Opinion Mining from Online Reviews in Bali Tourist Area

Puteri Prameswari, Isti Surjandari, Enrico Laoh
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Depok 16424, Indonesia
puteri.prameswari@yahoo.com, isti@ie.ui.ac.id, enrico.laoh@ui.ac.id

***Abstract*—Bali Island is the most popular tourist destination in Indonesia. Bali needs to make continuous quality improvements of its tourism industry by devoting particular attention to the hotel as an integral part of tourism. Through hotel user reviews, hotel managers gained insight about the hotel condition that was perceived by the users. based on online reviews in Tripadvisor.com, this study used text mining approach and aspect-based sentiment analysis to obtain hotel user opinion in the form of sentiment. Aspect-based sentiment analysis is able to provide information that is not provided by the typical sentiment analysis. To perform these tasks, this study tries to apply the Recursive Neural Tensor Network (RNTN) algorithm, which was commonly used for classifying sentiment in sentence level. With the average accuracy of 85%, the proposed algorithm performed well in classifying the sentiment of words or aspects. Moreover, the output can be used for evaluation in improving the quality of the hospitality industry as well as supporting the tourism industry in Indonesia.**

Keywords—aspect-based sentiment analysis; opinion mining; hospitality industry; text mining; RNTN

I. INTRODUCTION

Bali Island in Indonesia is a famous tourist destination, which renowned for its beautiful beaches, cultural, culinary and tourist objects. In 2016, Bali was awarded as the best island in Asia by one of the world's largest travel website, Tripadvisor.com. The assessment for the award was based on the presence of all tourism components that world travelers love such as hotels, restaurants, and tourist attractions.

The number of foreign tourists visiting Bali has been increasing for years as seen in Fig. 1. It contributes nearly 50% of the total foreign tourists visiting Indonesia. A lot of tourist arrivals was also supported by the adequate infrastructure and hotel facilities therein. Bali has 1,113 hotels as seen from Tripadvisor.com.

Although it has been known as the icon of world travel, Bali needs to make continuous improvements to the quality of its tourism industry by devoting particular attention to the hotels as an integral part of tourism. According to Linda [1], hotels does not only symbolize a city or simply the accommodation unit, but it also represents the attraction in a tourist. Hotels can also be regarded as a tourist attraction because of its main products, which are services and facilities, had a role in determining the whole traveler experiences. In line with the government's plan to bring in more tourists to Indonesia in 2019, hotels in tourist areas need to be prepared for fulfilling tourists' needs. To improve and maintain the quality of service and user loyalty, hotels must provide services that met or exceeded user's expectations [2], thus they need hotel user opinions. Hotel user opinions are needed to provide an overview for hotel managers about the current condition of the hospitality industry in Bali.

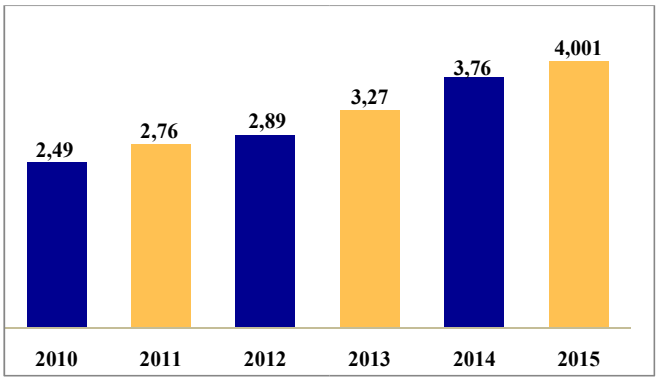

Fig. 1. The growth of foreign tourists visiting Bali.

The development of communication technology and website in tourism domain represented by the emergence of online tourism forum. These forums serve as the primary tools for the traveler to search for travel information [3]. A number of online reviews, which shared through travel forums are increasing every day due to the growing number of travelers who are willing to share their travel experiences. There are several websites that act as a discussion forum as well as a place for tourists to view and write online reviews such as Tripadvisor, Yelp, Citysearch, and Virtualtour. These forums gained popularity among worldwide traveler as a leading source of information in the field of tourism and hospitality [4]. As a data source of this research, Tripadvisor.com is one of the biggest travel forums that provide freely accessible online reviews, accommodating more than 25,000 reviews for all hotels in Bali.

Considering the very large number of text reviews, it needs an efficient way to gain knowledge from these texts [5]. Text mining approach became more useful since it offers potential solutions for handling unstructured data in textual form with

978-1-5090-5864-8/17/$31.00 ©2017 IEEE

large volume [6]. The application of text mining in tourism domain has been done recently. Li et al. used the same data source to identify user preferences today towards aspects related to service and facility of the hotel by using Emerging Pattern Mining (EPM) approach. The other studies were seeking for the dimension of words that are considered as sensitive and important factors that affect hotel user satisfaction level [8-9]. The other technique proposed by [3] combined sentiment analysis and text summarization approach to obtain summarization from 900 hotel user reviews. This study used text-mining approach to automatically process a large amount of text data so it can be treated as numerical data. Furthermore, this study utilized sentiment analysis to obtain the hotel user opinion about the services and facilities of the hotel. Sentiment analysis or known as opinion mining is a discipline that combines the process of information retrieval, text mining, and computational linguistics to detect the opinions expressed in the text. The main tasks in sentiment analysis include the identification and classification of opinions to positive, negative, or neutral [5].

Sentiment analysis has three research directions, which are document level, sentences level, and aspect level. Based on Hu and Liu's research, both document and sentences level of sentiment analysis, only showing the orientation of sentiment for each document or sentences but could not find the feature that writers like or dislike in detail [10]. Whereas the aspect level (aspect-based) sentiment analysis, which used in this study, is able to provide information that is not provided by the other levels of sentiment analysis. Taylor et al. used the same technique in the same field of the hospitality industry [11]. A coding scheme was established to obtain user opinion towards hotel aspects. Their work extended [10] using Python programming language.

This study aims to obtain hotel user opinion in the form of sentiment towards the services and facilities of the hotel in Bali through Recursive Neural Tensor Network (RNTN) algorithm, which mainly used for sentence-based sentiment analysis [12]. The main contribution of this study is to provide inputs for the hospitality industry in Bali, Indonesia, through a unique method that has never been applied in the field before.

II. Research Methodology

This section describes the data collection process to conclusion. This study divided into four principal steps including data collection, text pre-processing, determining hotel aspects, and analyzing review texts using aspect-based sentiment analysis.

A. Data Collection

Hotel user reviews in Bali were downloaded via Tripadvisor.com, a travel website which displays millions of online reviews from all travelers worldwide. The entire reviews were collected using an automated program. The program extracted all data from a web page or referred as web scraping. Web scraping is the process of retrieving a semi-structured document from the Internet, usually in the form of web pages in a markup language like HTML or XHTML, and analyzes these documents to obtain data from a specific page. The scraping framework on the web scraper application was constructed to learn HTML document and navigation techniques on the respective websites [13]. Python programming language was used to create scripts in this study.

B. Text Pre-processing

The text reviews that have been downloaded require a series of text pre-processes before they can be analyzed using text mining. Text pre-processing consists of several stages, which applied in accordance with the needs of the research. Specifically, the stages of text pre-processing include spelling normalization, filtering, case folding, lemmatization, and sentence boundary detection for this study. All stages were conducted through a text-mining program that was created for the purposes of this study. Table I described the stages in the text pre-processing.

TABLE I. Text Pre-processing Phase

No.	Phases
1	*Spelling Normalization,* the process of repairing misspelled words.
2	*Filtering,* the process of removing meaningless words and punctuations.
3	*Case Folding*, the process of changing letters in the document into the same form, e.g. upper cases to lower cases or vice versa.
4	*Sentence Boundary Detection,* the separating process of sentences in a paragraph or document.

C. Determining Hotel Aspects

This stage aims to obtain the hotel's services and facilities aspects which written in hotel user reviews, still with text mining approach. This study adopted the rules from previous studies by Hu and Liu, which only used the noun to represent hotel services and facilities that appear more than 1% from the overall reviews [10]. Table II described the stages in determining aspects of the hotel.

TABLE II. Stages of Determining Aspects

No.	Phases
1	*POS Tagging,* the process of determining the grammatical functions of words in a sentence. This stage aims to find nouns.
2	*Word Indexing,* the process of indexing nouns contained in the review texts. The selected nouns are those that appear more than 1% (minimum support) of the review sentences.
3	*Lemmatization*, the process of finding the basic form of the noun. This phase also synchronizes words that have same the meaning into one diction, e.g. meal and food.
4	*Taxonomy Formulation,* the grouping of nouns into categories that represent hotel services and facilities. The taxonomy consist of 8 categories namely (1) Accessibility, (2) Activities & Entertainment, (3) Food & Beverages Operations, (4) Guests' Perspective, (5) Human Resource, (6) Room Amenities, (7) Transportation Services, and (8) Physical Environment.

D. Aspect Based Sentiment Analysis

The next step is to determine the sentiment of every aspect that has been defined in the previous stages using opinion mining approach. The aspect-based sentiment classification was

done with the help of text mining program called Prameswari v1.4.0. This program used the principle of Sentiment Treebank created by Stanford University and worked based on RNTN model by Socher et al. [12].

Socher et al. used RNTN algorithm to classify sentiment orientation from movie reviews by Pang and Lee [14]. Whereas in this study, the same algorithm was developed so it can be applied to the Indonesian hospitality industry. The classification function was also upgraded to classify the sentiment in aspect level.

When processing a text review, the program will break down the words from a sentence and compile them into a tree, as shown in Fig. 2.

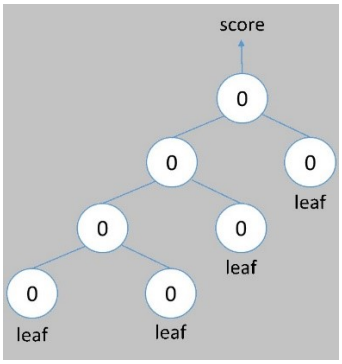

Fig. 2. Components in Parse Tree.

The database on Sentiment Treebank was specifically created for sentiment analysis. Thus, there were lists of words making up positive or negative sentiment, which called opinion words inside the database. The parse tree structure detects opinion word that carries positive or negative sentiment. Then the opinion word will move towards the root, which is the hotel aspect that has been defined in the previous stage. Fig. 3 showed the sentiment calculations performed by the model.

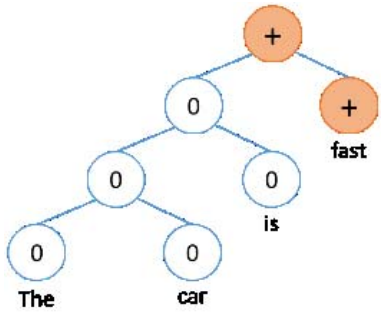

Fig. 3. Parse Tree from Sentence "The car is fast".

III. Result and Analysis

Aspects of hotel services and facilities that were selected are grouped into eight categories as shown in Table III and Table IV. Furthermore, every aspect will be going through sentiment classification. The results of aspect-based sentiment analysis were presented in the form of bar chart, where blue color indicates positive sentiment and red color otherwise, as shown on Fig. 4. Data were displayed in percentage to simplify the analysis process.

TABLE III. Services and Facilities Aspects of Hotel in Bali

Categories			
Food & Beverage Operations	***Accessibility***	***Human Resource***	***Activities & Entertainments***
food	location	front desk	swim
drink	distance	staff	travel
cocktail	main road	owner	shop
buffet	beach	reception	rest
menu	town	manager	spa
fruit	airport	chef	sunset
breakfast		driver	massage
lunch		management	club
dinner		hospitality	yoga
restaurant		service	
bar			
café			

TABLE IV. Services and Facilities Aspects of Hotel in Bali (Cont'd)

Categories			
Guests' Perspective	***Transportation Service***	***Room Amenities***	***Physical Environment***
experience	car	bedroom	lobby
atmosphere	taxi	bathroom	sea view
accommodation	shuttle	bed	sea front
value		wifi	place
privacy		air conditioner	garden
variety		balcony	swimming pool
selection		water	place
style		towel	lounge
price		fridge	toilet
expectation		kitchen	
security			

The sentiment graphs were obtained by summing up all the results of the sentiment analysis. As shown in Fig. 4, activities and entertainment, food and beverage operations, room amenities, transportation services, and physical environment of the hotel were the categories where the negative reviews outnumbered the positive reviews. Overall, the hotel users in Bali were not satisfied with the five categories of the hotel services and facilities.

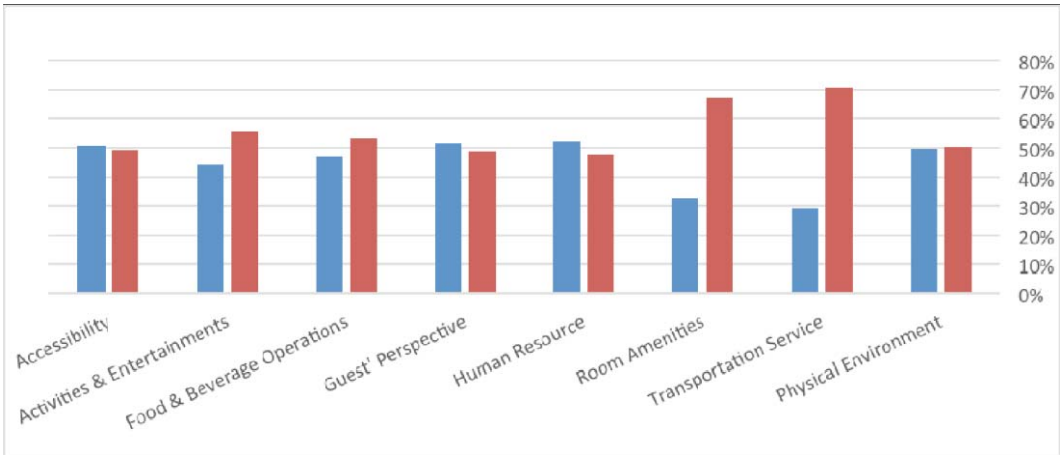

Fig. 4. Sentiment Analysis Results from Hotel User Reviews in Bali.

Each category consists of aspects, in which these aspects have an influence on the resulting sentiment. The more occurrence of the aspect in the document, the greater the contribution to the analysis results. The high frequency of the word as an aspect also indicated that these aspects were important by the hotel users.

In the category of activities & entertainment, the most mentioned aspect in reviews was "rest", while in the category of food & beverage operations was "restaurant". The category of room amenities was largely influenced by aspect "bed". In text reviews, aspect "bed" was often associated with insect harassment. Most hotel users found this problem can be overcome by placing the mosquito net on the bed. Moreover, the most negative aspect in physical environment category was aspect "place". For the last category that is transportation services, hotel users in Bali often complained about the aspect "car". The predominance complaints came from the traffic jam especially in the peak season. The number of tourists visiting Bali, especially in the southern part of Bali was already exceeded the capacity or can be regarded as overload.

IV. Conclusion

As shown in Table V, the classifier model used in this study had an average accuracy of 85%, while the average value of F1 Measure is 77%. The positive category has the highest value of F1 Measure that is 90%. This suggests that the model works best in categorizing aspects to positive sentiment. Meanwhile, the average value obtained for categorizing negative sentiment was 64%. According to Guo, Barnes, and Jia, a text-based classification accuracy is affected by several factors, including the size of the identified text fragments, the amount of training data, classification features, algorithms, and the similarity of the word [9].

TABLE V. Confusion Matrix of The Classifier Model

Observed	**Predicted**		**Total**
	Positive	***Negative***	
Positive	74	16	90
Negative	0	14	14
Total	74	30	104
Positive	Precision	100%	
	Recall	82%	
	F1	90%	
Negative	Precision	47%	
	Recall	100%	
	F1	64%	
F1 Measure		77%	
Accuracy		85%	

The success of Bali as a world tourism icon was inseparable from the role of the hotel as the main tourism superstructures. Therefore, the hotels need to be maintained by continuously evaluating the quality, taking into account the voice of customers from hotel users towards the services and facilities that they perceived.

This study utilized RNTN algorithm, which usually used for sentence-level sentiment analysis in previous studies, yet the result showed that RNTN functioned properly in classifying the sentiment of words or aspects.

Based on the computation performed in this study, five of the eight categories related to hotel services and facilities in Bali reaped negative sentiment. Negative sentiment showed dissatisfaction, disappointment, as well as incompatibility of hotel services and facilities with user expectations. Consequently, hotel managers need to prioritize improvements to the categories with the predominance of negative sentiment.

The data collection process through web scraping has not considered the upload time of the online reviews, so there are no visible changes from time to time. Future research may limit the upload time of the reviews to gain deeper analysis.

ACKNOWLEDGEMENT

The authors would like to express their gratitude and appreciation to Universitas Indonesia for financing this study through Thesis Research Grant for Indexed International Publication No. 2134/UN2.R12/HKP.05.00/2016 and Sangadji Prabowo for the development of text mining software, Prameswari v1.4.0.

REFERENCES

[1] Y.H. Hu, Y.L. Chen, and H.L. Chou, "Opinion mining from online hotel reviews – A Text Summarization Approach," *Information Processing and Management*, vol. 53, pp. 436–449, 2017.

[2] Z. Liu and S. Park, "What makes a useful online review? Implication for travel product websites," *Tourism Management,* vol.4, pp. 140-151, 2015.

[3] K. Khan, et al., "Mining opinion components from unstructured reviews: a review," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, pp. 258-275, 2014.

[4] N. Ur-Rahman and J. Harding, "Textual data mining for industrial knowledge management and text classification: a business oriented approach," *Expert Systems with Applications*, vol. 39 (5), pp. 1-11, 2013.

[5] G. Li, et al., "Identifying emerging hotel preferences using emerging pattern mining technique," *Tourism Management*, vol.46, pp. 311-321, 2015.

[6] Z. Xiang, et al., "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.

[7] Y. Guo, S. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist Satisfaction Analysis Using Latent Dirichlet Allocation," *Tourism Management,* vol. 59, pp. 467-483, 2017.

[8] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *The 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2004, pp. 168-177.

[9] E.M. Taylor, J. D. Velasquez, and F. B. Marquez, "A novel deterministic approach for an aspect-based opinion mining in tourism product reviews," *Expert System with Application,* vol. 41, pp. 7764-7775, 2014.

[10] R. Socher, et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Empirical Methods in Natural Language Processing*, Palo Alto, 2013. [Online]. http://nlp.stanford.edu/sentiment/.

[11] A. Josi, L. A. Abdillah, and Suryayusra, "Penerapan teknik w*eb scraping* pada mesin pencari artikel ilmiah," in *Jurnal Sistem Informasi (SISFO),* vol. 5, pp. 159 – 164, 2014.

[12] B. Pang and L. Lee, "Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales," in *The 43rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, Pennsylvania, 2005, pp. 115-124.

[13] R.A. Linda, "Hotels and their impact on the tourism of Transylvania," *GeoJournal of Tourism and Geosites*, vol. 12, pp. 163174, 2013.

[14] P. Kotler and K.L. Keller, *A Framework for Marketing Management*, 4th ed. New Jersey: Pearson Education Inc., 2009.