J Inf Process Syst, Vol.17, No.3, pp.586~598, June 2021
https://doi.org/10.3745/JIPS.04.0217

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

# Analyzing Customer Experience in Hotel Services Using Topic Modeling

Van-Ho Nguyen[*] and Thanh Ho[**,***]

## Abstract

Nowadays, users' reviews and feedback on e-commerce sites stored in text create a huge source of information for analyzing customers' experience with goods and services provided by a business. In other words, collecting and analyzing this information is necessary to better understand customer needs. In this study, we first collected a corpus with 99,322 customers' comments and opinions in English. From this corpus we chose the best number of topics (K) using Perplexity and Coherence Score measurements as the input parameters for the model. Finally, we conducted an experiment using the latent Dirichlet allocation (LDA) topic model with K coefficients to explore the topic. The model results found hidden topics and keyword sets with high probability that are interesting to users. The application of empirical results from the model will support decision-making to help businesses improve products and services as well as business management and development in the field of hotel services.

## Keywords

Customer Experience, Hotel Services, LDA, Text Mining, Topic Modeling

## 1. Introduction

Everyday users post or review a large volume of comments, reviews, blog posts, or online news [1,2]. This data is stored in unstructured texts [3]. The problem is that with a significant amount of textual content, not all of it can be read. Therefore, an approach is needed to "summarize" this data into profound characteristics, such as the subject of a comment or an opinion, or online reviews of the product or the service being discussed, in other words, the "topics" users are referring to.

The topic modeling approach has been used by many researchers to analyze data and extract information in many fields such as tourism, bioinformatics, accommodation, education, and online sales [2-10]. Topics of opinion when making comments or providing feedback on the company's products or services are highly diverse, they provide information and capture the habits and behavior of individuals or online communities. However, owing to the features of online networks, the subjects of the content of messages have not been created beforehand or, in other words, the topic being discussed on the network forum is implicit [4,7]. Therefore, discovering the topic and understanding the content of users' exchanged messages is a challenging problem [8].

Manuscript received May 14, 2020; first revision July 20, 2020; accepted August 4, 2020.

**Corresponding Author:** Thanh Ho (thanhht@uel.edu.vn)
* School of Business Information Technology, University of Economics Ho Chi Minh City, Vietnam (honguyen39.ck28@st.ueh.edu.vn)
** Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City, Vietnam (thanhht@uel.edu.vn)
*** Vietnam National University, Ho Chi Minh City, Vietnam (thanhht@uel.edu.vn)

In this study, we were able to successfully detect hidden topics that our guests were referring to about hotel services so that we could capture the issues of interest to our users. With the problems identified, we can retain customers, improve existing customers' satisfaction, or increase conversion rates when applying a business strategy that matches target products and services with the right customers.

The rest of this paper is organized as follows: Section 2 discusses related studies, focusing on analyzing text data and user opinions in the field of hotel services, latent Dirichlet allocation (LDA) topic modeling, and labeling topic techniques. Specific research methods applied are presented in Section 3 and the results are discussed in Section 4. Finally, Section 5 presents the conclusions with an outline of future work.

# 2. Related Work

In previous studies [1,2,4-8], the authors studied and experimented on unstructured data. This data mainly comprised of comments from customers' feedback on the company's products and services through online channels. Analyzing users' comments in the field of hotel services has also been studied [1,2,5,11-13]. The topic model [14,15] is a useful tool for identifying hidden topics from online customer reviews or social networking data [6,9,11,12,15-20]. Labeling the topics, which is a difficult problem, was mentioned in [7] and [21].

## 2.1 Analyzing Customer Experience Approaches to Text Mining

Text data mining refers to the process of extracting high-quality information from text. This is an important part of data mining and knowledge discovery [8]. It handles a set of documents and extracts meaningful information from the text, and creates the implicit information using various data mining algorithms such as statistics and machine learning.

Among the tasks of document data mining, opinion mining is a difficult task that has attracted much attention recently. This is one of the most active research topics and has been widely studied in data, web, and text mining [13]. User experience is the feeling that customers receive when using a company's products or brands. User opinion analysis focuses on opinions, feelings, reviews, attitudes, and emotions from written language.

Previously, due to time and technology limitations, customer opinions often had to be surveyed by questionnaire. However, with the development of the Internet and information technology, customer feedback can be quickly collected online and this reduces the time and manpower required to better understand customer experience. The growing importance of analyzing user comments coincides with the growth of social media, such as reviews, forum discussions, blogs, and social networks such as Facebook and Twitter. In particular, in the era of digital development, we now have a significant amount of data recorded in digital form for analysis [2,13,20].

## 2.2 Analyzing Customer Experience in Hotel Services

Currently, the hotel and restaurant industry has undergone continuous growth and profound development throughout the world, which is recognized by international organizations such as the World Bank and the World Tourism Organization [2]. A number of articles have emerged on several methods used in research to extract comments from hotel reviews. Most of these approaches are based on natural language processing techniques, resources, and vocabulary approaches based on machine learning [1,5]. In [13], the authors developed a hybrid model to implement opinion mining on multilingual social media texts and explore differences of opinions of cross-cultural customers. The results show that there is a difference in preferential hotel services between Vietnamese and American customers. Based on the results of this study, hotels should improve customer satisfaction by diversifying services according to customers from different cultures.

In other studies [2,5,11], the authors also analyzed customer experience in tourism by analyzing opinions according to the topic model [2,5] and characteristics extraction techniques including knowledge-based and aspect-based approaches [11]. With these methods, the author proposes a number of important tools for tourists when deciding which hotel to stay in and which restaurants or places to visit. These articles focus on discovering opinions applied to data from travel rating websites. The studies [2,5,11] describe how the emotion analysis results of text reviews can be visualized using Google Maps, enabling users to easily discover good hotels and places to stay. More advanced features are also provided, such as visual images and interactions [11]. Some applications from the research [5] also need to be experimented and applied, particularly the applications of (1) forecasting and proposing ratings of products and services, (2) topic analysis and interpretation, and (3) proposal ratings for reviewers.

## 2.3 LDA Topic Modeling

In machine learning and natural language processing, the topic model is a type of statistical model to explore abstract "topics" that occur in a collection of documents. The topic model is a text-mining tool frequently used to explore the semantic structures hidden in the body of documents [9,10]. In the topic-classification problem, the LDA algorithm [9] is one of the more favored model classes because of its fast calculation efficiency, providing accurate results based on the generative statistical model (Fig. 1). The results returned by LDA include textual and each word distribution. Textual distribution is a combination of a mixed distribution of a fixed number of topics, and each word distribution represents the level of contribution to the document through its representation in topics.

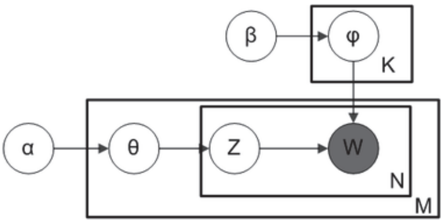


**Fig. 1.** Plate notation for LDA with Dirichlet-distributed topic-word distributions. Adapted from [9].

Definitions used in the LDA model:

- Word is the most basic unit of the LDA. A word is determined by an index in the dictionary with values of 1,2, ..., V. $w_i$ is expressed as a one-hot vector $w_i \in \mathbb{R}^V$ such that the *i*-th element of the vector is 1 and the remaining elements are 0.
- The document is a combination of N words denoted by $\mathbf{w} = (w_1, w_2, ..., w_N)$. Each vector represents one word in a sentence.
- The corpus is a combination of M documents denoted by $\boldsymbol{D} = \{w_1, w_2, ..., w_M\}$.

- Latent topics are determined based on the distribution of words and mediate the representation of documents by topic. The number of topics was predefined, and the symbol was **K**.

A formal description of the LDA model is shown in Table 1.

In the LDA model, there are two continuous iterative processes: the topic selection process and the word selection process. The initial parameters of the process were *α* and *β*. Then, we calculate the joint distribution of topic *θ* and the distribution of words according to topic *w*. To infer topics in the corpus, the model creates an imaginative process whereby texts are created in a mechanism that can infer and reverse that process. The probability distribution of the document is created in a random mix of topics, in which each topic is determined by the distribution across all words. LDA [9] assumes a generative process for a corpus ***D*** consisting of ***M*** documents, as follows:

(1) Each document has a length $N_i$, with N~Poisson (ξ).
(2) Choose θ~Dir(α).
(3) Choose φ~Dir(β).
(4) For each of the $w_{ij}$ of document:
  (a) Choose a topic $z_{ij}$–multinomial ($\theta_i$).
  (b) Choose a word $w_{ij}$ from $p(w_{ij}|z_{ij}, \beta)$, a multinomial conditional probability on the topic $z_{ij}$.

**Table 1.** Description of used variables in LDA model

| Symbol | Description | Consonant |
|---|---|---|
| $\alpha$ | the parameter of the Dirichlet prior on the per-document topic distributions | |
| $\beta$ | the parameter of the Dirichlet prior on the per-topic word distribution | |
| $\theta_i$ | the topic distribution for document *i* | |
| $\varphi_k$ | is the word distribution for topic *k* | |
| $z_{ij}$ | the topic for the *j*-th word in document *i* | $z_{ij}$ integer, between 1 and K |
| $w_{ij}$ | the *j*-th word in the document *i* | $w_{ij}$ integer, between 1 and K |

First, we assume that the number of hidden topics is known and equal to **K**; therefore, **K** will also specify the number of dimensions of the Dirichlet distribution. Second, we determine the probability of the word parameterized by $\beta \in \mathbb{R}^{KxV}$, with $\beta_{ij} = p(w_j\!=\!1|z_i\!=\!1)$. Then, these parameters are fixed. Finally, we calculate the probability density function of topics for each document once we know the *α* parameter according to the Dirichlet distribution:

$$p(\theta\,|\,\alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_i^{\alpha_i-1} \tag{1}$$

Then, the general probability distribution of the topic mixture *θ* with the set **N** topic *z* and the set **N** words *w* with *α*, *β* is known as

$$p(\theta,\mathbf{z},\mathbf{w}\,|\,\alpha,\beta) = p(\theta\,|\,\alpha)\prod_{i=1}^{N} p(z_n\,|\,\theta)p(w_n\,|\,z_n,\beta)\,. \tag{2}$$

The component $p(\theta|\alpha)$ is a mixture topic probability distribution corresponding to the document when the Dirichlet distribution *α* parameter is known in advance. The other side $\prod_{i=1}^{N} p(z_n|\theta)p(w_n|z_n,\beta)$ is the probability distribution when determining the distribution mixture topic *θ* and Dirichlet distribution parameter *β*. If we take the marginal probability of a document by integrating over *θ* and taking the sum over *z*, we obtain

$$p(\mathbf{w} \mid \alpha,\beta) = \int p(\theta \mid \alpha) \prod_{i=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n,\beta) d\theta \,. \tag{3}$$

Finally, we calculate the probability of the corpus based on the marginal probability from a single document:

$$p(D \mid \alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \prod_{i=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n,\beta) d\theta_d \tag{4}$$

From the probability equation on the whole document, the expectation maximization (EM) method is applied to estimate the parameters $\alpha$ and $\beta$ and then calculate $\theta$, $z$, $\varphi$, $w$.

## 2.4 Labeling of Topics

Because a word can be distributed among different topics, it is difficult to identify the label of each topic and distinguish the differences among topics. Research in [7] and [21] has shown that there are two main ways to label topics: (1) asking experts in specific fields to label the topics (this is very difficult to achieve because of time and cost constraints) and (2) automatic labeling of topics. In [7], a method of automatic labeling of topics was used based on training data. First, the authors created a list of topics, including labels for product marketing. They then manually chose documents that belonged to only one topic; they selected approximately 500 documents for each topic. Finally, they used an LDA model with this corpus. As a result, the list of distributions of each word belonging to a specific topic has already been labeled. In the next step, they calculated similar measures between the training topics and the discovered ones, including *cosine similarity*, *overlapping similarity*, *mutual similarity*, and *dice similarity* plus *Tanimoto* and *Jaccard* coefficients. These measures helped to compare topics and obtain labels for detected topics. They then chose the maximum value for the results.

# 3. Research Methods

## 3.1 Overview of Research Model

Opinion mining is a classification process with three main levels: document level, sentence level, and aspect level [13]. Document-level opinion mining aims to classify a document that expresses an opinion on a specific topic [4]. At this level, a document is considered a basic unit of information. To analyze the model in this study, customer reviews or comments from online websites were collected, and then those written in English were extracted. Data preprocessing was performed carefully using a Python package. After obtaining the clean data set, the model was trained until the optimal number of topics (K) were selected, and K is the input for building the LDA model. The output of the LDA model is the set of topics and keywords, and a model is built to label the extracted topics. Finally, hidden topics were identified and visualized using the *pyLDAvis* package [22]. The flow in Fig. 2 proposes an experimental model for collecting, processing, and analyzing customer experience.

## 3.2 Data Collection

The experimental corpus in this study was collected from websites of companies in hotel services, including https://www.agoda.com/ and https://www.booking.com/. To collect data, we researched application programming in Python to access the website's API and collect user reviews and articles on the review tabs. The JSON data were converted to a CSV data frame format, and then topic extraction analysis was performed on the collected corpus. Some extracted attributes for analysis include *hotel_id*, *review_comments*, *rating*, and *review_date*.

A total of 99,322 user comments were collected and used as input data for analyzing the model, further details are listed in Table 2.

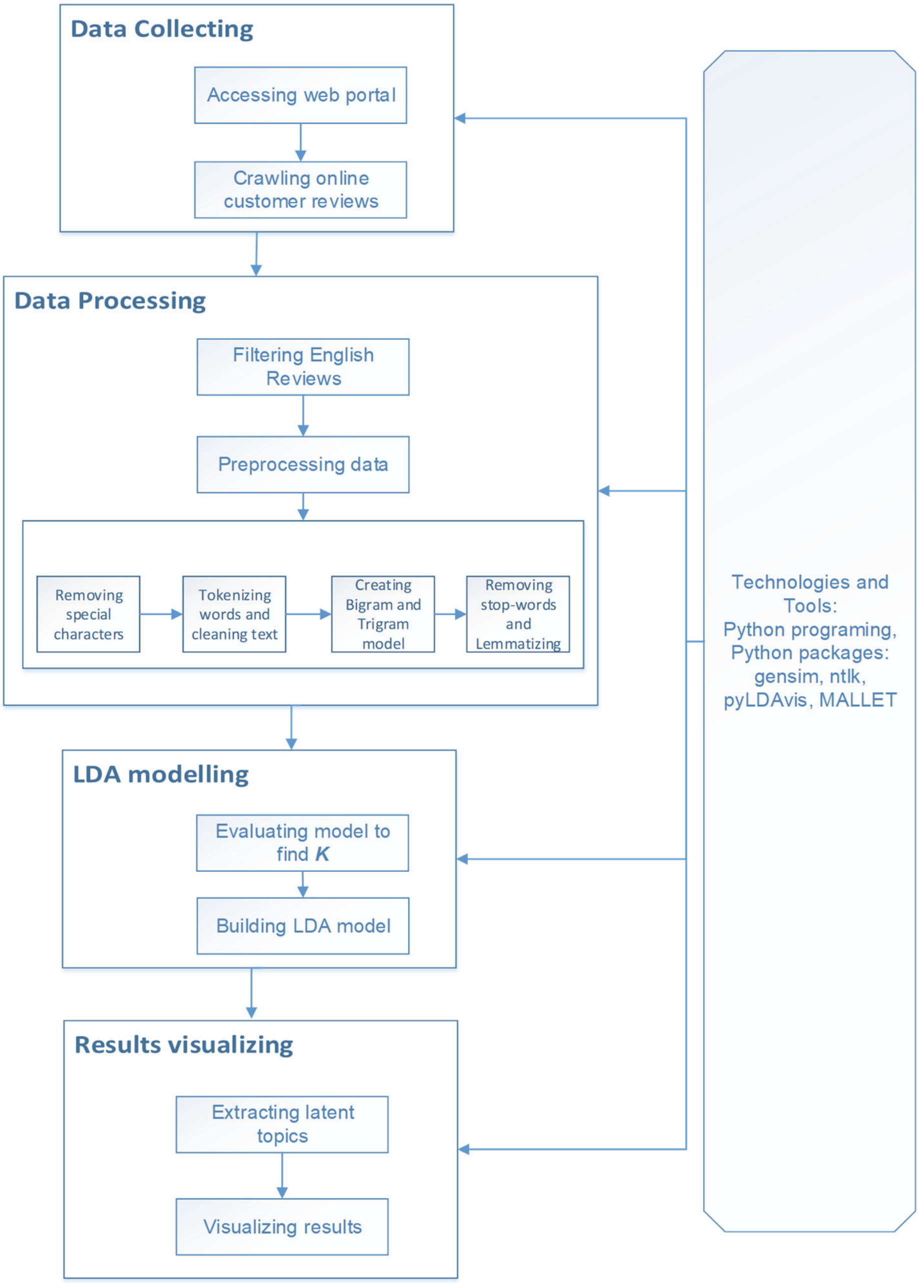

Data Collecting
Accessing web portal
Crawling online customer reviews
Data Processing
Filtering English Reviews
Preprocessing data
Removing special characters
Tokenizing words and cleaning text
Creating Bigram and Trigram model
Removing stop-words and Lemmatizing
LDA modelling
Evaluating model to find *K*
Building LDA model
Results visualizing
Extracting latent topics
Visualizing results
Technologies and Tools: Python programing, Python packages: gensim, ntlk, pyLDAvis, MALLET

**Fig. 2.** Proposed LDA topic modeling for analyzing customers experience in hotel services.

**Table 2.** Experimental corpus

| Sources | Number of reviews | Number of sentences |
|---|---|---|
| Agoda.com | 59,382 | 273,528 |
| Booking.com | 39,940 | 239,640 |
| Total | 99,322 | 513,168 |

## 3.3 Data Preprocessing

Preprocessing data is one of the most important steps in data mining, especially in the extraction of text data, because there are many differences in text-based content on electronic media channels on the Internet.

In online services, when providing feedback, users often repeat a number of special characters to emphasize their messages. These words can make it difficult to analyze reviews, to avoid this problem, these special characters should be removed. Punctuation marks that are not meaningful in the corpus were also deleted. Uppercase characters are converted to lowercase characters, numbers, and spaces, and stop words are removed. Users on social networks and online services often use a mix or match of misspelled words and unexpected or intentional mistakes. In particular, there have been recent users accessing online services via mobile phones and may ignore grammar and spelling rules, use of abbreviations, emoticons, and more concise sentences. Therefore, the data preprocessing stage is critical and determines the accuracy of the model. Fig. 3 shows the preprocessing data flow before being built into the LDA model.

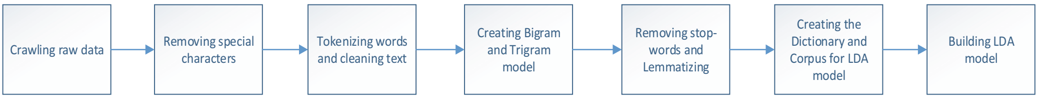

Crawling raw data
Removing special characters
Tokenizing words and cleaning text
Creating Bigram and Trigram model
Removing stop-words and Lemmatizing
Creating the Dictionary and Corpus for LDA model
Building LDA model

**Fig. 3.** Data preprocessing flow.

## 3.4 Building the LDA Model

### 3.4.1 Creating the bigram and trigram for the model

The LDA model uses the input as a co-occurrence matrix of words (n-grams). Bigrams are two words that frequently occur together in documents. Trigrams are three frequently occurring words. To calculate the frequency of copper appearing on these matrices, bigrams and trigrams were created. The function *phrases()* in Gensim is useful for building bigrams and trigrams. In the coding process, *min_ counts* the minimum frequency for a given word included in the grams, and the threshold value is input. In the next step, the stop-words are removed and only the words that are tagged with the category [‘NOUN’, ‘ADJ’, ‘VERB’, ‘ADV’] are filtered out. These stop words in the English language are already included in the *nltk* package.

### 3.4.2 Creating the dictionary and corpus

The dictionary and corpus are the two main inputs for the LDA model. The Gensim package was used to create the data. After processing, we obtained a corpus, that is; a set of pairs (index, frequency) that encodes the documents about the index specified in the dictionary along with the frequency of their occurrence in the text. To convert back from index to vocabulary, we used the dictionary *id2word*.

### 3.4.3 Building a LDA model

LDA describes documents as a mixture of topics with certain probabilities. For the purpose of paragraphs (documents) is represented by a number of topics, which are then represented by a small set of words (with the weight corresponding to each descending word). The main parameters specified in the Gensim LDA model are the number of topics (*num_topics*), the number of documents to be included in each training batch (*chunksize*), and the number of training sessions (*passes*). After the model is trained, the results can be saved to the folder for later use.

The Perplexity and Coherence Score are indicators that describe the quality of the model. It can be used in the best number of topics, which is consistent with the data. Perplexity is based on the logarithm of the maximum likelihood function (MLE), which lowers the perplexity of the model quality as much as possible. On the contrary, the model with the higher coherence score will be a better model. In this study, the authors experimented with the topic number 14 (with the corresponding Perplexity = -9.0373 and Coherence Score = 0.6392) as input parameters for the model. Fig. 4 shows the correlation between the coherence score and the number of topics in the experimental process.

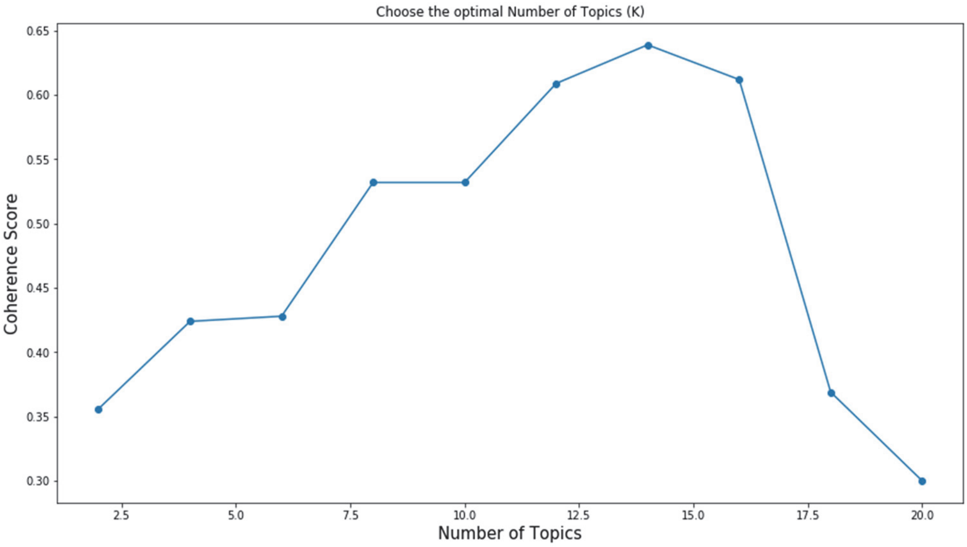


**Fig. 4.** The correlation between Coherence Score and number of topics (K).

## 4. Experimental Results

### 4.1 Identifying the Topics and Dominant Topics

After training the LDA model, we determined the distribution of documents by topics and represented topics according to the distribution of words. Table 3 shows keywords with probability of Topics 0, 2, 5, 9, and 13.

For each of the above topics, the keywords are expressed according to the probability of appearing from high to low, for example in Topic 0 the word ***"bus"*** has the highest probability of 0.1642, similar to Topic 9 the word ***"room"*** appears the most frequently with a probability of 0.0529.

**Table 3.** Topic analysis result for five topics out of a total of 14

| Topic 0 | | Topic 2 | | Topic 5 | | Topic 9 | | Topic 13 | |
|---|---|---|---|---|---|---|---|---|---|
| **Word** | **Pr** | **Word** | **Pr** | **Word** | **Pr** | **Word** | **Pr** | **Word** | **Pr** |
| bus | 0.1642 | money | 0.1937 | people | 0.2425 | room | 0.0529 | good | 0.2260 |
| hostel | 0.0795 | value | 0.1691 | option | 0.0930 | hotel | 0.0485 | nice | 0.1335 |
| car | 0.0590 | English | 0.0964 | single | 0.0775 | staff | 0.0427 | breakfast | 0.0959 |
| station | 0.0521 | extra | 0.0931 | double | 0.0603 | stay | 0.0371 | food | 0.0803 |
| tell | 0.0477 | whole | 0.0602 | group | 0.0527 | clean | 0.0277 | restaurant | 0.0497 |

After identifying the topic along with the corresponding probabilities of the word in that topic, we identify the predominant topics in the documents, i.e., the ones with the highest probability rate. The *format_topics_sentences()* function is used to determine the dominant topic in a document. Table 4 shows the dominant topics with percentages in the document.

**Table 4.** Dominant topics with percentage in the document of four topics out of a total of 14

| Dominant topic | Topic keywords | Number of documents | Percentage in document |
|---|---|---|---|
| 8 | hotel, good, restaurant, food, walk, staff | 1,027 | 0.0516 |
| 7 | room, hotel, staff, stay, night, bad, book | 3,781 | 0.1900 |
| 0 | bus, day, book, check, early, arrive, get | 3,679 | 0.1848 |
| 4 | stay, staff, place, would, recommend, time | 1,419 | 0.0713 |

## 4.2 Inferring the Label of Topics

LDA is a generative approach, in which for each topic, the model simulates the probabilities of word occurrences as well as the probabilities of topics within the document. The top 10 words for each topic are displayed in Fig. 5.

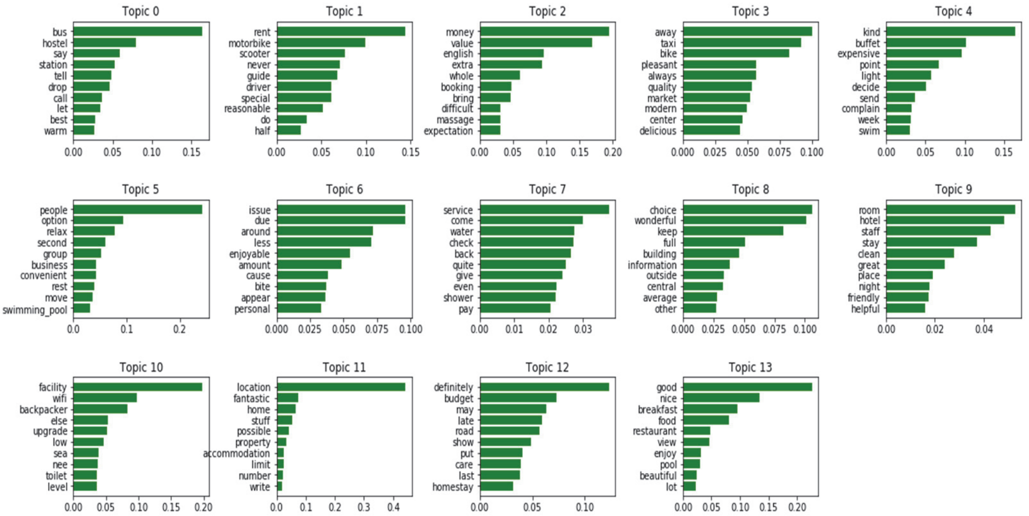


**Fig. 5.** The topics with descending probability keywords.

We can observe the following meaningful matches. Topic 0 deals with customers' transportation to arrive or leave the hotel. Topic 2 appears to be related to the price and cost of renting a room at the hotel. Topic 13 appears to be about the quality of meal services in restaurants and hotels.

An open source library, MALLET, which is a topic analysis toolkit based on the LDA model, was developed by McCallum and used to analyze the reviews not only chosen from the corpus but also those already labeled [7]. After the detailed list of training topics were collected, we executed the topic labeling technique according to the similarity measures mentioned in Section 2.4. The labels for the discovered topics were identified. For instance, in Fig. 6, the label of Topic 5 is "*people*" because it has the highest values. Equally, the label of Topic 9 is "*room services*." From the results of the keywords extracted from Fig. 5, we can infer the labels of the corresponding topics, as shown in Fig. 6.

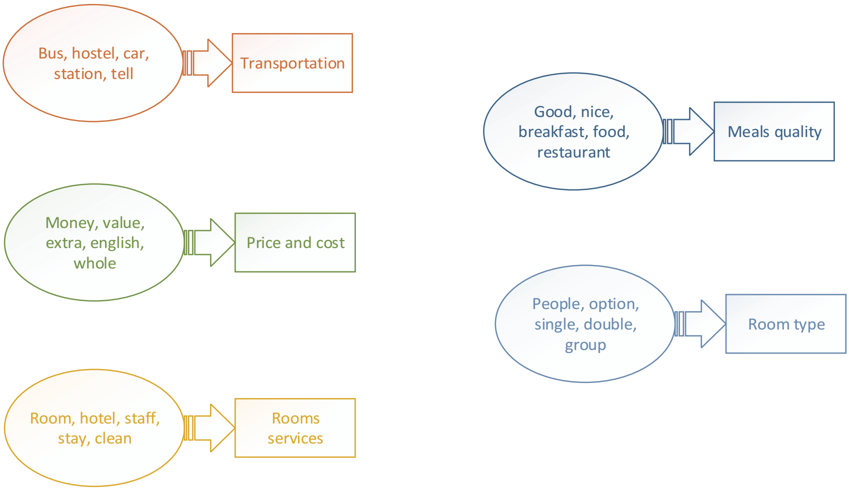
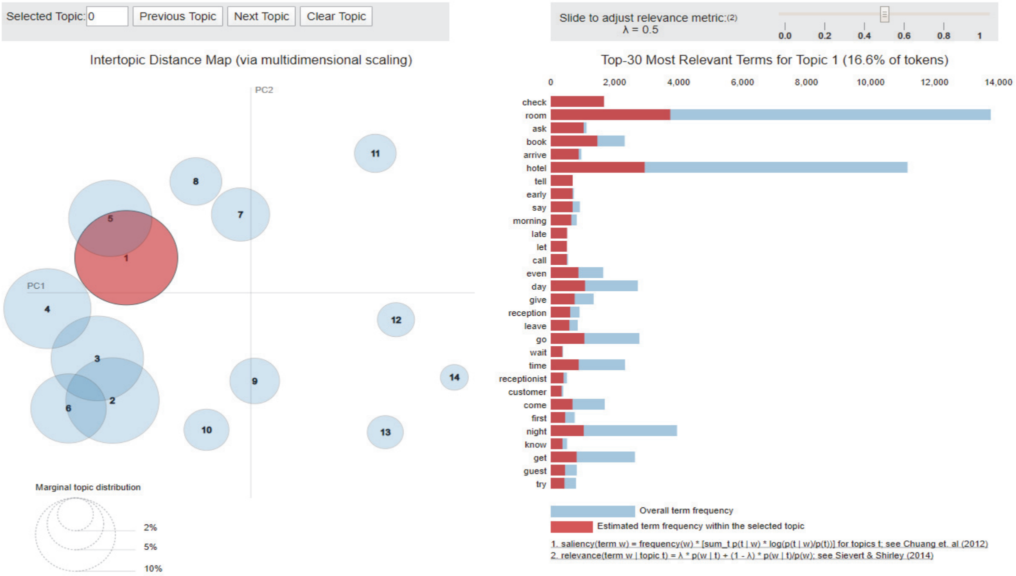

Bus, hostel, car, station, tell
Transportation
Good, nice, breakfast, food, restaurant
Meals quality
Money, value, extra, english, whole
Price and cost
People, option, single, double, group
Room type
Room, hotel, staff, stay, clean
Rooms services

**Fig. 6.** Inferring topic from keywords.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

**Fig. 7.** Visualizing the topics-keywords by *pyLDAvis* package.

### 4.3 Visualizing and Interpreting Results

The *pyLDAvis* package was developed to work effectively with Jupyter notebooks and to be useful in visualizing the LDA model results with interactive charts [22], as shown in Fig. 7. Each bubble on the left side represents the topic. The larger the bubble, the more popular the topic is. A good topic model will consist of fairly large, non-overlapping bubbles scattered in the chart instead of being grouped in a quadrant. A model with too many topics often has many small bubbles overlapping. The words and bars on the right-hand side are updated when one of the bubbles is moved. These words are prominent keywords that constitute the chosen topic.

By using the *pyLDAvis* package that is designed to support interactively: (1) we can manually select each topic to identify the most frequently and/or relevant terms and use different values for the λ parameter. This can be helpful when attempting to assign a human interpretable name or "meaning" to each topic; (2) discovering the intertopic gap plot can help to learn how topics relate to each other, including potential higher-level structures among topic groups.

## 5. Conclusion and Future Work

The main contribution of this article is that a novel method was presented for analyzing customer experience using a topic model based on a corpus for the hotel sector. The results show that the proposed research method is effective with the goal of applying the topic model to discover user comments. This study used data collected from e-commerce sites such as Agoda and Booking from 2012 to 2019, and can act as a representative to experiment with the model. The extracted keywords corresponded to each topic. The latent topic set was found to reflect the issues that customers are often interested in relating to hotel services. Establishing these topics is one of the best approaches to better understand customer experience. Thus, the proposed technique is more closely aligned with human intuition on topics and keywords. The results of this work can be applied to support decision-making systems for improving service quality as well as business development in the field of hotel services.

Future work will concentrate on proposing a model for collecting and classifying customer comments in real time, and the classification results will be forwarded to real-time analysis systems in which each comment will be recorded with the temporal factor. The user opinion analysis system will have the ability to analyze negative and positive changes. Customer issues will be addressed over time to help businesses quickly develop appropriate strategies, deal with crises in a timely manner, or identify and improve factors that enhance customer satisfaction.

## References

[1] C. Bucur, "Using opinion mining techniques in tourism," *Procedia Economics and Finance*, vol. 23, pp. 1666-1673, 2015.

[2] I. Putri and R. Kusumaningrum, "Latent Dirichlet Allocation (LDA) for sentiment analysis toward tourism review in Indonesia," *Journal of Physics: Conference Series*, vol. 801, article no. 012073, 2017. https://doi.org/10.1088/1742-6596/801/1/012073

[3] J. L. Boyd-Graber, Y. Hu, and D. Mimno, *Applications of Topic Models*. Hanover, MA: Now Foundations and Trends, 2017.

[4] D. T. Santosh, K. S. Babu, S. D. V. Prasad, and A. Vivekananda, "Opinion mining of online product reviews from traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet," *International Journal of Education and Management Engineering*, vol. 6, no. 6, pp. 34-44, 2016.

[5] M. Rossetti, F. Stella, and M. Zanker, "Analyzing user reviews in tourism with topic models," *Information Technology & Tourism*, vol. 16, no. 1, pp. 5-21, 2016.

[6] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, article no. 1608, 2016. https://doi.org/10.1186/s40064-016-3252-8

[7] M. Nguyen, T. Ho, and P. Do, "Social networks analysis based on topic modeling," in *Proceedings of the 2013 RIVF International Conference on Computing & Communication Technologies: Research, Innovation, and Vision for Future (RIVF),* Hanoi, Vietnam, 2013, pp. 119-122, 2013.

[8] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147-153, 2015.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[10] D. M. Blei, "Probabilistic topic models," *Communication of the ACM*, vol. 55, no. 4, pp. 77-87, 2012.

[11] E. Bjorkelund, T. H. Burnett, and K. Norvag, "A study of opinion mining and visualization of hotel reviews," in *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services (IIWAS)*, Bali, Indonesia, 2012, pp. 229-238.

[12] I. Sutherland, Y. Sim, S. K. Lee, J. Byun, and K. Kiatkawsin, "Topic modeling of online accommodation reviews via latent Dirichlet allocation," *Sustainability*, vol. 12, no. 5, article no. 1821, 2020. https://doi.org/10.3390/su12051821

[13] H. S. Le, J. H. Lee, and H. K. Lee, "Hotel services preferences across cultures: a case study of applying opinion mining on Vietnamese and American online reviews," in *Proceedings of the Korean Management Information Society Conference*, Busan, Korea, 2016, pp.137-149.

[14] H. H. Kim and H. Y. Rhee, "An ontology-based labeling of influential topics using topic network analysis," *Journal of Information Processing Systems*, vol. 15, no. 5, pp. 1096-1107, 2019.

[15] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on big data in marketing: a text mining and topic modeling based literature analysis," *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1-7, 2018.

[16] A. F. Hidayatullah and M. R. Maarif, "Road traffic topic modeling on Twitter using latent Dirichlet allocation," in *Proceedings of 2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, Malang, Indonesia, 2017, pp. 47-52.

[17] K. Park and L. Peng, "A development of LDA topic association systems based on Spark-Hadoop Framework," *Journal of Information Processing Systems*, vol. 14, no. 1, pp. 140-149, 2018.

[18] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169-15211, 2019.

[19] B. Ma, D. Zhang, Z. Yan, and T. Kim, "An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews," *Journal of Electronic Commerce Research*, vol. 14, no. 4, pp. 304-314, 2013.

[20] T. Ho and P. Do, "An integrated model for discovering, classifying and labeling topics based on topic modeling," *Science and Technology Development Journal*, vol. 17, no. 2, pp. 73-85, 2014.

[21] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR, 2011, pp. 1536-1545.

[22] C. Sievert and K. Shirley, "LDAvis: a method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Baltimore, MD, 2014, pp. 63-70.

**Van-Ho Nguyen** **https://orcid.org/0000-0001-6706-0276**

He received B.S. degree in Management Information System from Faculty of Information Systems, University of Economics and Law, VNU–HCM, Vietnam in 2015. Since September 2018, he has been with the School of Business Information Technology in University of Economics Ho Chi Minh City, Vietnam as a Master student. His current research interests include business intelligence, data analytics, and machine learning.

**Thanh Ho** **https://orcid.org/0000-0002-9033-3735**

He received M.S. degree in Computer Science from University of Information Technology, VNU–HCM, Vietnam in 2009 and Ph.D. degree in Computer Science from University of Information Technology, VNU-HCM, Vietnam in 2018. He is currently lecturer in Faculty of Information Systems, University of Economics and Law, VNU–HCM, Vietnam. His research interests are data mining, data analytics, business intelligence, social network analysis, machine learning and big data.