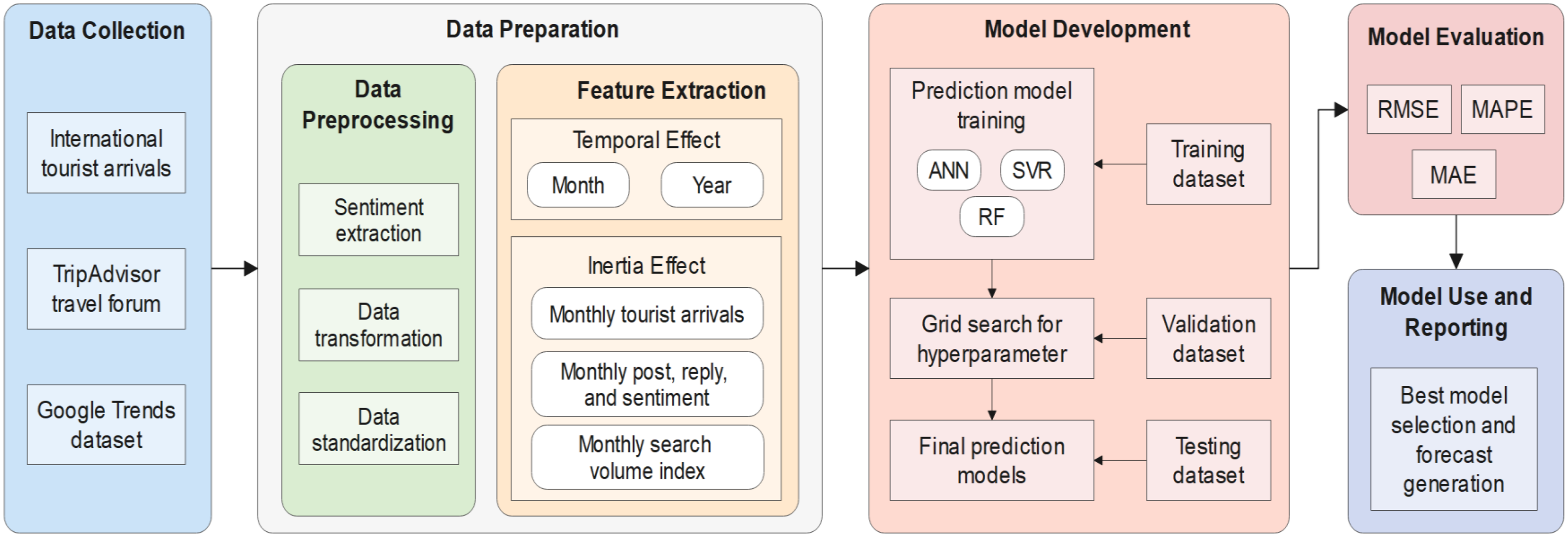
Research Philosophical Position



Note(s): Black boxes indicate the research philosophy, approach to theory development, methodological choice, research strategy, and time horizon used in this study.

- **Pragmatism**
- **Deduction**
- **Mixed method complex**
- **Case study**
- **Longitudinal**
- **Quantitative and qualitative data collection**

Research Framework



Data Collection

Indonesian Statistical Bureau

Data : Monthly International tourist arrivals in Indonesia by entry points

Period : January 2017 – January 2022

TripAdvisor travel forum

Variable	Data type	Data example
Forum	String	Bali
Post topic	String	Do I need to quarantine in Bali if fully vaccinated?
Post message	String	Hi,Can some one please confirm if there is a requirement for foreign nationals to quarantine in Bali if fully vaccinated and show negative covid result?
Post link	String	https://www.tripadvisor.com/ShowTopic-g294226-i7220-k13794978-Do_I_need_to_quarantine_in_Bali_if_fully_vaccinated-Bali.html
Author	String	travelbug
Location	String	Toongabbie, Australia
Author's profile link	String	https://www.tripadvisor.com/Profile/not-toself?tab=forum
Last post date	Date	Jan 24, 2022
Number of replies	Integer	8
Last reply by user	String	MangoMunky
Replier's profile link	String	https://www.tripadvisor.com/Profile/MangoMunky?tab=forum

Data : Daily posts and replies

Period : January 2017 – January 2022

Google Trends

Topic	Keyword
Main entry point	Ngurah Rai International Airport
	Soekarno-Hatta International Airport
	Batam Center Point International Ferry Terminal
	Bali
	Jakarta
	Batam
International travel requirement	Passport Indonesia
	Visa Indonesia
Indonesian tourism	Indonesia hotel
	Indonesia resort
	Indonesia restaurant
	Indonesia travel

Data : Monthly search volume index

Value range : 0-100, with a value of 100 as the peak popularity

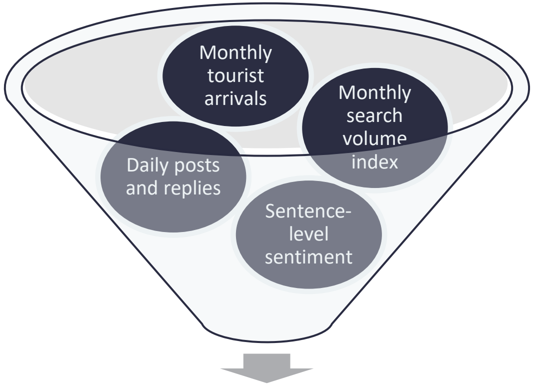
Period : January 2017 – January 2022

Data Preparation

Sentiment Extraction

- Data cleaning
- Lemmatization
- Tokenization
- Stop words removal
- Sentiment extraction
  Lexicon-based approach

Data Transformation and Feature Extraction



Monthly data

Features:

- **Temporal effects**
  Month, Year
- **Inertia effects**
  Tourist arrivals, search volume index, posts, replies, and sentiment score in the previous month

Data Standardization

This process **standardizes the scale** of the features closer to the **standard normal distribution**

$$X_{transformed} = \frac{X - \bar{X}}{\sigma}$$

Where: X is the original value, $\bar{X}$ is the mean, and σ is the standard deviation

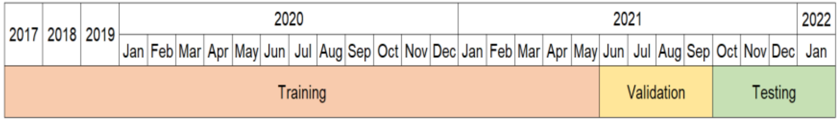
It aims to provide **a single scale** of features without distorting the variations in the value range

Model Development

Model Specification

Construct	Feature	Function	Model			
			1	2	3	4
Temporal	Month	Predictor	v	v	v	v
	Year	Predictor	v	v	v	v
TripAdvisor	Number of posts	Predictor		v		v
	Number of replies	Predictor		v		v
	Sentiment score	Predictor		v		v
Google Trends	Main entry point	Predictor			v	v
	International travel requirement	Predictor			v	v
	Indonesian tourism	Predictor			v	v
Tourist arrivals	International tourist arrivals in the previous month	Predictor	v			v
	Monthly international tourist arrivals	Predicted	v	v	v	v

Data Splitting



Divide the data during COVID-19 with a ratio of :

- 70% for **training** (March 2020 – May 2021)
- 15% for **validation** (June 2021 – September 2021)
- 15% for **testing** (October 2021 – January 2022)

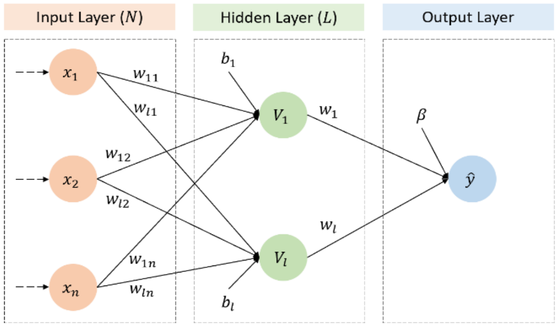
Parameter Optimization

Hyperparameter grid search:

- SVR : Regularization parameter, Kernel, and epsilon
- RF : Number of variables randomly sampled at each split, the number of trees, and maximum nodes
- ANN : Learning rate and the number of hidden layers

Model Development

01 Artificial Neural Networks

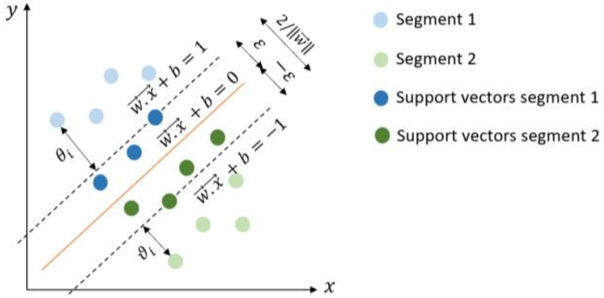


Hidden neuron: $$V_l = \sum_{n=1}^{N} h(w_{ln}\, x_n + b_l)$$

Output neuron: $$\hat{y} = \sum_{l=1}^{L} h(w_l\, V_l + \beta)$$

Where: w_{ln} is the input weight, x_n is the input neurons, b_l is the hidden layer threshold, w_l is the output weight, V_l is the output of hidden neurons, β is the output layer threshold, $h(x)$ is the activation function, $\hat{y}$ is the output neuron

02 Support Vector Regression

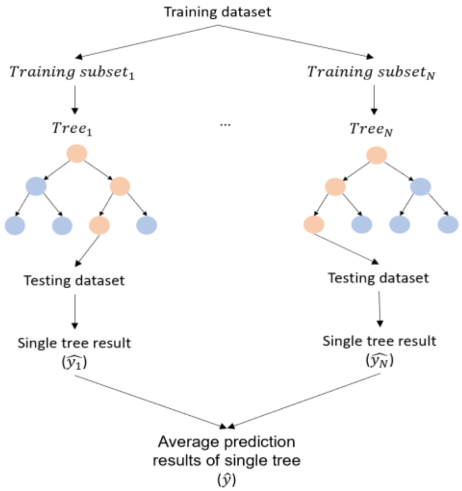


The model seeks a hyperplane to fit the training data points with the fitting function $f(\vec{x}) = \vec{w}.\vec{x} + b$ by minimizing $\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}(\theta_i + \vartheta_i)$

Output function: $$\hat{y} = f(\vec{x}) = \sum_{i=1}^{N}(\alpha_i - \beta_i)K(\vec{x_i}, \vec{x}) + b$$

Where: C is the regularization parameter, θ_i and ϑ_i are distances from the actual value (y_i) to the boundary values of ε, $\hat{y} = f(\vec{x})$ is the mapping function, $\vec{x_i}$ is the training data vectors, α_i and β_i are Lagrange coefficients, $K(\vec{x_i}, \vec{x})$ is the Kernel function, b is the constant

03 Random Forest



Output function: $$\hat{y} = \frac{1}{N_{trees}} \sum_{N=1}^{N_{trees}} \widehat{y_N}$$

Where: $\hat{y}$ is the final output, N_{trees} is the number of trees, $\widehat{y_N}$ is the output of a single tree

SBMITB School of Business & Management

Model Evaluation, Model Use, and Reporting

Model Evaluation

To evaluate the out-of-sample prediction/testing data

Scale-dependent errors:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Percentage error:

$$MAPE\ (\%) = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$$

Where: y_i is the actual value of international tourist arrivals, and $\hat{y}_i$ is the predicted value of international tourist arrivals

Model Use and Reporting

01 Best model selection based on evaluation results

02 Forecast the explanatory variables using Exponential Smoothing and evaluate the predictions

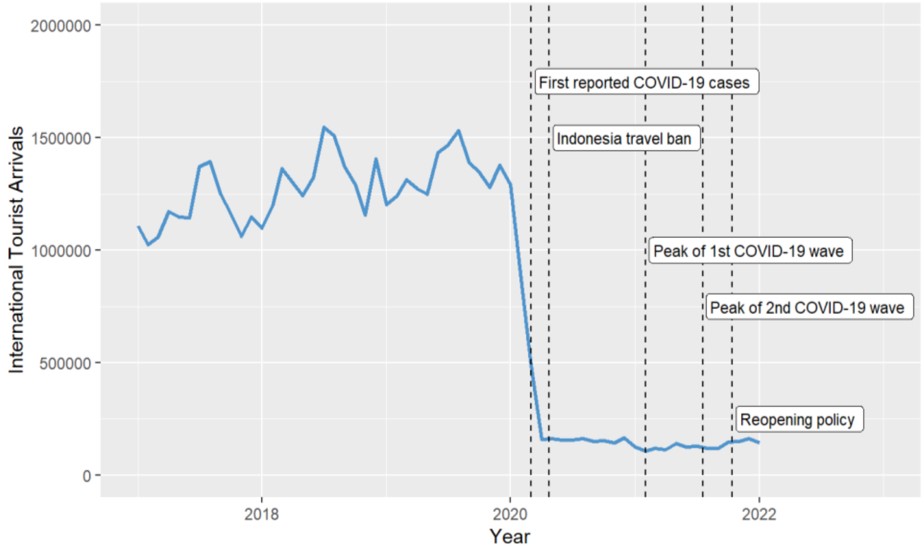
Jan 2017 – Jan 2022	Feb – Dec 2022
Predictions	Forecasts

03 Forecast the international tourist arrivals

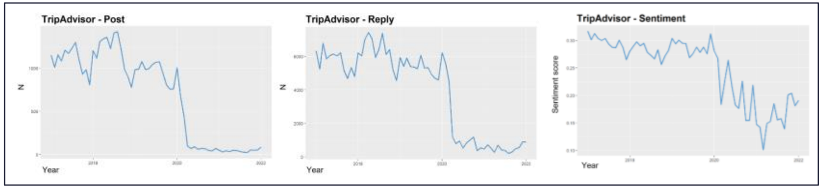
Jan 2017 – May 2021	Jun – Sep 2021	Oct 2021 – Jan 2022	Feb – Dec 2022
Training	Validation	Testing	Forecast
Model development			Model use

04 Reporting

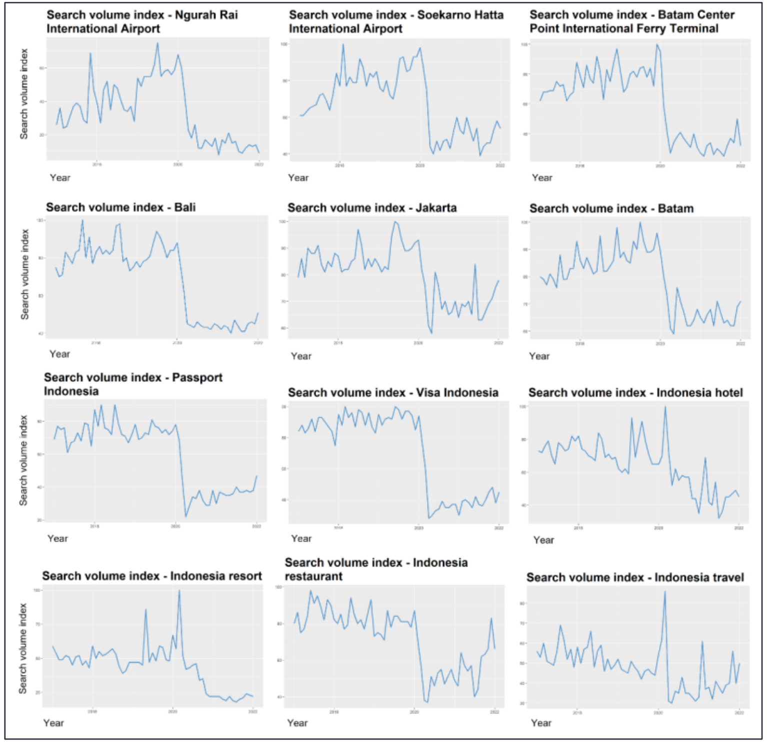
Descriptive Analysis



TripAdvisor



Google Trends



SBMITB School of Business & Management

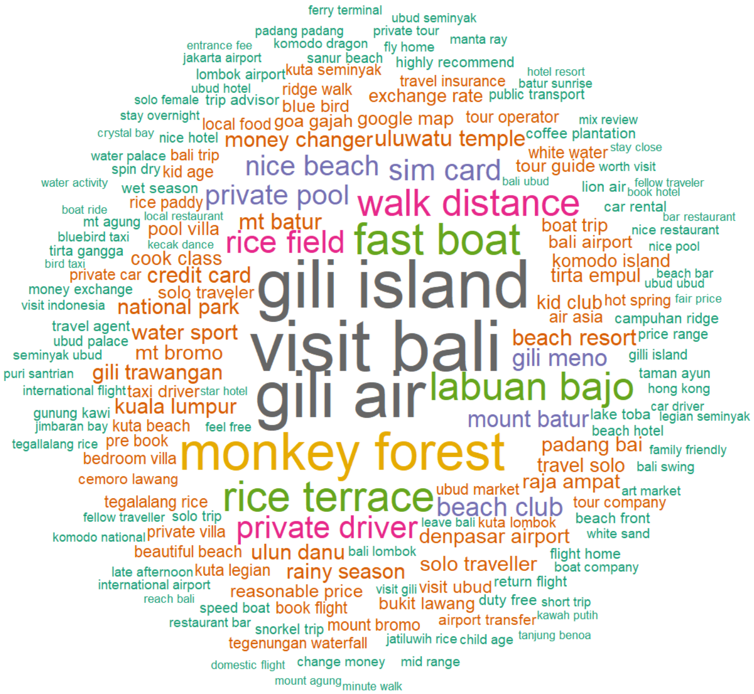
Descriptive Analysis

Data	Variable	Count	Mean	Std. Dev	Min	Max
Indonesian Statistical Bureau Indonesia	International tourist arrivals	61	847,725.49	552,242.05	105,788	1,547,231
TripAdvisor	Post	1,857	22.76	18.36	0	136
	Reply	1,857	128.88	107.32	0	793
	Sentiment	1,857	0.25	0.14	-0.46	1.04
Google trends	Ngurah Rai International Airport	61	34.26	18.31	8	75
	Soekarno-Hatta International Airport	61	68.08	16.80	39	100
	Batam Center International Ferry Terminal	61	60.52	22.86	25	100
	Bali	61	68.30	19.28	40	100
	Jakarta	61	80.26	10.23	58	100
	Batam	61	78.34	11.05	59	100
	Passport Indonesia	61	59.52	19.78	22	90
	Visa Indonesia	61	70.84	26.06	28	100
	Indonesia hotel	61	64.84	14.74	32	100
	Indonesia resort	61	43.92	16.08	18	100
	Indonesia restaurant	61	71.89	16.21	37	98
	Indonesia travel	61	47.85	10.65	30	86

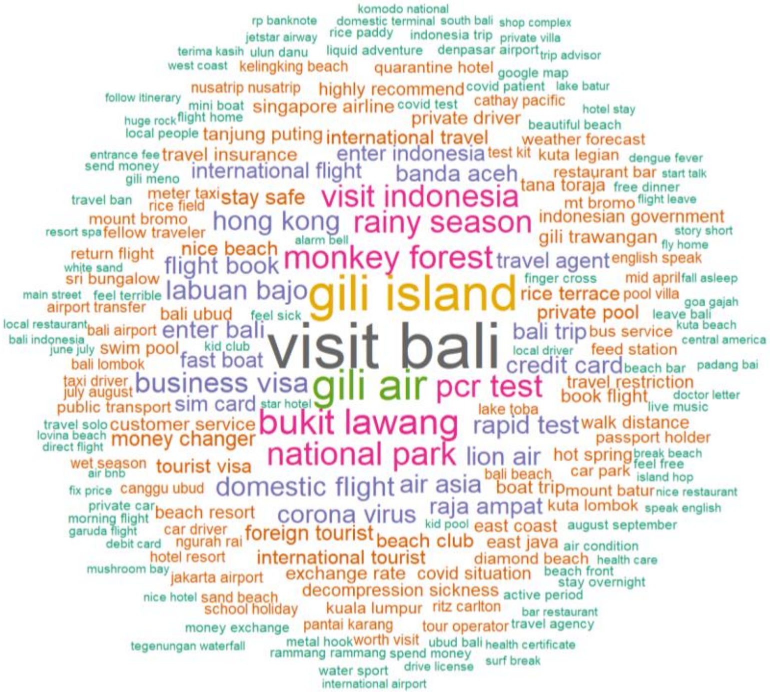
- During the pandemic, tourist arrivals reached their **lowest level**, which is much lower than the average for the last five years
- COVID-19 outbreak **disrupted** tourism industry, as well as **changed** public interest, attention, interaction, and sentiment on online platforms

Textual Analysis

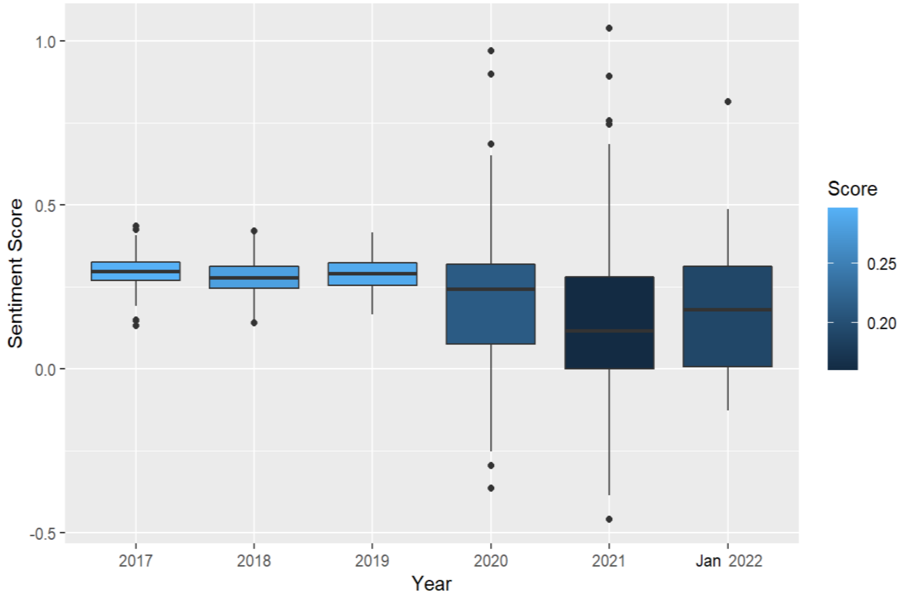
Wordcloud of bigrams in the pre-pandemic period

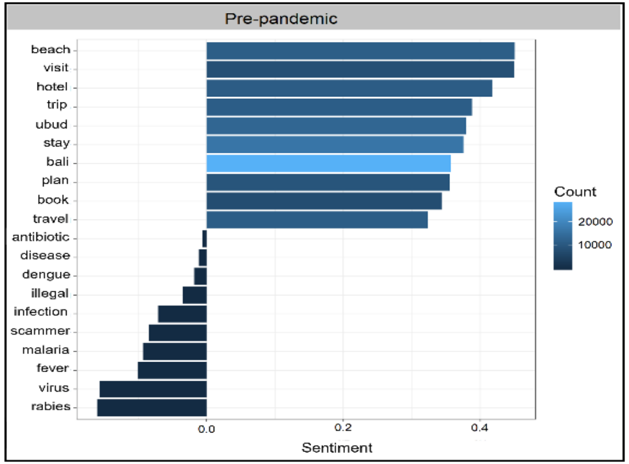


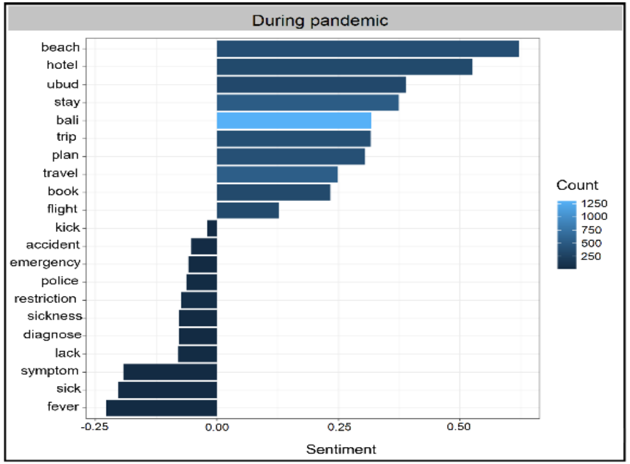
Wordcloud of bigrams in the pandemic period



Textual Analysis







The negative sentiments became more prevalent during the pandemic, resulting in a lower monthly sentiment score than pre-pandemic

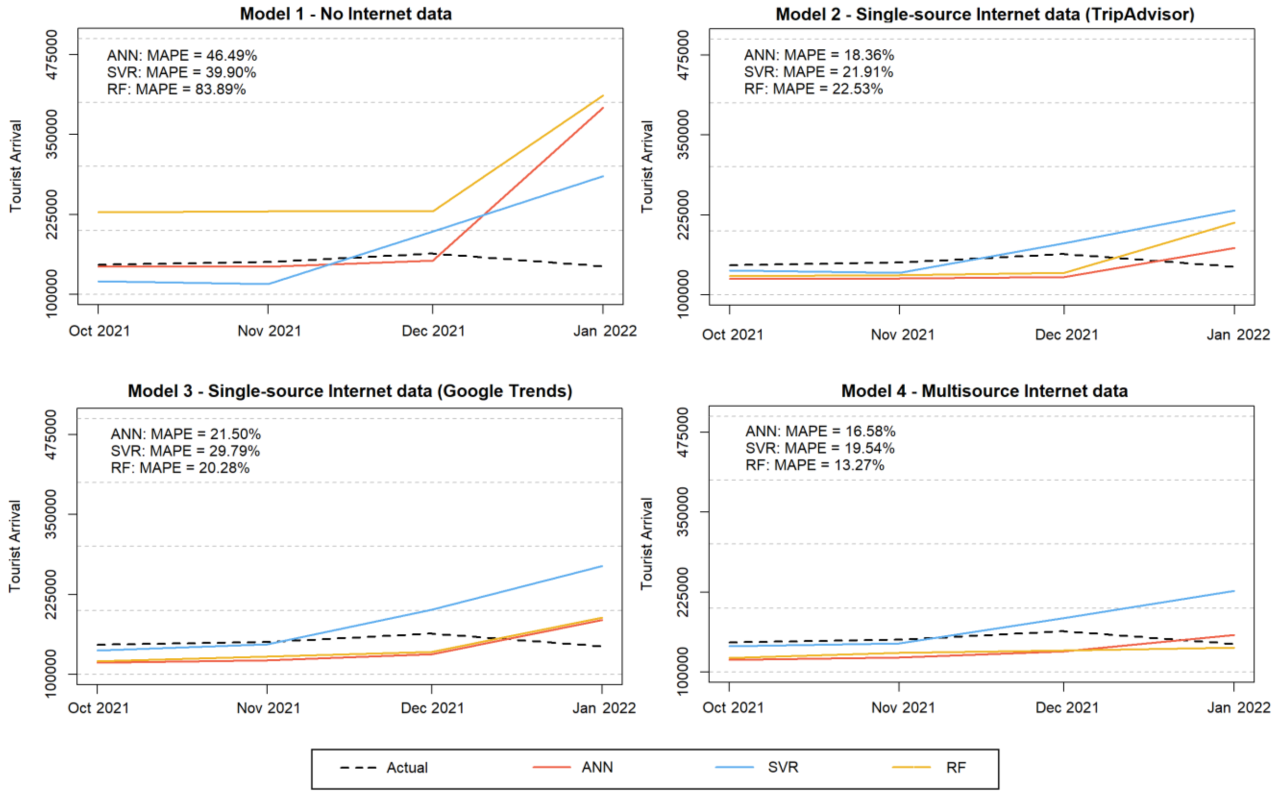
SBMITB School of Business & Management

Prediction Result and Analysis

Model	Predictors	Evaluation metric					
		RMSE	Rel. RMSE	MAPE (%)	Rel. MAPE	MAE	Rel. MAE
ANN	1. Temporal + Previous arrivals	124,504.68	1.00	46.49	1.00	67,239.69	1.00
	2. Temporal + TripAdvisor	28,344.90	0.23	18.36	0.39	27,848.93	0.41
	3. Temporal + Google Trends	32,770.99	0.26	21.50	0.46	32,310.60	0.48
	4. Temporal + Previous arrivals + TripAdvisor + Google Trends	**26,110.94**	0.21	**16.58**	0.36	**25,222.45**	0.38
SVR	1. Temporal + Previous arrivals	75,645.59	1.00	39.90	1.00	58,894.91	1.00
	2. Temporal + TripAdvisor	45,702.95	0.60	21.91	0.55	32,227.40	0.55
	3. Temporal + Google Trends	65,806.17	0.87	29.79	0.75	44,052.59	0.75
	4. Temporal + Previous arrivals + TripAdvisor + Google Trends	**42,963.41**	0.57	**19.54**	0.49	**28,800.96**	0.49
RF	1. Temporal + Previous arrivals	149,024.33	1.00	83.89	1.00	123,841.89	1.00
	2. Temporal + TripAdvisor	39,571.11	0.27	22.53	0.27	33,562.40	0.27
	3. Temporal + Google Trends	31,537.42	0.21	20.28	0.90	30,378.88	0.25
	4. Temporal + Previous arrivals + TripAdvisor + Google Trends	**22,278.41***	0.15	**13.27***	0.16	**20,334.13***	0.16

Note(s): The bold figures indicate the best performing model within a similar prediction model, and * indicates the best performing model across different predictors sets and prediction models.

Prediction Result and Analysis



- By utilizing single-source Internet data, the prediction results of all models improve compared to the first model
- The results **significantly improved** when we applied **multisource Internet data**
- This study confirms the **usefulness** of multisource Internet data for **increasing the accuracy** of tourist arrival predictions

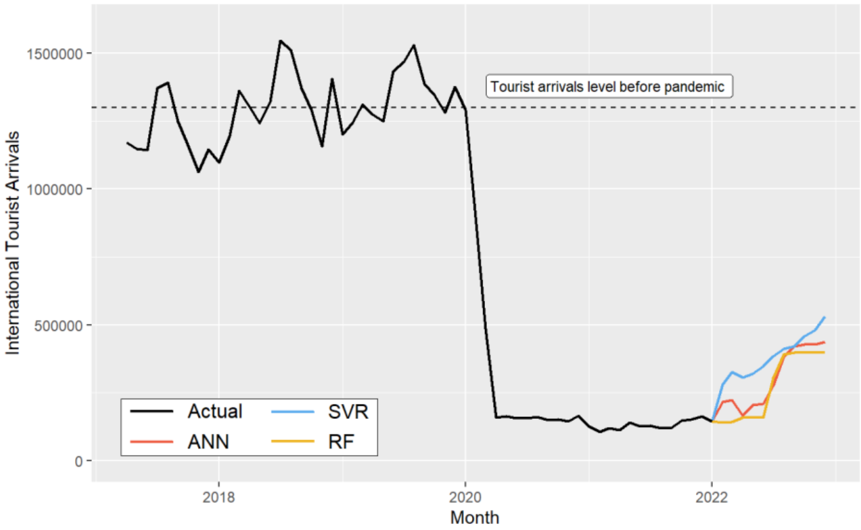
Forecast Result and Analysis

Performance evaluation of explanatory variables prediction

Data sources	Explanatory variables / Predictors	RMSE	MAPE (%)	MAE
Indonesian Statistical Bureau	Previous international tourist arrivals	141,645.70	16.59	99,396.66
TripAdvisor	Post	12.63	37.96	9.68
	Reply	10.68	13.40	8.44
	Sentiment	14.36	20.33	10.98
Google Trends	Ngurah Rai International Airport	8.68	9.73	6.51
	Soekarno-Hatta International Airport	8.62	8.35	6.62
	Batam Center International Ferry Terminal	7.12	6.68	5.19
	Bali	10.54	15.27	7.92
	Jakarta	9.61	13.92	7.43
	Batam	11.55	12.85	8.20
	Passport Indonesia	12.86	16.72	7.97
	Visa Indonesia	11.72	13.93	9.15
	Indonesia hotel	11.88	17.95	8.04
	Indonesia resort	157.81	35.59	108.85
	Indonesia restaurant	999.78	39.45	696.90
	Indonesia travel	0.03	9.36	0.02

Good and reasonable predictive power*

Forecasting result using the best prediction model of each method



- The tourist arrivals level at the end of 2022 is still **far below** the level before the COVID-19 pandemic
- The **accurate tourism demand prediction** is helpful for understanding **tourism demand recovery**

*Lewis (1982)

SBMITB School of Business & Management

General Conclusion

RQ1 *"How to develop a predictive analytics model of international tourist arrivals in Indonesia amid the COVID-19 pandemic using multisource Internet data?"*

STEP 01 | Collect the data from the Indonesian Statistical Bureau, TripAdvisor travel forum, and Google's search engine

STEP 02 | Prepare the data by performing data preprocessing (i.e., sentiment extraction, data transformation, and data standardization) and feature extraction

STEP 03 | Define model specification, split the data into training-validation-test dataset, and develop the prediction model using machine learning method

STEP 04 | Evaluate the out-of-sample prediction based on RMSE, MAPE, and MAE

STEP 05 | Compare the accuracy against all models and use the best prediction models to forecast the international tourist arrivals in Indonesia

RQ2 *"Does utilizing multisource Internet data lead to a more accurate tourist arrivals prediction during the COVID-19 pandemic than single-source Internet data and historical tourist arrivals data?"*

- This study shows the positive impact of combining **multisource Internet data** to **improve prediction accuracy** in terms of RMSE, MAPE, and MAE
- The prediction models trained using **multisource Internet data** are **consistently more accurate** in predicting international tourist arrivals than those trained using single-source Internet data and historical tourist arrivals data
- The superiority of the **multisource Internet data** is **consistent across different methods**

Research and Practical Implications

Research Implications

01 | This study pioneers the practice of a multisource Internet data approach in predicting tourist arrivals amid the COVID-19 pandemic

02 | This study has validated the use of multisource internet data to improve prediction performance during the COVID-19 pandemic

03 | This study is one of the few research to provide perspectives on the current state of Indonesia's tourism demand

Pratical Implications

01 | The proposed predictive model reinforces the foresight capabilities that can help the government to better formulate the corresponding policy for the tourism industry

02 | The tourism managers can use the proposed predictive model to support their resource allocation planning, pricing strategies, and contingency plans formulation during and after the COVID-19 pandemic

03 | The fast-growing Internet data allows tourism practitioners to analyze visitor interests, attentions, and sentiment in the online platforms

Limitations and Further Research

Limitations

This study focuses on the prediction of international tourist arrivals in Indonesia. The selected keywords are limited and solely represent this country's public interests and attention.

This study only combines two Internet data: social media data from TripAdvisor and search volume data from Google Trends.

This study examines the proposed multisource Internet data using three well-explored machine learning models.

Suggestions

- Investigate other data sources relevant to the specific contexts and explore the application of multisource Internet data for different contexts.
- Extending the forecast horizon to gain more practical insights and explore other methods to forecast the explanatory variables.

Other Internet data sources or external factors can be further examined as input for the prediction model to enrich the training data during model development.

Other advanced methods can be investigated in further research.

Publications

Main publication:

Andariesta, D.T. and Wasesa, M. (2022), "Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach", *Journal of Tourism Futures*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JTF-10-2021-0239 **(Published)**

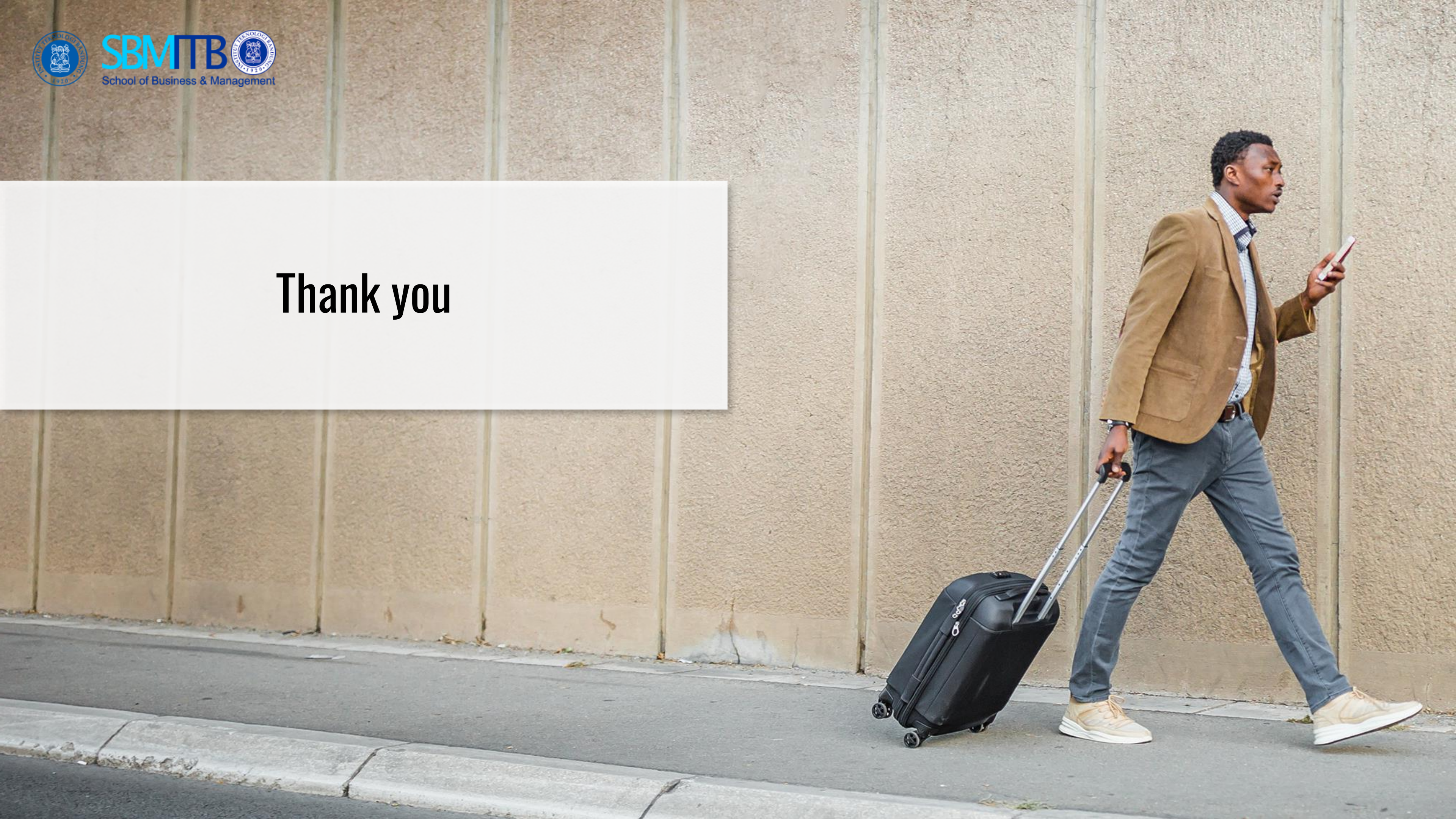
Andariesta, D.T. and Wasesa, M., "Using a mixed-method of social media analytics to analyze travel behaviour during the COVID-19 pandemic: Evidence from Indonesia" **(Draft to be submitted to Annals of Tourism Research)**

Additional publication:

Wasesa, M., **Andariesta, D. T.**, Afrianto, M. A., Haq, I. N., Pradipta, J., Siallagan, M., Leksono, E., Iskandar, B. P., & Putro, U. S. (2022). Predicting Electricity Consumption in Microgrid-Based Educational Building Using Google Trends, Google Mobility, and COVID-19 Data in the Context of COVID-19 Pandemic. *IEEE Access*, 10, 32255–32270. https://doi.org/10.1109/ACCESS.2022.3161654 **(Published)**

Wasesa, M., Hidayat, T., **Andariesta, D.T.**, Natha, M.G., Attazahri, A.K., Afrianto, M.A., Mubarok, M.Z., Zulhan, Z., Putro, U.S., "Economic and Environmental Assessment of an Integrated Supply Chain System for Lithium-Ion Battery Waste Recycling: A Hybrid Simulation Approach". *Journal of Cleaner Production.* **(Submitted)**

Andariesta, D.T. and Wasesa, M. (2021), "Machine Learning Models to Predict the Engagement Level of Twitter Posts: Indonesian E-commerce Case Study", *In Proceedings of the 6th International Conference on Computer Science and Computational Intelligence (ICCSCI)* **(Published)**



Thank you