A Machine-Learning-Based Approach for Tourist-Arrival Trend Prediction

I Putu Edy Suardiyana Putra*  
Department of Digital Business  
Institut Pariwisata dan Bisnis Internasional, Bali, Indonesia  

Denok Lestari  
Department of Digital Business  
Institut Pariwisata dan Bisnis Internasional, Bali, Indonesia  

Komang Ratih Tunjungsari  
University of Otago, Dunedin, New Zealand  
* edy.suardiyana@ipb-intl.ac.id

Abstract  
This study proposes a machine-learning-based technique to predict trend in tourist arrivals based on online news headlines and the number of previous tourist arrivals. Tourist arrivals prediction is important to give information to destinations’ local governments and businesses to prepare their services. We use Logistic Regression and Support Vector Machine to create a model to predict the increase in tourist arrivals monthly. News headlines from three online Indonesian news portals are used. A total of 47,298 online news headlines were collected. The results show that Logistic Regression can achieve up to 67.4% of F-score while Support Vector Machine can achieve up to 62.9% of F-score. These results show that adding online news headlines and machine-learning algorithms can give significantly better results in predicting tourist arrivals.

Keywords  

© 2023 Authors. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0). All writings published in this journal are personal views of the authors and do not represent the views of this journal and the author’s affiliated institutions.
Introduction

Tourism is one of the fastest growing industries in the world without a doubt. Forecasting tourism demand is important to reduce the risk of the tourism service vendors losing their business since the tourism service or product is short-lived. Although there are many studies have been conducted to show tourism demand, a common metric that is used to measure tourism demand is tourist arrival (Song & Li, 2008). Forecasting tourist arrival is important to the destination’s local government and numerous tourism service vendors to develop their strategic planning so that they can get benefit from foreign exchange income and other potential economic benefits (Yuan, 2020). Thus, predicting tourist arrivals is crucial for the tourism industry.

Studies from (Li, 2022; Purnaningrum & Athoillah, 2021; Xie et al., 2021; N. Yu & Chen, 2022; Yuan, 2020), try to develop techniques to predict tourism demand using machine learning. Data from Twitter, the Baidu search engine, the Central Bureau of Statistics of Indonesia, and past tourism arrival data are used by these studies to train the machine learning algorithms. Their research shows that machine-learning-based algorithms are able to predict tourist arrival relatively well. Another study from (Park et al., 2021), shows that including online news data can improve the performance of the algorithm to predict tourist arrival. Park et al. also mention that although using data from online news can give a better result, this data source remains unexplored.

In his study, shows that news readers usually scan the headlines, and rarely read the entire text (Dor, 2003). Studies from (Liu et al., 2018; Oncharoen & Vateekul, 2018), show that using news headlines can produce relatively good results in predicting stock price. Therefore, this study uses online news headlines together with the number of previous tourist arrivals as the main source of data to train the machine learning algorithm.

The main contribution of this study is to propose a machine-learning-based approach to predict the trend of tourist arrivals (whether it increases or decreases) using news headlines and the number of previous tourist arrivals as the data source. We use Bali (one of the provinces in Indonesia) as a case study because its economy is highly impacted by the tourism industry (Antara & Sumarniasih, 2017). When the pandemic hit the world, Bali was highly impacted in a negative way (Subadra & Hughes, 2022). We hope that the results of this study can help tourism service providers and the Balinese local government to prepare when another threat comes in the future.

Our experiment shows that Logistic Regression can get up to 67.4% F-score while the Support Vector Machine algorithm can get up to 62.9% F-score. We also show that using both the number of tourist arrivals and online
news headlines can significantly increase the performance (in terms of recall and F-score) of the classifier when the Logistic Regression is used.

The rest of the paper is organized as follows. Section II explains our related work, while Section III shows the methodology. Section IV provides results and analysis and Section V provides discussions. Our conclusion and future work can be found in Section VI.

Related Work
Since predicting tourist arrivals (especially the its trend) is important for the tourism industry, many studies have proposed techniques to predict tourist arrivals. Some of those studies use machine learning techniques, both supervised and unsupervised, to create an accurate predictor to predict tourist arrivals.

In 2011 worked on a system that can analyse tourism information in local city (Yuan, 2020). They implemented a sentiment analysis in their system to classify opinions about tourism location based on Twitter’s data. They use the Naïve Bayes algorithm to classify the Twitter data. They claim that their sentiment analysis system can classify the opinion with an accuracy of 89%.

In 2019, propose a deep-learning-based technique to predict tourist arrival in Macau Region, China (Law et al., 2019). They use keywords that people type in Baidu and Google Search Engine as the data to feed their deep-learning-based technique. The data collected are from January 2011 to August 2018. They show that the deep-learning-based technique can outperform support-vector-regression (SVR) and artificial neural network (ANN) techniques. This study shows that using machine learning in tourist arrival prediction has promising results.

In 2020 also show that machine-learning and Google keywords can be relatively effective to increase the accuracy of tourist arrivals prediction (Höpken et al., 2021). Their study focuses on predicting tourist arrivals in Denmark based on Google keywords and past arrivals. Höpken et al.’s study shows that using an Artificial Neural Network (ANN) can outperform Autoregressive Integrated Moving Average (ARIMA).

In 2020 investigates whether machine learning techniques can predict a hotel booking cancelation in Grand Canaria (Spain) (Sánchez-Medina & C-Sánchez, 2020). They treat their study as a binary classification (“canceled” and not canceled). In their study, they used four machine learning algorithms: Random Forest, Support Vector Machine (SVM), Decision Tree C5.0, and Artificial Neural Network (ANN). They show that ANN can outperform other machine learning algorithms by achieving a 98% of accuracy and a 97.9% of F-score. Inspired by Sánchez-Medina et al., this study treats the problem as
a binary classification, where we split the class into two: increase (class 1) and decrease (class 0).

Use the least squares support vector regression model with gravitational search algorithm (LSSVR-GSA) to predict Chinese Cruise tourist demand (Xie et al., 2021). They use Baidu search queries as the basis for the prediction. They show that LSSVR-GSA algorithm can outperform other algorithms such as autoregressive integrated moving average (ARIMA), back propagation neural network (BPNN), Radial Basis Function (RBF), and LSSVR-CV.

Purnaningrum and Athoillah [5] use Support Vector Machine (SVM) in predicting tourist arrivals in East Java. They use past tourist arrival as the variable to train their SVM-based technique. Although they mention that their technique receives relatively good results, there is not any comparison between their SVM-based technique with other machine learning algorithms.

Abdurahman, recently published an article about their study on predicting the number of visitors who come to visit protected areas in Sarawak, Malaysia in 2022 (Cavnar et al., 1994). They compare five machine learning algorithms including k-Nearest Neighbor (k-NN), Naïve Bayes, and three different types of Decision Tree. They show that machine learning algorithms can achieve up to 84.35% accuracy.

Yu and Chen, published a paper that proposes a technique called auto-encoder long short-term memory (AE-LSTM) network, that is used to predict tourism demand (N. Yu & Chen, 2022). This technique is an improvement of long short-term memory (LSTM) technique. They found that AE-LSTM can achieve lower mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean square error (RMSE) average values compared to LSTM.

From the studies mentioned above, it can be seen that using machine learning algorithms for tourist arrivals prediction can give relatively good results. Thus, in this study, we focus on using machine learning techniques to predict tourist arrivals.

To feed the machine learning algorithms, we decide to add online news headlines and the number of previous tourist arrivals (one-month window). This is inspired by studies that are conducted by Wang (Brylla & Hughes, 2004; L. Yu & Liu, 2004). In their study, Wang et al. try to give a perception of Singapore to residents who live in Hongkong and never travel to Singapore. Their study shows that the participants’ perception of Singapore is based on the type of news (positive or negative) that they have been told. This means that adding news headlines as an additional feature can give a high impact on people’s perceptions.
Using online news as an additional feature to improve the classification prediction have been done by some researchers in other fields. For example, (L. Yu & Liu, 2004) use news headlines and contents to predict stock price. They algorithm can achieve an up to 90% of accuracy when news headlines and contents are used together as the features.

Use financial news headlines as the features to predict stock movement (Liu et al., 2018). As the learning algorithm, they used long short-term memory (LSTM) and support vector machine (SVM) algorithms. They use convolutional neural network (CNN) as the feature extraction technique. Their technique is able to achieve 55.44% of accuracy, where they claim that this result is better than traditional machine learning techniques.

Talking about the impact of news on a real estate industry, (Hausler et al., 2018) shows that a news-based sentiment has a significant relationship with real estate market movements. To classify the sentiment of the news, (Hausler et al., 2018) implement a support-vector-machine-based classifier. The classifier categorised the news into three: positive, negative, and neutral. They claim that their study is the first study that capture textual sentiments and show their relationship to US real estate markets.

Use real-time news data to predict investors’ buying behavior (Yadav et al., 2019). To extract features, they use a vector space model (VSM) with binary weight. The VSM converts the unstructured news headlines to structured feature vector. To predict the sentiment of the news, Yadav et al. implements Naïve Bayes and soft-margin Support Vector Machine (SVM). Their experiment shows that their classifier can achieve an up to 0.31 of precision, 0.33 of recall, and 0.32 of F-score.

Lingwu and Ow, uses supervised machine learning algorithms to predict a sentiment from news. Then, this study analyses the impact of the news sentiment to stock market prices (Flach & Kull, 2015). They implement several supervised machine learning algorithms such as Stochastic Gradient Descent (SGD), Gaussian Naïve Bayes, Multinomila Naïve Bayes, Complement Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine, and Random Forest. The highest accuracy is achieved by using Stochastic Gradient Descent (SGD) with a 70% of accuracy.

Studies above have shown that using both machine learning algorithms and news (especially their headline) can give benefits in tourism industry and other industries as well.

Referring to the studies above, we use machine learning to build the classifier model, where online news headlines and the number of previous tourist arrivals are used as the features. This methodology in this study is slightly different from the above studies. The studies above intend to analyse the relationship between the sentiment of news and the other factors (for
example stock market price). Meanwhile, our study tries to use the news headlines as a feature without knowing their sentiment. We implement two machine learning algorithms including Logistic Regression and Support Vector Machine (SVM).

Method
Dataset
Table 1 shows the numbers of data that we collected from three Indonesian news portals namely Detik, Kompas, and CNN. To crawl the headline news data from these news portals, we use a Python library called Beautiful Soup\(^1\). News headlines data from 2014 until 2019 were collected. We choose 2014 as the starting point because we cannot collect data from the year 2013 backward from CNN. We excluded data from 2020 onward because many travel restrictions have been applied from 2020 until 2021 because of the COVID-19 pandemic, where these restrictions affected traveling activities.

Table 1. Dataset

<table>
<thead>
<tr>
<th>News portal</th>
<th>URL</th>
<th>Number of news headlines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detik Travel (General)</td>
<td><a href="https://travel.detik.com">https://travel.detik.com</a></td>
<td>27969</td>
</tr>
<tr>
<td>Detik Travel Bali (with “Bali” keyword)</td>
<td><a href="https://travel.detik.com">https://travel.detik.com</a></td>
<td>6164</td>
</tr>
<tr>
<td>Kompas Travel (with “Bali” keyword)</td>
<td><a href="https://travel.kompas.com">https://travel.kompas.com</a></td>
<td>4384</td>
</tr>
<tr>
<td>CNN Travel (General)</td>
<td><a href="https://travel.cnn.com">https://travel.cnn.com</a></td>
<td>8781</td>
</tr>
</tbody>
</table>

After crawling the data from their respective resources, we combine and cluster them on a monthly basis. Each row of our data set represents a collection of news headlines from the three news portals. Thus, we have a total of 71 rows starting from February 2014 until December 2019. We exclude January 2014, because we could not get news headlines data from December 2013. Table 2 shows the number of news headlines from the three news portals clustered by the year they are published.

Table 2. Total number of news headlines based on the year they are published

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of news headlines</th>
</tr>
</thead>
</table>

---

\(^1\) https://beautiful-soup-4.readthedocs.io/en/latest/
To label each row, we use an increase and a decrease in the number of tourist arrivals in Bali Province. We label the row as "1" if the number of the respective month is higher than the previous month. This we call it a positive trend. On the other hand, we label the row as "0" if it's a decrease in the number of tourists arriving compared to the previous month. We call this label as a negative trend. We achieved the number of tourist arrival in Bali Province from the Central Statistics Bureau of Indonesia. Since the label consists of two categories, this study uses binary classification. Table 3 shows the composition of our dataset's label.

<table>
<thead>
<tr>
<th>Label</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;0&quot;</td>
<td>36</td>
</tr>
<tr>
<td>&quot;1&quot;</td>
<td>35</td>
</tr>
</tbody>
</table>

### Data Pre-Processing

Before doing the training and testing for machine learning, we do some data pre-processing tasks. At first, we remove duplicated news headlines from our data set. The second step that we do is removing stop words from the news headlines. For this task, we use NLTK library that is written in Python programming language.

When our news headlines are cleaned from the stop words, the next step that we do is decomposing the news headlines using N-gram, where $N = \{1,2,3,4,5,6\}$. Figure 1 shows an illustration of the N-gram technique where $N = \{1,2,3\}$. Based on Cavnar, since N-gram decomposes a string into smaller parts, any errors that appear in one part may not affect the other parts (Cavnar et al., 1994). Our complete pre-processing steps are shown in Figure 2.

---

Feature Extraction

Features that are used in this research are categorized into two: (1) the number of tourist arrivals from the previous month and (2) online news headlines. Since online news headlines are in string format, we use TFIDF (term-frequency times inverse document-frequency) to extract the features. TFIDF is a common technique that is used as a weighting mechanism in document classification [28]. We use TFIDF to scale down the impact of tokens (a word or combination of words) that occur very frequently, where these tokens are empirically less useful than tokens that occur less frequently. TFIDF is calculated using the following formula

$$ \text{tfidf}(t, d) = \text{tf}(t, d) \times \text{idf}(t), $$

and we calculate the idf($t$) with the formula below

$$ \text{idf}(t) = \log\left[\frac{n}{\text{df}(t)}\right] + 1 $$

where $n$ is the total number of documents in the document set and df($t$) is the number of documents in the document set that has the term $t$. We use the ready-to-use library for TFIDF from Scikit-learn for this study namely TfidfVectorizer().

Figure 1. N-gram example, where $N=\{1,2,3\}$

Figure 2. Data pre-processing
Feature Selection

In theory, using more features can make the classifier’s detection rate increase [29]. However, using more features can make the learning process slower or even can reduce the classifier detection rate, as there may be some irrelevant or redundant features. Moreover, extracting more features can affect the system’s computational cost by increasing it [30]. Therefore, the number of features needs to be reduced by using a feature-selection technique.

In general, feature-selection techniques consist of three categories: wrapper, filter, and embedded. Both the wrapper and embedded techniques must use a learning algorithm for selecting features. The wrapper methods [31] use a learning algorithm (machine-learning algorithms) to choose a subset of features based on the machine-learning performance. On the other hand, the embedded methods choose features during the training process of the machine-learning algorithm [32].

Different from wrapper and embedded where those techniques involve machine learning algorithms, the filter methods [33] do not have any dependency on machine-learning algorithms. To select a subset of features, the filter methods need less computation than the wrapper methods. This gives the filter-based technique an advantage compared to wrapper techniques. However, the filter methods overlook features that can provide better information when they are combined together [31]. This weakness can affect the accuracy of the classifier by reducing it. The main disadvantage of the wrapper methods is that they have a higher computational cost compared to the filter-based methods [31] because the wrapper methods are required to test all possible feature subsets and select a subset of features that can give the highest accuracy. By doing an exhaustive search in a feature space of N features, the embedded techniques are required to assess \(2^n\) possible combinations of feature [33].

This study uses a filter-based feature selection technique to eliminate the number of features that are generated by the TFIDF technique. We use SelectKBest which is a library for feature selection from Scikit-learn using the Python programming language [34], where this technique is categorized as a filter-based selection technique. This technique gives a score to each feature, then non-increasingly ranks them. The number of selected features (K) is defined by the user. A score function is used to measure each feature. This study uses the chi-square test (\(\chi^2\)) as the score function for this study.

Machine Learning Algorithms

Logistic Regression
Logistic Regression (LR) is a machine-learning technique that gives the classifier the ability to estimate categorical outcomes (can be 2 or more categories) from different predictors, where those predictors can be either categorical, continuous, or both [35]. This estimation can be implemented by using the following formula,

$$P(Y|x) = \frac{1}{1+e^{-(\omega_0 + \omega_1 x_1 + \ldots + \omega_n x_n)}}$$

Where $P(Y|x)$ is a function to estimate the probability of class Y given x, $x_n$ is the predictor, and $\omega_n$ is a weight (sometimes called a regression coefficient).

For the two-class (or binary class) case, if $P(Y|x) > 0.5$, then this outcome is classified as 1 (or True) by the classifier. On the other hand, if $P(Y|x) \leq 0.5$, then the classifier classified the outcome as 0 (False). For this study, Y is a binary class that represents an increase in tourist arrival (True) and a decrease in tourist arrival (False), while $X = \{x_1, x_2, \ldots, x_n\}$ is the set of features used. For the Scikit-learn library, regularisation is implemented to avoid the model/classifier remembering the training data (called data overfitting). To get the value of $\omega_i$ for a feature $x_i$, LR with L2 regularisation minimises the following cost function

$$\min_{\omega + c} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{n} \log(exp(-y_i(X_i^T \omega + c)) + 1),$$

where n, c, and C are the number of instances, an intercept of the LR, and an inverse of the regularisation strength (or penalty parameter [36]), respectively. The value of C remains unchanged during the training process because C is a hyper-parameter. Higher value for C indicates a weaker regularization while a lower C means a stronger regularization. This regularization parameter is useful to reduce data overfitting.

More detailed information about the library can be found in Scikit-learn [34] and Fan et al. [36]. For this study we use a ready-to-use Logistic Regression library from the Scikit-learn [34].

**Support Vector Machine**

A support vector machine (SVM) does its training process by maximizing the margin between the classes in the training set [37]. Training examples that are near the decision boundary are called support vectors and the decision
boundary is called a hyperplane. Given \( n \) training inputs \( X = x_1, x_2, x_3, \ldots, x_n \) with labels \( Y = \{-1, 1\} \),

\[
\begin{align*}
\text{where } Y_k &= 1, \text{ if } x_k \in \text{ class } A \\
Y_k &= -1, \text{ if } x_k \in \text{ class } B
\end{align*}
\]

then the decision function \( D(x) \) is given by

\[
D(x) = \sum_{k=1}^{n} \alpha_k K(x_k, x) + b,
\]

where \( K(x_k, x), b, \alpha_k \) are the pre-defined kernel, the error bias, and coefficients that need to be adjusted during the training process. For this study we use the Gaussian radial basis function (RBF) as the kernel for the SVM, with the following formula \( K(x, x') = \exp(-\|xx'\|^2/\gamma) \), as the kernel for the SVM, where \( \gamma \) is the inverse of the radius of influence of samples selected by the model as support vectors.

**Method**

Figure 3 shows our method starting from data pre-processing until training and testing for the machine learning algorithm. The first step that we do is data pre-processing followed by feature extraction. We do a feature selection after the feature extraction step to reduce the dimensionality of the features. Reducing the number of features sometimes can give better results and shorten the training process.
Figure 3. The research method

For the training and testing process, we use k-fold cross-validation to measure the performance of the machine learning algorithms, where $k = 10$. Figure 4 shows an illustration of the k-fold cross-validation.

![k-fold cross validation illustration](image)

Figure 4. k-fold cross validation

We calculate the precision, recall, and F-score since this is a binary classification and these metrics are robust against data imbalance [38]. The precision, recall, and F-score are calculated using the following formulas.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}},$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}},$$

$$F - \text{score} = \frac{2 \times \text{True Positive}}{2 \times \text{True Positive} + \text{False Positive} + \text{False Negative}}.$$

In result and analysis section, we divide this section into two: (1) a performance comparison between the Logistic-regression-based classifier and the Support-vector-machine-based classifier when the feature used is number of previous tourist arrivals is used, and (2) a performance comparison between the Logistic-regression-based classifier and the Support-vector-based classifier when both news headlines and number of previous tourist arrivals are used as the feature.
RESULTS AND DISCUSSION

Result
For this sub-section, we split into two parts. Firstly, we compare the Logistic Regression with the SVM algorithms using only the previous tourist arrivals as the feature. In the second part, we compare the performance of the two machine-learning algorithms when the previous tourist arrivals and online news headlines are combined as features.

Results when only The Previous Tourist Arrivals are Used as The Features
Table 4 shows precision, recall, and F-scores achieved by Logistic Regression and SVM algorithms when only the number of previous arrivals is used. Logistic Regression is able to achieve a 72.0±19.9% precision, a 62.5±5.9% recall, and a 57.0±11.2% F-score. On the other hand, the SVM can achieve a 73.5±19.2% of precision, a 66.7±13.0%, and a 62.1±16.6% of F-score. From the precision, recall, and F-score values, we can see that Support Vector Machine achieves slightly better results.

Table 4. Precision, recall, and F-score produced by using only previous arrivals as the feature

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Logistic Regression</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (%)</td>
<td>72.0±19.9</td>
<td>73.5 ±19.2</td>
</tr>
<tr>
<td>Recall (%)</td>
<td>62.5±5.9</td>
<td>66.7±13.0</td>
</tr>
<tr>
<td>F-score (%)</td>
<td>57.0±11.2</td>
<td>62.1±16.6</td>
</tr>
</tbody>
</table>

To measure the performance difference between the two algorithms, we use the Wilcoxon Signed Rank Test. The Wilcoxon Signed Rank Test shows the difference between the F-score values that are achieved by the Logistic Regression algorithm and the SVM algorithm is not significant (p-value= 0.4). Similar to the F-score values, the precision and recall values between the two algorithms are not significantly different, with p-values equal to 0.4 and 1, respectively.

Results when the Tourist Arrivals and On-Line News Headlines are Used as The Features
Precision, recall, and F-score values from the two machine learning algorithms, when both the tourist arrivals and online news headlines are used, are shown in Table 5. Logistic Regression is able to get slightly better precision, recall, and F-score compared to SVM. The Logistic Regression algorithm can get a 78.7±9.1% of precision, a 70.4±10.1% of recall, and a 67.4±11.6% of F-score. Meanwhile, the SVM algorithm is able to achieve a 74.6±15.0% precision, a 66.7±14.0% recall, and a 62.9±14.7% F-score. Based on the Wilcoxon Signed Rank Test, we found that the precision, recall, and F-score between the two machine learning algorithms are not significantly different, with p-values equal to 0.6, 0.6, and 0.2 respectively.

Table 5. Precision, Recall, and F-score produced by using news headlines and previous arrivals as the features

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Logistic Regression</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (%)</td>
<td>78.7±9.1</td>
<td>74.6±15.0</td>
</tr>
<tr>
<td>Recall (%)</td>
<td>70.4±10.1</td>
<td>66.7±14.0</td>
</tr>
<tr>
<td>F-score (%)</td>
<td>67.4±11.6</td>
<td>62.9±14.7</td>
</tr>
</tbody>
</table>

Discussion

The main focus of this study is to see whether online news headlines can increase the performance of the machine learning algorithms when they are used as features together with tourist arrivals. To see the improvement, we use the Wilcoxon Signed Rank Test to test the significance of the improvement.

For the Logistic Regression, adding online news headlines can increase the recall and F-score of the classifier model significantly with p-values equal to 0.04 for both metrics. Precision is increased when online news headlines are added. However, the increase in precision is not significant. Figure 5 shows the performance comparison when using only previous tourist arrivals and using both previous tourist arrivals and online news headlines for Logistic Regression.
On the other hand, adding online news headlines as the features does not increase the performance of the SVM-based classifier model, in terms of precision, recall, and F-score, significantly with p-values equal to 0.8, 0.8, and 1, respectively. Figure 6 shows the performance comparison when using only previous tourist arrivals and using both previous tourist arrivals and online news headlines for the Support Vector Machine.
From the results above, we can see that adding online news headlines can increase the predictor performance (in terms of precision, recall, and F-score) significantly. Though, not all machine learning algorithms can be affected significantly. Since this study only use headline news from a specific category (only from the travel category), the future work of this study will consider using news headline from more categories such as politics and economy.

This study has an unorthodox approach to use news headlines as a feature directly. Other studies that we mention in related work section do not use the news headline as a feature directly, instead only using the news’ sentiment. Although this study do not involve sentiments of the news, we show that our approach can give promising results. We also show that adding news headlines directly as a feature can give a significantly better results for Logistic-regression-based classifier.

In this study, we only focus on tourist arrivals in Bali Province. This becomes the limitation of this study. In future work, we will use more provinces to see if the headline news also affect the number of tourist arrival in other provinces in Indonesia.
Conclusion

This study proposes a machine-learning-based technique to predict tourist-arrival trend in Bali Province based on the number of previous tourist arrivals and online news headlines that are published in Indonesian news portal. Two machine learning algorithms are used including Logistic Regression and Support Vector Machine. The results show that the Logistic Regression can achieve up to 78.7% precision, 70.4% recall, and 67.4% F-score. The Support Vector Machine can achieve up to 74.6% precision, 66.7% recall, and 62.9% F-score. In this study, we show that using both the number of previous tourist arrivals and online news headlines can significantly increase the recall and F-score when Logistic Regression is used to build the model.

This study shows that adding online headline news and machine learning can increase performance significantly. Based on our knowledge, adding only the headline of the news together with the number of previous tourist arrivals to feed the machine learning algorithms to predict tourist arrivals remains unexplored. Thus, this study can be a pilot study for future work.

The future work of this study will include using news headlines from categories other than travel, using the body of the news, and incorporating other features to increase precision, recall, and F-score.
References


Acknowledgment
None

Funding Information
None

Conflicting Interest Statement
There is no conflict of interest in the publication of this article.

Publishing Ethical and Originality Statement
All authors declared that this work is original and has never been published in any form and in any media, nor is it under consideration for publication in any journal, and all sources cited in this work refer to the basic standards of scientific citation.

Publisher
The Journal published by Institut Pariwisata dan Bisnis Internasional (IPB Internasional), Denpasar Bali, Indonesia. IPB Internasional is committed to become one of the leading universities in hospitality and tourism in Indonesia.