Annals of Tourism Research 90 (2021) 103273

Contents lists available at ScienceDirect

Annals of Tourism Research

journal homepage: https://www.journals.elsevier.com/annals-of-tourism-research

Tourism demand forecasting with online news data mining

Eunhye Park [a], Jinah Park [b], Mingming Hu [c,*]

[a] Department of Food Nutrition, Gachon University, South Korea
[b] Hospitality and Tourism Research Centre, School of Hotel and Tourism Management, The Hong Kong Polytechnic University, Hong Kong, China
[c] Business School, Guangxi University, Nanning 530004, China

ARTICLE INFO

Article history:
Received 10 February 2021
Received in revised form 28 June 2021
Accepted 30 June 2021
Available online xxxx

Associate editor: Gang Li

Keyword:
News discourse
Topic modeling
Tourism demand forecasting
Hong Kong

ABSTRACT

This study empirically tests the role of news discourse in forecasting tourist arrivals by examining Hong Kong. It employs structural topic modeling to identify key topics and their meanings related to tourism demand. The impact of the extracted news topics on tourist arrivals is then examined to forecast tourism demand using the seasonal autoregressive integrated moving average with the selected news topic variables method. This study confirms that including news data significantly improves forecasting performance. Our forecasting model using news topics also outperformed the others when the destination was experiencing social unrest at the local level. These findings contribute to tourism demand forecasting research by incorporating discourse analysis and can help tourism destinations address various externalities related to news media.

© 2021 Elsevier Ltd. All rights reserved.

Introduction

Tourism demand fluctuates significantly due to internal and external volatilities (Song et al., 2019), including today's frequent social, economic, political, and technological changes and a range of risks (Ritchie & Jiang, 2019). The factors affecting tourism demand exert both direct and indirect impacts; thus, modeling and forecasting must examine a diverse array of determinants in a broad context. The tourism demand model is traditionally based on economic theory (Song & Witt, 2000). However, researchers have expanded the economic framework by incorporating new determinants on a massive scale (Hu & Song, 2020; Wen et al., 2020). For example, amid the rapid advance of Internet technologies, web data have been acknowledged as a valuable indicator of tourism demand (Law et al., 2019). Accordingly, numerous tourism demand forecasting studies have employed search query data to predict tourist arrivals accurately (Dergiades et al., 2018; Law et al., 2019; Wen et al., 2020; Yang et al., 2015).

These efforts have contributed to the development of tourism demand forecasting models that analyze tourists' web search behaviors by isolating their intentions and preferences. Researchers have tended to identify tourists' preferences and experiences by quantifying the volume and sentiments embedded in their social media posts (Li, Hu, & Li, 2020). Existing studies on big data and tourism demand forecasting have focused on the information tourists share. Empirical research on public discourse, formed by various forms of media at the structural level, has lagged behind in the tourism demand modeling and forecasting field. Examining public media discourse is critical for understanding how groups of people form collective images of the world beyond their borders.

The impact of news coverage on public perceptions has been studied extensively across the social sciences (Mainil et al., 2011; Mutz & Soss, 1997). News-framing effects have been identified in the context of tourism destinations. For example, Kapuściński and Richards (2018) found that news media representations of travel sites had a powerful influence on the tourist decision-

* Corresponding author.
E-mail addresses: epark@gachon.ac.kr, (E. Park), jinah.park@polyu.edu.hk, (J. Park), mingming.hu@gxu.edu.cn. (M. Hu).

https://doi.org/10.1016/j.annals.2021.103273
0160-7383/© 2021 Elsevier Ltd. All rights reserved.

making process. This media influence has become even greater with the Internet streamlining media distribution and improving reader accessibility. In the online media environment, the delivery of news through increasingly personalized sources may cause homophily, the human tendency to construct echo chambers comprising compatible voices and beliefs (Dubois & Blank, 2018). This complex media environment and its many significant influences make tourism demand forecasting based on public discourse in online media essential. News media representations of tourism destinations allow tourists to link familiar locales with distant destinations across borders (Cai & Wang, 2018; Chilton, 2004). Online media provide images and descriptions that ultimately generate visits to a destination (Salazar, 2012). By layering this crucial phenomenon onto the concept of discourse, this study deepens the research on media discourse and its impact on tourist behaviors.

This study quantifies the key topics in news media discourse and their impact on tourist arrivals through an empirical investigation of news discourse on and tourism demand in Hong Kong. Hong Kong, as a global destination, derives great benefit from its tourism-related industries, but its extremely high population density (6940 people per square kilometer) can lead to conflicts between residents and tourists (Cramer et al., 2004; Qiu et al., 2020). The asymmetric market structure of tourism in Hong Kong also generates social discourses in and outside Hong Kong. Specifically, among Hong Kong's tourist arrivals in 2019, those from Mainland China accounted for 78.3%, 61.2% of whom were same-day visitors (HKTB, 2020). Large-scale and frequent visitation from a single source market can generate hostility toward tourism and tourists from Hong Kong residents (Chen et al., 2018; Tse & Qiu, 2016; Zhang et al., 2018). However, this phenomenon is difficult to pin down, as its prevalence may be exaggerated by hostile media and political interests. As Yu et al. (2020) indicated, political animosity may be exemplified in the Hong Kong protests, which have caused a significant decline in tourism. However, little is known about how social and political discourse influences tourism revenues, particularly the impact on tourist arrivals in Hong Kong. This study performs tourism demand forecasting based on online news data mining to answer three questions. First, what are the relevant topics in major market news discourse? Second, are there positive or negative associations between key topics and tourist arrivals? Third, does the inclusion of news data significantly improve forecasting performance?

The study answered these questions by identifying relevant topics in major market news discourse using structural topic modeling (STM). The study used news data mining to identify topics related to Hong Kong as a tourism destination and compared them between two major markets: Mainland China and the United States (US). Second, the study explored the positive and negative associations between the key topics and tourist arrivals. The study then quantified the topic weights and interpreted the meanings behind the topics discussed by public media in tourist origin areas. Finally, the study employed selected news topics and tourist arrival data to forecast tourism demand using the seasonal autoregressive integrated moving average with exogenous variables (SARIMAX) method. The study assessed the value of the news topic data for tourism demand forecasting by comparing the SARIMAX models under two scenarios: with and without news data used as explanatory variables. The SARIMAX model was then compared with basic time series models such as the seasonal autoregressive integrated moving average (SARIMA), exponential smoothing state space (ETS), and seasonal Naïve models.

The remainder of this paper is organized as follows. The second section provides an academic foundation. The third section presents the study's data collection and methods such as topic modeling and tourism demand forecasting. The fourth section outlines the empirical results and the fifth section concludes.

Literature review

Discourse analysis in tourism

Discourse is an analytical concept that reflects power, knowledge, and changes through text and representation (Livholts & Tamboukou, 2015). The goal of discourse analysis is to identify topics and segments of discourse conveying a coherent aggregate of thoughts (Chafe, 2001). Discourse analysis in tourism is a complex and evolving topic that is playing an increasingly significant role in broader social science research (Ateljevic & Doorne, 2002; Pritchard & Jaworski, 2005). Discourse studies in tourism examine how distant places and people look at and interact with one another beyond their borders. As Chilton (2004) noted, discourse in the global context involves great complexity and enormous changes, with increasing information, capital, and human flows. Beyond the conventional hegemonic business-centered research paradigm, discourse analysis also examines more decentralized and multi-actor tourism phenomena (Fazito et al., 2016). By aligning this complex interplay with multilayer representations, various social discourses may show how tourism development affects and has been affected by society at large (Jordan et al., 2019).

Media representations of a tourism destination can be linked to the social construction of places and images (Fusté-Forné, 2020). Salazar (2012) conceptualized tourism imaginaries and their diffusion in the form of images and discourse. His study found that tourism imaginaries derived from various information sources reduce the uncertainty and risk of tourism in spaces apart from the mundane world of tourists (Salazar, 2012). Beyond destination spaces, collective sentiments and perceptions of tourism itself (i.e., pro- or anti-tourism) have also emerged. As tourism is an interactive process between guests and hosts, discourse in tourism also resonates with host communities' responses to tourism and tourists (Gustafson, 2002). Public sentiment is influenced by positive and negative media coverage, which reinforces or mitigates responses to the environmental, societal, and economic impacts of tourism on local systems (Milano et al., 2019; Seraphin et al., 2018). This study examined the former case to identify what leading media in local markets were discussing and how those discussions may have influenced actual and future visitation to the destination.

Many information sources such as mass and social media, news articles, and word of mouth have a great effect on tourism markets, especially in shaping images of tourism destinations and their associated visit intentions (Cai & Wang, 2018;

Chaulagain et al., 2019). Destination images formed by various sources mediate tourists' decision making by raising awareness of potential destinations (Jalilvand, 2017). Thus, tourism destination organizations should focus on media discourse about destination images and narratives relating to tourism (McLennan et al., 2017). Cai and Wang (2018) argued that news articles were more carefully considered than other sources and had more credibility. Identifying how news coverage intentionally or unintentionally influences public perceptions and behaviors has been a longstanding concern in the social sciences (Mainil et al., 2011; Mutz & Soss, 1997). However, tourism studies tend to use social media to capture a representation of tourists' experiences (Giglio et al., 2019; Månsson, 2011; Tussyadiah & Fesenmaier, 2009).

Although news coverage may deliver an overamplified frame (Oliver & Maney, 2000), considerable attention has been paid to the media in tourism and its leading role in promoting tourism destinations (Kim et al., 2014). As Viken (2016) emphasized, discourse in media involves power relations not only in politics but also in tourism studies that address spaces and people. Hannam and Knox (2005) underlined that discourse from various sources can show how people interpret their own and other parts of the world. Hence, tourism researchers need to utilize discourse analysis to investigate image formation, representations of tourism experiences, sentiment, and behavioral intentions (Hannam & Knox, 2005). However, this requires further efforts toward a rigorous methodological approach to an interdisciplinary perspective on discourse analysis (Pritchard & Jaworski, 2005). For instance, in the recent econometrics literature, Shapiro et al. (2020) provided a valid application of news-based sentiment analysis to nowcasting models of economic indices. Their study argued that to improve predictive accuracy, nowcasting and forecasting should rely on not only objective and quantifiable variables (e.g., employment and income) but also more subjective variables.

Big data and online news media in tourism demand forecasting

Big data analytics can extract new information with which to measure and forecast tourist arrivals (Colladon et al., 2019). Numerous tourism scholars and practitioners have acknowledged that a macro perspective is needed to address tourism demand modeling and forecasting, urging the collection of new indicators beyond traditional economic variables, such as search engine data from the Google and Baidu indices (Dergiades et al., 2018;Wen et al., 2020; Yang et al., 2015), user-generated content, and reviews from online travel forums (Colladon et al., 2019; Li et al., 2015). Law et al. (2019) conceptually and empirically demonstrated that using the most popular search engine data in a major market can make tourism demand forecasting more precise and timelier.

Given the benefits of big data, a growing body of the literature is examining the use of text- and news-based measures as sources of information for forecasting and assessing the economy (Ardia et al., 2019; Shapiro et al., 2020; Thorsrud, 2018). Several researchers have examined their use in opinion forecasting in the social and political sciences (Lerman et al., 2008). Soft (as opposed to hard) information from texts, opinions, and expressions of sentiment can improve predictive power (Shapiro et al., 2020). In particular, in the knowledge domains of economics and computer science, topic mining from large-scale online news data has been widely employed (Hurtado et al., 2016; Schumaker & Chen, 2009). For instance, applying a dynamic topic model, Morimoto and Kawasaki (2017) extracted topics from financial news that helped improve the accuracy of market volatility forecasting. Tobback et al. (2018) argued that a forecasting model combining news article data has greater accuracy, especially when addressing significant uncertainty and risk during tumultuous periods.

Demand forecasting studies within the tourism context have integrated analyses of tourism-related big data and new sets of variables into traditional models to improve their accuracy (Song et al., 2019). Web data from search engines and online review platforms have become valuable predictors because they represent tourists' intentions and preferences (Wen et al., 2020; Yang et al., 2015). User-generated content from online travel communities such as TripAdvisor also works for tourism demand forecasting when coupled with search query data (Colladon et al., 2019). Li, Hu, and Li (2020) and Li, Li, et al. (2020) indicated that Internet big data have significantly improved forecasting performance and that tourism demand forecasting based on multiple sources outperforms other single-source models. While many tourism demand forecasting studies have utilized web data to improve accuracy, only a handful have used news media data to create forecasting models.

By reviewing 165 publications on big data applications in tourism studies, Li et al. (2018) concluded that online news and opinions should be incorporated into tourism research and practices more deeply because they uniquely identify various externalities impacting tourism market fluctuations and policy responses. Examining tourism demand modeling and forecasting, Stepchenkova and Eales (2011) found that news coverage was a significant factor in destination demand, with online versions of newspapers making media even more influential for potential tourists. Önder et al. (2019) capitalized on online news media coverage to predict tourist arrivals to European cities. Stepchenkova and Eales (2011) quantified the volume, topic, and favorability of news media messages and their significant impact on destination choice (i.e., tourist arrivals). However, the quantification of media discourse was criticized by Colladon et al. (2019), who warned about integrating online discourse into forecasting models without explaining how the new data influenced the forecasts. Going beyond quantifying the volumes of reviews or search queries, identifying topics in media discourse may also explain particular variations in destination choice.

This study brings together two distinctive methodological approaches: discourse analysis and tourism demand forecasting. Combining them allows this study to adopt a broader view and identify both *what* the news media in relevant major markets (China and the US) are saying about the target destination (Hong Kong) and *how* such knowledge could improve forecasting demand in the target destination.

Topic modeling in tourism

Recent tourism and hospitality studies have used topic modeling to analyze online customer reviews of various tourism-related service settings such as hotels (Hu et al., 2019), restaurants (Park et al., 2020), and accommodation sharing (Ding et al., 2020). Going beyond its dominant use for online review data from the demand sides, Wang et al. (2020) applied latent Dirichlet allocation (LDA) to examine tourism destination image. Unlike a structured or item-based survey used to compare image changes before, during, and after travel, LDA enabled the researchers to explore the transformed representations of a tourism destination image from the content (Wang et al., 2020). Stamolampros et al. (2020) applied STM, a topic modeling algorithm, to identify employee feedback in tourism and hospitality firms. These novel methods are becoming critical tools in tourism studies for the analysis of text data that convey rich information on the perceptions and sentiments of the supply and demand sides.

Among various topic modeling algorithms, STM has several advantages over other topic-modeling algorithms. First, STM can generate models that have a predictive power greater than that of competing algorithms such as LDA (Roberts et al., 2016). In addition to improving predictive performance, STM is designed to generate readily interpretable topic models and offer quantitative metrics for evaluating the interpretability of topics such as semantic coherence and topic exclusivity (Roberts et al., 2016). In the tourism literature, text mining has typically been applied to user-generated content from social media or online review sites (Jia, 2020; Vu et al., 2019); few attempts have been made to use news as a text-mining data source. Therefore, there is little understanding on the value of this technique. Hence, this study applied STM to infer information from online news sources to discover the hidden thematic structure with less subjective biases to improve tourism demand forecasting.

Methodology

Data and preprocessing

This study focused on forecasting tourist arrivals in Hong Kong from two tourism markets: the largest short-haul market, Mainland China (78.29% of all arrivals in 2019), and the largest long-haul market, the US (28.09% of the long-haul market in 2019). Data on monthly tourist arrivals from Mainland China and the US between January 2011 and November 2019 were collected from the Hong Kong Tourism Board (see Fig. 1). This study posits that information related to a destination published in a tourism market's news creates associations and images of the destination among the locals, affecting tourists' choices. Although not all travelers read newspapers or online news, the news addresses pressing economic, cultural, political, and social issues (Dehler-Holland et al., 2021; Schudson, 2002). Studies have found that mainstream news media in different countries reported on the same event using different vocabulary and with different tones, indicating that news content tends to reflect each country's prevailing culture and interests (Roberts et al., 2016; Schudson, 2002).

One mainstream online news source was selected from each tourism market to explore how trending events and topics were depicted and disseminated. The study chose *China Daily* as the online news source with which to forecast tourist arrivals from Mainland China for two reasons. The first was to avoid language bias. *China Daily* is published in English, allowing an unbiased comparison with US news sources. Second, *China Daily* enjoys the largest circulation among English-language online news in Mainland China (He et al., 2020). To forecast tourist arrivals from the US, the study chose CNN, a widely known mainstream US news source that covers global issues (Zhang & Hellmueller, 2017). The keyword "Hong Kong" was used to search for destination-related online news articles from online news sources. This generated 163,571 related news items from *China Daily* and 9886 CNN news items published between January 2011 and November 2019. The distribution of news in *China Daily* and CNN is shown in Fig. 2.

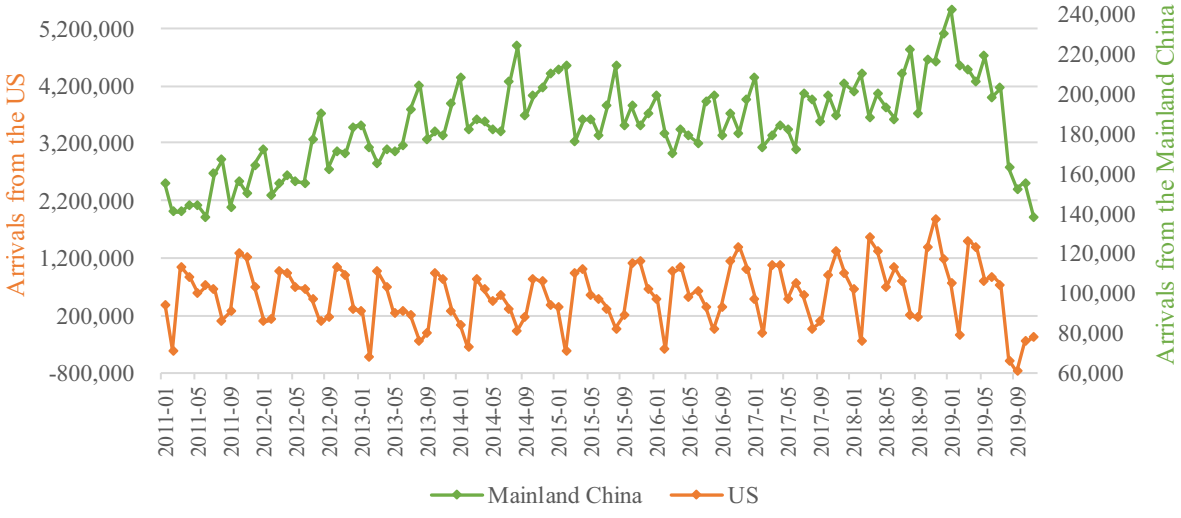

Fig. 1. Monthly tourist arrivals from Mainland China and the US.

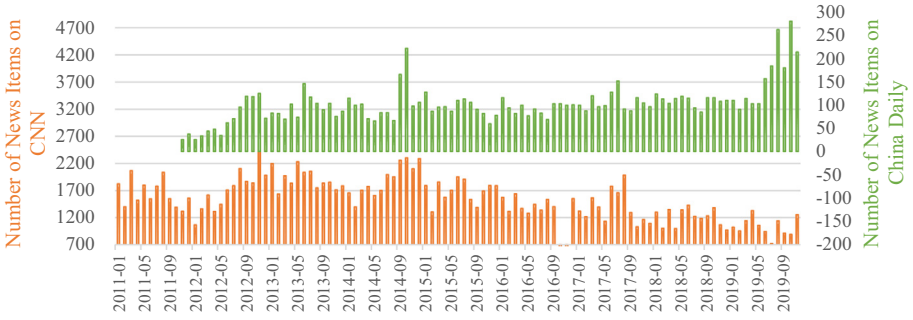

Fig. 2. News distribution for China Daily and CNN.

Data cleaning is essential before analyzing a corpus of text to improve the text mining results. Before data mining, the study performed a series of text preprocessing steps using the Python Natural Language Tool Kit (NLTK) package, including tokenization, non-alphanumeric character removal, and lowercase conversion. Part-of-speech tagging was used to select nouns, verbs, adjectives, and adverbs; identify bigram and trigram words; and lemmatize all terms. In addition to NLTK stop words (i.e., frequently appearing but not meaningful), the study included custom stop words (e.g., "also," "could," "get") for each news dataset. Bigram and trigram words (i.e., those with a sequence of two or three tokens) were also used to capture frequently appearing multiword expressions such as "Hong Kong" and "special administrative region."

Topic modeling

With the input of preprocessed online news texts, STM automatically extracts latent topics from the text data and set of keywords associated with each topic. This produces two outputs: (1) topic prevalence (θ), which indicates the probability that any document (d) is associated with multiple topics, and (2) topic term proportion (β), which captures the probability that each term from the corpus belongs to a certain topic. Because STM adopts the mixed membership approach, which assumes that a document contains a mixture of multiple topics, the likelihood that each document is associated with multiple topics can be calculated. For single-membership models, a document is assigned solely to the most notable theme. As a news article is likely to address multiple topics (i.e., the impact of "diplomatic issues" on "the tourism industry"), mixed membership models were deemed appropriate for this study's data.

The generative process for computing the topic models was as follows. First, topic prevalence was inferred from a logistic normal distribution:

$$\theta_d \sim \text{LogisticNormal}_{K-1}(\Gamma x_d, \Sigma),$$

where Γ represents a matrix of the coefficients of the topic prevalence covariates and Σ is a covariance matrix. The use of logistic normal distributions makes it possible to estimate topic correlations so that the relationships among topics can be considered to build the models and ultimately improve topic interpretability (Blei & Lafferty, 2007). Based on the topic prevalence (θ_d), a topic (Z_{dn}) was assigned to each term within any document d inferred from a multinomial distribution as follows:

$$z_{dn} \sim \text{Multinomial}(\theta_d).$$

The probability that a term within the corpus was assigned to each topic was derived from a multinomial logistic regression that included the baseline word's usage rate across all documents as well as the word usage rate within the particular topic, a document-level covariate, and the interaction between the topic and covariate:

$$\beta_{dkv} \propto \sim \exp\left(m_v + k_{k,v}^{(\text{topic})} + k_{yd,v}^{(\text{cov})} + k_{yd,k,v}^{(\text{int})}\right).$$

Then, an observed word in a document (W_{dn}) was sampled from the following multinomial distribution:

$$W_{d,n} \sim \text{Multinomial}(\beta_{dk1}, \ldots, \beta_{dkV}).$$

Although STM automatically extracts latent topics, researchers must provide the number of topics (K) needed for the final model. To make this "K" decision, this study relied on two metrics: (1) the held-out likelihood that evaluates predictive performance and (2) semantic coherence. After comparing the values of the two metrics over a series of topics with different K, the optimal K for the *China Daily* dataset was determined to be 100 and that for CNN was determined to be 30. The number of topics for both data sources was within the range of topic number recommendations of five to 50 topics for small datasets (those with fewer than 10,000 documents) and 100 topics for large datasets (with over 100,000 documents) (Roberts et al., 2019).

After building the topic models, the researchers reviewed the top terms strongly associated with each topic to ensure the topic quality generated from the algorithm and identify the types of topics frequently covered in the two news datasets. Topic prevalence can reflect the main topics covered in each sampled article. To isolate the news topics reported most frequently each month, the monthly average of document-level topic prevalence was calculated separately for both the *China Daily* and the CNN datasets.

Forecasting and evaluation models

A framework (see Fig. 3) was utilized to forecast tourism demand based on news in the market region. This framework had four aspects: online news sources, topic data mining, estimation and forecasting, and forecast accuracy evaluation. For the first aspect, as reported above, *China Daily* was sampled for news coverage in Mainland China and CNN was sampled for the US. Second, topic modeling was utilized to generate topics from *China Daily* and CNN. The monthly distribution of topics was calculated using topic modeling. Third, both the Boruta algorithm (Kursa & Rudnicki, 2010) and the stepwise algorithm (Venables & Ripley, 2002) were used to select the explanatory variables from the lagged monthly distribution of topics. The selected lagged monthly distribution of topics was used to forecast tourism demand. One time series model with explanatory variables (SARIMAX) was utilized as the forecast model based on the topics from the online news sources. Three pure time series models were taken as the benchmarks (automatically selected ARIMA, automatically selected ETS, and seasonal naïve), which are typically used to test forecasting methods (Hyndman, 2018). Finally, MAE, MAPE, RMSE, and RMSPE were used as the accuracy evaluation models. When comparing two models, an improvement index was used to calculate the relative performance.

SARIMA/SARIMAX

ARIMA is a classical time series model and frequently utilized to derive benchmarks in tourism demand forecasting because of its strong forecast performance (Song et al., 2019). It integrates seasonality, autoregressive, and moving average components, allowing it to make otherwise unstable time series more stable. The general model of SARIMA $(p, d, q)(P, D, Q)$ is specified as

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = \Theta(B^m)\theta(B)\varepsilon_t$$

where y_t is tourism demand at time t (i.e., the number of tourist arrivals in month t). $\phi(B) = (1 - \sum_{i=1}^{p}\phi_i B^i)$ is the autoregressive component, $\theta(B) = (1 + \sum_{i=1}^{q}\theta_i B^i)\varepsilon_t$ is the moving average component, $(1 - B)^d$ is the d times difference indicator, $(1 - B^m)^D$ is the seasonal D times difference indicator, and $\Phi(B)$ and $\Theta(B)$ represent the seasonal AR(P) and seasonal MA(Q) components, respectively. The "Forecast Package" (Hyndman et al., 2020) in R was used to estimate the SARIMA model and forecast the future value. d, p, q, P, D, and Q are automatically determined using the function auto.arima () according to the AICc and return the fitted model. forecast () helps generate the future value.

The explanatory variables in SARIMA elevate the model to SARIMAX. When the topic variables generated from online news sources are incorporated into the model, SARIMAX can be specified as

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d TD_t = \mu + \sum_{k=1}^{n}\beta^k \cdot Topic_t^k + \Theta(B^m)\theta(B)\varepsilon_t$$

where the SARIMA parameter takes the same meaning as described above, $Topic_t^k$ is the selected monthly distribution series of topic k by the Burora algorithm (Kursa & Rudnicki, 2010) and stepwise algorithm (Venables & Ripley, 2002) to forecast tourism demand, and β^k is the corresponding coefficient.

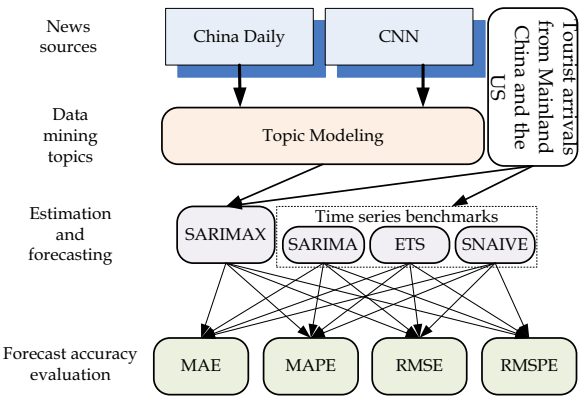

Fig. 3. Study framework.

In the modeling process, the AIC and BIC are generally utilized to determine the explanatory variables and their lags (Gunter et al., 2019; Önder et al., 2020). This study involved a large number of topics. Addressing all the study's topic variables would thus have led to overfitting, rendering the AIC and BIC useless. Thus, this study used a combination of the Boruta algorithm (Kursa & Rudnicki, 2010) and stepwise algorithm (Venables & Ripley, 2002) as the feature selection method. The Boruta algorithm was designed to be a wrapper fitting around a random forest classification algorithm. First, it iteratively removes the features proved by a statistical test to be less relevant than random probes. Then, the AIC is used to double check the significance of the variables and remove the less relevant ones. This study used the Boruta and Stats packages in R to conduct the feature selection. The forecast package (Hyndman et al., 2020) in R was used to estimate the SARIMAX model and forecast future values.

ETS model

The ETS method is also automatically selected. This method develops a simple exponential smoothing method by including level, trend, seasonality, and smoothing (Hyndman & Athanasopoulos, 2018). The general ETS model can be written as ETS (e, t, s), where *e* denotes the error type, *t* denotes the trend type, and *s* denotes the seasonality type. ets ( ), as included in the Forecast Package (Hyndman et al., 2020) in R, can be used to select the model automatically according to the AICc among the different error types, trend types, seasonality types, and estimate parameters. forecast ( ) can be used to derive future values.

Seasonal naïve

Seasonal naïve is a typical benchmark model for seasonal data. It uses the value of the historical period to forecast future values. To forecast monthly tourism demand, the general model is

$$\widehat{y_t} = y_{t-12}$$

where $\hat{y}_t$ denotes the forecast value of tourist arrivals in month t and y_{t-12} is delayed tourist arrivals in month 12.

Forecast evaluation

To evaluate the forecast accuracy of the above models and show the role of topic variables in tourism demand forecasting, the MAE, RMSE, MAPE, and RMSPE were calculated by the following series:

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

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

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

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

where y_i is the real value of tourist arrivals and $\hat{y}_i$ is the forecast value of tourist arrivals. To compare the two models, the relative improvement of Model 1 was compared with that of Model 2 using the MAPE as follows:

$$RI_{Model_2}^{Model1} = \frac{MAPE(Model\ 2) - MAPE(Model\ 1)}{MAPE\ (Model\ 2)}$$

Results

Topic selection and analysis

Employing the data-mining process, 100 topics from *China Daily* and 30 topics from CNN were extracted. The Boruta algorithm (Kursa & Rudnicki, 2010) and stepwise algorithm (Venables & Ripley, 2002) were used to conduct the feature selection and select the most relevant topics related to tourist arrivals from Mainland China and the US. The Boruta algorithm estimated the importance of each independent variable by taking tourist arrivals from Mainland China or the US as a dependent variable and using a lagged series of different topics extracted from *China Daily* or CNN as independent variables. The result was used to identify the

Table 1
Selected topics illustrating tourist arrivals from Mainland China.

Lag	Topic	Top words	Coefficient	Significance
1	Topic 20: Economy	growth, increase, rise, decline, economy	−11,713,599	0.0774*
	Topic 39: Industry issues	company, brand, product, market, sale	−22,670,482	0.0614*
	Topic 46: Societal issues	child, family, woman, old, parent	−62,072,318	0.00001***
	Topic 50: Protest	police, protest, protester, city, violence	−8,900,996	0.0000***
	Topic 65: China's economy	economy, china, market, financial, reform	27,509,372	0.0008***
	Topic 88: Online market	online, company, alibaba, internet, platform	12,135,557	0.0911*
	Topic 93: Guangzhou	province, guangzhou, guangdong_province, guangdong, local	32,828,677	0.0316**
2	Topic 7: Foreign regulation	foreign, regulation, enterprise, authority, policy	17,194,570	0.1843
	Topic 11: Stock	point, hang_seng, end, wall street, stocks	−29,721,270	0.0082***
	Topic 21: HK SAR	hong_kong, special_administrative, region, chief_executive, central_government	−6,948,677	0.1950
	Topic 34: Foreign relations	american, canada, community, canadian, asia	28,057,262	0.0591*
	Topic 63: Politics	political, election, government, member, vote	5,840,715	0.1385
3	Topic 9: Figures/Celebrities	wong, lee, cheung, lam, kwok	−56,735,768	0.0026***
	Topic 36: Art	art, artist, painting, auction, collector	−9,903,689	0.0644*
	Topic 96: International event	international, event, forum, exhibition, fair	37,500,698	0.0002***
	Topic 99: Investment	company, group, investment, acquisition, chairman	−11,993,099	0.0347**

Note:
*** Denotes statistical significance at 1% level.
** Denotes statistical significance at 5% level.
* Denotes statistical significance at 10% level.

most important topics by the AIC based on the results of the Boruta algorithm. The most important topics explaining tourist arrivals are shown in Table 1 for *China Daily* and in Table 2 for CNN.

The 100 topics extracted from *China Daily* highlighted diverse issues, including society, economy, and politics, which were found to be important in explaining Mainland Chinese tourist arrivals. Topic 46, labeled "societal issues," had a negative association with tourist arrivals from Mainland China. Specifically, the top words (i.e., "child," "family," "woman," "old," "parent") of Topic 46 showed that news addressing gender or family issues in Hong Kong had a negative influence on Mainland Chinese tourist arrivals. Turning to sociopolitical issues, Topic 50 was related to the Hong Kong protest movement. This topic also had a negative association with Chinese tourist arrivals, displaying two topic weight peaks with the Umbrella Movement in 2014 and the Hong Kong protest in 2019–20. Topic 11 was on stock prices and business conditions, while Topic 20 was on capital investments and mergers and acquisitions, which had a negative impact on Mainland Chinese tourist arrivals.

Topic 96 (international events), Topic 93 (Guangzhou), and Topic 65 (China's economy) were positively associated with Mainland Chinese tourist arrivals. Topic 96 included events such as exhibitions, expos, and conferences. Topic 93 was related to issues in Guangzhou, the capital of Guangdong province, which shares a border with Hong Kong. This topic reflected Guangzhou's progress in becoming a financial hub and its efforts to develop using transportation infrastructure investments. For Topic 65, events such as economic development in China and the Mainland's reform efforts were mentioned as well as news on how Chinese economic development has influenced the Hong Kong economy.

Topics 2 and 7 were found to be positively associated with tourist arrivals from the US. Topic 2 consisted of news related to massive Chinese tech companies such as Alibaba and Huawei as well as the Asian market overall. The rapid development of companies such as Alibaba and Huawei has led to fierce competition for market share, leading to issues such as the US–China trade war. Topic 7 comprised attributes that describe Hong Kong as a tourism destination. Both its natural attractiveness (e.g., island, beach) and tourism infrastructure (e.g., hotel, bar, pool) were mentioned in relation to Topic 7. Similar to the *China Daily* results, CNN's coverage of the Hong Kong protests proved to have a negative association with tourist arrivals from the US.

Cointegration test

The study conducted a cointegration test using the Engle–Granger two-step method to confirm the long-term equilibrium relationship between tourist arrivals and the selected topics from the targeted online news sources (Huang et al., 2017). The augmented Dickey–Fuller (ADF) unit root test results in Table 3 show that all the variables except Topic 36 in Mainland China and Topic 7 in the US were first-order integrated. These first-order integrated variables were used to conduct a cointegration test. The

Table 2
Selected topics illustrating tourist arrivals from the US.

Lag	Topic	Meaning	Coefficient	Significance
2	Topic 2: Chinese tech companies	company, apple, alibaba, huawei, market	101,006	0.0263**
3	Topic 7: HK destination	hotel, bar, room, city, island	252,327	0.0019***
2	Topic 9: Protest	police, protester, protest, hong_kong, government	−111,403	0.0060***

Note:
*** Denotes statistical significance at 1% level.
** Denotes statistical significance at 5% level.

Table 3
ADF unit root test results.

Variable	ADF value	Prob.	Conclusion	Variable	ADF value	Prob.	Conclusion
Mainland China							
y_t	−0.55	0.58	Non-stationary	Topic 46	−1.11	0.27	Non-stationary
Δy_t	−10.73	0.00	Stationary	ΔTopic 46	−12.70	0.00	Stationary
Topic 7	−1.01	0.314	Non-stationary	Topic 50	0.77	0.44	Non-stationary
ΔTopic 7	−10.58	0.00	Stationary	ΔTopic 50	−7.65	0.00	Stationary
Topic 9	−1.13	0.26	Non-stationary	Topic 63	−0.696	0.49	Non-stationary
ΔTopic 9	−9.91	0.00	Stationary	ΔTopic 63	−8.18	0.00	Stationary
Topic 11	−1.66	0.10	Non-stationary	Topic 65	0.93	0.35	Non-stationary
ΔTopic 11	−13.47	0.00	Stationary	ΔTopic 65	−10.72	0.00	Stationary
Topic 20	−1.36	0.18	Non-stationary	Topic 88	−1.47	0.15	Non-stationary
ΔTopic 20	−12.85	0.00	Stationary	ΔTopic 88	−11.89	0.00	Stationary
Topic 21	−1.47	0.15	Non-stationary	Topic 93	−0.90	0.37	Non-stationary
ΔTopic 21	−11.37	0.00	Stationary	ΔTopic 93	−11.34	0.00	Stationary
Topic 34	−1.324	0.18	Non-stationary	Topic 96	−1.29	0.20	Non-stationary
ΔTopic 34	−12.32	0.00	Stationary	ΔTopic 96	−10.00	0.00	Stationary
Topic 36	−2.55	0.01	Stationary	Topic 99	−1.45	0.15	Non-stationary
ΔTopic 36	\	\	\	ΔTopic 99	−11.84	0.00	Stationary
Topic 39	−1.51	0.14	Non-stationary				
ΔTopic 39	−12.39	0.00	Stationary				
US							
y_t	−0.91	0.37	Non-stationary	Topic 7	−2.157	0.03	Stationary
Δy_t	−10.90	0.00	Stationary	ΔTopic 7	\	\	\
Topic 2	−1.17	0.24	Non-stationary	Topic 9	−1.07	0.29	Non-stationary
ΔTopic 2	−10.39	0.00	Stationary	ΔTopic 9	−7.07	0.00	Stationary

Note: 1% critical value is −2.58, 5% critical value is −1.95, and 10% critical value is −1.62 for the variables relating to Mainland China; 1% critical value is −2.6, 5% critical value is −1.95, and 10% critical value is −1.61 for the variables relating to the US; Δy_t indicates that y_t is integrated of order 1.

ADF value of the residuals of the fitting variables relating to Mainland China was −6.78 ($P = 8.49 \times 10^{-10}$) and the value relating to the US was −8.81 ($P = 8.72 \times 10^{-14}$). This means that tourist arrivals from Mainland China and the US showed a long-term equilibrium relationship with the topics selected from both *China Daily* (as shown in Table 1) and CNN (as shown in Table 2).

The Granger causality test was used to confirm the causal relationship between tourist arrivals and each selected topic. The results in Table 4 show that the Granger causality relationships of most of our selected topics from *China Daily* and all the topics selected from CNN were confirmed. This finding indicates that the selected topics from *China Daily* and CNN led to enhanced tourist arrivals from Mainland China and the US. They can thus be used to forecast tourist arrivals.

Table 4
Granger causality test results.

Null hypothesis	F-statistic	Prob.	Null hypothesis	F-statistic	Prob.
Mainland China					
Topic 7 does not Granger cause tourist arrivals y_t	3.6768	0.0580*	Topic 46 does not Granger cause y_t	9.4499	0.0027***
y_t does not Granger cause Topic 7	2.7071	0.1030	y_t does not Granger cause Topic 46	0.8466	0.3597
Topic 9 does not Granger cause y_t	3.0372	0.0845*	Topic 50 does not Granger cause y_t	2.5269	0.1151
y_t does not Granger cause Topic 9	2.9795	0.0874*	y_t does not Granger cause Topic 50	2.5916	0.1106
Topic 11 does not Granger cause y_t	3.7293	0.0563*	Topic 63 does not Granger cause y_t	0.0714	0.7898
y_t does not Granger cause Topic 11	7.1521	0.0087***	y_t does not Granger cause Topic 63	0.2816	0.5969
Topic 20 does not Granger cause y_t	26.7170	0.0000***	Topic 65 does not Granger cause y_t	1.9687	0.1637
y_t does not Granger cause Topic 20	5.7816	0.0180**	y_t does not Granger cause Topic 65	0.0016	0.9686
Topic 21 does not Granger cause y_t	0.0609	0.8056	Topic 88 does not Granger cause y_t	3.3622	0.0697*
y_t does not Granger cause Topic 21	0.0711	0.7902	y_t does not Granger cause Topic 88	12.3153	0.0007***
Topic 34 does not Granger cause y_t	3.3124	0.0718*	Topic 93 does not Granger cause y_t	2.2028	0.1409
y_t does not Granger cause Topic 34	3.8338	0.0530*	y_t does not Granger cause Topic 93	6.8966	0.0100**
Topic 36 does not Granger cause y_t	1.2978	0.2573	Topic 96 does not Granger cause y_t	6.5964	0.0117**
y_t does not Granger cause Topic 36	1.6094	0.2075	y_t does not Granger cause Topic 96	1.8364	0.1784
Topic 39 does not Granger cause y_t	10.8786	0.0013***	Topic 99 does not Granger cause y_t	0.2007	0.6551
y_t does not Granger cause Topic 39	4.6038	0.0343**	y_t does not Granger cause Topic 99	0.0439	0.8345
US					
Topic 2 does not Granger cause tourist arrivals y_t	5.8826	0.0173**	Topic 9 does not Granger cause y_t	8.826	0.0038***
y_t does not Granger cause Topic 2	1.8485	0.1774	y_t does not Granger cause Topic 9	0.6297	0.4295
Topic 7 does not Granger cause y_t	12.433	0.0007***			
y_t does not Granger cause Topic 7	1.913	0.1701			

Note:
*** Denotes statistical significance at 1% level.
** Denotes statistical significance at 5% level.
* Denotes statistical significance at 10% level.

Table 5
Forecast performance.

	Model	MAE	RMSE	MAPE	RMSPE
Mainland China					
Forecast error	SARIMAX	555,651	655,897	0.1623	0.2076
	SARIMA	611,167	816,256	0.1961	0.2862
	ETS	741,649	878,114	0.2327	0.3083
	SNAIVE	1,099,889	1,363,247	0.3802	0.5550
Improvement of *SARIMAX* compared with	SARIMA	9.08%	19.65%	17.24%	27.44%
	ETS	25.08%	25.31%	30.26%	32.66%
	SNAIVE	49.48%	51.89%	57.32%	62.59%
US					
Forecast error	SARIMAX	7754	11,121	0.1030	0.1539
	SARIMA	8656	12,015	0.1157	0.1710
	ETS	18,112	23,048	0.2309	0.3234
	SNAIVE	15,274	24,378	0.2023	0.3291
Improvement of *SARIMAX* compared with	SARIMA	10.42%	7.45%	10.97%	10.03%
	ETS	57.19%	51.75%	55.39%	52.42%
	SNAIVE	49.24%	54.38%	49.07%	53.24%

In addition, the Granger causality test result of Topic 36 from *China Daily* was not significant, while that of Topic 7 from the CNN was significant. This indicates that the Granger causality relationship between Topic 36 from *China Daily* and tourism demand from Mainland China was not confirmed, while the relation between Topic 7 from CNN and tourism demand from the US was confirmed. Therefore, we removed Topic 36 of *China Daily* from the exogenous variable datasets to forecast tourism demand from Mainland China. Topic 7 of CNN was used as an exogenous variable to forecast tourism demand from the US.

Forecast performance

By incorporating the topic variables generated from *China Daily* and CNN, as shown in the proposed framework in Fig. 3, SARIMAX was used to forecast tourism arrivals from Mainland China and the US. SARIMA, ETS, and seasonal naïve were used as the benchmark models. The full dataset used to forecast tourism demand from Mainland China spanned 107 months, from January 2011 to November 2019, while the dataset used to forecast tourism demand from the US started in November 2011. The most recent 12 months were used to test the forecast accuracy of the models.

The average forecast errors of the 12-month one-step-ahead rolling forecast are listed in Table 5. The results indicate three critical points. First, comparing SARIMAX with SARIMA shows that the topics from *China Daily* and CNN both play positive roles in improving forecasting accuracy when estimating tourist arrivals from Mainland China and the US. This result indicates that news reported in a source market affects the image of a destination and plays a role in determining tourism demand. Additionally, comparing *China Daily* with CNN shows that Chinese news coverage played a greater role in determining tourism demand than did US coverage. There are two possible explanations for these findings. First, Hong Kong is a special administrative region in China. *China Daily*, as a mainstream online news source and the first national daily English-language newspaper in China, necessarily pays more attention to Hong Kong than CNN can pay to any particular tourist destination. Second, short-haul tourists from Mainland China to Hong Kong tend to be more malleable to news coverage than are long-haul US tourists. Short-haul tourists thus have a greater capacity to change their travel plans based on recent news reporting, while long-haul travel plans may be “locked in” and unaffected by

Table 6
Forecast performance during the protest (August to November 2019).

	Model	MAE	RMSE	MAPE	RMSPE
Mainland China					
Forecast error	SARIMAX	658,021	793,554	0.2693	0.3135
	SARIMA	1,043,809	1,257,605	0.4141	0.4737
	ETS	1,143,979	1,254,478	0.4682	0.5037
	SNAIVE	2,045,633	2,105,590	0.8809	0.9374
Improvement of *SARIMAX* compared with	SARIMA	36.96%	36.90%	34.95%	33.83%
	ETS	42.48%	36.74%	42.47%	37.77%
	SNAIVE	67.83%	62.31%	69.42%	66.56%
US					
Forecast error	SARIMAX	18,431	18,756	0.2612	0.2628
	SARIMA	20,012	20,142	0.2882	0.2913
	ETS	32,952	34,157	0.4874	0.5191
	SNAIVE	39,141	41,870	0.5444	0.5677
Improvement of *SARIMAX* compared with	SARIMA	7.90%	6.88%	9.38%	9.79%
	ETS	44.07%	45.09%	46.42%	49.37%
	SNAIVE	52.91%	55.20%	52.03%	53.71%

such an influence. Finally, comparing the forecasting performance of SARIMAX with that of the time series benchmarks shows that SARIMAX with topic variables drawn from source market news sources consistently proved to be the best model.

Given the recent social unrest in Hong Kong and its severe impact on tourism, the study investigated the role of news coverage in improving forecasting accuracy, especially during turbulent times at the destination level. To this end, the study extracted the forecast performance of each model between August and November in 2019. These data are shown in Table 6. Two findings are notable. First, SARIMAX with news source topics again consistently outperformed the benchmark models during Hong Kong's social unrest. This indicates that such topics remained relevant and improved forecasting accuracy. Second, the performance of the model with topics from *China Daily* during the social unrest improved significantly over the one-year test results, while that of the model using CNN topics remained almost unchanged. There are two possible reasons for this difference. First, news coverage of the social unrest on *China Daily* might have attracted more attention in China. This could cause destination image (re)formation and ultimately influence travel intentions. Second, as mentioned, short-haul tourists from Mainland China have the flexibility required to change their travel itineraries in response to news reports on social unrest.

Conclusion and implications

Tourist arrivals are highly dependent on the external environment of the destination, including newsworthy phenomena such as social changes and political events. However, most studies have built their forecasting models on quantitative historical data that may ignore the influence of unforeseen but significant social occurrences. Li, Hu, and Li (2020) suggested the need for a tourism demand forecasting model that reflects real-time changes using Internet big data. Accordingly, this study presented a tourism demand forecasting model for Hong Kong that combines news topics from two online news sources derived from probabilistic topic modeling.

The STM results revealed the key news topics that are most frequently and prominently exposed by two major online news sources. For instance, the topic weights related to the Hong Kong protests surged twice, in 2014 and 2019, when mass protests escalated there. These sudden increases during social unrest were found to be significantly associated with tourist arrivals in Hong Kong. This study discovered that the inclusion of a monthly timeframe of topic weights can improve the performance of forecasting models and reduce forecasting errors. Specifically, this study identified the lag effects (i.e., one-, two-, and three-month lags) that show when news topics display the best predictive performance. These findings indicate that some news topics may have immediate or short-term impacts on tourism demand, while others may have long-term effects.

By comparing the number of online news articles and topic structures regarding Hong Kong published in *China Daily* and CNN, the study found that the extent and diversity of social discourses about a tourism destination differed between tourism markets. Moreover, including online news as exogenous variables was found to be more effective if the geographical distance between the tourism market and tourism distance was short and bilateral relations were established in various areas. The impact of online social discourses on tourist behavior is largely dependent on geographical proximity and international relationships. Therefore, tourism stakeholders should consider including news sources to improve tourism demand forecasting, especially for tourism markets that are close and feature high interaction levels.

This finding makes an important contribution to the literature. Although text-based sets of big data contain rich information corresponding to major events and market trends, they have rarely been utilized to formulate tourism demand forecasting models. Moreover, online news releases are underexplored as a means of forecasting tourist arrivals relative to other types of big data sources such as search query data, online reviews, and website traffic data (Li, Hu, & Li, 2020; Li, Li, et al., 2020; Pan et al., 2012; Sun et al., 2019). Online news sources not only convey the latest information at a faster pace but are also accessible immediately and readily from anywhere. Therefore, this study proposes that online news is a novel source of volume data that can serve as an indicator of significant social changes.

This study has several limitations that suggest potential avenues for future studies. First, this study has demonstrated the potential value of including key news topics based on STM for tourism demand forecasting. As different news channels may have different perspectives of the same event, the latent sentiment or tone with which news is delivered may vary. Hence, future studies should combine sentiment analysis with topic modeling to assess the association between latent sentiment toward a particular news topic and potential tourist behaviors. Second, while online news is clearly a key medium for information dissemination, this study included only a single type of online data source. Future studies may consider including other types of unstructured texts from additional information sources such as articles from blogs, Internet forums, and user-generated content (Chen & Chen, 2019). Third, this study demonstrated that the forecasting model with news topics outperformed the others not only in the normal period but also at a time of social unrest in the destination at the local level. As impacts of crisis can be determined by geographical scale (Ritchie & Jiang, 2019) and the devastating global impact of the COVID-19 pandemic now confronting the world, tourism demand forecasting studies should endeavor to conduct innovative and robust forecasts (Qiu et al., 2021). Future studies on tourism demand forecasting with discourse analysis at the global level are encouraged when tourism relaunches and tourism statistics recover after the pandemic. Fourth, this study evaluated model performance based on 12-month one-step-ahead forecasting for a limited sample size. Future research should test the model's forecasting performance over two- or three-month forecast horizons by incorporating more samples.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (71761001, 71761002), the Guangxi Key Research and Development Plan (Guike-AB20297040), and Hong Kong Scholars Program.

References

Ardia, D., Bluteau, K., & Boudt, K. (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting, 35*(4), 1370–1386.
Ateljevic, I., & Doorne, S. (2002). Representing New Zealand: Tourism imagery and ideology. *Annals of Tourism Research*, *29*(3), 648–667.
Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35.
Cai, L., & Wang, S. (2018). The US tourists' perceptions of destination China over two transformative periods. *Asia Pacific Journal of Tourism Research*, *23*(3), 217–230.
Chafe, W. (2001). The analysis of discourse flow. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 673–687). Massachusetts, USA: Blackwell Publisher.
Chaulagain, S., Wiitala, J., & Fu, X. (2019). The impact of country image and destination image on US tourists' travel intention. *Journal of Destination Marketing & Management*, *12*, 1–11.
Chen, M. -Y., & Chen, T. -H. (2019). Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena. *Future Generation Computer Systems*, *96*, 692–699.
Chen, N., Hsu, C. H., & Li, X. R. (2018). Feeling superior or deprived? Attitudes and underlying mentalities of residents towards Mainland Chinese tourists. *Tourism Management*, *66*, 94–107.
Chilton, P. (2004). *Analysing political discourse: Theory and practice.* London: Routledge.
Colladon, A. F., Guardabascio, B., & Innarella, R. (2019). Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. *Decision Support Systems, 123*, Article 113075.
Cramer, V., Torgersen, S., & Kringlen, E. (2004). Quality of life in a city: The effect of population density. *Social Indicators Research*, *69*(1), 103–116.
Dehler-Holland, J., Schumacher, K., & Fichtner, W. (2021). Topic modeling uncovers shifts in media framing of the German renewable energy act. *Patterns*, *2*(1). https://doi.org/10.1016/j.patter.2020.100169.
Dergiades, T., Mavragani, E., & Pan, B. (2018). Google trends and tourists' arrivals: Emerging biases and proposed corrections. *Tourism Management*, *66*, 108–120.
Ding, K., Choo, W. C., Ng, K. Y., & Ng, S. I. (2020). Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation. *International Journal of Hospitality Management*, *91*. https://doi.org/10.1016/j.ijhm.2020.102676.
Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729–745.
Fazito, M., Scott, M., & Russell, P. (2016). The dynamics of tourism discourses and policy in Brazil. *Annals of Tourism Research*, *57*, 1–17.
Fusté-Forné, F. (2020). What do New Zealand newspapers say about food tourism? *Tourism and Hospitality Research*, *20*(1), 82–92.
Giglio, S., Bertacchini, F., Bilotta, E., & Pantano, P. (2019). Using social media to identify tourism attractiveness in six Italian cities. *Tourism Management*, *72*, 306–312.
Gunter, U., Önder, I., & Gindl, S. (2019). Exploring the predictive ability of LIKES of posts on the Facebook pages of four major city DMOs in Austria. *Tourism Economics*, *25*(3), 375–401.
Gustafson, P. (2002). Tourism and seasonal retirement migration. *Annals of Tourism Research*, *29*(4), 899–918.
Hannam, K., & Knox, D. (2005). Discourse analysis in tourism research: A critical perspective. *Tourism Recreation Research*, *30*(2), 23–30.
He, Y., Zhang, G., & Chen, L. (2020). Analysis of news coverage of haze in China in the context of sustainable development: The case of China daily. *Sustainability*, *12*(1), 386.
HKTB (2020). Research & statistics. Visitor arrivals to Hong Kong. Retrieved from https://partnernet.hktb.com/en/research_statistics/latest_statistics/index.html.
Hu, M., & Song, H. (2020). Data source combination for tourism demand forecasting. *Tourism Economics*, *26*(7), 1248–1265.
Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, *72*, 417–426.
Huang, X., Zhang, L., & Ding, Y. (2017). The Baidu Index: Uses in predicting tourism flows–A case study of the Forbidden City. *Tourism Management*, *58*, 301–306.
Hurtado, J. L., Agarwal, A., & Zhu, X. (2016). Topic discovery and future trend forecasting for texts. *Journal of Big Data*, *3*(1), 7.
Hyndman, R. J. (2018). A forecast ensemble benchmark. Retrieved from: https://robjhyndman.com/hyndsight/benchmark-combination/.
Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice.* OTexts.
Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., ... Wang, E. (2020). Package 'forecast'. Retrieved from https://cran.r-project.org/web/packages/forecast/forecast.pdf.
Jalilvand, M. R. (2017). Word-of-mouth vs. mass media: Their contributions to destination image formation. *Anatolia*, *28*(2), 151–162.
Jia, S. S. (2020). Motivation and satisfaction of Chinese and US tourists in restaurants: A cross-cultural text mining of online reviews. *Tourism Management*, *78*. https://doi.org/10.1016/j.tourman.2019.104071.
Jordan, E. J., Moran, C., & Godwyll, J. M. (2019). Does tourism really cause stress? A natural experiment utilizing Arcgis Survey123. *Current Issues in Tourism* (pp. 1–15).
Kapuściński, G., & Richards, B. (2018). Destination risk news framing effects—The power of audiences. *The Service Industries Journal*, 1–24.
Kim, T. -K., Lee, C. -K., Mjelde, J. W., & Lee, H. -M. (2014). The carryover effect of newspaper reports on a mega event: Ex post analysis of the 2012 Expo Yeosu Korea. *Asia Pacific Journal of Tourism Research*, *19*(9), 1009–1022.
Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*(11), 1–13.
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.
Lerman, K., Gilder, A., Dredze, M., & Pereira, F. (2008). Reading the markets: Forecasting public opinion of political candidates by news analysis. *Paper presented at the Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK.*
Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. R. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management*, *46*, 311–321.
Li, H., Hu, M., & Li, G. (2020). Forecasting tourism demand with multisource big data. *Annals of Tourism Research*, *83*. https://doi.org/10.1016/j.annals.2020.102912.
Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323.
Li, X., Li, H., Pan, B., & Law, R. (2020). Machine learning in Internet search query selection for tourism forecasting. *Journal of Travel Research*. https://doi.org/10.1177/0047287520934871.
Livholts, M., & Tamboukou, M. (2015). *Discourse and narrative methods: Theoretical departures, analytical strategies and situated writings.* London: Sage.
Mainil, T., Platenkamp, V., & Meulemans, H. (2011). The discourse of medical tourism in the media. *Tourism Review*, *66*(1/2), 31–44.
Månsson, M. (2011). Mediatized tourism. *Annals of Tourism Research*, *38*(4), 1634–1652.
McLennan, C. -l. J., Becken, S., & Moyle, B. D. (2017). Framing in a contested space: Media reporting on tourism and mining in Australia. *Current Issues in Tourism*, *20*(9), 960–980.
Milano, C., Novelli, M., & Cheer, J. M. (2019). Overtourism and degrowth: A social movements perspective. *Journal of Sustainable Tourism*, *27*(12), 1857–1875.
Morimoto, T., & Kawasaki, Y. (2017). Forecasting financial market volatility using a dynamic topic model. *Asia-Pacific Financial Markets*, *24*(3), 149–167.
Mutz, D. C., & Soss, J. (1997). Reading public opinion: The influence of news coverage on perceptions of public sentiment. *Public Opinion Quarterly*, *61*(3), 431–451.
Oliver, P. E., & Maney, G. M. (2000). Political processes and local newspaper coverage of protest events: From selection bias to triadic interactions. *American Journal of Sociology*, *106*(2), 463–505.
Önder, I., Gunter, U., & Gindl, S. (2020). Utilizing Facebook statistics in tourism demand modeling and destination marketing. *Journal of Travel Research*, *59*(2), 195–208.

Önder, I., Gunter, U., & Scharl, A. (2019). Forecasting tourist arrivals with the help of web sentiment: A mixed-frequency modeling approach for big data. *Tourism Analysis*, *24*(4), 437–452.
Pan, B., Wu, D. C., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, *3*(3), 196–210.
Park, E. O., Chae, B. K., & Kwon, J. (2020). The structural topic model for online review analysis: Comparison between green and non-green restaurants. *Journal of Hospitality and Tourism Technology*, *11*(1), 1–17.
Pritchard, A., & Jaworski, A. (2005). Discourse, communication and tourism dialogues. In A. Jaworski, & A. Pritchard (Eds.), *Discourse, communication, and tourism*. Clevedon, UK: Channel View Publications.
Qiu, R. T., Park, J., Li, S., & Song, H. (2020). Social costs of tourism during the COVID-19 pandemic. *Annals of Tourism Research*, *84*. https://doi.org/10.1016/j.annals.2020.102994.
Qiu, R. T., Wu, D. C., Dropsy, V., Petit, S., Pratt, S., & Ohe, Y. (2021). Visitor arrivals forecasts amid COVID-19: A perspective from the Asia and Pacific team. *Annals of Tourism Research*, *88*. https://doi.org/10.1016/j.annals.2021.103155.
Ritchie, B. W., & Jiang, Y. (2019). A review of research on tourism risk, crisis and disaster management: Launching the Annals of Tourism Research curated collection on tourism risk, crisis and disaster management. *Annals of Tourism Research*, *79*. https://doi.org/10.1016/j.annals.2019.102812.
Roberts, M., Stewart, B., Tingley, D., Benoit, K., Stewart, M. B., Rcpp, L., ... KernSmooth, N. (2019). Package 'stm'. Retrieved from http://nbcgib.uesc.br/mirrors/cran/web/packages/stm/stm.pdf.
Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003.
Salazar, N. B. (2012). Tourism imaginaries: A conceptual approach. *Annals of Tourism Research*, *39*(2), 863–882.
Schudson, M. (2002). The news media as political institutions. *Annual Review of Political Science*, *5*(1), 249–269.
Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The Azfin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 1–19.
Seraphin, H., Sheeran, P., & Pilato, M. (2018). Over-tourism and the fall of Venice as a destination. *Journal of Destination Marketing & Management*, *9*, 374–376.
Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2020.07.053.
Song, H., Qiu, R. T., & Park, J. (2019). A review of research on tourism demand forecasting: Launching the annals of tourism research curated collection on tourism demand forecasting. *Annals of Tourism Research*, *75*, 338–362.
Song, H., & Witt, S. F. (2000). *Tourism demand modelling and forecasting: Modern econometric approaches*. Routledge.
Stamolampros, P., Korfiatis, N., Chalvatzis, K., & Buhalis, D. (2020). Harnessing the "wisdom of employees" from online reviews. *Annals of Tourism Research*, *80*. https://doi.org/10.1016/j.annals.2019.02.012.
Stepchenkova, S., & Eales, J. S. (2011). Destination image as quantified media messages: The effect of news on tourism demand. *Journal of Travel Research*, *50*(2), 198–212.
Sun, S., Wei, Y., Tsui, K. -L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, *70*, 1–10.
Thorsrud, L. A. (2018). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, *1-17*.
Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E. J., & Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. *International Journal of Forecasting*, *34*(2), 355–365.
Tse, T. S., & Qiu, H. (2016). Issues arising from the rapid growth of Mainland Chinese visitors to Hong Kong: Implications for tourism marketing. *Journal of China Tourism Research*, *12*(3–4), 291–312.
Tussyadiah, I. P., & Fesenmaier, D. R. (2009). Mediating tourist experiences: Access to places via shared videos. *Annals of Tourism Research*, *36*(1), 24–40.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
Viken, A. (2016). Destinations discourses and the growth paradigm. *Tourism destination development* (pp. 35–60). Routledge.
Vu, H. Q., Li, G., & Law, R. (2019). Discovering implicit activity preferences in travel itineraries by topic modeling. *Tourism Management*, *75*, 435–446.
Wang, J., Li, Y., Wu, B., & Wang, Y. (2020). Tourism destination image based on tourism user generated content on internet. *Tourism Review*, *76*(1), 125–137.
Wen, L., Liu, C., Song, H., & Liu, H. (2020). Forecasting tourism demand with an improved mixed data sampling model. *Journal of Travel Research*. https://doi.org/10.1177/00472875209062.
Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, *46*, 386–397.
Yu, Q., McManus, R., Yen, D. A., & Li, X. R. (2020). Tourism boycotts and animosity: A study of seven events. *Annals of Tourism Research*, *80*. https://doi.org/10.1016/j.annals.2019.102792.
Zhang, J., Wong, P., & Lai, P. (2018). A geographic analysis of hosts' irritation levels towards mainland Chinese cross-border day-trippers. *Tourism Management*, *68*, 367–374.
Zhang, X., & Hellmueller, L. (2017). Visual framing of the European refugee crisis in Der Spiegel and CNN International: Global journalism in news photographs. *International Communication Gazette*, *79*(5), 483–510.