Received: December 17, 2021. Revised: April 2, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Sentiment Analysis User Regarding Hotel Reviews by Aspect Based Using Latent Dirichlet Allocation, Semantic Similarity, and Support Vector Machine Method

Moch Deny Pratama[1] Riyanarto Sarno[1]* Rachmad Abdullah[1]

[1]*Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology*
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
* Corresponding author's Email: riyanarto@if.its.ac.id

Abstract: The Hotel website always provide sentiment reviews of the service, then users can easily choose the desired facility according to the best advice results from the previous user. In this study, the sentiment of the reviews are determined based on five hotel aspects, those are: food, service, location, comfort and cleanliness. Every data hotel reviews is pre-processed to produce a term list. Extract related hidden topics using Latent Dirichlet Allocation (LDA). The extracted documents are categorized by matching for similarity to determine the term documents into five hotel aspects. Each term document is expanded using synonyms to increase the similarity value to LDA with 100% expanded document using Cosine Similarity produces the highest performance value of 0.856. Labeling the sentiment of each review based on the aspect and sentiment classification using the Support Vector Machine method gets an average value of 0.940 for all fifth aspects. The ranking of the most important aspects shows that users talk a lot about the cleanliness aspect, having the highest positive sentiment value of 39.450 and having a negative sentiment of 3.861, indicating that each review sentiment is influenced by certain aspects.

Keywords: Sentiment analysis, Hotel reviews, Latent dirichlet allocation, Term, Hidden topic, Aspect based, Cosine similarity, Support vector machine, Most important aspect ranking.

1. Introduction

In the current era of the industrial revolution, we can use technological sophistication in various ways, such as for businesses like an e-commerce business [1]. One of the e-commerce businesses that utilize information technology is an online travel agent website that allows ordering needs for traveling or traveling such as train tickets, airline tickets, tourist attractions tickets, hotel reservations, ordering food at restaurants, and other similar facilities [2]. Travel websites always provide features to provide reviews of these services so that users can easily choose the facilities they want with the best suggestions based on the results of reviews from previous users. The results of a review of reviews from these facilities can also be used as supporting data for these service providers to assess reactions and feedback experiences from consumers on the services provided. Sentiment Analysis can be called as a computational technique for automating the extracted of subject information such as customer opinions with respect to a product [3]. Based on the results of the review, they can also evaluate their performance.

Aspect-Based Sentiment Analysis (ABSA) is a type of sentiment analysis that can see all sentiments in each predetermined aspect [4]. ABSA is intended to better understand reviews by aspect category than traditional sentiment analysis. In particular, it can infer data on the sentiment polarity of the aspect category or target entity in the text [5]. Extracting information from text, commonly known as text mining, text or document categorization, sentiment analysis, search engine searches for more precise themes, and spam screening are some examples. [6]. As a result of customer reviews of the service getting, positive reviews increase the popularity and rating of the service. Results of negative reviews can

serve as a self-evaluation mechanism to improve services.

One of the things that is closely related to hotel reservations is hotel reviews from several previous users. Reviews posted on the web by every user who has had a hotel stay experience become a powerful information sharing tool for digital customer interaction. User reviews and opinions have always been valuable information that may considerably impact people's judgments, with over 95% of people reading user-generated hotel reviews online before making a booking decision since they are very crucial in making hotel decisions [7]. What's more, about 86% see reviews as a fundamental asset when deciding between hotels and it shows that around 66% of respondents worldwide trust review posts on the web [8]. Hotel reviews are useful to find what needs to be service improved [9]. From the results of reviews by users, service managers can conduct reviews that can allow assessing consumer reactions to the services provided so that they can evaluate the services so that they can understand what they need to improve for future experiences.

In this study, using Latent Dirichlet Allocation (LDA) extraction of term topic modeling to obtain hidden topics that are interrelated with frequency weights and inter-topic interrelationships. Applying to hotel reviews that can extract topics according to the desired number of topics followed by category aspect classification of five hotel aspects. The five aspects of the hotel are food, service, location, comfort and cleanliness. LDA is an unsupervised learning method that is most widely used in topic modeling to help determine the topic of a text that is able to determine the hidden topic of a document [10]. Using the LDA unsupervised learning approach to extract sentiment aspect pairs from hotel reviews in order to create a sentiment analysis system at the aspect level. [11]. Combining the results of topic extraction by calculating the similarity in each word obtained compared to the term list in each specified aspect [12]. Calculation of similarity between terms using Cosine Similarity on the average value of the Glove vector and expand term using Wordnet Synset Synonyms [13].

First Pre-processing, where the results will be continued in topic modeling. The results of the topic term extraction are then matched for similarity values. Then the results of the extraction are carried out by synonym expansion so that the more words that will be calculated similarity, the value of similarity will increase, and the results are better. The results of similarity are categorized into the five aspects that are related to each extracted topic. The aspect categorization with three kinds of categorization process to increase the document similarity value. In aspect Categorization 1 (AC1), match the extraction terms from Latent Dirichlet Allocation (LDA) + semantic similarity using cosine similarity based on the data list terms in Table 2. Aspect Categorization 2 (AC2), match the extraction terms from (LDA) + semantic similarity using cosine similarity + 20% synonym expansion. and In Aspect Categorization 3 (AC3) perform a match on the extraction term resulting from (LDA) + semantic similarity using cosine similarity + 100% synonym expansion.

After performing the aspect category stage, the next stage is the classification sentiment category stage. In this stage, a term list for each aspect that has been categorized has been collected based on the relevant aspect, the sentiment is determined on all data related to the term list on five aspects. The labeling process is carried out on each data regarding positive and negative sentiments in all aspects. Then will be classified sentiment on the aspect based on the SVM method is carried out for the classification [14]. The performance of the classification model created with the Support Vector Machine approach. After that, the ranking will be carried out on each existing aspect. Determine which aspects are currently being discussed by customers based on the results of the reviews. In this stage, using the ranking method to sort the reviews with the largest number to the smallest number of reviews based on aspects that can show the importance of the level of customer reviews in the relevant aspect.

2. Related theory

It contains several explanations of theories relevant to the research carried out as supporting theories.

2.1 Pre-processing

Pre-processing is the first step in normalizing data so that all data has the same form and in the form of fractions word (token). The techniques in pre-processing as follow : [15]

2.2 Term list hotel aspect

Term list on each aspect of the term that has been determined based on several previous studies, [12, 16, 17]. After determining the aspect to be made into the 5 Aspect Term, categorize each of the reviews into predetermined aspect based on the results of topic extraction. The results of some terms

Table 1. Illustration of pre-processing process

Pre-processing	Detail Process
Case Folding	Convert random word structure uppercase or capital into lowercase.
Tokenization	Sentence reviews are broken down from sentences into the arrangement of each word in the form of arrays.
Stop Word Removal	Remove unnecessary words, such as punctuation, to be, conjunctions and the process of removing less important words that often appear on documents it can eliminate stop words.
Stemming / Lemmatization	Change the words that are broken from the tokenizing results into basic words and converting into a word or root word for each word.

Table 2. The keyword of aspect term

Aspect Term	Keyword Term
Food	delicious, food, dish, wine, salad, cafe, drink, spicy, meal, restaurant, breakfast dinner, lunch, brunch, coffee, item, cup, menu, bagel, tea, buffet, bar, waffle.
Service	polite, helpful, friendly, reliable, pool, facility, reliable, fast, gym, desk, quick, parking, conference room, fee, convenient, good, staff, internet, wifi.
Location	convenient, train, place, mall, metro airport, far, close, distance, location, view, station, railway.
Comfort	activity, bedroom, comfort, feel, sleep, meeting, charge, connection.
Cleanliness	furniture, housekeeping, toilet, wall, cobweb, carpet, laundry, smoke, ventilation, smell, cleanliness.

that most appear on hotel review for each aspect are determined based on previous research : [12]

2.3 Latent dirichlet allocation (LDA)

David Blei, Andrew Ng, and Michael I. Jordan first introduced LDA as a topic discovery graphical model in 2003, and it is now widely used in topic modeling. [18]. To discover the underlying structure of a document collection, the topic model is employed as an aspect extraction tool. Other topic modeling approaches, such as Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA), have flaws that LDA may overcome. [19]. LDA is a text mining technique for identifying representative topics in documents. To explain the frequency or appearance of terms together in each text, LDA discovers subjects shared by documents in the corpus. The probability distributions across all potential words are used to describe topics. The LDA findings describe each document with a set of topic probabilities in this fashion. Each result of the topic document content with the highest probability of indicating the document content is given a value [20]. The primary premise of LDA is that documents contain a variety of topics. The bag-of-words topics are defined formally as a distribution over a specified vocabulary. The generative method is provided to summarize how the topics are inferred and to produce an LDA model from the documents generated, as follow : [18]

1. Determine the most likely words for each topic.
2. For each document:

a. Determine the proportions of topics that should be included in the document,

b. For each word:

i. Choose topic,

ii. Choose a likely term based on the topic (Step 1).

LDA is defined formula as : [12]

$$p(w,z \,/\, \alpha,\beta) \;=\; p(w \,/\, \alpha,\beta)\, p(z \,/\, \alpha) \quad (1)$$

These model parameters are α and β, the word is w, topic is z. $p(z \,/\alpha)$ is the probability that topic z has founded.

2.4 Expanding term list with wordnet synonyms

WordNet is a lexicon or lexical database created by Princeton as part of the NLTK corpus in Python. It was using Synonyms in wordnet to expand each term in the hotel aspect. This research was conducted using WordNet to find synonyms of words from the LDA document extraction. The expansion process is done by looking for words that have the same meaning in each extracted term. Identify synonyms for each word to expand the query using WordNet. This is done to increase the value of the semantic similarity in the category aspect[21].

2.5 Semantic similarity

Semantic similarity is a measure for comparing the semantic similarity of words and phrases that establish values based on semantic relationships. Semantic similarity metrics are used to compare words and terms in natural language texts and are calculated to compute the similarity between concepts [22]. The semantic relation with each term

in the LDA extracted term and the keyword phrase, as well as the expansion results. Applying Cosine Similarity to estimate the score of similarity between words (w_i, wj) based on the calculation of the number of similar words, where the Cosine similarity is calculated using the word vector.[23] The equation of Semantic Similarity formula is described as follow : [16]

$$Similarity = \frac{\sum_{m-1}^{k} {w_i}^m wj^m}{\sqrt{\sum_{m-1}^{k} ({w_i}^m)^2 \sqrt{\sum_{m-1}^{k} wj^m)^2}}} \quad (2)$$

Cosine Similarity determines the degree of similarity between sentences 1 (S1) and 2 (S2) by counting the number of similar terms in both. Where word vectors are used to measure the Cosine Similarity formula equation as follows : [23]

$$Cosine(S1, S2) = \frac{S1 * S2}{||S1|| * ||S2||} = \frac{\sum_{i=1}^{k} S1i\ S2i}{\sqrt{\sum_{i=1}^{k} S1i^2 \sqrt{\sum_{i=1}^{k} S2i^2}}} \quad (3)$$

2.6 Support vector machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm in machine learning that separates class data by finding the most optimal hyperplane. Classification with the SVM method is carried out in the python programming language using the sklearn.SVM library for the data classification process using the SVM method, in addition to using the sklearn.metrics library to measure the performance of the classification model or determine the accuracy of learning values with SVM. [14] The general concept of SVM is to discover a hyperplane in the sample space that has the greatest margin on the training set with the following as : [24]

$$hyperplane = \omega Tx + b = 0 \quad (4)$$

where ω = (ω1; ω2; ⋯; ωd) is the normal vector that determines the hyperplane's direction. The distance between the hyperplane and the coordinate origin is determined by the displacement term b.

2.7 Performa evaluation

Perform the performance evaluation using Confusion Matrix and Classification Report. Using the formula Accuracy, Precision, Recall, and F1-Score. In the Confusion Matrix, there are four types of outcomes from the model such as: The number of documents accurately detected or predicted on the relevant documents is referred to as TP (True positive). The amount of documents accurately identified on irrelevant documents are referred to as TN (True Negative). The number of faculty papers in the relevant documents is referred to as FP (False Positive). The quantity of wrong documents on irrelevant papers is referred to as FN (False Negative). The classification report shows performance evaluation metrics in the machine learning model. It is used to show the precision, recall, F1 score, and support of trained classification model. Some formulas for calculating performance using in the model as follows :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1 - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (8)$$

3. Research method

The data which are obtained in the form of text reviews from customers. Next, do the pre-processing process, which removes and cleans the noise data as shown in Table 1. The results of the pre-processing data are tokens, then this result will be trained using the LDA method to get the results of document extraction obtained. After that, the term results obtained from the LDA extraction are matched one by one to the keywords in Table 2. using the Cosine Similarity. The similarity is calculated in each term per existing aspect to determine customer satisfaction with hotel facilities based on aspects. The process of carrying out the categorization aspect has three kinds of stage Aspect Categorization 1 (AC1), then Aspect Categorization 2 (AC2) and Aspect Categorization 3 (AC3). Where AC 1 calculates the similarity value in the data purely without doing synonym expansion. While AC 2 uses LDA + 20% expansion and AC 3 uses LDA + 100% expansion to increase the similarity value in order to determine the best performance.

3.1 Data collection

Data collection describes the source of the data obtained. In the form of reviews from various customers on hotels with datasets + crawling data review with webharvy on the tripadvisor.com

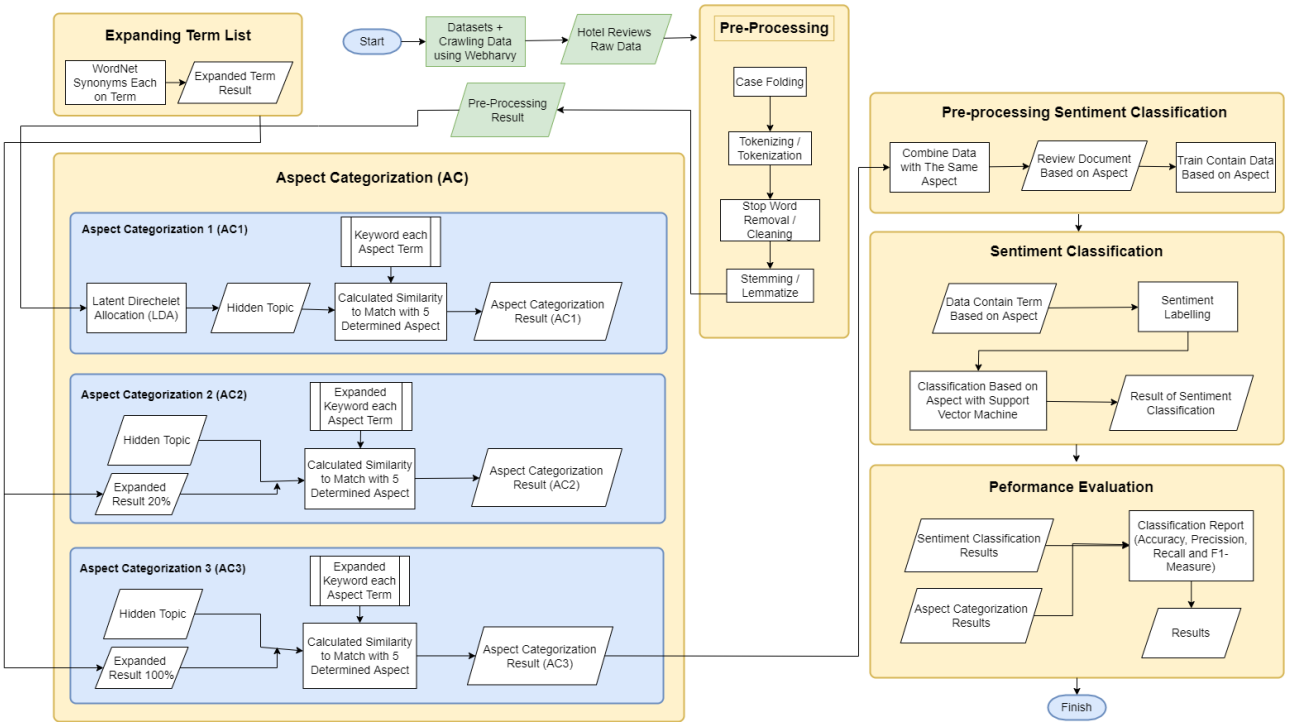

Figure. 1 Research method

Table 4. Example data collection

No.	Review
1.	…4-6 **times** years **stayed** hotel, time period travel extensively domestically internationally favorite hotel, comfortable **beds** efficient workspaces, **breakfasts** Lola just outstanding assaggio terrific Italian place dinner **street** excellent…
2.	…great place, **nice service**, took part union square walked easy, **location** great walking union **square**, stayed **night** beat central valley heat, certainly return…
3.	…water bottles **room** complimentary, helpful staff restaurants transportation, told **staff** worked 20 years, **small** room **hotel** totally re-furbished 04 great bed linens small **clean** really need…
4.	…great **place** decent rate arrived check room ready, **rooms** small **comfortable**, **good** location **union** square 4 blocks bart…
5.	…room perfectly small **bathroom** great, **great** thanksgiving buffet dinner street sazerac **restaurant** located hotel, delicious, lovely way spend holiday certain, **like** energy ambience old hotels adore hot baths excellent customer service **stay** executive hotel pacific downtown Seattle…

website[25]. Contains the review sentence given by the customer on the results of the review.

3.2 Pre-processing

Perform pre-processing use NLTK. Performing the steps in doing text pre-processing as shown in Table 2. which produces as follow :

Table 5. Result of pre-processing

Document Review	Pre-processing Result
4-6 times years stayed hotel, time period travel extensively domestically internationally favorite hotel, comfortable beds efficient workspaces	'time','year','stay', 'hotel','time','period', 'travel','extensive','domestic','international', 'favorite','hotel', 'comfort', 'bed','efficient','workspace'
great place, nice service, took bart union square walked easy, location great walking union square	'great','place','nice', 'service','take','bart', 'union','square','walk', 'easy','location','great','walk', 'union','square'

Table 6. Example of expand term synonyms

Term	Expand Result
connection	'connection','connexion', 'joining' 'connectedness','link','connector', 'connecter','connective','association',
sleep	'sleep', 'slumber', 'sopor', 'nap', 'rest', 'quietus', 'kip', 'slumber'
meeting	'meeting','encounter','merging','contact' 'confluence','meet','see','meet', 'converge', 'fill', 'fulfill', 'fit', 'match', 'gather', 'assemble', 'forgather', 'receive', 'suffer', 'touch', 'adjoin',

3.3 Expanding term list with wordnet sinonyms

Expansion of the document on the terms contained in the keywords in each of the five aspects in Table 2. The terms from LDA then be followed by a similarity matching process. The results of the calculation of the similarity in each aspect will be

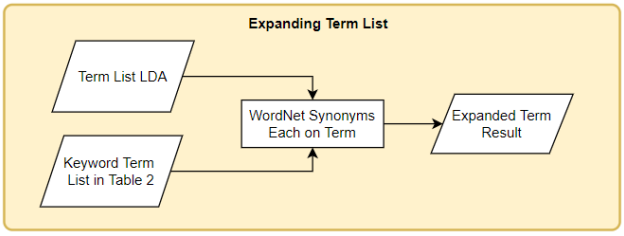

Figure. 2 Expanding term

used to determine which aspects will be included in the terms of the extracted document.

An explanation of expanding documents shown in Fig. 2 as follows :

3.4 Aspect categorization

In this section, we conduct 3 kinds approach to determine which is the best aspect categorization performance. The aspect categorization process is carried out by calculating using the Semantic

Similarity method to categorize into 5 hotel aspects that have been determined in Table 2. Aspect Categorization 1 (AC1) uses the Latent Dirichlet Allocation (LDA) method to extract hidden topic data in documents. The results of the LDA in the form of hidden topic data are calculated for similarity to categorize aspects. While AC2 = AC1 + Document Expand Synonym 20% and AC3 = AC1 + Document Expand Synonym 100%

3.4.1. Aspect Categorization 1 (AC1)

Aspect Categorization 1 (AC1), it has the same process as AC2 and AC3. But only using the semantic similarity between the LDA data and the data in Table 2, without document expansion.

3.4.2. Aspect categorization 2 (AC2)

Aspect Categorization 2 (AC2), it has the same process as AC1. Not only using the semantic similarity between the LDA data and the data in Table 2, but also adding a 20% document expansion.

3.4.3. Aspect categorization 3 (AC3)

Aspect Categorization 3 (AC3) is using the Latent Dirichlet Allocation (LDA) method is used to extract hidden topic documents and then calculate the LDA extraction results with expanded each term 100%. with semantic similarity use the data in Table 2. to perform aspect categories for each review. Using Cosine Similarly to calculate similarity and perform categorization.

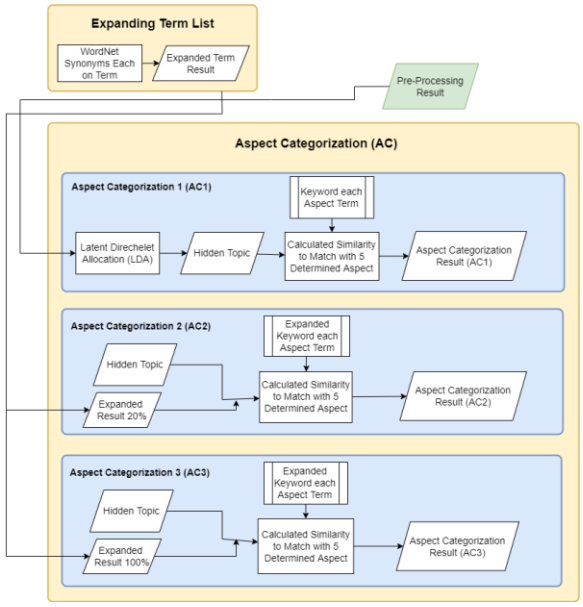

Figure. 3 Aspect categorization

3.4.4. Latent dirichlet allocation (LDA)

Results obtained from the LDA process are hidden topics and the frequency value of the possibility of words appearing from the document. Calculated the occurrence of each word in the document using the bag-of-word LDA corpus. Then build a train LDA model that calculates the possible proximity of related topics. The LDA produces hidden topics, which are key words or phrases that appear frequently in documents. The hidden topic extraction results from each term list document using this approach are as follows :

The results of the LDA are hidden topics on documents, results and terms in Table 2. in each aspect as a whole are calculated using Semantic Similarity + expanded term 100%. The results of the Semantic Similarity match to categorize each review

Table 7. Results of latent dirichlet allocation

Term Topic	Hidden Topic
Term List Topic 1	(0, '0.022*"stayed" + 0.020*"street" + 0.019*"breakfast" + 0.019*"bed" + 0.017*"time"')
Term List Topic 2	(1, '0.035*"location" + 0.034*"nice" + 0.032*"service" + 0.026*"night" + 0.021*"square"')
Term List Topic 3	(2, '0.165*"hotel" + 0.121*"room" + 0.041*"staff" + 0.029*"clean" + 0.028*"small"')
Term List Topic 4	(3, '0.038*"rooms" + 0.033*"place" + 0.030*"good" + 0.020*"comfortable" + 0.020*"union"')
Term List Topic 5	(4, '0.067*"great" + 0.054*"stay" + 0.029*"like" + 0.025*"bathroom" + 0.014*"restaurant"')

Table 8. LDA result + semantic similarity expand 100% (AC3)

Term Topic	Aspect 1	Aspect 2	Aspect 3	Aspect 4	Aspect 5
	Cleanliness	Comfort	Service	Food	Location
Topic 1	0.871318936	0.871318698	0.871318815	0.871318816	0.871318935
Topic 2	0.965874195	0.965874314	0.965874254	0.965874196	0.965874255
Topic 3	0.977102399	0.977102337	0.97710222	0.97710226	0.97710228
Topic 4	0.856841862	0.856841683	0.856841685	0.856841743	0.856841624
Topic 5	0.974544883	0.974544764	0.974544882	0.974545002	0.974544942

on five specified hotel aspects, cosine similarity calculations are used.

3.4.5. Aspect categorization with semantic similarity (cosine similarity)

Similarity calculation using Cosine Similarity, based on the data extracted from LDA hidden topics as in Table 7. Formula Cosine Similarity refer to Eq. (3). The result of the Semantic Similarity on each topic that produces the highest similarity value will be categorized as the relevant aspect. The Cosine Similarity value scale ranges between 1 and 0.

Based on Table 8. found that term list 1 gets the highest similarity calculation results on the " Cleanliness " aspect, so it can be concluded that topic 1 is included in the Cleanliness category. Other term topics the same applies to the following topics, based on the results of the highest similarity in the related aspects.

3.5 Pre-processing aspect based sentiment analysis

In this process, a description of the resulting data from the categorization aspect term list for each aspect that has been categorized collected based on the relevant aspect. Categorize each term on each topic based on the same aspects. After that, the sentiment is determined on all data related to the term list on 5 aspects. The labeling process is carried out on each data regarding positive and negative sentiments in all aspects. The term list on the categorization of each aspect is as follows :

3.6 Sentiment classification based in aspect

In this stage, a term list for each aspect that has been categorized has been collected based on the relevant aspect. After that, the sentiment is determined on all data related to the term list on every in five aspects. The labeling process is carried out on each data regarding positive and negative sentiments in all aspects. Sentiment classification is determined based on the relevant aspect, for example, such as data containing sentiment regarding the cleanliness aspect. Classification will be carried out based on data that has been carried out by training which has been categorized as cleanliness aspect. Then the data containing the term list will be classified sentiment on an aspect based.

Table 9. Aspect categorization result data reviews

Topic	Term List	Aspect
Topic 1	‘stayed’, ‘street’, ‘breakfast’, ‘bed’, ‘time’	Cleanliness
Topic 2	‘location’, ‘nice’, ‘service’, ‘night’, ‘square’	Comfort
Topic 3	‘hotel’, ‘room’, ‘staff’, ‘clean’, ‘small’	Cleanliness
Topic 4	‘rooms’, ‘place’, ‘good’, ‘comfortable’, ‘union’	Cleanliness
Topic 5	‘great’, ‘stay’, ‘like’, ‘bathroom’, ‘restaurant’	Food

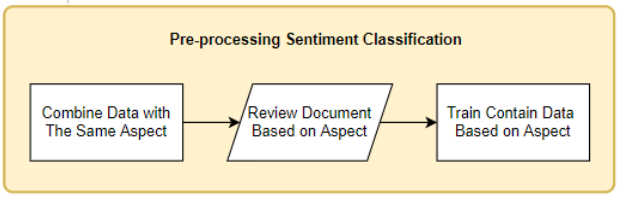

Figure. 4 Pre-processing aspect based sentiment analysis

Table 10. Aspect categorization result on cleanliness

No.	Term List	Aspect
1.	'stayed', 'street', 'breakfast', 'bed', 'time'	Cleanliness
2.	'hotel', 'room', 'staff', 'clean', 'small'	Cleanliness
3.	'rooms', 'place', 'good', 'comfortable', 'union'	Cleanliness

Table 11. Aspect categorization result on comfort

No.	Term List	Aspect
1.	‘location’, ‘nice’, ‘service’, ‘night’, ‘square’	Comfort

Table 12. Aspect categorization result on food

No.	Term List	Aspect
1.	‘great’, ‘stay’, ‘like’, ‘bathroom’, ‘restaurant’	Food

3.7 Performa evaluation

Performing performance evaluation using Semantic Similarity Result Score, Confusion Matrix, and Classification Report. Using the formula

Table 13. Sentiment classification based on aspect cleanliness

No. Term List	Aspect	Sentiment	Sentiment Score
1.	Cleanliness	Positive	0.449318
2.	Cleanliness	Positive	0.335897
3.	Cleanliness	Positive	0.245101

Table 14. Sentiment classification based in aspect comfort

No. Term List	Aspect	Sentiment	Sentiment Score
1.	Comfort	Positive	0.288288

Table 15. Sentiment classification based in aspect food

No. Term List	Aspect	Sentiment	Sentiment Score
1.	Food	Positive	0.263340

Accuracy, Precision, Recall, F1-Score, and similarity score. The calculated performance is the performa of the aspect categorization and sentiment classification. The formula for calculating the performance can be seen referring to: Eqs. (5) to (8) in this stage.

4. Results and discussion

At this stage, there will be a discussion of the results of the research methodology that has been carried out, as follows :

4.1 Approach of aspect categorization reviews

Based in the data used, carry out the classification process on five predetermined aspects using 3 Aspect Based Categorization Approaches, namely AC1, AC2 and AC3. Several approaches to categorization aspects are aimed to increase the similarity value so that the results are more accurate, shown in Table 16. Explaining the approach to the categorization aspect and performance, as follows :

Table 17. shows that the approach of Aspect Categorization 3 (AC3) has the best value. The aspect categorization process uses three approaches (AC1-AC3) aimed at increasing the similarity value.

4.2 Approach of aspect categorization reviews

Review data from hotel customers, labeling the sentiment whether it is in the category of positive or negative sentiment. After labeling the sentiment data on each aspect, the data will be divided into ratio 75:25 of split data where 75% of training data and25% of testing data. Get the review data findings,

Table 16. Approach of aspect categorization reviews

Approach	Detail Information
AC1	Aspect Categorization 1 (AC1) to extract hidden topic data in documents, the Latent Dirichlet Allocation (LDA) approach is used.. The results of the LDA in the form of hidden topic data are calculated for similarity using Cosine Similarity to categorize in aspects into five hotel aspects that have been determined in Table 2.
AC2	AC1 + adding 20% expanding term list using WordNet Synonyms from all the total synonym.
AC3	AC1 + adding 100% expanding term list using WordNet Synonyms from all the total synonym.

Table 17. Performance of aspect categorization

Approach	Metrics	Score
AC1	LDA + Cosine Semantic Similarity with each keyword term in Table 2.	0.750
AC2	LDA + Expanded WordNet Synonyms(20%) + Cosine Semantic Similarity	0.811
AC3	LDA + Expanded WordNet Synonyms(100%) + Cosine Semantic Similarity	0.856

as well as the aspects of each review and aspect, after classifying the aspects (AC3). Perform performance evaluation using Confusion Matrix and Classification Report shown in Figure. 5. Using the formula Accuracy, Precision, Recall and F1-Score. The formula for calculating the performance can be seen referring to: Eqs. (5) to (8) in this stage.

In Table 19. Describes the results of the evaluation sentiment classification on each Aspect Term: Food, service, location, comfort and cleanliness using the Support Vector Machine algorithm, which is compared with the results of several previous studies. Perform performance

Table 18. Approach of sentiment classification

Approach	Detail Information
Results of AC3	Based on the results of the categorization of aspect 3 (AC3), conduct training and testing based on each aspect of the categorization.
Sentiment Labelling	It was labeling the data review using TextBlob to calculate the sentiment value whether it is in the positive or negative category.
Classification Model	They are classifying labeled data using the method Support Vector Machine algorithm.

Table 19. Sentiment classification peformance

Prior Paper	Accuracy	Precision	Recall	F1-Score
This Paper	0.940	0.956	0.980	0.966
[16]	-	0.932	0.960	0.946
[12]	-	0.906	0.960	0.932

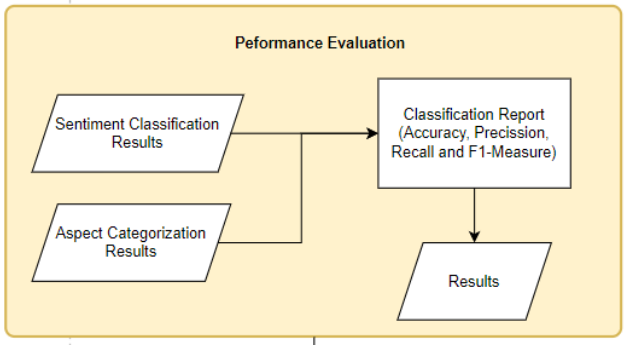

Figure. 5 Performance evaluation

evaluation using Confusion Matrix and Classification Report using Accuracy, Precision, Recall, and F1-Score formulas. This gets the highest score of accuracy, precision, and F1-Score of 0.990 and recall of 1.000 as follows :

4.3 Most important aspect ranking

This section determines which aspects are currently being discussed by customers based on the results of the reviews. Fig. 6 shows that most of the customers discuss the cleanliness aspect, can be seen that the value of the cleanliness aspect has the largest percentage of 43% then followed by the food aspect at 16%, then the comfort aspect at 15% and followed by the service and location aspects having a value of 13% of the total. It can be concluded that the results of the overall customer review of the hotel discuss more about the cleanliness aspect, as follows :

4.4 Sentiment evaluation results based on aspect

Table 20. shows the percentage of the overall positive and negative sentiment from all aspects. It shows that the cleanliness aspect has the highest percentage value of positive sentiment at 39%, followed by the food aspect at 14%, then from the comfort aspect by 13% and in the location and service aspect, it has a positive sentiment percentage value of 12%. The lowest negative of sentiment is in the location aspect, with a percentage value of 1% of the total. Table 20. shows the conclusion that positive sentiment has the highest value in the "Cleanliness" aspect with a value of 39.450.

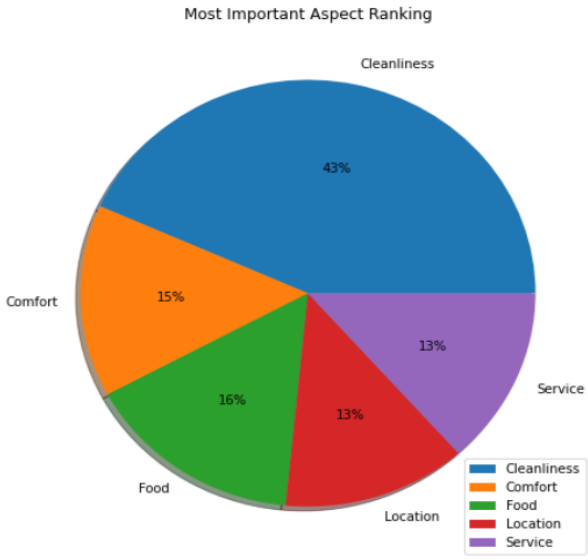

Figure. 6 Most important aspect

Table 20. Sentiment evaluation results based on aspect

Results of Sentiment Based Each Aspect		
Aspect	**Sentiment**	**Results Evaluation (Percentage)**
Food	Positive	14.140
	Negative	1.430
Service	Positive	12.250
	Negative	1.239
Location	Positive	12.075
	Negative	1.000
Comfort	Positive	13.346
	Negative	1.207
Cleanliness	Positive	39.450
	Negative	3.861
Total		100.000

Meanwhile, from negative sentiment, the highest value lies in the "Cleanliness" aspect, with a value of 3.861. Then for positive sentiment with the lowest value is the "Location" aspect with a value of 12.074, as follows:

5. Conclusion

The conclusion is that sentiment-based reviews may be divided into five categories. There are three techniques of categorizing aspects in the process, namely AC1 - AC3, which is intended to increase the similarity score. The highest value was obtained by Aspect Categorization 3 (AC3) using LDA + 100% expand document. The average accuracy value of sentiment classification was 0.940. The highest positive sentiment value was in the cleanliness aspect of 39.450, and the highest negative sentiment value was 3.850. Meanwhile, the lowest positive sentiment value was in the location aspect of 12.074, and the lowest negative sentiment

value was 1.000 shows that most of the customers discuss things on the aspect of cleanliness.

Acknowledgments

This research was funded by Lembaga Pengelola Dana Pendidikan (LPDP) under Riset Inovatif-Produktif (RISPRO) Invitation Program managed, the Indonesian Ministry of Education and Culture under Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Program, and Institut Teknologi Sepuluh Nopember (ITS) under project scheme of the Publication Writing and IPR Incentive Program (PPHKI)

References

[1] N. H. Gabriela, R. Siautama, C. I. A. Amadea, and D. Suhartono, "Extractive Hotel Review Summarization based on TF/IDF and Adjective-Noun Pairing by Considering Annual Sentiment Trends", *Procedia Comput. Sci.*, Vol. 179, No. 2020, pp. 558-565, 2021.

[2] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews", *Procedia Comput. Sci.*, Vol. 179, No. 2020, pp. 728-735, 2021.

[3] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Semantic features for optimizing supervised approach of sentiment analysis on product reviews", *Computers*, Vol. 8, No. 3, pp. 1-16, 2019.

[4] S. Cahyaningtyas, D. H. Fudholi, and A. F. Hidayatullah, "Deep learning for aspect-based sentiment analysis on Indonesian hotels reviews", Vol. 4, No. 3, 2021.

[5] W. Xue and T. Li, "Aspect Based Sentiment Analysis with Gated ConVolutional Networks", *arXiv Prepr. arXiv1805.07043*, 2018.

[6] I. E. Tiffani, "Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review", *Joscex*, Vol. 1, pp. 1-7, 2020.

[7] I. Perikos, K. Kovas, F. Grivokostopoulou, and I. Hatzilygeroudis, "A system for aspect-based opinion mining of hotel reviews", *WEBSITE 2017 - Proc. 13th Int. Conf. Web Inf. Syst. Technol.*, No. Website, pp. 388-394, 2017.

[8] E. M. Grey, "Opinion Mining With Hotel Review using Latent Dirichlet Allocation-Fuzzy C-Means Clustering (LDA-FCM) Turkish Journal of Computer and Mathematics", Vol. 12, No. 11, pp. 3001-3007, 2021.

[9] R. Annisa, I. Surjandari, and Zulkarnain, "Opinion mining on mandalika hotel reviews using latent Dirichlet allocation", *Procedia Comput. Sci.*, Vol. 161, pp. 739-746, 2019.

[10] K. R. Nastiti, A. F. Hidayatullah, and A. R. Pratama, "Discovering Computer Science Research Topic Trends using Latent Dirichlet Allocation", *J. Online Inform.*, Vol. 6, No. 1, p. 17, 2021.

[11] E. Ekinci, "An Aspect-Sentiment Pair Extraction Approach Based on Latent Dirichlet Allocation", *Int. J. Intell. Syst. Appl. Eng.*, Vol. 3, No. 6, pp. 209-213, 2018.

[12] R. A. Priyantina and R. Sarno, "Sentiment analysis of hotel reviews using Latent Dirichlet Allocation, semantic similarity and LSTM", *Int. J. Intell. Eng. Syst.*, Vol. 12, No. 4, pp. 142-155, 2019, doi: 10.22266/ijies2019.0831.14.

[13] F. Nurifan, R. Sarno, and K. R. Sungkono, "Aspect based sentiment analysis for restaurant reviews using hybrid ELMo-Wikipedia and hybrid expanded opinion lexicon-senticircle", *Int. J. Intell. Eng. Syst.*, Vol. 12, No. 6, pp. 47-58, 2019, doi: 10.22266/ijies2019.1231.05.

[14] S. Fransiska and A. I. Gufroni, "Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method", *Sci. J. Informatics*, Vol. 7, No. 2, pp. 2407-7658, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji.

[15] A. A. Farisi, Y. Sibaroni, and S. A. Faraby, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier", *J. Phys. Conf. Ser.*, Vol. 1192, No. 1, 2019.

[16] D. A. K. Khotimah and R. Sarno, "Sentiment analysis of hotel aspect using probabilistic latent semantic analysis, word embedding and LSTM", *Int. J. Intell. Eng. Syst.*, Vol. 12, No. 4, pp. 275-290, 2019, doi: 10.22266/ijies2019.0831.26.

[17] F. A. Bachtiar, W. Paulina, and A. N. Rusydi, "Text Mining for Aspect Based Sentiment Analysis on Customer Review : a Case Study in the Hotel Industry", *5th Int. Work. Innov. Inf. Commun. Sci. Technol.*, No. March, 2020.

[18] S. Yang and H. Zhang, "Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis", *Int. J. Comput. Inf. Eng.*, Vol. 12, No. 7, pp. 525-529, 2018.

[19] A. Pradhan, M. R. Senapati, and P. K. Sahu, "Improving sentiment analysis with learning concepts from concept, patterns lexicons and negations", *Ain Shams Eng. J.*, No. xxxx, 2021.

[20] Z. Li, J. C. White, M. A. Wulder, T. Hermosilla, A. M. Davidson, and A. J. Comber, "Land cover harmonization using Latent Dirichlet

Allocation", *Int. J. Geogr. Inf. Sci.*, Vol. 35, No. 2, pp. 348-374, 2021.
[21] Y. S. L. Afuan and A. Ashari, "A study: query expansion methods in information retrieval Lasmedi", *J. Phys. Conf. Ser.*, Vol. 1367, 2019.
[22] M. Kulmanov, F. Z. Smaili, X. Gao, and R. Hoehndorf, "Semantic similarity and machine learning with ontologies", *Brief. Bioinform.*, Vol. 22, No. 4, pp. 1-18, 2021.
[23] A. Mahmoud and M. Zrigui, "Semantic similarity analysis for paraphrase identification in Arabic texts", *PACLIC 2017 - Proc. 31st Pacific Asia Conf. Lang. Inf. Comput.*, No. Pacific 31, pp. 274-281, 2019.
[24] Y. Liu, X. Wang, L. Li, S. Cheng, and Z. Chen, "A Novel Lane Change Decision-Making Model of Autonomous Vehicle Based on Support Vector Machine", *IEEE Access*, Vol. 7, pp. 26543-26550, 2019.
[25] M. H. Alam, W. J. Ryu, and S. K. Lee, "Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews", *Inf. Sci. (NY).*, Vol. 339, pp. 206-223, 2016.