DISCOVERING TOURISM TOPICS FROM SOCIAL MEDIA: A CASE STUDY OF JAPAN

Valentinus Roby Hananto[1,2], Uwe Serdült[1,3], Victor V. Kryssanov[1]

[1] *Graduate School of Information Science and Engineering, Ritsumeikan University, Japan*
[2] *Department of Information System, Universitas Dinamika, Indonesia, valentinus@dinamika.ac.id*
[3] *Center for Democracy Studies Aarau (ZDA), University of Zurich, Switzerland*

ABSTRACT

In this digital age, tourism data on the Internet grows massively. The huge amount of data can be utilized to gain value propositions in smart tourism. Japan, as a major tourism destination worldwide, has a number of organizations that actively promote tourism through various social media sites. Discovering emerging topics of trends in tourism from social media platforms is a challenging task, especially in an unsupervised manner. The presented research aims to discover important tourism topics from social media in Japan using a topic model. The data was obtained from twelve tourism agencies' Twitter accounts. 21,766 tweets and retweets were collected for the period of four years from 2016 to 2019. A topic model was built using the Latent Dirichlet Allocation (LDA) method. The popular topics obtained reveal most discussed issues posted by tourism agencies in Japan. These topics include, for example, trip guides, culinary experience, and the cherry blossom season. The topic classification from this study provides with insights of Twitter usage promoting tourism across Japan.

Key words: topic modeling, Twitter, tourism.

1. INTRODUCTION

The growth of social media activities and user-generated content on the Internet creates an opportunity for the tourism industry. With the support of information technology, smart tourism can be implemented by transforming huge amounts of data into value propositions (Gretzel *et al*., 2015). The "smart experience" element puts the emphasis on technology-mediated tourism improvement through personalization, context-awareness, and real-time monitoring (Buhalis and Amaranggana, 2015). Implementation of smart tourism relies on the abundance of data and systems able to transform this data into value propositions. Brandt *et al.* (2017) identified three types of value propositions from smart tourism ecosystems: presence, environmental engagement, and topical engagement. For instance, discovering popular tourism spots, attractions, and events can be done with the help of social media analysis.

Japan is one of the most popular tourist destinations in the world. Based on The Travel & Tourism Competitiveness Report 2019 (World Economic Forum, 2019), Japan is ranked 4th out of 141 countries overall, after Spain, France, and Germany. As the country is going to host the Olympic Games in Tokyo, it would expect a growing number of tourists in the near future. The smart tourism concept in Japan can be related to the Japanese word "Omotenashi", which stands for the "Japanese style of providing hospitality". Implementing smart tourism to support "Omotenashi" would be a success factor in improving the performance of the tourism industry (Lim *et al*., 2017).

There are many organizations or agencies promoting tourism in Japan through various activities. The Japan National Tourism Organization (JNTO) is one of the leading tourism agencies. Established in 1964, it was initially a government-affiliated corporation before it was changed to become an independent agency (JNTO, 2006). Its main goal is to encourage international tourists worldwide to visit Japan. One way of promoting Japanese tourist attractions is to publish relevant information on the Internet. JNTO therefore maintains a website that can be accessed in multiple languages. JNTO also regularly posts tourism information using social media, such as Facebook, Twitter, and Instagram.

Analyzing tourism information data stemming from social media helps to get a better understanding about tourism trends. As one of the most popular social media networks, Twitter is a real-time microblogging service that allows users to post a status update (tweet) in 280 characters. JNTO's official Twitter account (@Visit_Japan) has been active since 2009 and has more than 100,000 followers. In addition to JNTO, there are other tourism organizations that utilize Twitter to promote tourism in Japan. The tremendous number of tweets yields various topical engagements of tourism. Although many studies explored the social media content, there is still limited research on topics of tourism, especially in Japan. Discovering important topics would provide valuable insights

IICST2020: 5th International Workshop on Innovations in Information and Communication Science and Technology, Malang, Indonesia

for people to illustrate tourism trends and images. In this study, a topic modeling method is proposed to obtain the most important and valuable topics dealing with tourism in Japan. The Twitter data retrieved from tourism agencies in Japan was used to build a topic model. Topics generated from the data, based on the model, are evaluated and discussed through the study. The goal of this study is to discover emerging topics of trends in tourism in Japan, using reviews available on Twitter. Discovering popular topics of trends from a text document (e.g. tweet) in an unsupervised manner is a challenging task. LDA is a method applied in this study to cluster text documents in an unsupervised manner.

The rest of the paper is organized as follows. Section 2 describes the related work on social media analytics and topic modeling. The proposed method is presented in Section 3. Section 4 specifies the data used in the research. Section 5 outlines the experiment, while results obtained are given in Section 6. Finally, Section 7 briefly discusses the results and draws conclusions.

2. RELATED WORK

The usefulness of social media analytics to support smart tourism was discussed in several studies. Marine-Roig and Clave (2015) analyzed more than 100,000 relevant online travel reviews with the goal of compiling an image of Barcelona. The authors' findings were useful for developing marketing strategies to improve branding among tourism agencies. A study on the smart tourism ecosystem using social media messages was conducted in San Francisco (Brandt *et al*., 2017). Using 600,000 geotagged tweets, a spatial and semantic analysis was done to present information for stakeholders in the tourism industry. The usage of Twitter data to analyze tourism was also discussed in Halim *et al*. (2019) for the case of Sabah, Malaysia. The data collected was processed in order for discovering new potential tourist attractions.

A relatively novel approach to analyze social media data is topic modeling. Topic modeling is a statistical method applied to discover latent themes and trends in a collection of text documents. Social media data, such as tweets, often can be characterized as text documents containing latent themes. Hidayatullah and Ma'arif (2017) proposed to create a topic model from traffic information originating from Indonesian Twitter messages. The topic model obtained revealed what topics were mostly posted by the Traffic Management Center in Java. A study of topic modeling was also done, using Twitter data in London, to classify distinctive and interpretive topic groupings. The classification was found to provide insights of the content and coverage of Twitter usage across inner London (Lansley and Longley, 2016).

3. PROPOSED METHOD

The concept of a system proposed to discover and analyze tourism topics from social media is illustrated in Figure 1.

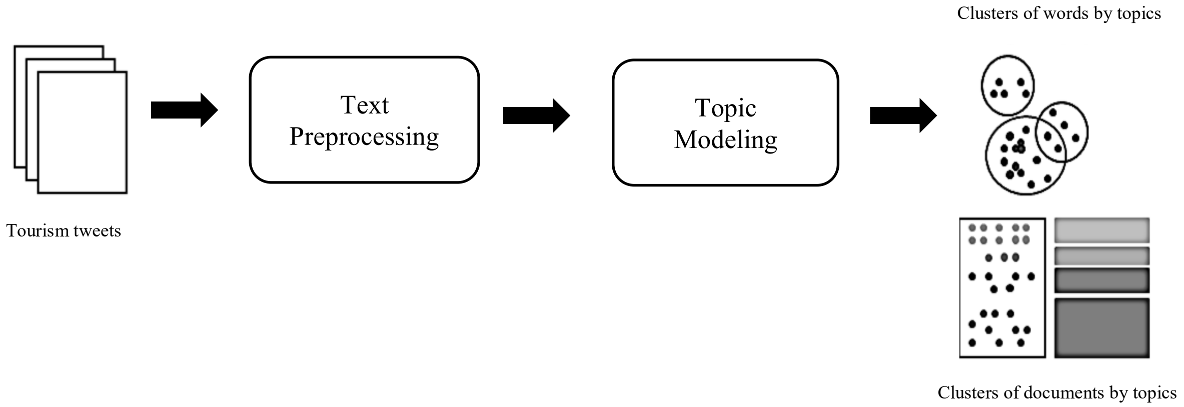

Fig. 1. The proposed system concept for analyzing tourism data from Twitter

First, relevant data (English texts) for is collected from Twitter accounts of Japan tourism agencies. The data is then combined into one dataset by aggregating tweets from all the relevant Twitter accounts. Before topic modeling is applied to cluster the documents, the data needs to be pre-processed to remove unnecessary elements (e.g. punctuation marks, numbers, Twitter symbols, URLs, and stop words). Non-English tweets are also removed. Each tweet is sliced into separate words (tokenization). Each tokenized word is then converted into its dictionary form (lemmatization). After finishing text preprocessing, the word dictionary is used to produce a document-term matrix. This document-term matrix is used as the input to build a topic model.

An LDA topic modeling method was deployed in this study. LDA builds on a generative probabilistic model to cluster discrete data, such as a text corpus (Blei *et al.*, 2003). This unsupervised machine learning method thus aims to model documents as emerging from multiple topics, where a topic is defined as a distribution over fixed vocabulary terms. There is a generative procedure of three steps in LDA. First, a topic is selected randomly from the distribution over topics for each document. Second, a word is selected from the distribution of words related to the chosen topic. Finally, the process is repeated for all words in the document. Applying LDA assumes inverting the actual generative process to extract latent topics from the documents.

4. DATA

Twitter data was collected from Twitter accounts of tourism agencies in Japan. The respective Twitter accounts were filtered, based on time and activity. Specifically, accounts that have been on Twitter for at least four years, have more than 1,000 followers, have posted more than 500 tweets over the last four years, and use English were selected. It is important to note that the selected accounts keep regularly posting tourism information, so that the analysis would reflect the current tourism trend. Some of the accounts are managed by JNTO, as it has several accounts for branch offices overseas (e.g. @visitjapanuk in London, @jntocanada in Canada). @JapanSafeTravel is JNTO's newest Twitter account, providing foreign visitors with safety tips and latest information in the case of a natural disaster. Although the latter account exists for only one year and the number of tweets is still less than 500, it already has more than 20,000 followers, and can provide for a variety of additional topics. This account was therefore selected as well.

Table 1. Twitter data sources used in the study.

No.	Twitter account	Managed by	Since	# followers	# tweets
1	@Visit_Japan	JNTO	2009	101,696	3,224
2	@JapanSafeTravel	JNTO	2018	21,358	443
3	@visitjapanuk	JNTO	2010	10,311	3,176
4	@jntocanada	JNTO	2010	2,446	1,151
5	@japantimes_life	Japan Times	2008	24,789	3,214
6	@japanguidecom	Japan Guide	2011	6,093	512
7	@JapanTravel	Japan Travel	2011	9,734	1,357
8	@MATCHA_global	Matcha Japan	2014	3,830	3,200
9	@tsunagu_Japan	Tsunagu Japan	2014	3,014	625
10	@tadaima_jp	Tadaima Japan	2014	4,741	1,020
11	@LIVEJAPANGuide	Live Japan	2015	14,500	3,201
12	@JapanTravelAdv	Japan Travel Advice	2012	17,853	643
Total					21,766

The data was collected using the Twitter Python API for the period of four years from 2016 to 2019. The Twitter accounts used, the number of followers, and the number of tweets obtained from each account are given in Table 1. The number of tweets collected from the twelve accounts sums up to 21,766. The frequency of tweets during this period was also analyzed over time (see Figure 2). As one can see from the figure, there is a decline in the number of tweets posted after Q4 2016. In 2017 and 2018, the number of tweets each quarter was less than 1,500, with the lowest number of 998 in Q4 2017. However, there is an increase again in 2019, reaching 1,685 tweets in Q4 2019.

After finishing text preprocessing, words that occurred many times, such as *Japan*, *visitjapan*, and *Japanese,* are identified by exploring the word frequency. The common words tend to be dominant, and would negatively affect the topic modeling, as they would appear in every cluster. Therefore, frequent words occurring in 5% of the documents and having no distinctive meaning for a specific topic were excluded from the dictionary. Each unique term in the dictionary and its frequency were included a document-term matrix. This document-term matrix is used as an input to train the LDA model.

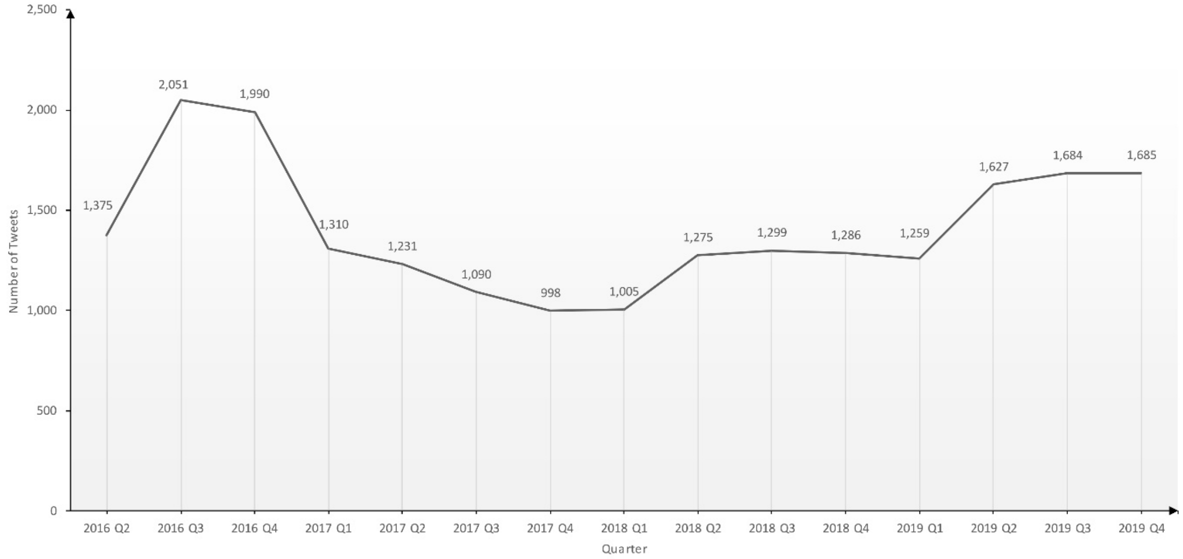

Fig. 2. Number of tweets posted from Q2 2016 to Q4 2019

5. EXPERIMENTS

The LDA modeling is implemented with the Gensim package in Python. In addition to the corpus and dictionary as inputs, the number of topics k is to be selected. This parameter was set, starting from 10 and, then, in increments of 10, i.e. k = 10, 20, 30, …, 120 topics. As a result, there were 12 topic models built that were ranked by a coherence score, $C_{v.}$ The latter measure is calculated, based on a sliding window that uses normalized pointwise mutual information and the cosine similarity (Röder *et al.*, 2015). A higher C_v value reflects a stronger correlation between all words within a topic and, therefore, stands for a better model. Figure 3 shows the change in the coherence score for the given data.

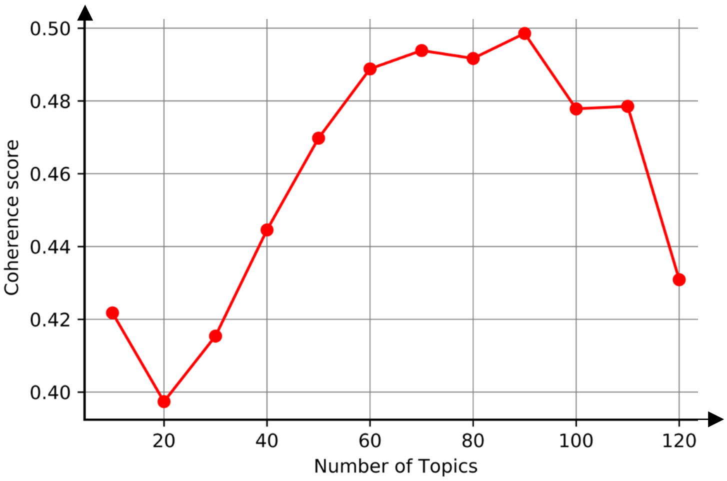

Fig. 3. Change in C_v over the number of topics

When C_v appears to keep increasing, the model that gave the highest coherence score before flattening out is selected. With the given data, the highest coherence score was achieved for the number of topics set at 90. Such a high number of topics would however be difficult to interpret. Too few topics (*e.g.* k = 20 or so), on the other hand, would result in non-cohesive clusters. In an experiment with 60 topics, the topic spread appeared to be satisfactory. A model with 90 topics (the highest coherence score) was also trained, but no significant improvement in the coherence of the topics was achieved with this model, compared to the case of k = 60. LDA results with 60 topics are shown in Figure 4.

The visualization of Figure 4 was obtained, using the LDAvis interactive chart for Python. The latter system was developed by Sievart and Shirley (2014) to better understand LDA modeling. On the left-hand side, the topics are presented as bubbles on a two-dimensional plane, whose centers are determined by computing semantic distances between the topics. The larger the bubble, the more prevalent the topic. An “ideal” topic model should, therefore, have only non-overlapping bubbles scattered uniformly throughout the chart. Selecting a bubble on the left panel interactively reveals the most relevant terms for the particular cluster on the right-hand side panel.

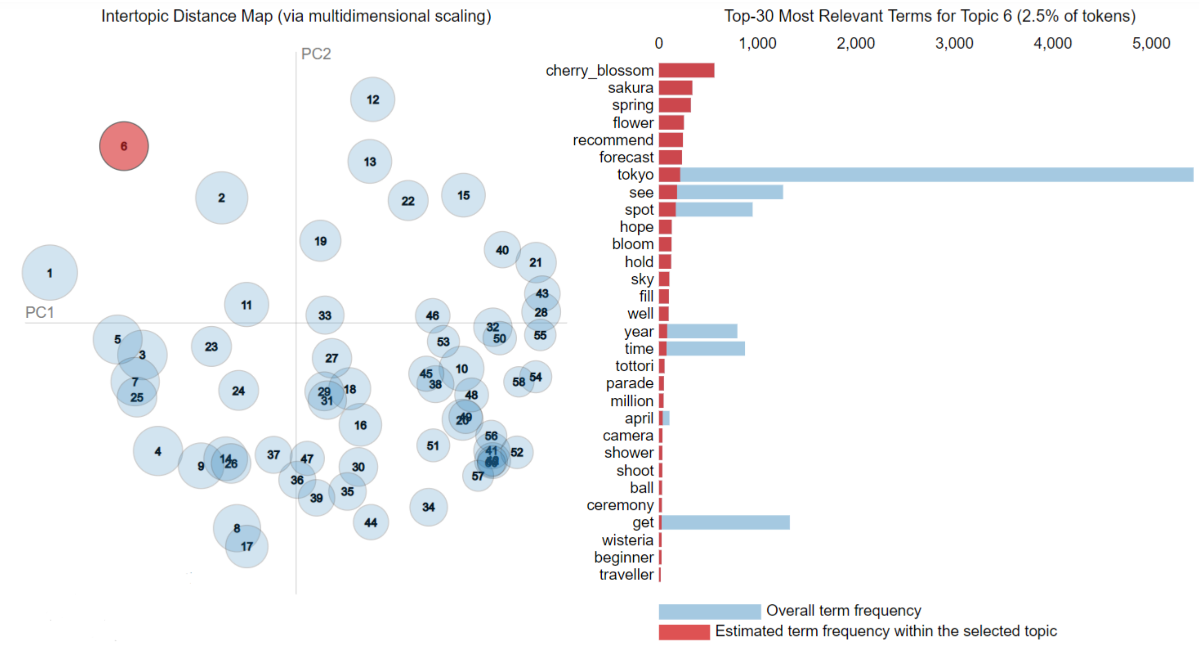

Fig. 4. LDA topic modeling visualization

6. RESULTS

Based on the output of the LDA model obtained in the experiments, the Twitter data was clustered into 60 topics. Each document has membership degree (or weight) of each topic. Each tweet is then classified into a topic with the highest membership degree. To illustrate the output of the LDA model, the top-10 popular topics with the top-5 probable words are given in Figure 5.

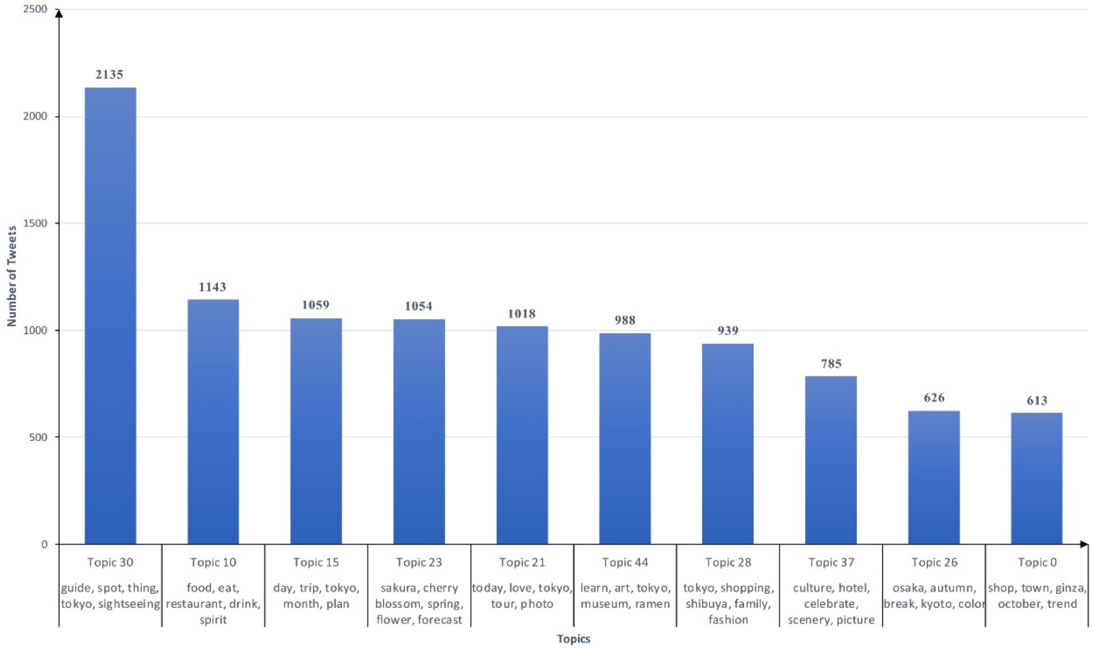

Fig. 5. Number of tweets by the most popular topics

Topic 30 is the most popular topic of the data that includes a total of 2,135 tweets. It has five keywords in the order of their associated weights: *guide, spot, thing, tokyo,* and *sightseeing*. From these keywords, the topic would be inferred as mostly discussing the guidance to enjoy sightseeing spots in Tokyo. An example of a unique topic of Japan tourism is Topic 23: *sakura, cherry blossom, spring, flower,* and *forecast*. Japan is famous for the sakura (cherry blossom) season, which typically takes place in early April. Thus, many tourism tweets posted are related to the forecast of the full bloom of sakura.

In addition to discovering the topic keywords, an analysis of each document for what topic it contributes mostly would help interpret the topic. By reading documents within a topic, the common theme would be inferred. Three tweets were selected per topic with the highest per-topic percentage contribution. Topics 13 and 15 are mainly about the tourist guides, Topic 10 is about food and drinks, and Topic 23 is about the sakura season. Examples of tweets from the dominant topics are given in Table 2.

Table 2. Top tweets by most dominant topics.

No.	Topic Number	Tweets	Account
1	Topic 30 (guide, spot, thing, tokyo, sightseeing)	"Top 4 free thing to see in Tokyo: https://t.co/CueWBWngsG via @YouTube"	@ JapanTravel
		"Stranger things: Weird ways to get festive in Japan https://t.co/FpJGILrVVf"	@ japantimes_life
		"Visiting Roppongi, Tokyo? With so much to see and do in Roppongi here's a guide to get you started! https://t.co/iIcW5uy2gM"	@ jntocanada
2	Topic 10 (food, eat, restaurant, drink, spirit)	"Where to Eat and Drink in Kyoto | via Food & Wine"	@Visit_Japan
		"The Best Food in Tokyo Is Eaten at the Counter https://t.co/SHCmvIzeZg via @CNTraveler"	@Visit_Japan
		"Where to Eat and Drink in Okinawa, Japan | via Food & Wine https://t.co/OAyv3jWCaj"	@Visit_Japan
3	Topic 15 (day, trip, tokyo, month, plan)	"Hiroyuki Ito, a photographer, spent two months documenting interesting moments across #Japan => https://t.co/fjcTcOLgIw via @nytimes https://t.co/gm1iR7uJH1"	@Visit_Japan
		"20 easy day trips to make from Tokyo https://t.co/7FQ9ySzMBU #Japan https://t.co/3qCS5TbH3U"	@visitjapanuk
		"Five of the best day trips from Tokyo | via Time Out Tokyo https://t.co/jE11tHK9IO"	@Visit_Japan
4	Topic 23 (sakura, cherry blossom, spring, flower, forecast)	"15 Cherry Blossom Spots In Tokyo That You Just Have To See! https://t.co/SJVDFfwJCx #Tokyo #spring #sakura #travel"	@MATCHA_global
		"Japan's Cherry Blossoms In 2019 - Forecast and Best Spots! https://t.co/8EikepUEA5 #sakura #cherryblossoms #spring"	@MATCHA_global
		Cherry blossoms are already blooming in Atami, just south of #Tokyo! Find out more about where and when you can see these and other flowers. https://t.co/kJlenVTTZI #sakura2019 #traveling #Japan"	@LIVEJAPANGuide

7. DISCUSSION AND CONCLUSIONS

In the presented study, topic modeling using LDA was conducted on Twitter messages from twelve Japan tourism agencies. In terms of tweet frequency within the investigated period, there is a downward trend from 2016 to 2017. However, the tweet number goes up again gradually in 2018 through 2019. This makes one wonder whether the recent increase in the number of tourism posts is related to any tourism issue. In terms of the contextual analysis, the topic model built generated 60 clusters, with three popular topics discovered being sightseeing trip guide, culinary, and sakura season. Using the clusters generated by LDA, analyzing tweets from each topic would be beneficial for business intelligence in tourism industry.

Deciding on the suitable number of topics is the most challenging part of topic modeling. If the number is set too low, some topics would be too broad, and would need splitting apart. On the other hand, if the topic number is set too high, some very similar topics that should be joined are found. Metrics, such as coherence score, can be used to help decide the optimal number of topics, but no perfect metrics exist. Research on advanced methods to select a suitable number of topics, and on semi-supervised topic modeling is left for the future.

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Brandt, T., Bendler, J., and Neumann, D. (2017). Social media analytics and value creation in urban smart tourism ecosystems. *Information and Management*, 54(6), 703–713.

Buhalis, D., and Amaranggana, A. (2015). Smart Tourism Destinations Enhancing Tourism Experience Through Personalisation of Services, In: *Information and Communication Technologies in Tourism 2015*, Tussyadiah, I., and Inversini, A. (Eds.), 377-389. Springer: Cham.

Gretzel, U., Sigala, M., Xiang, Z., and Koo, C. (2015). Smart tourism: foundations and developments. *Electronic Markets*, 25(3), 179–188.

Halim, M. A., Saraf, N. M., Hashim, N. I., Rasam, A. R. A., Idris, A. N., and Saad, N. M. (2018). Discovering new tourist attractions through social media data: A case study in Sabah Malaysia, In: *2018 IEEE 8th International Conference on System Engineering and Technology*, 157–161. IEEE: New York NY.

Hidayatullah, A. F., and Ma'Arif, M. R. (2017). Road traffic topic modeling on Twitter using latent dirichlet allocation, In: *2017 International Conference on Sustainable Information Engineering and Technology*, 47–52. IEEE: New York NY.

JNTO. (2006). JNTO-What We Do. https://www.jnto.go.jp/eng/about/pdf/about_JNTO_20060925.pdf (last accessed on December 20, 2019)

Lansley, G., and Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58, 85–96.

Lim, C., Mostafa, N., and Park, J. (2017). Digital Omotenashi: Toward a Smart Tourism Design Systems. *Sustainability*, 9(12), 2175.

Marine-Roig, E., and Anton Clavé, S. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing and Management*, 4(3), 162–172.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures, In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 399–408. ACM: Shanghai.

Sievert, C., and Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics, In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. Association for Computational Linguistics: Baltimore.

World Economic Forum. (2019). The Travel & Tourism Competitiveness Report 2019 Travel and Tourism at a Tipping Point. http://www3.weforum.org/docs/WEF_TTCR_2019.pdf (last accessed on October 10, 2019)