ISSN 2959-6157

Hotel review analysis based on LDA model and KeyBert model

Jizhao Zhang

School of Data Science (Computer Science),The Chinese University of Hong Kong,Shenzhen,Guangdong,518172,China
Email: kentlapa@outlook.com

Abstract:

The popularity of Internet applications and the continuous advancement of technology have made it easier for people to choose hotels. More and more people not only tend to book hotels through travel websites or software but also leave online reviews during their stay or after leaving the hotel. Almost all online bookers will carefully refer to reviews from previous customers of different hotels before making an order and then making a choice. In the past few years, many language models in the fields of machine learning and artificial intelligence have been widely used to analyze various texts. For reviews in areas such as hotels, most studies focus on sentiment analysis, that is, analyzing whether the sentiment in the reviews is positive or negative. Most previous studies used the LDA model to divide the topic words in the reviews or the KeyBert model to extract keywords from the text and compare and analyze them with other models. Some of these studies also used the hotel dataset compiled by Tan Songbo for analysis. Based on the theories of previous studies, this study uses the LDA model to extract and analyze different topic words of hotel reviews based on the hotel review data set compiled by Professor Tan Songbo to find out the advantages and disadvantages of the hotel. This study also uses the KeyBert model to extract the number of occurrences of specific keywords estimate the number of repeat customers by analyzing the keywords in the positive reviews, and estimate the number of non-target customer groups based on the keywords in the negative reviews. The research results are of great significance to the management and marketing decisions of the hotel, and can also provide better help to the hotel's booker.

Keywords: positive reviews, negative reviews, keywords, advantages, disadvantages, repeat customers ,non-target customer groups

1. Introduction

Due to the integration and development of Internet technology and the tourism industry, the online platform economy is extremely popular and attracts an astonishing number of customers. It can be said that most people have formed a strong dependence on booking hotels through Internet platforms, and a large number of customers are also accustomed to leaving comments on hotels on these online platforms. These comments include customers' views on hotel services, environment, facilities, and other conditions. To improve the service quality of the hotel and better meet customer needs, hotel management needs to analyze customer reviews.

Customer reviews of hotels can be roughly divided into positive reviews and negative reviews. In recent years, research on language models has been able to use language models to perform sentiment analysis on hotel reviews, determine whether the sentiment in the comments is positive or negative, or retrieve high-frequency keywords in the comments. In research on hotel marketing, it has been proposed to judge whether a customer is a repeat customer based on some phrases.

This study is different from previous studies. This study no longer focuses on analyzing whether the sentiment of hotel reviews is positive or negative. Instead, it uses the LDA model to cluster and analyze the high-frequency keywords in the positive and negative reviews based on the classification of positive and negative reviews to find out the highlights and disadvantages, which is beneficial for hotels to optimize the highlights and improve the disadvantages to improve service quality, and can also provide useful help for customers to choose hotels more accurately. Based on past research on hotel marketing, this study uses the KeyBert model to extract specific keywords. It roughly estimates the number of repeat customers and non-target customer groups based on the number of occurrences of particular keywords, to make the hotel's marketing work more accurate and efficient.

2. Literature review

There have been many studies on review analysis and sentiment analysis, and many language models have been used.

Cenni analyzed the negative reviews of hotels in his study, proposed the importance of analyzing negative reviews,and analyzed the communication methods of customers and the use of different languages.

Farkhood used models such as LDA in his study to analyze the emotions of users on products, services, and social events reflected in customer reviews on the Internet and obtained positive topic division results.

In the study of Giarelis,N.,Kanakaris,N.,& Karacapilidis,N.,KeyBert and RAKE,YAKE, TextRank, SingleRank, and other models were used to analyze keywords and compare and analyze several models.

In Kim,Y. J.,& Kim,H. S.'s 2022 study, sentiment analysis was conducted on hotel reviews posted by customers online, and it was concluded that customer emotions and evaluations depend on specific aspects such as catering and services.

In the study of Zhang,B.,Zhang,H.,Shang,J.,& Cai,J. ,multiple data sets such as Tan Songbo's hotel reviews were used to analyze multiple language models such as LSTM. And it is concluded that this kind of text dataset can enable the language model to perform better.

3. Research Questions

The goal of this study is to enable hotel management to optimize advantages and improve weaknesses based on customer reviews of the hotel, and to conduct marketing more accurately based on the number of repeat customers and non-target customer groups. This study is mainly divided into the following four parts: (1) How to find out the obvious advantages of a hotel through customer reviews; (2) How to find out the disadvantages that need to be improved most based on hotel customer reviews; (3) How to determine the number of repeat customers based on hotel customer reviews; (4) How to determine the number of non-target customer groups based on hotel customer reviews.

4. Methods and Models

4.1 . Research ideas

If the hotel management wants to understand the advantages and disadvantages of the hotel, it needs to analyze the customer comments on the hotel. Negative comments can seriously damage the business, while on the contrary, positive comments can greatly improve the prospects of the business. The Internet field has brought extensive changes to various industries and has also profoundly affected the hotel industry (Cenni,2024). Customer comments on hotels can be divided into positive reviews and negative reviews. In positive reviews, most customers will mention the advantages of the hotel, and the comments often contain a large number of positive subject words and emotional keywords. In negative reviews, most customers will mention the disadvantages of the hotel, and the comments often contain a large number of negative emotional keywords.

Some words or phrases that appear frequently in the text can often represent the theme and characteristics of the text, so the subject words that appear more frequently in positive reviews can be understood as the advantages of the hotel. The subject words that appear more frequently in negative reviews can be understood as the disadvantages of the hotel. The number of occurrences of some specific emotional keywords in positive reviews can reflect the number of repeat customers. The number of occurrences of some specific emotional keywords in negative reviews can reflect the number of non-target customer groups.

Based on the above analysis, from a business perspective, the research path of this article is mainly to extract information at the product level and customer relationship level from customer reviews. At the product level, this study uses the keywords in the positive reviews as advantages to continue to improve the service and extracts the keywords in the negative reviews as disadvantages to improve the service. At the customer relationship level, this study extracts specific emotional keywords. Depending on the customer's situation, the customer experience will be remembered positively or negatively, which will lead to the sustainability of customer satisfaction, and customers who make positive comments on the experience can be said to be satisfied (Kim,2022). Customers who are satisfied with the experience are often likely to become repeat customers. Therefore, specific emotional keywords such as“ 还会 ”, “ 第二次 ”, “ 又 ”, “ 再 ”(“still will”,“second time”,“again”, “once again”). should be extracted from the positive reviews, and the sum of the number of occurrences of these words should be determined to determine the number of potential repeat customers. Similarly,“ 很差 ”, “ 极差 ”, “ 糟糕 ”, “ 再也不会 ”(“very bad”, “extremely bad”, “terrible”, “never again”). should be extracted from the negative reviews, and the sum of the number of occurrences of these words should be determined to determine the number of non-target customer groups.

4.2 . Research Path

Extracted from customer reviews

① : Product level

Positive review keywords → Advantages → Continue to

improve
Negative review keywords → Disadvantages → Improve service
② : Customer relationship level
Emotional keywords “ 还 会 ”, “ 第 二 次 ”, “ 又 ”, “ 再 ”(“still will”, “second time”,“ again”,“once again”)→ Potential repeat customers
Emotional keywords “ 很 差 ”, “ 极 差 ”, “ 糟 糕 ”, “ 再 也 不 会 ”(“very bad”, “extremely bad”,“terrible”,“never again”)→ Non-target customer groups

4.3 . Specific Implementation

4.3.1 . Dataset Selection

The first step of the study is to find a data set. Zhang,B.,Zhang,H.,Shang,J.,& Cai,J. et al. found that the current Tan Songbo hotel data set shows that the extraction model can achieve the best results (zhang,2022),so this study selected and downloaded the hotel review corpus compiled by Tan Songbo. The comments in the corpus are all in Chinese and have been divided into positive and negative comments. It contains 5358 positive comments and 1172 negative comments, which are stored in two excel files respectively. This dataset is a large-scale hotel review corpus collected and compiled by Tan Songbo. It has been divided according to id, text, label, etc. in the excel table, and can also be read directly through pandas. (Data source: https://www.aitechclub.com/data-detail?data_id=29) (Download source: https://aistudio.baidu.com/datasetdetail/106431),so the conclusions drawn based on this data are relative)

4.3.2 . Product level implementation

According to the research ideas and research paths, the implementation of the product level is to find out the advantages and disadvantages. The principles of the methods used to find out the advantages and disadvantages in this study are the same. To improve the quality, we used the LDA (latent Dirichlet allocation) model, which performs very well when mixing distributed documents from multiple topics and within them (Farkhod,2001).
First, the positive and negative review texts were preprocessed by jieba’s part-of-speech tagging tool and regular expressions, removing some modal particles and some meaningless part-of-speech combinations.
After preprocessing, finding out the advantages is roughly divided into the following four steps. First, the text of the positive reviews is clustered by the LDA model. Then, the definition of the phrase search length in the code is two two-character or three-character word phrases. Then, according to the code running results, the 15 most frequent topic phrases in the positive reviews are obtained. Finally,2-5 of the most frequent phrases are selected in the running results as the advantages of the hotel.
A similar method is used to find out the disadvantages of the hotel. First, the text of the negative reviews is clustered by the LDA model. Then, the definition of phrase search length in the code is two two-character or three-character keyword phrases. According to the code running results, the 15 most frequently appearing keyword phrases in the negative reviews are obtained. Finally,2-5 of the most frequently appearing phrases are selected from the running results as hotel disadvantages.

4.3.3 customer relationship implementation

What needs to be achieved at the customer relationship level is to estimate the number of repeat customers and non-target customer groups by finding and calculating the total number of occurrences of specific emotional keywords. The extraction of keywords can be achieved through the keyword extraction enhancement model (KeyBert model).
First, the KeyBert model is used to find the number of occurrences of comments containing specific emotional keywords such as “ 还会 ”, “ 第二次 ”, “ 又 ”, “ 再 ”(“still will”,“second time”,“ again”, “once again”) in positive reviews and these numbers are summed up to estimate the number of repeat customers of the hotel. Similarly, the KeyBert model is used to find the number of occurrences of specific emotional keywords such as“ 很差 ”, “ 极差 ”, “ 糟糕 ”, “ 再也不会 ”(“very bad”, “extremely bad”, “terrible”, “never again”) in negative reviews and these numbers are summed up to estimate the number of non-target customer groups.

5. Model Overview

5.1 . LDA Model

The LDA model is a topic model for text data that can depict the topic distribution in a document collection and then discover the semantic relationship hidden behind the text. The LDA model is widely used in text mining, information retrieval, recommendation systems, social network analysis, and other fields. The basic assumption of the LDA model is that each document is generated by a mixture of multiple topics, and each topic is composed of a mixture of multiple words. By continuously iterating and optimizing parameters, the LDA model can infer the hidden topic structure in each document based on the word distribution of the document, thereby revealing the semantic relationship behind the text.

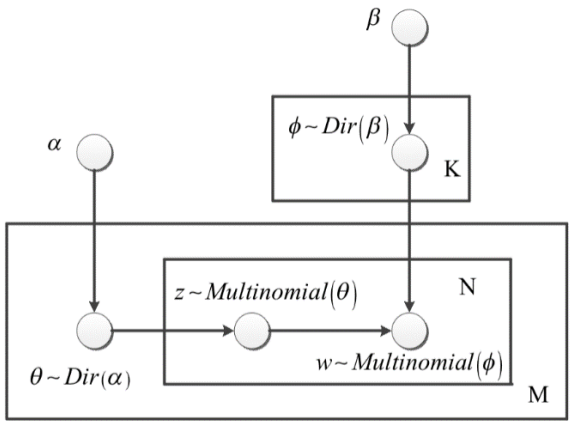

Figure 1 Schematic diagram of LDA model

Figure 1 is a rough schematic diagram of the specific principle of the LDA model.

In this study, LDA model clustering will be used to find high-frequency keywords in positive and negative reviews.

5.2 . KeyBert Model

KeyBERT relies on a BERT-based pre-trained word embedding model to enhance the quality of extracted key phrases (Giarelis,2021). KeyBert is a minimal and easy-to-use keyword extraction technique that leverages BERT embeddings to create keywords and key phrases that are most similar to a document.

First, document embeddings are extracted using BERT to get a document-level representation. Then a word embedding model is used to extract N-gram words/phrases. Finally, cosine similarity is used to find the words/phrases that are most similar to the document. The most similar words can then be identified as the words that best describe the entire document. KeyBert is by no means unique, it is a quick and easy way to create keywords and key phrases.

In this study, the KeyBert model was used to find the number of occurrences of specific keywords in good and bad reviews. Extract and calculate the number of comments containing words such as "还会", "第二次", "又", " 再 "("still will", "second time"," again", "once again") from the positive reviews, and extract and calculate the number of comments containing words such as " 很差 ", " 极差 ", " 糟糕 ", " 再也不会 "("very bad", "extremely bad", "terrible", "never again") from the negative reviews. The final result section will add up these times to determine the number of repeated customers and non-target customer groups. The result section will show the situation before and after the summation using the KeyBert model.

6. Results

6.1 . Code running results of LDA model to determine advantages and disadvantages

6.1.1 . Code running results of LDA model to determine advantages

Figure 2 Keywords in positive reviews

As shown in Figure 2,the result is the first 15 most common phrases of different subject words in the good reviews calculated by using jieba's part-of-speech tagging tool and regular expressions to exclude most meaningless phrases, and then applying LDA model clustering. The left side of "/" represents the number of occurrences, and the right side of "/" represents the total number of reviews. The specific code for finding the advantages is detailed in the appendix.

Based on the above results,4 (2-5 can be selected) most representative subject word phrases are selected as advantages. The following table shows the results after screening.

Table 1. Screened advantages

Advantages	Number of occurrences/total number of comments
交通方便 (Convenient transportation)	294/5358
房间干净 (Clean room)	261/5358
环境不错 (Nice environment)	258/535
早餐不错 (Good breakfast)	123/5358

As shown in Table 1,the above four items have been screened out to obtain different results that best represent the hotel's advantages. The main advantages of the hotel listed in the table above include " 交通方便 (convenient transportation)"," 房间干净 (clean rooms)"," 环境不错 (nice environment)"," 早餐不错 (good breakfast)" and other advantages. The hotel can continue to optimize these aspects and use them as the hotel's characteristics to provide customers with better services.

6.1.2 . Code running results of LDA model to determine disadvantages

Figure 3 Keywords in negative reviews

As shown in Figure 3,the result is the first 15 most common keywords in negative reviews calculated by using jieba's part-of-speech tagging tool and regular expressions to exclude most meaningless phrases, and then applying LDA model clustering. The left side of "/" represents the number of occurrences, and the right side of "/" represents the total number of reviews. For specific codes, please see the appendix.

Based on the above results,2-5 most representative keyword phrases can be selected as disadvantages. The following table shows the results after screening.

Table 2 Screened disadvantages

Disadvantages	Occurrence/Total number of reviews
房间很小 (The room is small)	25/1172
服务一般 (Service is average)	25/1172
设施简陋 (Facilities are simple)	21/1172

As shown in Table 2,the above three items are the most representative results of the hotel's disadvantages. The disadvantages of the hotel listed in the table above are mainly " 房间很小 (The room is small)", " 服务一般 (Service is average)", and " 设施简陋 (Facilities are simple)". The hotel management should consider improving these disadvantages.

6.2 . KeyBert operation results

6.2.1 . Results of Identifying the Repeated Customer Group

This study uses the KeyBert model to process specific sentiment keywords in positive and negative reviews.

In the positive reviews, words such as "还会", "第二次", " 又 ", " 再 "("still will", "second time"," again", "once again") often represent that the customer has been to the hotel more than once or will come back again. These customers can be identified as the hotel's repeat customers. Therefore, when determining the number of repeat customers, keyBert should be used to extract and calculate the number of comments containing words such as " 还会", "第二次", " 又", "再"("still will", "second time", " again", " once again"), and then add up these occurrences to estimate the number of repeat customers. For the specific code for estimating the repeat customer group, please refer to the appendix.

```
# List of specified words to search for
specified_words = ['还会', '第二次', '又', '再']
```

Figure 4 Specific emotional keywords in

positive reviews

Figure 4 shows the sentiment keywords in the positive reviews defined in the code.

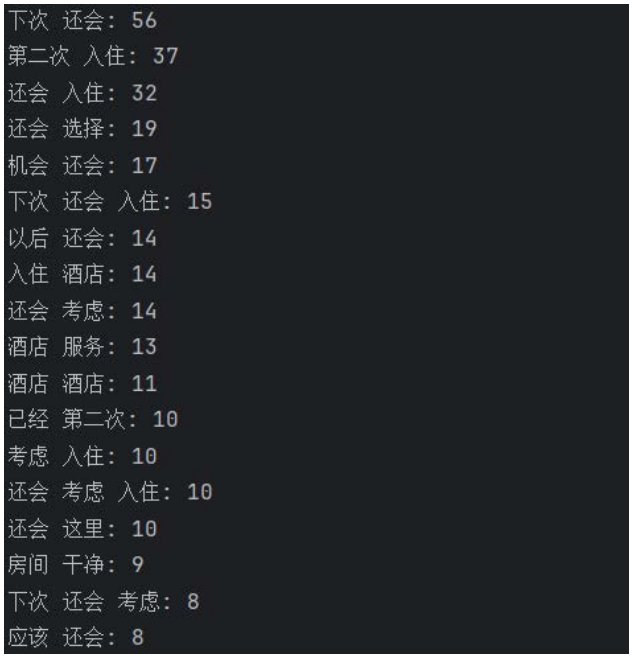

Figure 5 The KeyBert model finds the number of occurrences of specific sentiment keywords in positive reviews

Since there are many types of comments containing specific sentiment keywords, Figure 5 only shows the number of occurrences of some comments containing words such as“第二次”, “还会”, “又”, “再次”(“second time”, “still will”, “ again”, “once again”) found by the KeyBert model. The numbers in the above figure represent the number of occurrences of the comments.

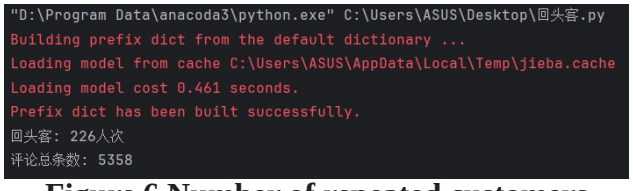

Figure 6 Number of repeated customers

Figure 6 shows the result of adding the above data. As shown in the figure, there are 226 repeat customers(回头客).

6.2.2 Results of Identifying Non-Target Customer Groups

In negative reviews,words such as“ 很差 ”, “ 极差 ”, “ 糟糕 ”, “ 再也不会 ”(“very bad” , “extremely bad”, “terrible” , “never again”)often represent that the customer has been to the hotel and has a very bad impression of the hotel or has expressed an attitude that he will never come to the hotel again. These customers can be determined as the non-target customer group of the hotel. Therefore,when determining the number of non-target customer groups,keyBert should be used to extract and calculate the number of comments containing words such as“ 很差 ”, “ 极差 ”, “ 糟糕 ”, “ 再也不会 ”(“very bad”,“extremely bad” , “terrible”,“never again”)from the negative reviews,and then add up these occurrences to roughly estimate the number of non-target customer groups. The code for estimating the number of non-target customer groups is detailed in the appendix.

```
# List of specified words to search for
specified_words = ['很差', '极差', '糟糕', '再也不会']
```

Figure 7 Specific emotional keywords in negative reviews

Figure 7 shows the sentiment keywords in the negative reviews defined in the code.

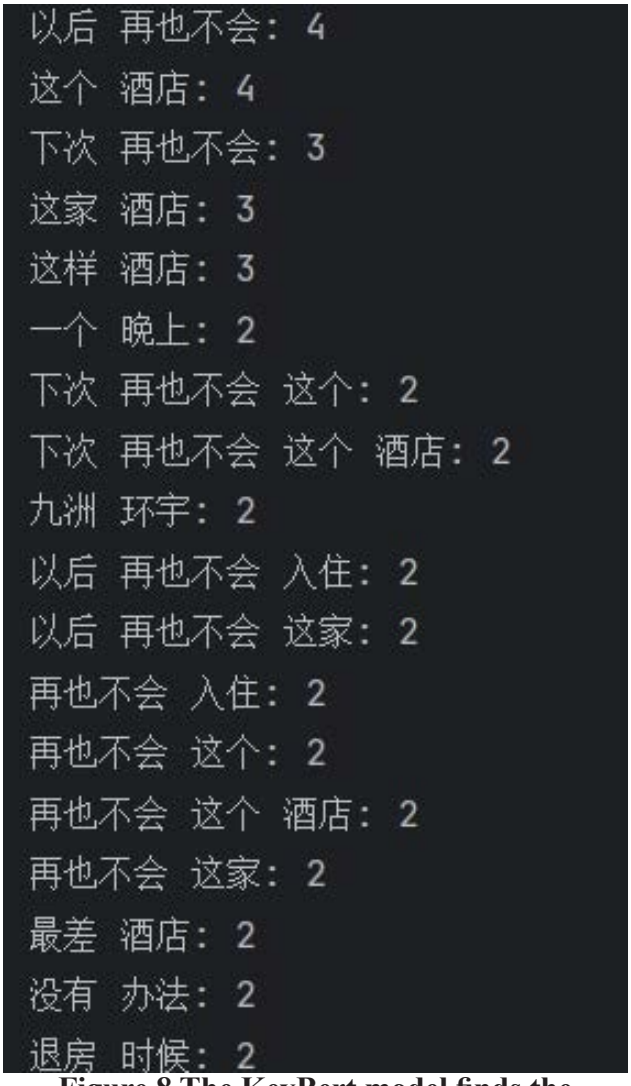

Figure 8 The KeyBert model finds the number of occurrences of specific sentiment keywords in negative reviews

Since there are many types of comments containing specific sentiment keywords,Figure 8 only shows the number of occurrences of comments containing words such as “很差”, “极差”, “糟糕”, “再也不会”(“very bad” , “extremely bad” , “terrible” , “never again”)found by the KeyBert model. The numbers in the figure above represent the number of occurrences of the comments.

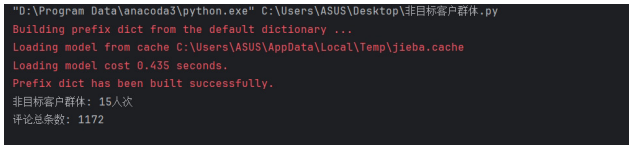

Figure 9 Number of non-target customer groups

Figure 9 shows the result of adding the above data. As shown in the figure, there are 15 non-target customers(非目标客户群体).

7. Accuracy and error analysis

The accuracy of the results of this study needs to be improved. There may be some errors in the running results of some codes. In subsequent research, the accuracy needs to be improved and the code ideas need to be improved.

After preprocessing with regular expressions and jieba word segmentation tools, some meaningless or repeated topic phrases can still be seen in the running results of the LDA clustering model. This situation may be caused by different users’ different ways of expressing the same advantages or disadvantages.

In the KeyBert model part, this study only selected 5 specific emotional keywords that are most likely to represent the repeated customer group and the non-target customer group based on the good and bad reviews and estimated the number of repeated customer groups and the number of non-target customer groups. In addition to the several specific emotional keywords mentioned in this article, there may be other phrases or topic words from the comments of the repeated customer group and the non-target customer group. Therefore, estimating the number of repeated customer groups and non-target customer groups by the number of occurrences of specific emotional keywords may have errors and omissions.

8. Conclusion

In this study, compared with the traditional hotel review sentiment analysis, this study did not simply study whether the sentiment of hotel reviews is positive or negative like most studies. This study used the already sorted customer evaluation data set for hotels, used the LDA model to analyze the good reviews and find out the advantages of the hotel, and also used the LDA model to analyze the bad reviews and find out the disadvantages of the hotel. In addition, this study used the KeyBert model to calculate the number of occurrences of specific keywords in the good reviews to estimate the number of repeat customers and also used the KeyBert model to calculate the number of occurrences of specific emotional keywords in the bad reviews to estimate the number of non-target customer groups.

For hotels, the results of this study can allow the hotel management to continue to make the advantages more attractive to attract more customers and improve and upgrade the shortcomings of the hotel to meet the needs of customers. At the same time, understand the number of repeat customers and non-target customers of the hotel, and conduct more accurate marketing.

For customers, the results of this study can allow customers to see the advantages, disadvantages, and the number of repeat customers and non-target customers of different hotels at a glance, so as to choose a hotel that is more preferred and more able to meet their needs according to their needs and can also improve the stay experience to a certain extent.

This study has not yet been perfect in terms of the accuracy of the results. There may be some omissions in finding out the strengths and weaknesses of the hotel, estimating the number of repeat customers, and the number of non-target customer groups. However, this can be improved by improving the model and code ideas.

References

[1] Cenni,I. (2024). Sharing travel experiences on TripAdvisor: A genre analysis of negative hotel reviews written in French,Spanish and Italian. Journal of Pragmatics,221,76-88.

[2] Farkhod,A.,Abdusalomov,A.,Makhmudov,F.,& Cho,Y. I. (2021). LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model. Applied Sciences,11(23),11091.

[3] Giarelis,N.,Kanakaris,N.,& Karacapilidis,N. (2021,June). A comparative assessment of state-of-the-art methods for multilingual unsupervised keyphrase extraction. In IFIP International conference on artificial intelligence applications and innovations (pp. 635-645). Cham: Springer International Publishing.

[4] Kim,Y. J.,& Kim,H. S. (2022). The impact of hotel customer experience on customer satisfaction through online reviews. Sust ainability,14(2),848.

[5] Zhang,B.,Zhang,H.,Shang,J.,& Cai,J. (2022). An Augmented Neural Network for Sentiment Analysis Using Grammar. Frontiers in Neurorobotics,16,897402.