KOMPUTASI: JURNAL ILMIAH ILMU KOMPUTER DAN MATEMATIKA
20 (2)(2023), 93-100 p-ISSN: 1693-7554, e-ISSN:2654-3990

Judul Modeling LDA and SVM in Sentiment Analysis of Hotel Reviews

Erniyati[1,*], **Prihastuti Harsani**[2], **Mulyati**[3], **Lutfi Dani Fahriza**[4]

[1,2,3,4]Department of Computer Science, Faculty of Mathematics and Natural Science, Pakuan University, Bogor, West Java, 16143, Indonesia

Abstract

The number of visitor comment review data that enters the TripAdvisor and Agoda sites continues to grow over time, this makes it difficult for the hotel to obtain overall information from all comment reviews. Therefore, the purpose of this study is to apply topic modeling and classifying in the analysis of hotel service sentiment. The data for comment reviews were obtained from 3 five-star hotels, namely 1-HTL, 2-HTL and 3-HTL. The hotel has a five-star rating and has the most comments compared to other hotels in Jakarta. The topic modeling method using Latent Dirichlet Allocation (LDA) in this study succeeded in dividing the comments into several topics that were often discussed from Indonesian and English comments regarding the hotel services provided. By using Support Vector Machine (SVM) obtained the number of positive, negative and neutral comments. The stages of this research include data selection, preprocessing (tokenization, cleansing, filtering and stemming), transformation and reduction, data mining, interpretation and evaluation. The research data used is comment reviews from the TripAdvisor site 2575 comments in Indonesian and 2313 comments in English and from the Agoda website 2795 data for comments in Indonesian and 4562 comments in English. The results of the sentiment analysis obtained an average value on the Tripadvisor site is comments in Indonesian with a precision of 70.37%, accuracy of 92.75%, recall of 99.49%, and comments in English with a precision 94.19%, accuracy 97.77%, and recall 89.28%, while on the Agoda site comments are in Indonesian with precision 19.65%, accuracy 93.37%, recall 90.66%, and comments in English with precision 91.19%, accuracy 97.56%, recall 85.70%.

Keywords: *Hotel; Sentiment Analysis; Topic Modeling; Latent Dirichlet Allocation; SVM*

1. Introduction

The hospitality industry has now been facilitated with many online services engaged in tourism that can assist in promotion and marketing. The number of online sites that provide services in the form of online bookings, both airline tickets and hotels, has become a driving force for the world of tourism, accommodation, and also other fields that require services from these sites. In line with the rapid development of online media or technology, guest comments or complaints can now be made and seen by many people through social media and other online media. World tourism managers in providing more detailed information about the tourism products offered. At this time before ordering, someone will check the opinion of the existing web reviews. Hotels are one of the products of tourism that are highly considered both in terms of facilities, services or travel distances [1].

[*]**Corresponding author:** ***E-mail adress:*** *neni_erniyati@unpak.ac.id*
Received: 5 May 2023, Accepted: 11 July 2023 and available online 31 July 2023
https://doi.org/10.33751/komputasi.v20i2.7604

The TripAdvisor and Agoda sites provide various facilities that can make it easier for visitors to obtain detailed information about the hotel to be visited, both location, number of rooms, and various other facilities available. In addition, this site is visited a lot because it provides information about the reviews of tourists. The number of visitor comment review data that enters the TripAdvisor and Agoda sites continues to grow over time, this makes it difficult for the hotel to obtain overall information from all comment reviews, because it will take a long time to read each review that goes on the page one by one. TripAdvisor and Agoda sites. In this study, the LDA (Latent Dirichlet Allocation) is one of the topic modeling methods implemented to find out the topic groups of words [2] that often appear regarding incoming comments and identify comments. The comments are classified into positive, negative and neutral sentiments with the Support Vector Machines (SVM) method. SVM proved to be the best algorithm for text categorization [3].

Research conducted about sentiment analysis [4] of hotel reviews using naïve bayes classifier algorithm. The test results show that the sentiment classification using the Naïve Bayes Classifier obtains an accuracy of 90.61%, a precision of 93.03%, a recall of 89.52% and an f-measure of 90.99% [4]. Another research by Baskoro, et al (2021) on hotel customer sentiment analysis in Purwokerto using Random Forest and TF-IDF Methods (Case Study: Customer Reviews on the TripAdvisor Site). The results showed that the accuracy of the model reached 87.23%. However, without the stemming process, the model accuracy is only 76.07% [5]. Another research by Setiawan, et al (2021) about Hotel Sentiment Analysis in West Nusa Tenggara Using the SVM Algorithm. This study uses existing data on the traveloka site, preprocessing is carried out using a dictionary in the sastrawi library which has been improved by the author. The results of this study indicate the number of positive sentiments is 84.97% and negative sentiments 15.03%, with an accuracy value of 92.32%, precision of 93.34% and a recall value of 92.32% [6]. Morama [7] also conducted research on Aspect-based Sentiment Analysis of Hotel Tentrem Yogyakarta Reviews using the Random Forest Classifier Algorithm. The test results prove that the greater the number of trees and the depth of the trees, the better the prediction results. The best classification results for the two parameters for the room aspect are 90% for the accuracy value and the f1 score [7]. Another research [8] about sentiment analysis on Oakwood residence Cikarang hotel reviews on TripAdvisor website using k-Nearest Neighbor algorithm. The test results related to sentiment analysis with the k-NN algorithm get an average accuracy of k = 3 of 90% [8]. Thomas [9] also conducted research on Sentiment Analysis of Hotel Reviews in Indonesian Using the Support Vector Machine and TF-IDF. The data are entered into the machine learning process using Support Vector Machine (SVM) and obtained the accuracy of the model by 85%. For testing scenarios if not using slang handling get F1-Score by 80% and if not using stop word get F1-Score by 82%. On the evaluation of the performance of the model using K-Fold obtained the best results on the Fold-7 with a precision value of 87%, recall 86%, F1-score 86%, and accuracy of 87% [9]. Another research [10]. About Topic modeling and sentiment analysis about Mandalika on social media using the latent Dirichlet allocation method. The test results show that the SVM algorithm can classify sentiment toward the Mandalika Circuit well, as indicated by the measurement of the performance of the SVM algorithm, namely 87% accuracy, 77% precision, 84.81% recall, and 98.52% specificity. These results also show that the F1 Score compares the average precision and recall, which is weighted at 80.72% [10]. Dewi, et al (2023) also conducted research on Sentiment analysis aspects based on Hotel Customer reviews in Bali using the Decision Tree Method. The Decision Tree model for aspects produces performance, accuracy, precision, recall, and F1-Score, respectively 82,5%, 80%, 90,9%, and 85,1%, model for service aspect sentiment produces accuracy, precision, recall, and F1-Score, respectively which are 75%, 72,7%, 80%, and 76,2%, while model for the sentiment of cleanliness aspect produces performance of accuracy, precision, recall, and F1-Score, respectively which is 81,8%, 87,5%, 77,8%, and 82,4% [11]. Another research [12] about Sentiment Analysis of Opinions on the Use of Devices in Students Using the Support Vector Machine (SVM) Method. The result show that average value of accuracy in the distribution of training data and test data using k-fold cross validation of 10-fold value of 85.3%. From the test result the highest accuracy value 83.3% of the data successfully classified according to the actual class [12].

2. Methods

Research Methods applied in comment sentiment analysis five-star hotel service uses data mining stages or also called data mining. Knowledge Discovery and Data Mining (KDD) Data mining is a process that using statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and knowledge assembled from various large databases or data warehouses [6]. The system flowchart can be seen in Figure 1.

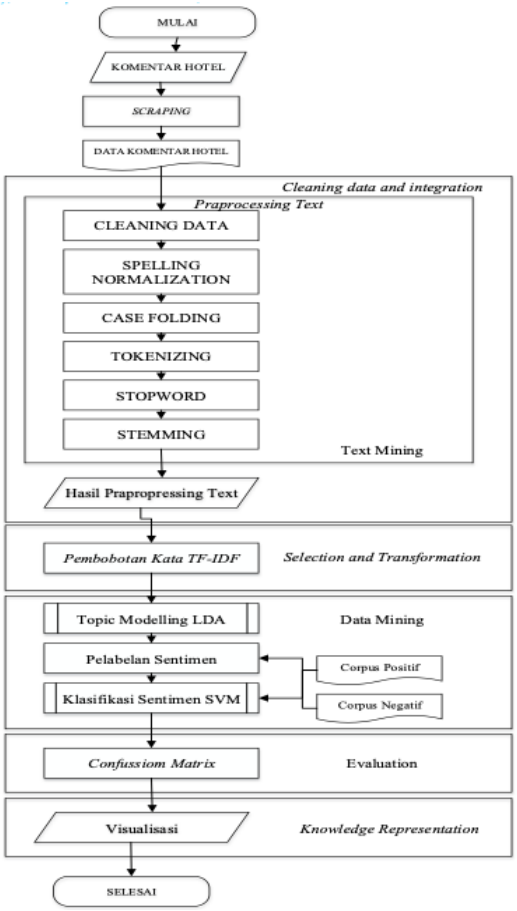

Figure 1. Flowchart System

2.1. Topic Modelling Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model of a collection of writings called corpus. The basic idea proposed by the LDA method is that each document is represented as a random mixture of hidden topics [14], where each topic has a character that is determined based on the distribution of the words contained in it [2].

LDA works with the input of individual documents and several parameters, to produce an output in the form of a model consisting of weights that can be normalized according to probability. This probability refers to two types, namely the type (a) the probability that a certain specific document produces a specific topic as well and the type (b) the probability that a certain specific topic produces specific words from a vocabulary collection. Probability of type (a), documents that have been labeled with a list of topics are often continued to produce probability of type (b), which produces certain specific words [15].

$$\text{Topic probability}(k, d, alpha = 0.1) = \frac{\text{Number of topics } k \text{ in document } d + alpha}{\text{document length } d + \text{number of topics} * alpha} \quad (1)$$

$$\text{Probability of a word } (t, k, beta = 0.1) = \frac{\text{Number of topics } t \text{ in document } k + alpha}{\text{document length } k + \text{number of distict} * alpha} \quad (2)$$

2.2. Support Vector Machine (SVM)

Support vector machine (SVM) is a classification technique. The basic principle of SVM is to find the best hyperplane or dividing plane based on the training data which can then classify the test data correctly. The dividing plane in the p-dimensional space is an affine subspace with p-1 dimension which divides the vector space into two classes, where p is the number of explanatory variables. The best dividing plane between the two classes can be found by finding the maximum point of the margin of the dividing plane. Margin is the distance between the dividing field and the closest observation of each class. The closest observation is referred to as a support vector. The steps in the SVM method according to [16]:

1. Determine the data point: $x_i = \{x_1, x_2, ..., x_n\} \in R_n$
2. Specifies the data class : $y_i \in \{-1, 1\}$
3. Data and class pairs : $\{(x_i, y_i)\}\, N_i = 1$
4. Maximize the following functions:
   $L_d = \sum \alpha_i n_i = 1 - \sum \sum \alpha_i \alpha_j . y_i y_j$ syarat $0 \leq a_i \leq C$ dan $\sum a_i n_i = 1 . y_i = 0$
5. Calculate the values of w and b:
   w $= \sum a_i n_i = 1 . y_i . x_i, b = -12(wx - 1)$
6. Classification decision function sign(f(x)):
   F(x)=w.x+b atau f(x)= $\sum a_i n_i = 1 . y_i K(x, y) + b$

3. Result and Discussion

This section describes the results discussed from Topic Modeling that can be used using the Topic Modeling Latent Dirichlet Allocation method obtained from hotel comments, and the results obtained from sentiment analysis on hotel comments are entered into positive, negative or neutral sentiments using the Support Vector Machine method.

3.1. Data Condition

Scraping data on comments from 3 five-star hotels in Jakarta named 1-HTL, 2-HTL, 3-HT obained from the Tripadvisor site from January 2006 – April 2020, while on the Agoda site from August 2008 – May 2020.

Table 1. Scraping data on comments from 3 five-star hotels

No	Hotel site	Hotel	Number of Comment	
			Indoensain language	English
1	Tripadvisor	1-HTL Jakarta	879	727
		2-HTL Jakarta	1044	960
		3-HTL Jakarta	652	626
		Total	2575	2313
2	Agoda	1-HTL Jakarta	55	721
		2-HTL Jakarta	72	665
		3-HTL Jakarta	93	863
		Total	220	2249
		Total	**2795**	**4562**

3.2. Prepossessing

This preprocessing stage aims to remove unnecessary parts contained in a document, where this will become noise in the next process. In the preprocessing process, the stages of cleaning data, spelling normalization, Case Folding, Tokenizing, Stop word, and Stemming.

3.3. LDA Modelling Topic Results

To find out the optimal number of topics, topic testing was carried out using the LDA is library, topic visualization with the LDA is library had many clusters that intersect, this indicates that clusters that are contiguous or adjacent can be one cluster. This research trial was conducted from hotel booking sites and divided into 2 languages, namely Indonesian and English can be seen in Table 2.

Table 2. Modelling topic results

1-HTL Jakarta (Tripadvisor)		**1-HTL Jakarta (Agoda)**	
Indonesia	English	Indonesia	English
3 topic modelling:	4 topic modelling	3 topic modelling	4 topic modelling
Topic 1 (Restoran)	Topic 1 (Star)	Topic 1 (Pelayanan)	Topic 1 (Room)
Topic 2 (Fasilitas)	Topic 2 (Staff)	Topic 2 (Kamar), dan	Topic 2 (Staff),
and Topik 3 (Staf)	Topik 3 (Room), and	Topik 3 (Staf)	Topik 3 (Restaurant)
	Topik 4 (Restaurant)		Topik 4 (Business)
2-HTL Jakarta (Tripadvisor)		**2-HTL Jakarta (Agoda)**	
Indonesia	English	Indonesia	English
3 topic modelling:	3 topic modelling	4 topic modelling	4 topic modelling
Topic 1 (Restoran)	Topic 1 (Restaurant)	Topic 1 (Room)	Topic 1 (Room)
Topic 2 (Fasilitas)	Topic 2 (Room), and	Topic 2 (Staff), dan	Topic 2 (Room),
and Topik 3 (Restaurant)	Topik 3 (Room), and	Topik 3 (Staf)	Topik 3 (Pool), and
		dan Topik 4 (Business)	Topik 4 (Business)
3-HTL Jakarta (Tripadvisor)		**3-HTL Jakarta (Agoda)**	
Indonesia	English	Indonesia	English
4 topic modelling:	3 topic modelling	3 topic modelling	4 topic modelling
Topic 1 (Kamar),	Topic 1 (Staff),	Topik 1 (Kamar),	Topic 1 (Restaurant)
Topic 2 (Layanan)	Topic 2 (Restaurant),	Topic 2 (Kamar), dan	Topic 2 (Facility)
Topik 3 (Staf), and	Topik 3 (Room)	Topik 3 (Ranjang)	Topik 3 (Room), and
Topik 4(Fasilitas)			Topik 4 (Staff)

In this study, topic testing was carried out using the LDAvis library to determine the number of topics obtained and topic models library to find out the distribution of words and model topics obtained. On the TripAdvisor site, 2-HTL Jakarta, the Indonesian language was tested with 1000 iterations and the topic testing was divided into 3 clusters and the results obtained were that there were no slices in each topic cluster. In addition, the results obtained are the distribution of words contained in each topic which will later be used to determine the topic according to the word distribution.

3.4. Sentiment Labeling Results

The classification process is carried out by studying the data pattern using training data. Training data in which there is positive review training data and negative review training data are used by the SVM Algorithm in studying patterns data based on the characteristics of the data in each class. The data used for training data and test data is data that already has a class label, with the amount of training data and test data having a ratio of 80%: 20%. Based on the Paretto Principle, the ratio commonly used is 80:20 for data sets training and testing. The ratio used in this study is 80:20.

3.5. Sentiment Analysis Results by Criteria

In accordance with the Ministry of Tourism and Creative Economy regulations in 2013 [11], regarding the criteria for five-star hotels, namely Product, Service and Management. The results of sentiment analysis in accordance with the topics that have been obtained from several Jakarta hotels from comments in Indonesian and English, can be seen in Table 3.

Table 3. Sentiment analysis results based on criteria

Criteria	Positive Sentiment Analysis	Negative Sentiment Analysis	Number of	
			Pos	Neg
1-HTL Jakarta (Tripadvisor)				
Product	Friendly, Helpful Staff, Strategic Location, Good facilities, Complete, Restaurants near, good.	Swimming Pool Too Small	195	16
Service	Business Area Good Meeting Place, good fit center, big	Disappointed in the restaurant staff	6	1
Management	Good location near shopping center, have shuttle bus	-	2	-
2-HTL Jakarta (Tripadvisor)				
Product	Complete facilities, fast internet, complete breakfast, friendly waiters, comfortable, spacious rooms. Friendly restaurant service, good service, friendly	Smell of cigarette smoke	223	1
Service	Staff service is friendly and polite	-	100	-
Management	Complete fitness center	-	1	-
3-HTL Jakarta (Tripadvisor)				
Product	The rooms are comfortable, nice, spacious, big, nice, luxurious and clean.	Old room	221	4
Service	Friendly, good and fast service. Nice comfortable atmosphere for office meetings	-	29	-
Management	Big, complete, good fit center. good spas	-	10	-
1-HTL Jakarta (Agoda)				
Product	Spacious, comfortable, nice, clean, big room. strategic location, good, clean, good staff	Small swimming pool	175	3
Service	Close to the mall, neat shuttle transportation service, good service	-	31	1
Management	Spacious fitness center	-	1	-
2-HTL Jakarta (Agoda)				
Product	The room is spacious, comfortable, big, nice, luxurious. good facilities, complete, luxurious. good restaurant food, polite staff, friendly. strategic location	The room smells of cigarettes	109	1
Service	The location is near the shopping center	-	8	-
Management	High standard business center, family business facilities	-	6	-
3-HTL Jakarta (Agoda)				
Product	Complete facilities	-	91	-
Service	Good service, varied breakfast, suitable for holding meetings with business partners	Lacking	22	1
Management	The health club lounge is complete, clean	Parking area is far	3	1

3.6. Sentiment analysis results Support Vector Machine

The results of the sentiment analysis obtained from the TripAdvisor and Agoda sites using SVM are: 3-HTL Jakarta from the TripAdvisor site Indonesian on topic 1 with 440 Positive, negative 32, and 17 neutral comments. English on topic 1 with the number of positive comments 342, negative 11, and neutral 9. The highest number of positive comments came from the Agoda site, in Indonesian comments, namely 1-HTL Jakarta from topic 1, with the number of positive comments 28, neutral 8, and negative 1. While in English comments, namely 2-HTL Jakarta on topic 2 with 326 positive comments, 13 neutral, and 3 negative comments.

4. Conclusion

From the research results obtained and presented, it can be concluded, using the LDA topic modeling method, this study succeeded in dividing the comments into several topics that were often discussed from the modeling topic obtained from Indonesian and English comments regarding the hotel services provided. The results of the sentiment analysis obtained in accordance with the criteria for five-star hotels, from the two sites and the three hotels, several comments on the same sentiment analysis were obtained on the two sites, including: Analysis of positive sentiment on 2-HTL Jakarta, namely good facilities, friendly, helpful staff, strategic location, spacious, comfortable, and clean rooms. At 3-HTL Jakarta, the facilities are complete, the staff is kind, polite and helpful. While 1-HTL Jakarta Complete facilities, nice, big swimming pool. the rooms are clean, comfortable and spacious. Good, friendly and fast service. As for the analysis of negative sentiments at 2-HTL Jakarta, customers complained about the small swimming pool, old room interiors. At 3-HTL Jakarta, customers complain about rooms that smell of cigarette smoke. while at 1-HTL Jakarta customers complain about the high prices for food. From the results of sentiment analysis obtained from the Tripadvisor website, the highest number of positive comments on Indonesian and English comments is on 3-HTL from the Tripadvisor website on topic 1 with the number of positive comments 440, 32 negative, and 17 neutral. topic 1 with 342 positive comments, 11 negative, and neutral 9. In addition to the results obtained from the agoda site, the highest number of positive comments on Indonesian comments is 1-HTL Jakarta from topic 1, with 28 positive comments, 8 neutral, and negative 1. While the English comments are 2-HTL Jakarta on topic 2 with a total of 326 positive comments, 13 neutral, and 3 negative comments. Future research n-gram and bi-gram method for lacking in defining the word negation in Indonesian and English comments.

5. Acknowledgement

Thank you to the Department of Mathematics for providing support and input in this research. Also, thank you very much to the Faculty of Mathematics and Natural Sciences (FMIPA) which has funded this research with the PNBP 2021 funding source, University of Riau.

References

[1] Marreese Taylor, E., Velasquez, J.D., Bravo-Marquez, F., Matsuo, Y. 2013. Identifying customer preferences about tourism products using an aspect - based opinion mining approach. Procedia Computer Science, 22. pp.182–191

[2] Chauhan, U., Shah, A. 2021. Topic modeling using latent Dirichlet allocation. ACM Computing Survey. 54 (7). Pp. 1-35.

[3] Rahat, AM., Kahir, A., Masum, AKM. 2019. Comparison of naive Bayes and SVM algorithm based on sentiment analysis using review dataset. In 8th International Conference on System Modeling and Advancement in Research Trends. pp. 266–270.

[4] Suryadi., Ridho A., Murhaban. 2021. Analisis Sentimen Review Hotel Menggunakan Algoritma Naïve Bayes Classifier. TECHSI. 13(2). Pp. 95-105. https://doi.org/10.29103/techsi.v13i2.5596.

[5] Baskoro, B., Susanto, I., Khomsah, S. 2021. Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs TRIPADVISOR). INISTA (Journal of Informatics Information System Software Engineering and Applications), 3(2), Pp. 21-29. https://doi.org/10.20895/inista.v3i2.218.

[6] Setiawan, RA., Estetikha, AKA., Nurharyanto, EMO, Asmara, Y., Wahyudi, A. 2021. Analisis Sentimen Hotel di Nusa Tenggara Barat Menggunakan Algoritma SVM. Prosiding. Seminar Multimedia and Artificial Intelligence: Optimalisasi Artificial Intelligence di Era Revolusi Industri 4.0 dan Society 5.0. Vol. 4. Pp. 149-155.

[7] Morama, HC.,Ratnawati, DE., Arwani, I. 2022. Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogyakarta menggunakan Algoritma Random Forest Classifier. Jurnal Pengembangan Teknologi Informasi dan Ilmu.. 6(4). Pp. 1702-1708.

[8] Nugraha, RN., Eviana, Trisnawati Y., 2022. Sentiment Analysis On Oakwood Residence Cikarang Hotel Reviews On Tripadvisor Website Using K-Nearest Neighbor Algorithm. Jurnal Inovasi Penelitian. 3 (6). Pp. 6495-6506.

[9] Thomas, VWD., Rumaisa, F. 2022. Analisis Sentimen Ulasan Hotel Bahasa Indonesia Menggunakan Support Vector Machine dan TF-IDF. Jurnal Media Informatika Budidarma. 6,(3). Pp. 1767-1774. https://doi.org/10.30865/mib.v6i3.4218.

[10] Hardita, VC., Hammad, R., Amrullah, AZ. 2022. Topic modeling and sentiment analysis about Mandalika on social media using the latent Dirichlet allocation method. Jurnal Manajemen Teknologi dan Informatika. 12(3). Pp. 109-116.

[11] Dewi, NPA., Sanjaya NA, Karyawati E., Mahendra IBM., Dwidasmara, IBG., Wibawa IGA. 2023 Sentiment analysis aspects based on Hotel Customer reviews in Bali using the Decision Tree Method. Jurnal Elektronik Ilmu Komputer Udayana. 11 (3).pp. 625-634.

[12] Zuhri, M., Qur'ania, A., Mulyati. 2023. Sentiment Analysis of Opinions on the Use of Devices in Students Using the Support Vector Machine (SVM) Method. KOMPUTASI: Jurnal Ilmiah Ilmu Komputer Dan Matematika Vol. 20 (1). Pp. 51-55

[13] Bambang Sugiantoro, Prasdika F. B. S. 2018. A review paper on big data and data mining. IJID International Journal on Informatics for Development. 7(1), Pp. 36-38 http://www.ieomsociety.org/detroit2020/papers/506.pdf

[14] D. M. Blei. 2003. Latent Dirichlet Allocation. Machine Learning Research 3, pp. 933-1022 http://www.ieomsociety.org/detroit2020/papers/506.pdf

[15] J. C. Campbell, A. Hindle and a. E. Stroulia. 2014. Latent Dirichlet Allocation: Extracting Topics.

[16] Suyanto. 2017. Data Mining Untuk Klasifikasi dan Klasterisasi Data. Informatika Bandung. ISBN: 978-602-6232-36-6. Bandung.