Received: March 25, 2019

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Sentiment Analysis of Hotel Aspect Using Probabilistic Latent Semantic Analysis, Word Embedding and LSTM

Dewi Ayu Khusnul Khotimah[1] **Riyanarto Sarno**[2]*

[1] *Department of Information Technology Management,*
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2] *Department of Informatics,*
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
* Corresponding author’s Email: riyanarto@if.its.ac.id

Abstract: In the industrial era 5.0, product reviews are necessary for the sustainability of a company. Product reviews are a User Generated Content (UGC) feature which describes customer satisfaction. The researcher used five hotel aspects including location, meal, service, comfort, and cleanliness to measure customer satisfaction. Each product review was preprocessed into a term list document. In this context, we proposed the Probabilistic Latent Semantic Analysis (PLSA) method to produce a hidden topic. Semantic Similarity was used to classify topics into five hotel aspects. The Term Frequency-Inverse Corpus Frequency (TF-ICF) method was used for weighting each term list, which had been expanded from each cluster in the document. The researcher used Word embedding to obtain vector values in the deep learning method from Long Short-Term Memory (LSTM) for sentiment classification. The result showed that the combination of the PLSA + TF ICF 100% + Semantic Similarity method was superior are 0.840 in the fifth categorization of the hotel aspects; the Word Embedding + LSTM method outperformed the sentiment classification at value 0.946; the service aspect received positive sentiment value higher are 45.545 than the other aspects; the comfort aspect received negative sentiment value higher are 12.871 than the other aspects. Other results also showed that sentiment was affected by the aspects.

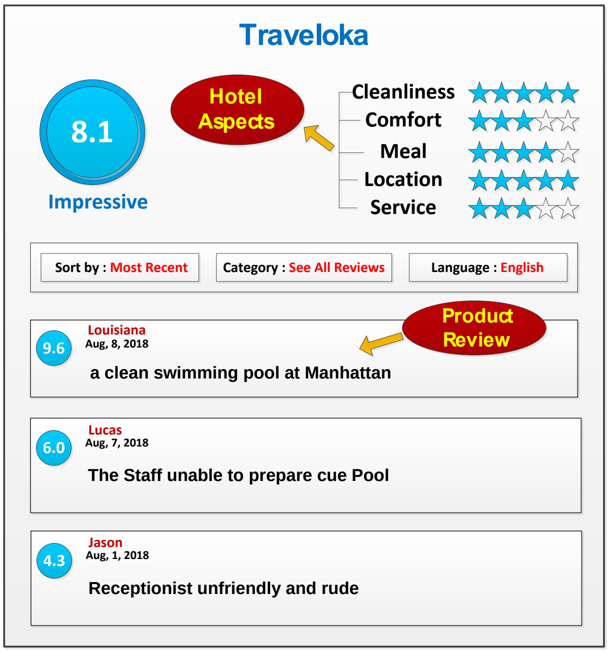
Keywords: Product reviews, Customer satisfaction, Hotel aspects, Term list document, Hidden topic, Semantic similarity, Aspect categorization, Word embedding, Deep learning, Sentiment classification.

1. Introduction

In the industrial era 5.0 (Society 5.0), product reviews are necessary for the sustainability of a company. Product reviews are a User Generated Content (UGC) feature which describes customer satisfaction. The impact of industry 5.0 can generate big data on social media which is the main point of business [1]. Social media becomes the central stage of assessment of human habits which can affect the decision making process [2]. Product reviews written by customers on social media will be a large, unstructured amount of data, which must be analyzed using appropriate techniques [3]. According to Xun Xu [4], product reviews can be analyzed using sentiment analysis technique. Another study [5] conducted a sentiment analysis to review customer opinions from movie, using SentiWordNet. Bagus et al. [6] also examined customer opinions from the Amazon website by addressing local and global contexts.

Previously, Dewi Ayu [7] conducted a study to measure hotel customer satisfaction, based on sentiment analysis using the Probability Latent Semantic Analysis (PLSA) method. Customer satisfaction is considered to be important for a decision related to product purchase [4]. Customer decisions often depend on the opinion and brand image of a product [8]. The researcher used product reviews of hotels to measure customer satisfaction based on its aspects.

Received: March 25, 2019



9.6 Louisiana Aug, 8, 2018
a clean swimming pool at Manhattan

6.0 Lucas Aug, 7, 2018
The Staff unable to prepare cue Pool

4.3 Jason Aug, 1, 2018
Receptionist unfriendly and rude

Figure. 1 Traveloka Website appearance

The brand image of a hotel can influence customer opinions measured by reviews alongside other affecting aspects. Aspects which can affect the hotel brand image are called hotel aspects [9]. The hotel aspects aim to predict opinions from each product review [10]. The researcher opts for Traveloka as the most ideal data source, since the website globally displays customer reviews [11].

The researcher conducted Aspect Categorization (AC) and Sentiment Classification (SC), based on product reviews on five hotel aspects, using Python programming language. The five hotel aspects taken from Traveloka include location, meal, service, comfort, and cleanliness. The statement matches the infographic described in the display of the Traveloka website on Fig. 1

Fig. 1 shows that Traveloka has a broad potential, with the five aspects used to assess review. Customer experience in assessing hotels needs to be further analyzed [9]. In the literature review [4], the researcher analyzed customer reviews of various types of hotels using the Latent Semantic Analysis (LSA) method. The fallback of the previous research was that it did not focus on aspects which could affect the sentiment value. The star rating on the five aspects in Fig.1 is considered not to be a good representation for customer satisfaction [4, 7]. Thus, the researcher measured customer satisfaction based on reviews of the hotel aspects. Customer satisfaction based on the hotel aspects is considered to be better in representing the overall customer sentiment.

The importance of knowing aspects is to predict the topic of each product review, for example, “Receptionist unfriendly and rude”; the term list of “**reception**” and “**unfriendly**” implies topics which enter into the “**service**” aspect. If aspect in the review is not captured properly, the perception of the customer regarding the “**service**” aspect cannot be identified. However, if the aspect is captured properly, it can be concluded that the customer review the service of the hotel.

Product review data on Traveloka were taken by crawling using the WebHarvy software [12]. The data were from Manhattan, New York hotel reviews. The researcher only used the product reviews in English. The next step is preprocessing [13]. Each product review was preprocessed into a term list document. In the preprocessing, the process consists of Convert into Lowercase, Tokenization, Stemming, Stopwords Removal, Remove Punctuation, and Spelling Correction into a term list. The term list document would be used as training and testing.

The researcher used the Probabilistic Latent Semantic Analysis (PLSA) method to produce hidden topic of the term list. Hidden topic was produced from words which contain many meanings and were grouped with the same words [14]. The algorithm used by the PLSA method to produce hidden topic is Expectation (E-Step) and Maximization (M-Step) [15]. PLSA is able to produce better hidden topics with a machine learning approach on Natural Language Processing. E-Step and M-step algorithms was able to handle words which contained polysemy and equipped with document training corpus. The topics found by the PLSA method were used for categorization of the five hotel aspects.

The researcher categorized the five aspects in three processes, namely Aspect Categorization 1 (AC1), Aspect Categorization 2 (AC2), and Aspect Categorization 3 (AC3). Aspect categorization was used to categorize hidden topics into the five hotel aspects using the semantic similarity method taken from hidden topic data and keyword term list for hotel aspect.

The researcher used the Term Frequency-Inverse Corpus Frequency (TF-ICF) method to weight each term list [16], which was expanded from each cluster in the document [17]. The TF-ICF was used as an extension of the term list, comparison and increasing data accuracy. TF-ICF was used in the AC2 and AC3 aspect categorization as a comparison of AC1. Aspect Categorization 2 (AC2) used 20% and Aspect Categorization 3 (AC3) used 100% of the TF-ICF result.

After the five aspects were categorized, the researcher conducted pre-process aspect-based sentiment classification. Pre-process aspect based sentiment classification was grouping each product review based on its aspect. The grouping of each aspect aims to make it easier for the researcher to classify sentiments based on its aspect.

The result of the pre-process aspect based sentiment classification was in the form of training data of each aspect. The training data of each aspect and AC3 was used in the Word embedding process to obtain word vector values. The researcher used the word vector values for sentiment classification using the Long Short-Term Memory (LSTM) method [18, 19]. The LSTM method was used as a concept of deep learning to improve performance in the sentiment classification [20, 21]. The researcher used the Global Vector model (GloVe) in Word embedding to help LSTM to model excellent term list dependencies [22-24].

The researcher used two parameters in the sentiment classification, namely positive and negative. If there is an adjective (Adj) that has a "neutral" sentiment value, the annotator will classify it as positive or negative. Positive or negative sentiment is determined by each word in the sentence in the product review. Customers are satisfied if they have a positive value of sentiment. Customers are dissatisfied if they have negative sentiment values. Sentiment values clearly describe the feelings, opinions, emotions of many customers, such as satisfied or dissatisfied [25].

The "receptionist unfriendly and rude" review will be detected as negative sentiment based on the "**unfriendly**" and "**rude**" term lists. The researcher performed sentiment classification with three processes, namely *Sentiment Classification* 1 (SC1), *Sentiment Classification* 2 (SC2), and *Sentiment Classification* 3 (SC3) as a comparison. The researcher chose the best performance from the evaluation of five aspects categorization, and sentiment classification as the advantage of the proposed method.

The result was evaluated using Precision, Recall, and F1-Measure. The result showed that the combination of the PLSA + TF ICF 100% + Semantic Similarity method was superior in the fifth categorization of the hotel aspects. The Word embedding + LSTM method outperformed the sentiment classification. The result showed that the "comfort" aspect received negative sentiment value higher than the other aspects. Other results also showed that sentiment could be influenced by hotel aspects.

The use of Business Intelligence and Analytics somewhat outperformed the importance of sentiment on aspects. We expected that this study is that could help companies to identify customer perceptions globally and appreciating timeliness in the service to increase net profit margins in the industrial era 5.0.

The researcher compiled this journal article in the following order: in the Literature Review, we reviewed several literature studies from several previous researchers about sentiment analysis and aspect extraction on hotels. The data set and method that we proposed were listed in the Research Method. In the Result and Analysis, we showed the analysis and experimental result. Conclusions and further study were explained in the Conclusions.

2. Related theory

This section describes numerous theories related to the study including a number of surveys from previous studies.

2.1 Keyword term list for hotel aspect

The keyword term list for hotels were taken from several studies. We identify the term list hotels based on five labels [7, 26, 27]. The hotel aspects include: 1) Location; 2) Meal; 3) Service; 4) Comfort; and 5) Cleanliness. The term list can be seen in Table 1.

2.2 Preprocessing

Fig. 2 describes the process of preprocessing data. The preprocessing technique was conducted in six stages, namely: Converting into Lowercase, Tokenization, Stemming, Stopwords Removal, Punctuation Removal, and Spelling Correction.

The first stage is "Convert into Lowercase" which functions to change the text in the customer review into lowercase. The second stage, Tokenization, is to break the text into tokens or words. The third stage is "Stemming" which functions to change the text on customer reviews into its root form. The fourth stage is "Remove Punctuation" which functions to remove punctuation in customer review. The last step is "Spelling Correction" which functions to refine the customer reviews which contained errors.

We used Natural Language Processing (NLP) for part-of-speech tagging (POS Tagging). Using POS tagging on the preprocessing can avoid using inappropriate words. We first conducted POS Tagging on preprocessing for tagging each term list

Table 1. Term list of hotel aspects

Hotel Aspects	Term List of Hotel Aspects
1 Location	railway, view, station, airport, distance, far, close, convenient, train, metro, times, square, central, park, fifth, rockefeller, grab, uber, near, access.
2 Meal	drink, breakfast, spicy, meal, tea, buffet, bar, restaurant, dinner, lunch, brunch, delicious, food, dish, wine, salad, pretzel, lobster, knish, honey, almond, juice, coocktail, milk, chiken, donut, lemon, vichyssoise, mocha, latte, matcha, beer, dish.
3 Service	desk, care, reliable, fast, convenient, check-in, check-out, good, staff, polite, helpful, funny, friendly, reliable, unable, quick, manager, guy, facility, smile, check, Wi-Fi, pool, information, help, poker, luggage, concierge, customer, bag, arrival, person, question, spa, connection, suggestion, notch, service, reception, impression, doorman, maid, attitude, request, employee, front desk, bartender.
4 Comfort	facility, Wi-Fi, cue, gym, business, internet, spa, connection, meeting, charge, activity, sleep, broken, bed, space, elevator, lift, scary, comfort.
5 Cleanliness	cleanliness, clean, smell, water, smoke, carpet, swimming, fragrance, toilet, complaint, heat.

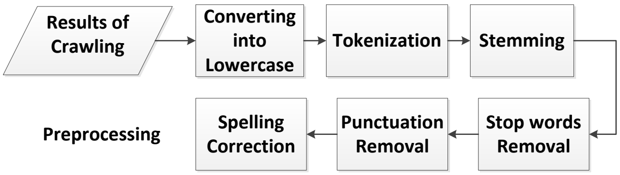


Figure. 2 The process of preprocessing data

[16]. POS tagging was done by assigning part-of-speech obtained from the Natural Language Toolkit Library (NLTK) [28].

The assigning using POS Tagging on each term list was combined with the calculation of similarity in the singular noun and infinitive form until it formed into a sentence [29].

2.3 Expanded term list (TF-ICF)

The researcher used the Term Frequency-Inverse Corpus Frequency (TF-ICF) method to obtain important terms from the weighting of each term to be an expanded term list of each cluster [16]. The TF-ICF formula is explained in Eq. (1) and (2).

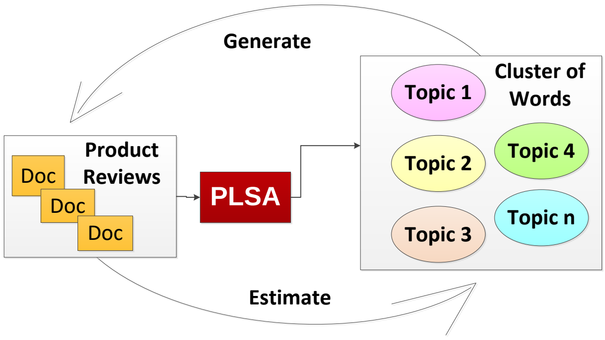


Figure. 3 Overall topic extraction process

$$TF - ICF = TF_{t,i} \times ICF_t \tag{1}$$

$$TF - ICF = TF_{t,i} \times log\left(\frac{N}{CF_t}\right) \tag{2}$$

$TF_{t,i}$	= Number of terms in i class
N	= Number of class
CF_t	= Number of class containing t term

The researcher conducted several stages for the expansion of the term list using the TF-ICF. The first step was calculating the frequency of each term list. The second stage was to determine the terms that had the highest score on the TF-ICF from each cluster.

2.4 Probabilistic latent semantic analysis (PLSA)

Fig. 3 describes the process of searching for hidden topics from the term list document. The first process was to collect the term list from each product review. Each document contained a product review labeled Identity Document (ID). The ID label was adjusted to the number in the product review.

The overall result would be expanded using WordNet for semantic syntax relations in English. The WordNet functioned to detect words which contained multiple meanings. The result would be used as a data set [30]. The annotator would label the training data and save them in the data set as a corpus. We used data sets in real-life, obtained from product reviews.

The final process was to use the WordNet expanded term list to be calculated using semantic probabilities from each word into several topics [31, 32]. The algorithm Expectation-Maximization (E-Step and M-Step) in the PLSA produced hidden topics from the relevance of words and similarity of topics in a document [15, 33].

The Probabilistic Latent Semantic Analysis (PLSA) method was chosen because it was able to handle words which contained polysemy and equipped with document training corpus from the E-

Step and M-step algorithms. PLSA is able to produce better hidden topics with a machine learning approach on Natural Language Processing [14].

The following is an algorithm for how PLSA works. The researcher determines the number of topics (z) and initializes the probability parameters in the topic, P (z), P (d|z) is the probability document containing the topic, P (w|z) is the probability of the words randomly contained in the topic. The calculation of words in the document is shown in Eq. (3).

$$P(d_i, w_j) = \sum_{k=I}^{K} P(Z_k)P(Z_k)P(d_i|Z_k)\, P(w_j|Z_k) \quad (3)$$

The next step is to calculate the probability value of words in each parameter using the Expectation step and Maximization step algorithms. Expectation step (E-step) calculates the probability of the topic on the document, which is described in Eq. (4).

$$P(Z_k|d_i, w_j) = \frac{P(w_j|Z_k)P(z_k|d_i)}{\sum_{k=1}^{k} P(w_j|Z_l)P(z_l|d_i)} \quad (4)$$

The last step is to calculate the Maximization step (M-step). M-step is used to calculate the value update from the document parameters shown in Eqs. (5) and (6).

$$P(w_j|Z_k) = \frac{\sum_{i=1}^{N} n(d_i|w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i|w_m)P(z_k|d_i,w_m)} \quad (5)$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{N} n(d_i|w_j)P(z_k|d_i,w_j)}{n(d_i)} \quad (6)$$

The result on Eqs. (5) and (6) is the words probability in the document which produces the hidden topics [14].

2.5 Calculation hidden topic with 5 hotel aspects using semantic similarity

Hidden topics generated by the PLSA method would be classified using semantic similarity for labeling in each document [33] into five hotel aspects. The five hotel aspects taken from Traveloka include location, meal, service, comfort, and cleanliness. The researcher used annotator to validate documents on training data, based on the five hotel aspects.

The similarity of the hidden topic in the five hotel aspects is determined by the semantic similarity value of the term list [16]. Each hidden topic document would be calculated using Semantic Similarity to see its similarity to the term list in Table 1, and added to the TF-ICF expanded term list. Each document was collected to form a cluster under the same aspect. This aims to correctly classify the topics in each term list document into five hotel aspects.

The Semantic Similarity method calculated the similarity of each cluster based on the thesaurus word in the corpus [16]. The following is the Semantic Similarity formula described in Eq. (7).

$$Sem\_Sim\,(w_i, w_j) : \frac{\sum_{m-1}^{K} w_i^m w_j^m}{\sqrt{\sum_{m-1}^{K}(w_i^m)^2}\,\sqrt{\sum_{m-1}^{K}(w_j^m)^2}} \quad (7)$$

Eq. (7) was used to measure the similarity between word 1 (w_i) and word 2 (w_j). $\sum_{\mathrm{m-1}}^{\mathrm{K}} =$ is the number of iterations from m to K word. The similarity has a value of 1 which means equal and -1 which means different. Semantic Similarity is considered to be better than TF-IDF weighting, which cannot give a value of -1 [16].

2.6 Aspect based sentiment classification using Word Embedding + LSTM

The sentiment analysis was carried out after the classification of the five hotel aspects [34]. The data used to conduct sentiment classification were data from the Aspect Classification Hotel Reviews.

Word embedding was used to obtain vector values from each term list. Vector values were used for sentiment classification using the Long Short-Term Memory (LSTM) method. Sentiment classification used the machine learning technique on TensorFlow [35]. TensorFlow is software used for sentiment training in the LSTM method [36].

In this module, each label of the term list document would be processed into a vector value as a feature value [16]. The feature value was obtained from the average vector value in the Word embedding representation in each term [33]. The representation of Word embedding is used for initialization of each term using Global Vectors (GloVe). The advantage of GloVe is that it can model vector which represent terms from global corpus statistics. It is due to the fact that Word embedding must be based on the probability ratio of the each term occurrence [37].

In the LSTM method, there was a thesaurus library to implement the Neural Networks model for synonyms, antonyms, definitions, pronunciations in English. This study used functional Application

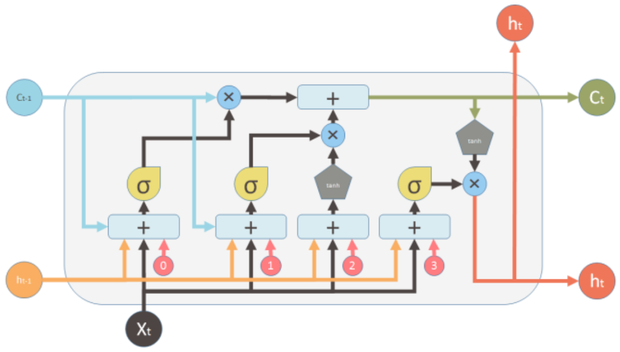


Figure. 4 Standard LSTM method

Programming Interface (API) on Keras to determine a more complex model. Keras is a library in Neural Networks, which can run on TensorFlow with the Python programming language.

Fig. 4 describes the standardization of the LSTM method which is then divided into four components, namely Input-Gate (*i*), to control the input current which enters the neuron, Forget-Gate (*f*), which makes the neuron being in the reset status of its current status, Output-gate (*o*), to control the effect of neuron activation on other neuron, and memory cell (*c*) [38]. The following is a function of LSTM described in Eqs. (8), (9), (10), (11), (12), and (13):

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) \quad (8)$$

$$o_t = \sigma(W^{(O)}x_t + U^{(o)}h_{t-1} + b^{(o)}) \quad (9)$$

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) \quad (10)$$

$$u_t = tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}) \quad (11)$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1} \quad (12)$$

$$h_t = o_t \odot tanh(c_t) \quad (13)$$

The matrix weight between two consecutive hidden layers was given the *Wk* and *Uk* symbol. Between inputs, hidden layers, and two sequential cell activations, each was connected to gate *k* (example: input → output → forget → cell). The vector value was symbolized by *bk*. The product value in each element of the two vectors was given the symbol ⊙. Activation of sigmoid value was a gate function, symbolized with *σ*. The *g* and *h* symbols are the activation of input and output cells of which value is *tanh*.

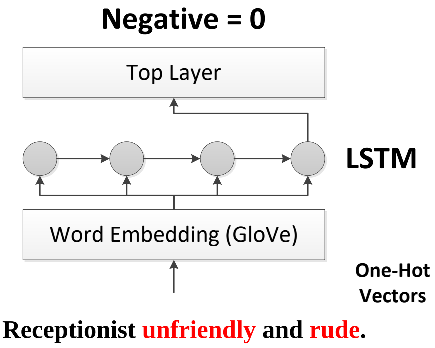


Figure. 5 LSTM method flow illustration

Fig. 5 describes the illustration of the LSTM method flow which is divided into 4 components by the sentiment classification [39]. Components in LSTM were used as embedding words with a binary softmax classifier which gives vector values for sentiment. The value of vectors in the binary softmax classifier would be multiplied by the matrix of other weight to produce values of 0 and 1 [40]. Effectively, it could give value parameters to positive sentiment with 1 and 0 to the negative sentiment.

The calculation on the sentiment classification produced Accuracy. Accuracy was used as a formula to evaluate the performance of sentiments from the deep learning method using the binary classification technique.

In this study, we would like to prove that Word embedding combined with the LSTM method was able to analyze the term list document, not only the adjectives. Each term would be determined by a vector value, based on the proximity of positive and negative words which have undergone through the training process. Thus, the Word embedding method and LSTM is most likely to have a higher performance than the techniques performed on SentiWordNet.

In the previous study [5], sentiment analysis was carried out using SentiWordNet. However, the accuracy of SentiWordNet tends to have a small value compared to the LSTM method. The technique performed by SentiWordNet is to calculate the similarity of each term, using a score in the SentiWordNet dictionary. The Dictionary of SentiWordNet includes positive, neutral, and negative. We use two parameters in this study, namely positive and negative in SentiWordNet as a comparison in sentiment classification. The following is the SentiWordNet formula described in Eqs. (14), (15), and (16).

$$pos\_score = \sum_{i=t}^{n} pos\_score\_Senti(i) \quad (14)$$

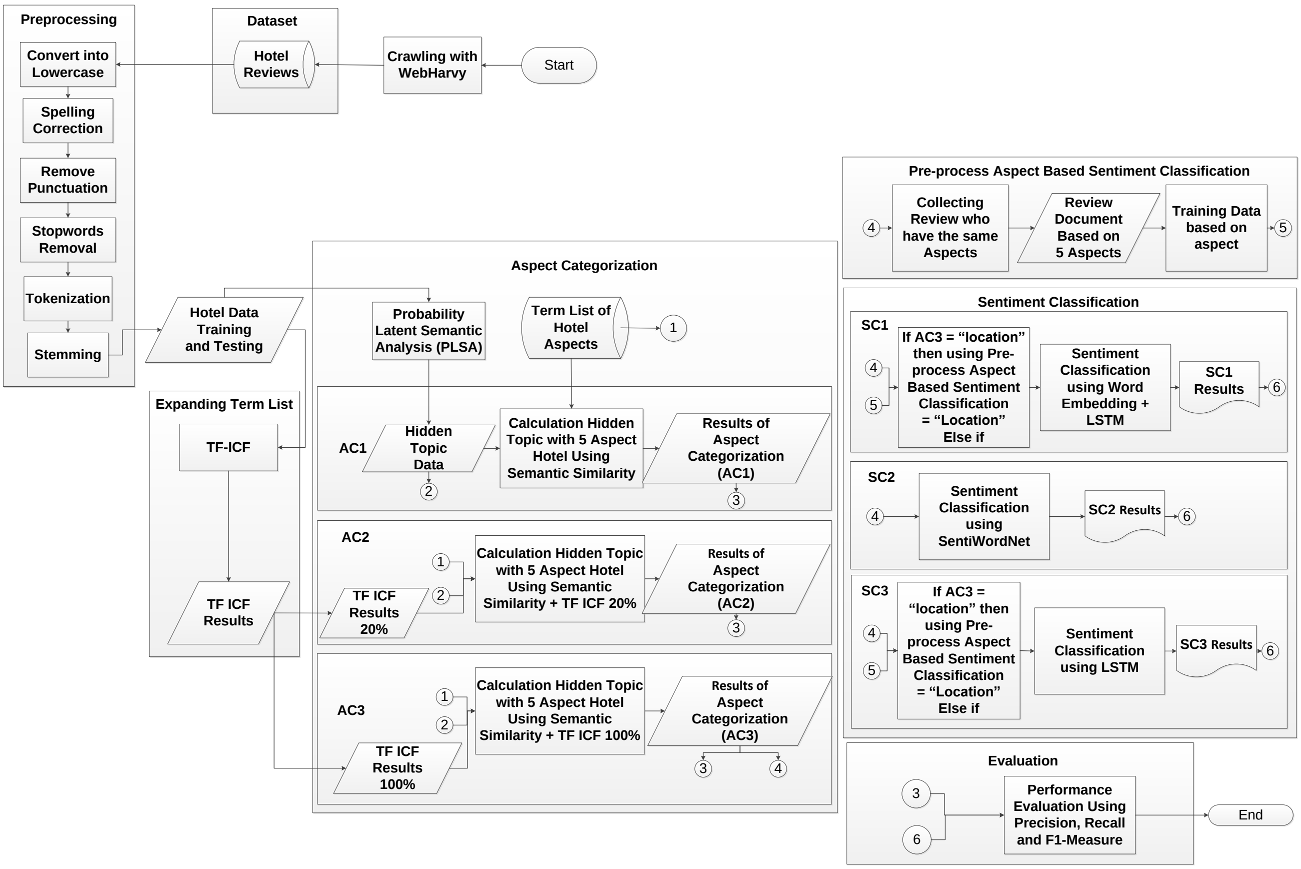

Start
Crawling with WebHarvy
Dataset
Hotel Reviews
Preprocessing
Convert into Lowercase
Spelling Correction
Remove Punctuation
Stopwords Removal
Tokenization
Stemming
Hotel Data Training and Testing
Expanding Term List
TF-ICF
TF ICF Results
Aspect Categorization
Probability Latent Semantic Analysis (PLSA)
Term List of Hotel Aspects
1
AC1
Hidden Topic Data
2
Calculation Hidden Topic with 5 Aspect Hotel Using Semantic Similarity
Results of Aspect Categorization (AC1)
3
AC2
1
2
TF ICF Results 20%
Calculation Hidden Topic with 5 Aspect Hotel Using Semantic Similarity + TF ICF 20%
Results of Aspect Categorization (AC2)
3
AC3
1
2
TF ICF Results 100%
Calculation Hidden Topic with 5 Aspect Hotel Using Semantic Similarity + TF ICF 100%
Results of Aspect Categorization (AC3)
3
4
Pre-process Aspect Based Sentiment Classification
4
Collecting Review who have the same Aspects
Review Document Based on 5 Aspects
Training Data based on aspect
5
Sentiment Classification
SC1
4
5
If AC3 = “location” then using Pre-process Aspect Based Sentiment Classification = “Location” Else if
Sentiment Classification using Word Embedding + LSTM
SC1 Results
6
SC2
4
Sentiment Classification using SentiWordNet
SC2 Results
6
SC3
4
5
If AC3 = “location” then using Pre-process Aspect Based Sentiment Classification = “Location” Else if
Sentiment Classification using LSTM
SC3 Results
6
Evaluation
3
6
Performance Evaluation Using Precision, Recall and F1-Measure
End

Figure. 6 Research method on the five hotel aspect categorization and sentiment classification

$$neg\_score = \sum_{i=t}^{n} neg\_score\_Senti(i) \quad (15)$$

$$total\_score = pos\_score - neg\_score \quad (16)$$

2.7 Performance evaluation

TP (True positive) is the number of documents correctly identified as a relevant document. TN (True Negative) is the number of documents correctly identified as an irrelevant document. FP (Positive False) is the number of incorrect documents and classified as a relevant document. FN (False Negative) is the number of incorrect documents and classified as an irrelevant document. The following is a formula for TP, TN, FP and FN shown in Eqs. (17), (18), and (19).

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

$$F1 - Measure = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (19)$$

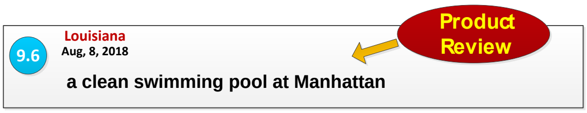
3. Research methods

Sentiment analysis based on the aspect of the hotel is a challenge for the researcher in measuring customer satisfaction. The researcher analyzed five aspects of the hotel which could affect sentiment.

Fig. 6 explains the overall performance of the research method. The overall performance included the PLSA method which produced hidden topics and categorization of the five hotel aspects using Semantic Similarity. We used TF-ICF as an extension to the term list. After the five aspects were identified, sentiment classification was conducted using the Word embedding + LSTM method. The researcher used several different approaches to prove the superiority of the proposed method.

3.1 Data collection

This study used online product review data on Traveloka as its data source. The researcher conducted data crawling using WebHarvy software.



9.6 Louisiana
Aug, 8, 2018

a clean swimming pool at Manhattan

Figure. 7 Customer review on Manhattan hotel, New York, 2018

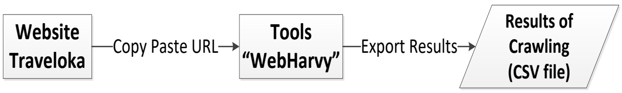


Figure. 8 The process of data crawling using WebHarvy

Table 2. Result of data crawling using WebHarvy

Product Reviews
a clean swimming pool at Manhattan
The Staff unable to prepare cue Pool
Receptionist unfriendly and rude
Poor breakfast.
Crowded lobby and scary corridors.
Don't be deceived by the nice smell in the lobby.
The hotel has a fab location and should be attracting great reviews.
The breakfast is als0 great!!!
A nice hotell,small but GOOD 👍👍👍.!!
the internet was VERY SLOWWW at night. make us not comfortable to stay at this hotel.

Fig. 7 shows that the product review obtained is in the form of text.

Fig. 8 explains the process of data crawling using the WebHarvy software [12]. The first process is to enter the address of the Manhattan at Times Square Hotel review from Traveloka using WebHarvy. In WebHarvy, we obtained the product reviews in the customer comment column in the form of text.

The last step is to conduct export results. The result of data crawling from export results would be saved in the excel file format (.csv), so that it would be easy to use for data processing using the python programming language. The result of data crawling would be used for the next stage namely preprocessing [13, 34].

3.2 Data crawling

Table 2 displays a number of data crawling result with document labeling. The researcher obtained 529 data from the product review of Manhattan Hotel, New York.

The data obtained would be used as training and testing data. The researcher used 80% of the data for training and the other 20% for testing; namely 423 training data and 106 data testing. Training and testing data were used for the entire process in the study.

Table 3. Result of crawling with document labeling

id_rev	id_doc	Product Reviews
1	1	a clean swimming pool at Manhattan
2	2	The Staff unable to prepare cue Pool
3	3	Receptionist unfriendly and rude
4	4	Poor breakfast.
5	5	Crowded lobby and scary corridors.
6	6	Don't be deceived by the nice smell in the lobby.
7	7	The hotel has a fab location and should be attracting great reviews.
8	8	The breakfast is als0 great!!!
9	9	A nice hotell,small but GOOD 👍👍 👍.!!
10	10	the internet was VERY SLOWWW at night.
11	10	make us not comfortable to stay at this hotel.

Table 4. Ilustration of precprocessing data

Product Review	Preprocessing	Results
A nice hotell,small but GOOD 👍👍👍.!!	Converting into Lowercase	a nice hotell,small but good 👍👍👍.!!
	Tokenization	a, nice, hotell, small, but, good, 👍👍👍,.,!!
	Stemming	nice, hotell, small, but, good, 👍👍👍,.,!!
	Stopwords Removal	nice, hotell, small, good, 👍👍👍,!!
	Punctuation Removal	nice, hotell, small, good
	Spelling Correction	nice, hotel, small, good

Table 3 displays the result of data crawling from Table 2, which were labeled Identity Document (ID). The ID label was adjusted to the order in the product review. The order of the product review was given the initial "id_rev", which indicated the document. The document would be broken down and given an ID each. The ID for each document would be given the initials "id_doc" as its unique number which represents part of one document.

Period (.) on each review was used as a new document, as in the "id_rev 10" review. The review would be represented in two id_review with one id_doc. Thus, the sentence on a long product review could be easily detected according to its aspect and sentiment.

Received: March 25, 2019

Table 6. Result of term list using TF-ICF

Id_term	term	TF1	TF2	TF3	TF4	TF5	ICF	TFICF1	TFICF2	TFICF3	TFICF4	TFICF5
1	the	6	1	14	12	4	0	0	0	0	0	0
2	hotel	2	0	3	2	0	0.511	1.022	0	1.532	1.022	0
3	very	0	0	2	1	0	0.916	0	0	1.833	0.916	0
4	well	1	0	0	0	0	1.609	1.609	0	0	0	0
5	locate	3	0	0	0	0	1.609	4.828	0	0	0	0

Table 5. Result of data precprocessing

id_rev	Data Preprocessing Result
1	‘clean’, ‘swimming’, ‘pool’, ‘manhattan’
2	‘staff’, ‘unable’, ‘prepare’, ‘cue’, ‘pool’
3	‘receptionist’, ‘unfriendly’, ‘rude’
4	‘poor’, ‘breakfast’
5	‘crowd’, ‘lobby’, ‘scary’, ‘corridor’
6	‘dont’, ‘deceive’, ‘nice’, ‘smell’, ‘lobby’
7	‘hotel’, ‘fab’, ‘locate’, ‘attract’, ‘great’, ‘review’
8	‘breakfast’, ‘great’

3.3 Preprocessing data

The preprocessing technique is done in six stages, namely Convert into Lowercase, Tokenization, Stemming, Stopwords Removal, Remove Punctuation, Spelling Correction as illustrated in Table 4.

A document as a result of preprocessing is called term list document. The term list document would be labeled by annotator. The label consists of two types, namely hotel aspect label and sentiment label. The hotel aspect label includes: 1) location; 2) meal; 3) service; 4) comfort; and 5) cleanliness. The sentiment label indicates that the number 0 has a negative value and number 1 has a positive value.

We took eight product review examples from Table 3 to be shown in Table 5.

The eight examples in Table 5 was used for the illustration in this study.

3.4 Expanded term list data (TF-ICF)

Term Frequency (TF) is the simplest method to weight each term. Each term is assumed to have an interest that is proportional to the number of term occurrences in the document. The following is the result of the formula described in Eqs. (1) and (2).

Table 6 explains that TFICF1 is the value of a term in cluster 1, TFICF2 shows the value of a term in cluster 2, TFICF3 shows the value of a term in cluster 3, TFICF4 shows the value of a term in cluster 4, and TFICF5 shows the value of a term in cluster 5.

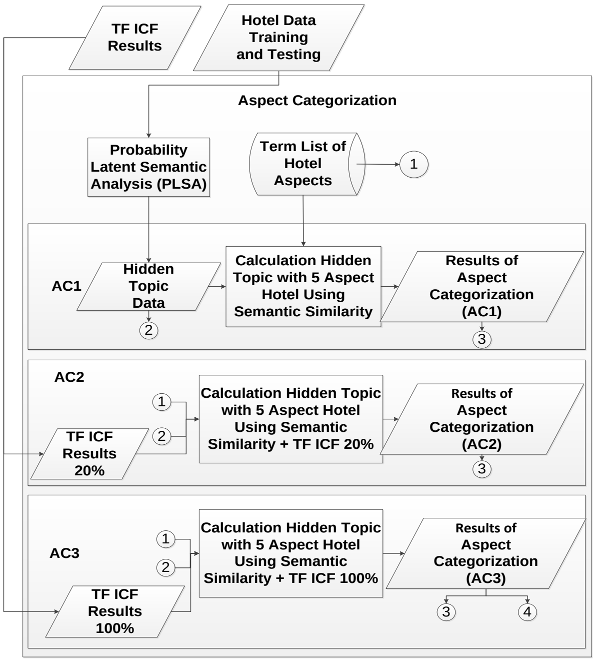

TF ICF Results
Hotel Data Training and Testing
Aspect Categorization
Probability Latent Semantic Analysis (PLSA)
Term List of Hotel Aspects
1
AC1
Hidden Topic Data
2
Calculation Hidden Topic with 5 Aspect Hotel Using Semantic Similarity
Results of Aspect Categorization (AC1)
3
AC2
1
2
TF ICF Results 20%
Calculation Hidden Topic with 5 Aspect Hotel Using Semantic Similarity + TF ICF 20%
Results of Aspect Categorization (AC2)
3
AC3
1
2
TF ICF Results 100%
Calculation Hidden Topic with 5 Aspect Hotel Using Semantic Similarity + TF ICF 100%
Results of Aspect Categorization (AC3)
3
4

Figure. 9 The process of aspect categorization

Table 7. Result of hidden topic data

id_rev	Hidden Topics Data Results
1	‘clean’, ‘swimming’, ‘pool’
2	‘staff’, ‘unable’, ‘prepare’, ‘cue’, ‘pool’
3	‘receptionist’, ‘unfriendly’, ‘rude’
4	‘poor’, ‘breakfast’
5	‘crowd’, ‘lobby’, ‘scary’
6	‘nice’, ‘smell’, ‘lobby’
7	‘hotel’, ‘locate’, ‘attract’, ‘great’, ‘review’
8	‘breakfast’, ‘great’

The frequency value of a term in a cluster can affect the inverse corpus frequency (ICF) value in generating the final TF-ICF value. The higher the frequency value of a term is, as long as the ICF value is not zero, the bigger the amount of the resulting TF-ICF value. Thus, the highest TF-ICF

Received: March 25, 2019

Table 8. Result of PLSA + Semantic Similarity

Document Term List	Hotel Aspects					Prediction PLSA + Semantic Similarity	Label Annotator
	1. Location	2. Meal	3. Service	4. Comfort	5. Cleanliness		
‘clean’	0.083	0.017	0.061	0.122	0.337	**5**	**5**
‘swimming’	0.020	0.000	0.030	0.122	0.200		
‘pool’	0.070	0.000	0.286	0.122	0.100		
Total	0.173	0.017	0.377	0.366	**0.637**		
‘staff’	0.002	0.002	0.286	0.002	0.010	**3**	**3**
‘unable’	0.000	0.016	0.277	0.070	0.007		
‘prepare’	0.000	0.007	0.097	0.000	0.000		
‘cue’	0.010	0.000	0.010	0.200	0.000		
‘pool’	0.070	0.000	0.286	0.070	0.100		
Total	0.082	0.025	**0.956**	0.342	0.117		
‘receptionist’	0.042	0.002	0.314	0.100	0.002	**3**	**3**
‘unfriendly’	0.042	0.002	0.314	0.095	0.002		
‘rude’	0.042	0.010	0.314	0.057	0.010		
Total	0.126	0.014	**0.942**	0.252	0.014		
‘poor’	0.007	0.072	0.140	0.122	0.020	**3**	**2**
‘breakfast’	0.002	0.274	0.233	0.020	0.000		
Total	0.009	0.346	**0.373**	0.142	0.020		

* 🟥 = a mistake in identification

value can represent and be used as an addition to the term in the cluster.

3.5 Aspect categorization

Fig. 10 describes the entire process on aspect categorization with a number of approaches including AC1, AC2, and AC3.

3.5.1. Aspect categorization (AC1)

Results Topic Categorization using PLSA

The result of data preprocessing in Table 5 was expanded using WordNet for semantic syntax relations in English. The expanded data were processed with the PLSA method using Eqs. (3), (4), (5), and (6). Therefore, the produced data were presented in Table 7.

The result of hidden topic in Table 7 was used for the classification process into five hotel aspects using Semantic Similarity.

Result of Hidden Topic with 5 Aspects using Semantic Similarity

Data in Table 7 were processed using the Semantic Similarity method with Eq. (7). Semantic Similarity was used to measure the similarity between the hidden topics (which had been identified using the PLSA method) and the term list in Table 1.

The calculation of Semantic Similarity was used to classify each term list document into the five hotel aspects. The result of the term list document was taken from Table 7 and was classified and presented in Table 8.

Table 8 shows the hidden topics generated by the PLSA method, and was classified in the five hotel aspects with Semantic Similarity.

The five hotel aspects were obtained from the Traveloka website. The hotel aspects on the Traveloka are: 1) Location; 2) Meal; 3) Service; 4) Comfort; and 5) Cleanliness. Errors in aspect detection would be highlighted in red as in the “**Poor Breakfast**”, which is detected in the “**Service**” aspect, while the green highlight is the correct answer. The review was labeled by the annotator in the “Meal” aspect and the system could not identify correctly. There have been not any studies which explore the categorization of five hotel aspects with few errors in the system identification.

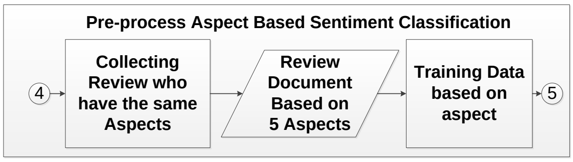

Pre-process Aspect Based Sentiment Classification
4
Collecting Review who have the same Aspects
Review Document Based on 5 Aspects
Training Data based on aspect
5

Figure. 10 The process of pre-process aspect based sentiment classification

Received: March 25, 2019

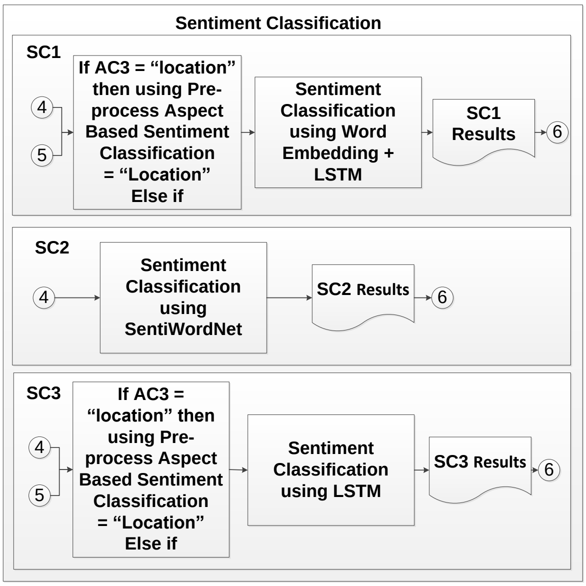

Sentiment Classification
SC1
4
5
If AC3 = "location" then using Pre-process Aspect Based Sentiment Classification = "Location" Else if
Sentiment Classification using Word Embedding + LSTM
SC1 Results
6
SC2
4
Sentiment Classification using SentiWordNet
SC2 Results
6
SC3
4
5
If AC3 = "location" then using Pre-process Aspect Based Sentiment Classification = "Location" Else if
Sentiment Classification using LSTM
SC3 Results
6

Figure. 11 The process of sentiment classification

3.5.2. Aspect categorization (AC2)

The process is the same with AC1, but the term list used would be added with 20% of the TF-ICF result. We used the 20% TF-ICF result to determine the accuracy of the data.

3.5.3. Aspect categorization (AC3)

The process is the same with AC1, but the term list used would be added with 100% of the TF-ICF result. We used the 100% TF-ICF result to increase the accuracy of the data.

3.6 Pre-process aspect based sentiment classification

Fig. 11 describes the whole process in the pre-process aspect based sentiment classification. We collected the term list document which had the same hotel aspects from AC3. The same hotel aspects were grouped according to the five hotel aspects. For example, a document that had a "Location" aspect would be categorized into the collection of location aspect. A term list document that had a "Meal" aspect would be categorized into the collection of meal aspect. A term list document that had a "Service" aspect would be categorized into the collection of service aspect. A term list document that had a "Comfort" aspect would be categorized into the collection of comfort aspect. A term list document that had a "Cleanliness" aspect would be categorized into the collection of cleanliness aspect.

The result of the pre-process aspect based sentiment classification is in the form of term list document based on the five hotel aspects. Moreover, we would use the data for training based on each aspect.

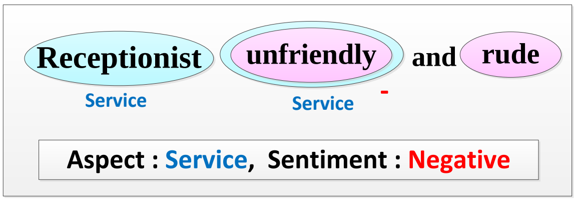


Figure. 12 Illustration of our sentiment classification using Word embedding + LSTM

Table 9. Illustration of aspect based sentiment classification

No	Document Term List	Aspects
1	'clean', 'swimming', 'pool'	Cleanliness
2	'staff', 'unable', 'prepare', 'cue', 'pool'	Service

3.7 Sentiment classification

Fig. 12 describes the entire process of sentiment classification with several approaches including SC1, SC2, SC3.

3.7.1. Sentiment classification (SC1)

Sentiment classification used the data in Table 7. to determine the positive and negative sentiment in each term list document. The illustration of sentiment classification is presented in Fig. 13

Fig. 13 illustrates how the word embedding + LSTM method works in the sentiment classification on the hotel aspect. The researcher used GloVe as word embedding to obtain vector value in increasing data accuracy.

Fig. 13 also explains that the result is categorized into the "**Service**" aspect and have Negative sentiment value, based on the word "**unfriendly**" and "**rude**". The sentiment classification can change depending on the result of the aspect. The example of the term list document in the aspect based sentiment classification is presented in Table 9.

The term list document in Table 9 would be extracted using Word embedding to obtain vector value. Result from vector value would be classified using the LSTM method to test in terms of sentiment classification.

Sentiment classification was conducted using the LSTM method with Eqs. (8), (9), (10), (11), (12), and (13) producing a vector score of 0 and 1. The number 0 indicates a negative sentiment (-) and 1 indicates a positive sentiment (+), which is explained in Table 10.

Table 10. Sentiment classification using Word Embedding + LSTM

Doc Term List	Similarity		LSTM	Label Annotator
	1 (+)	0 (-)		
'clean', 'swimming', 'pool', 'manhattan'	**0.673**	0.437	**1**	**1**
'staff', 'unable', 'prepare', 'cue', 'pool'	0.319	**0.824**	**0**	**0**
'receptionist, 'unfriendly', 'rude'	0.244	**0.717**	**0**	**0**
'poor', 'breakfast'	0.449	**0.539**	**0**	**0**

Table 10 used the term list document from the result of word embedding.

3.7.2. Sentiment classification (SC2)

The process of Sentiment Classification 2 (SC2) is the same with the process of SC1, but using SentiWordNet. SentiWordNet was used to compare the performance of the used method.

3.7.3. Sentiment classification (SC3)

The process of Sentiment Classification 3 (SC3) is the same with the process of SC1, but using LSTM. LSTM was used to compare the performance of the used method.

3.8 Evaluation

Evaluation was conducted using Precision, Recall, and F-1 Measure, each of which formula can be seen in Eqs. (17), (18), and (19).

4. Results and discussion

This section discusses the results and evaluation of the study.

4.1 Approach for aspect categorization

The researcher conducted performance test on each approach of the research method. The test was conducted to determine the best performance of the five hotel aspect categorization.

Table 11 discusses the approaches conducted by the researcher in performance evaluation of the five aspect categorization. The evaluation result can be seen in Table 12.

Table 11. Approach for aspect categorization

Aspect Categorization (AC)	Aspect Categorization Steps
Approach AC1	• PLSA method determines the hidden topic and produces term list document. Term list document is categorized into five hotel aspects using the Semantic Similarity method. Semantic Similarity is used to identify the similarity of the term list on the hotel aspects presented in Table 1.
Approach AC2	• AC1 + Expanded term list using TF-ICF, 20% from the total synonym.
Approach AC3	• AC1 + Expanded Term List using TF ICF, 100% from the total synonym.

Table 12. Aspect categorization performance evaluation

ASPECT CATEGORIZATION PERFORMANCE		
Aspect Categorization (AC)	**Evaluation**	**F1-Measure**
Approach AC1	PLSA + Semantic Similarity	0.736
Approach AC2	PLSA + TF ICF 20% + Semantic Similarity	0.777
Approach AC3	PLSA + TF ICF 100% + Semantic Similarity	**0.840**

Table 12 indicates the approaches evaluation result of the aspect categorization using PLSA + Semantic Similarity and PLSA + TF-ICF 100% + Semantic Similarity. The evaluation is reviewed from Precision, Recall, F1-Measure, using Eqs. (17), (18), and (19). Approach AC3 earns the best value. The highest score is 0.840, which indicates that the term list is descriptive, with good extraction aspect.

4.2 Approach for sentiment classification

The term list document compiled on each aspect would be classified using sentiment. Sentiment classification used two parameters: positive and negative. If there is an adjective (Adj.) of which sentiment value was neutral. the annotator will classify the word into positive or negative, based on each sentence. The system runs automatically from the training data labeled by the annotator. It is due to the fact that we would like to emphasize the result into "satisfied" or "dissatisfied".

Table 13. Approach for sentiment classification

Sentiment Classifiction (SC)	Sentiment Classification Steps
Approach SC1	• Data set of AC3 is used for SC1 approach. Data set is classified using Word Embedding + LSTM, which results in hotel sentiment.
Approach SC2	• SC1, but the classification using SentiWordNet.
Approach SC3	• SC1, but the classification using LSTM.

Table 14. Sentiment classification performance

SENTIMENT CLASSIFICATION PERFORMANCE			
Evaluation	Precision	Recall	F1-Measure
Approach SC1	**0.932**	**0.960**	**0.946**
Approach SC2	0.909	0.928	0.918
Approach SC3	0.919	0.929	0.924

Table 13. describes approaches conducted by the researcher in sentiment classification. The evaluation result is presented in Table 14.

Table 14 describes the evaluation result of the sentiment classification using SentiWordNet and Word Embedding + LSTM. The sentiment classification evaluation is reviewed from Precision, Recall, and F1-Measure elaborated in Eqs. (17), (18), and (19). Approach SC1 earns the best value. The system accuracy earns the highest score of 0.932; the score of system success or Recall is 0.960; and the F1-Measure score is 0.946.

4.3 Selection of a research approach

We used the best performance from each approach which had been evaluated. For approach for aspect categorization, it is found that the AC3 approach has a superior value. Thus, we used the AC3 approach to classify each term list document in five hotel aspects.

The approach for sentiment classification (SC1) has a superior value. So, the SC1 approach was used for sentiment classification. The selection result of each approach is presented in Table 15.

4.4 Evaluation of sentiment analysis based on hotel aspect

The researcher conducted evaluation using the best approach. The result of five aspect categorization and sentiment classification is presented in Table 16.

Table 15. Aspect based sentiment classification

Aspect Based Sentiment Classification	
Approach AC3	Approach SC1
PLSA would identify hidden topic and produce term list document. Term list document is categorized into five hotel aspects using Semantic Similarity. Semantic Similarity is used to identify the similarity of term list to hotel aspects presented in Table 1 + Expanded term list using TF-ICF, 100% of total synonym.	Data set of AC3 is used for SC1 approach. Data set is classified using Word Embedding + LSTM, which results in hotel sentiment.

Table 16. Sentiment classification based on hotel aspects

Hotel Aspects	Sentiment	Evaluation Results
Location	Positive	18.416
	Negative	0.594
Meal	Positive	1.782
	Negative	0.594
Service	Positive	45.545
	Negative	3.762
Comfort	Positive	11.683
	Negative	12.871
Cleanliness	Positive	3.168
	Negative	1.584

Table 16 shows a systemic conclusion that the highest positive sentiment lies in the “Service” aspect, with the value being 45.545. The highest negative sentiment lies in the “Comfort” aspect, with the value being 12.871. The lowest positive sentiment is the “Meal” aspect, with the value being 1.782 and the lowest negative sentiment in the “Location” and “Meal” aspects, with the value being 0.594.

4.5 Effect of sentiment on aspects

The sentiment classification of aspects is considered to be highly important. Errors in aspect categorization can affect the result of sentiment. In this study, we found examples of product reviews regarding the importance of sentiment on aspects.

Table 17. Effect of sentiment on aspects

Effect of Sentiment on Aspects			
Reviews	**Aspects**	**Sentiments**	**Results**
a clean swimming **pool** at Manhattan	**Cleanliness**	**Positive**	Customers are satisfied with hotel cleanliness
The Staff unable to prepare cue **Pool**	**Service**	**Negative**	Customers are dissatisfied with the hotel service.

The following is an example of the importance of sentiment on aspects presented in Table 17.

The result of the effect of sentiment on the five hotel aspects in Table 17 is obtained from Table 8 and Table 10. The “a clean swimming **pool** at Manhattan” and “The Staff unable to prepare cue **Pool**” contained the same word, “**pool**”. The same word in different sentences can mean differently in terms of aspect and sentiment. The sentence on “a **clean swimming pool** at Manhattan” as a whole, is detected as “Cleanliness” aspect and the review has a positive sentiment. The sentence on “The Staff **unable** to prepare **cue Pool**” is detected as “Service” aspect and the review has negative sentiment.

Both reviews contain word of which meaning can change, depend on the context. Thus, sentiment can change due to aspects. Errors in detecting aspects can lead to different perceptions for consumers and business managers in understanding customer needs.

5. Conclusion

This study contributes to other studies in understanding the importance of sentiment on aspects.

The result showed that the combination of the PLSA + TF ICF 100% + Semantic Similarity method was superior are 0.840 in the fifth categorization of the hotel aspects (AC3); the Word Embedding + LSTM method outperformed the sentiment classification (SC1) at value 0.946. The result of this study indicates that the “Service” aspect received positive sentiment value higher are 45.545 than the other aspects; the “Comfort” aspect has a higher are 12.871 negative sentiment value than the other aspects.

Good hotel service does not guarantee the customers’ convenience. So, hotel management can improve deficiencies to make customers more comfortable. Other results also showed that sentiment was affected by the aspects. Errors in detecting aspects can lead to different perceptions for consumers and business managers in understanding customer needs. Such a fact can contribute to business practitioners in improving hotel quality and services. Business practitioners can easily understand customer needs and appreciating timeliness in the service for company sustainability in the industrial era 5.0. In addition, this study contributes to the Business Intelligence and Analytics by proposing a new approach to represent customer perceptions.

Future study is expected to adopt a semantic approach in categorizing the five hotel aspects, and a better sentiment classification, since wellconducted study is one which can provide opportunities for other researchers to develop the topic further. Further research is also expected to be able to see more details about which aspects can affect customer satisfaction (For example, more aspects or see from different aspects).

References

[1] V. Özdemir and N. Hekim, “Birth of Industry 5.0: Making Sense of Big Data with Artificial Intelligence, “The Internet of Things” and Next-Generation Technology Policy”, OMICS: *International Journal of Integrative Biology*, Vol.22, No.1, pp.1-12, 2018.

[2] K. Berezina, A. Bilgihan, C. Cobanoglu, and F. Okumus, “Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews”, *International Journal of Hospitality Marketing & Management*, Vol.25, No.1, pp.1-24, 2016.

[3] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics”, *International Journal of Information Management*, Vol.35, No.2, pp.137-144, 2015.

[4] X. Xu, X. Wang, Y. Li, and M. Haghighi, “Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors”, *International Journal of Information Management*, Vol.37, No.6, pp.673-683, 2017.

[5] Suhariyanto, A. Firmanto, and R. Sarno, “Prediction of Movie Sentiment Based on Reviews and Score on Rotten Tomatoes Using SentiWordNet”, In: *Proc. of International Conf. on Application for Technology of Information and Communication*, pp.202-206, 2018.

[6] B. S. Rintyarna, R. Sarno, and C. Fatichah, “Enhancing the performance of sentiment

analysis task on product reviews by handling both local and global context", *International Journal of Information and Decision Science*, Vol.11, 2018.

[7] D. A. K. Khotimah and R. Sarno, "Sentiment Detection of Comment Titles in Booking.com Using Probabilistic Latent Semantic Analysis", In: *Proc. of International Conf. on Information and Communication Technology*, pp.514-519, 2018.

[8] R. P. Schumaker, A. T. Jarmoszko, and C. S. Labedz, "Predicting wins and spread in the Premier League using a sentiment analysis of twitter", *International Journal of Information and Decision Support Systems*, Vol.88, pp.76-84, 2016.

[9] C. C. Hnin, N. Naw, and A. Win, "Aspect Level Opinion Mining for Hotel Reviews in Myanmar Language", In: *Proc. of International Conf. on Agents*, pp.132-135, 2018.

[10] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment", In: *Proc. of IEEE Second International Conf. on Big Data Computing Service and Applications*, pp.52-57, 2016.

[11] Z. Xiang, Z. Schwartz, J. H. Gerdes, and M. Uysal, "What can big data and text analytics tell us about hotel guest experience and satisfaction?", *International Journal of Hospitality Management*, Vol.44, pp.120-130, 2015.

[12] L. P. Kaspa, V. N. S. S. Akella, Z. Chen, and Y. Shi, "Towards Extended Data Mining: An Examination of Technical Aspects", In: *Proc. of International Conf. on Computer Science*, Vol.139, pp.49-55, 2018.

[13] A. R. Baskara, R. Sarno, and A. Solichah, "Discovering traceability between business process and software component using Latent Dirichlet Allocation", In: *Proc. of International Conf. on Informatics and Computing*, pp.251-256, 2016.

[14] F. Revindasari, R. Sarno, and A. Solichah, "Traceability Between Business Process and Software Component using Probabilistic Latent Semantic Analysis", In: *Proc. of International Conf. on Informatics and Computing*, pp.3-8, 2016.

[15] D. Aliyanto, R. Sarno, and B. S. Rintyarna, "Supervised Probabilistic Latent Semantic Analysis (sPLSA) for Estimating Technology Readiness Level", In: *Proc. of International Conf. on Information and Communication Technology and System*, pp.79-83, 2017.

[16] B. Rintyarna, R. Sarno, and A. L. Yuananda, "Automatic ranking system of university based on technology readiness level using LDA-Adaboost.MH", In: *Proc. of International Conf. on Information and Communications Technology*, pp.495-499, 2018.

[17] G. Anugrah and R. Sarno, "Business Process model similarity analysis using hybrid PLSA and WDAG methods", In: *Proc. of International Conf. on Information & Communication Technology and Systems*, pp.231-236, 2016.

[18] C. Gao and J. Ren, "A topic-driven language model for learning to generate diverse sentences", *International Journal of Neurocomputing*, Vol.333, pp.374-380, 2019.

[19] O. Araque, I. Corcuera-Platas, J. F. Sánchez-RadaCarlos, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications", *International Journal of Expert Systems with Applications*, Vol.77, pp.236-246, 2017.

[20] B. K. Reddy and D. Delen, "Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology", *International Journal of Computers in Biology and Medicine*, Vol.101, pp.199-209, 2018.

[21] S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks", In: *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conf. on Natural Language Processing*, pp.1556-1566, 2015.

[22] Pennington, R. Socher, and C. D. Manning, "Glove: global vectors for word representation", In: *Proc. of International Conf. on Empirical Methods in Natural Language Processing*, pp. 1532-1543, 2014.

[23] D. Pham and A. Le, "Exploiting multiple word embeddings and one-hot character vectors for aspect-based sentiment analysis", *International Journal of Approximate Reasoning*, Vol.103, pp.1-10, 2018.

[24] P. Wang, B. Xu, J. Xu, G. Tian, C. L. Liu, and H. Hao, "Semantic Expansion using Word Embedding Clustering and Convolutional Neural Network for Improving Short Text Classification", *International Journal of Neurocomputing*, Vol.174, pp.806-814, 2016.

[25] E. Ekinci and S. I. Omurca, "An Aspect-Sentiment Pair Extraction Approach Based on Latent Dirichlet Allocation", *International Journal of Intelligent System and Applications in Engineering*, Vol.6, No.3, pp.209-213, 2018.

[26] G. Chen and L. Chen, “Augmenting service recommender systems by incorporating contextual opinions from user reviews”, *International Journal of User Modeling and User-Adapted Interaction*, Vol.25, No.3, pp.295-329, 2015.

[27] S. Yoo, J. Song, and O. Jeong, “Social media contents based sentiment analysis and prediction system”, *International Journal of Expert Systems with Applications*, Vol.105, pp.102-111, 2018.

[28] R. A. Stein, P. A. Jaques, and J. F. Valiati, “An analysis of hierarchical text classification using word Embeddings”, *International Journal of Information Sciences*, Vol.471, pp.216-232, 2019.

[29] N. Akhtar, N. Zubair, A. Kumar, and T. Ahmad, “Aspect based Sentiment Oriented Summarization of Hotel Reviews”, In: *Proc. of International Conf. on Computer Science*, Vol.115, pp.563-571, 2017.

[30] X. Wang, M. Chang, L. Wang, and S. Lyu, “Efficient algorithms for graph regularized PLSA for probabilistic topic modeling”, *International Journal of Pattern Recognition*, Vol.86, pp. 236-247, 2019.

[31] X. Wang, M. Chang, Y. Ying, and S. Lyu, “Co-regularized PLSA for multi-modal learning”, In: *Proc. of International Conf. on Artificial Intelligence*, pp. 2166-2172, 2016.

[32] Blokh and V. Alexandrov, “News Clustering based on similarity Analysis”, In: *Proc. of International Conf. on Computer Science*, Vol.122, pp.715-719, 2017.

[33] P. Kaspa, V. N. S. S. Akella, Z. Chen, and Y. Shi, “Towards Extended Data Mining: An Examination of Technical Aspects”, In: *Proc. of International Conf. on Computer Science*, Vol.139, pp.49-55, 2018.

[34] Wang, B. Peng, and X. Zhang, “Using a stacked residual LSTM model for sentiment intensity prediction”, *International Journal of Neurocomputing*, Vol.322, pp.93-101, 2018.

[35] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, “Sentiment analysis based on improved pre-trained word embeddings”, *International Journal of Expert Systems with Applications*, Vol.117, pp.139-147, 2019.

[36] R. Fernandez-Beltran and F. Pla, “Latent topics-based relevance feedback for video retrieval”, *International Journal of Pattern Recognition*, Vol.51, pp.72-84, 2016.

[37] G. Rao, W. Huang, Z. Feng, and Q. Cong, “LSTM with sentence representations for document-level sentiment classification”, *International Journal of Neurocomputing*, Vol.308, pp.49-57, 2018.

[38] X. Fu, X. Sun, H. Wu, L. Cui, and J. Z. Huang, “Weakly supervised topic sentiment joint model with word embeddings”, *International Journal of Knowledge-Based Systems*, Vol.147, pp.43-54, 2018.

[39] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, “Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review”, *International Journal of Expert Systems with Applications*, Vol.118, pp.272-299, 2019.

[40] A. Chaudhuri and S. K. Ghosh, “Sentiment Analysis of Customer Reviews Using Robust Hierarchical Bidirectional Recurrent Neural Network”, *International Journal of Artificial Intelligence Perspectives in Intelligent Systems*, Vol.464, pp.249-261, 2016.