Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Sentiment Analysis of Hotel Reviews Using Support Vector Machine

Alexander Romian Simarmata[1], Muhammad Zakariyah[2]
romian37@gmail.com, muhammad.zakariyah@staff.uty.ac.id
Informatic, Yogyakarta University of Technology

Article Information

Submitted : 22 Sep 2023
Reviewed: 5 Oct 2023
Accepted : 30 Oct 2023

Keywords

Sentiment Analysis, Hotel Reviews, Support Vector Machine, Naïve-Bayes, Logistic Regression.

Abstract

With technology nowadays, everyone can leave their review about a hotel on the internet. This creates a new issue for the hotel itself because the reviews can come in in thousands amount. This will consume a lot of time to handle these reviews manually. In this study, a sentiment analysis model will be made to overcome the issue. The data in this study is collected from Kaggle website. This data contains 20,491 reviews about a hotel. The data will then be preprocessed and given a label for each data point. Then, the model is trained using the clean data. The model will use Naïve-Bayes, Logistic Regression, and Support Vector Machine algorithm. From the result performed, it's concluded that Support Vector Machine performed more accurately with 94% rate.

A. Introduction

With the easy utility of the internet nowadays, everyone can write a review about the hotel easily. Online hotel review is the way customers express their experience while visiting a hotel [1]. Usually, the review is written on a website, such as hotel booking site, or social media. The reasons why people tend to write a review are to help other consumers, personal motivation, or to improve the service quality [2]. With these reviews, the hotel can acquire visitors' feedback towards the hotel's performance, and the hotel can improve their performance towards the customers based on the visitor's feedback. Online hotel reviews have substantial impact on consumers decision in booking the hotel. When the online reviews are positive, consumers tend to book the hotel of a familiar brand rather than unfamiliar brand; by contrast, when the reviews are negative, consumers have little intention to book the hotel regardless the brand is familiar or not. Same thing happens to price variable, when the reviews are positive, lower price leads to higher chance of consumer to book the hotel, while negative reviews lower the chance of the consumer to book the hotel regardless the price of the hotel [3].

When the reviews are in dozens amount, hotel can still handle these reviews without difficulty. What if the reviews are in hundreds or thousands amount? Surely this creates a new issue for the hotel, as this will require a lot of time and resources to handle this many reviews. If the hotel decides to ignore the reviews, it will lead to poor commercial performance [4], this means that these reviews have significant business benefit to the hotel.

In this study, a sentiment analysis model will be made to overcome the issue. Sentiment analysis is a method to understand and process textual data automatically [5]. This means sentiment analysis is to identify and extract subjective information from text data. This method can help to understand the public's sentiment towards a product, or service. The algorithms that will be used for this model are Support Vector Machine, Naïve-Bayes, dan Random Forest.

Similar study was conducted by [6] using Linear Support Vector Machine to do sentiment analysis to amazon product reviews, the algorithm had the highest value of accuracy among the other algorithms, which the ratio was 93.70%. The next study was done by [7], this study was using sentiment analysis to analyze Twitter community sentiment towards 2019 Indonesian presidential candidates. The study found that Naïve-Bayes performed more accurately at 80.90% than Support Vector Machine and K-Nearest Neighbor algorithm. Similar study was also performed by [8], this study performed sentiment analysis to analyze customer reviews about restaurants in Karachi, Pakistan. The algorithms used in this study were Support Vector Machine, Naïve-Bayes, Logistic Regression, and Random Forest. It was found in this study that Random Forest performed better than the other 3 algorithms, the algorithm made 95% accuracy. The aim of this study is to help hotels in choosing which algorithm should be applied to do sentiment analysis toward online reviews, so it will be more efficient to handle the reviews coming to their hotel.

B. Research Method

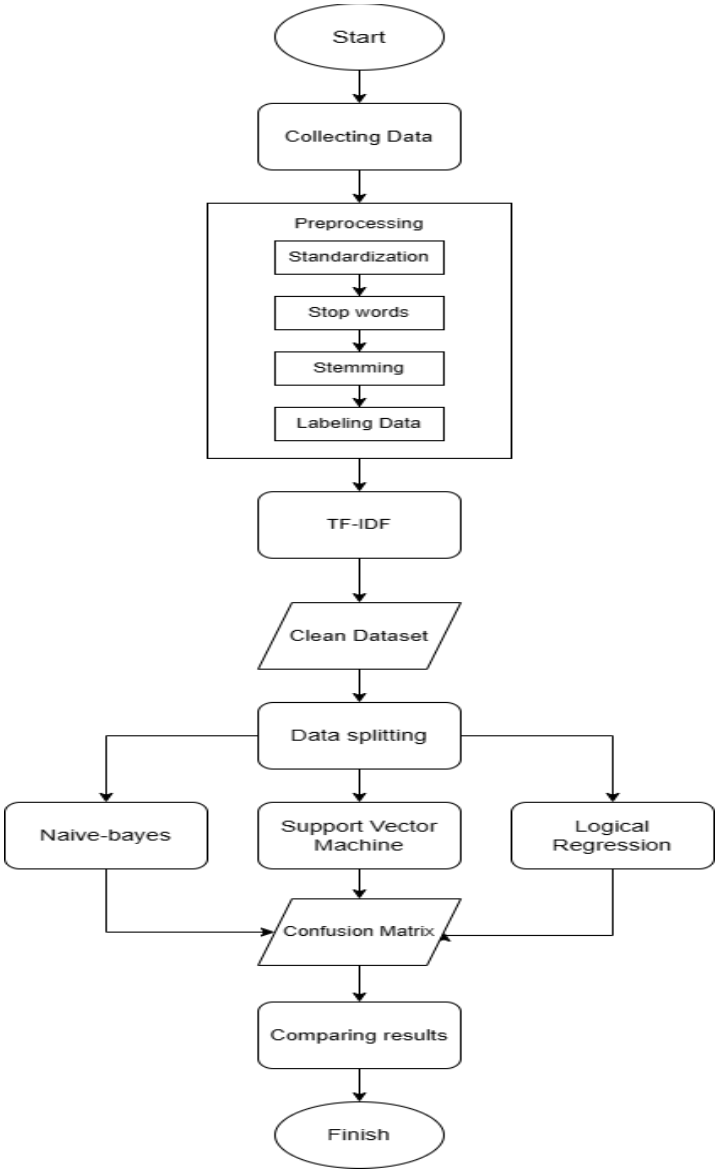

Figure 1. Research Flowchart

This research starts with collecting the data which will be used to train and test the sentiment analysis model. The data is acquired from Kaggle site [9], this data contains 2 columns, the first column is 20,491 reviews about a hotel which crawled from Trip Advisor (an online hotel booking site), and the second column is visitor rate of the hotel visitor visited, the value of the rate is between 1 to 5. After the data is collected, preprocessing will be conducted. From [10], it is shown that doing preprocessing to the data will increase the sentiment analysis model accuracy. The preprocessing in this study involves 5 stages. All stages of this process are:

1. Standardization
   This sub-process is set of another sub-process, such as:
   a. Removing non-alphabetical characters
   b. Converting all text into lowercase

2. Stop words
   This process is to remove unnecessary words that often appear in the data, for example pronouns, and prepositions.
3. This process is to remove unnecessary words, but they often appear in the data, for example pronouns, and prepositions. This process can increase the performance of sentiment analysis model [11].
4. Stemming
   This process changes an affixed word into its root or base form [12].
5. Labeling Data
   From each review data point will be given label column next to star rating column. The labeling is based on star rating of each review data, if the star is greater than or equal to 3, the label will be 1 or positive review, but if the rating is less than 3, the label will be 0 or negative review.

After the data has been preprocessed, feature selection will be done using TF-IDF method. The number presented is based on relevancy of that word itself to the whole document. The mathematical function of this process can be seen below [13]:

$$WDt = tfdt * Idft \qquad (1)$$

Where:
WDt = the weight of *d*-document toward *t*-term.
tfdt = the frequency of term *t* within document *d*.
Idft = the inverse of document frequency (log N/df)
N = the total number of documents in the corpus
Df = the total number of documents in the corpus that contain the term.

Now, the data has been preprocessed, it will create clean data. This clean data will then be split into 80:20 ratio. 80% of the data will be used to train the model, and the other 20% will be used to test the model. The distribution of the split data can be shown from the table below.

Table 1. Distribution of Split Data

	Training Data	**Testing Data**
Positive Data	13798	3479
Negative Data	2594	620
Total	16392	4099

After the data is split, 3 models will be made, which the algorithms used are Support Vector Machine, Naïve-Bayes, and Logistic Regression. These models were trained using training data and tested using testing data to see how each of these algorithms performed. A confusion matrix of each model will be made to assist this study in evaluating these algorithms. Confusion matrix is a machine learning tool that consist of information about the actual classification and prediction abilities of machine learning algorithms, this tool can evaluate performance of machine learning algorithm, making it very useful for this study [14]. Confusion matrix has four outputs that measure the model performance, these are accuracy, recall, precision, and F1-score. Accuracy measures overall model accuracy in doing prediction. Recall measures the ability of the model to correctly identify positive

class out of the total of positive class. Precision measures how many model's positive predictions made are correct. F1-score is a combination of recall and precision. It is usually known as harmonic mean of recall and precision. The mathematical function of these four can be seen below:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

C. Result and Discussion

The data in this study was acquired from Kaggle site. This data contains 2 columns, the first column is filled with reviews of a hotel, and the second column is rating of a hotel from 1 to 5. The total amount of reviews in this dataset is 20,491, and this was crawled from Trip Advisor, an online hotel booking site. This data then will be preprocessed, the result of this process can be represented in **Table 1**:

Table 2. Preprocessing Result

Stage	Result		
	Review	**Rating**	**Sentiment**
raw data	The hotel was excellent! The staff was friendly, and the room was clean, I like this hotel!	5	
standardization	the hotel was excellent the staff was friendly and the room was clean i like this hotel	5	
stop words	hotel excellent staff friendly room clean like hotel	5	
stemming	hotel excel staff friendly room clean like hotel	5	
labeling data	hotel excel staff friendly room clean like hotel	5	1

From **Table 1**, it shows what preprocessing stage did to the data. In standardization stage, it removed non-alphabetical characters and converted all text to lowercase. In stop words stage, it removed all the unnecessary words such as pronouns, prepositions, and words that often appear. The process continued with stemming all the text, this process stemmed all the words to its root form, e.g., the word of running will be stemmed to run. At the end of the process, a new column was made called Sentiment to label all the data. The data was labeled 1, which means positive, if the rating is greater or equal to 3, the data got 0 label, which means negative, if the rating was below 3.

Now the data has been preprocessed, the next step was doing feature selection using TF-IDF. Then, the data will be split into 80:20 ratio, this means that 80% of the data will be the training data for the model, and the rest will be the testing data for the model. After this stage, the model will be trained using the training data, 3 models will be made using Support Vector Machine, Naïve-Bayes, and Logistic

Regression algorithms. The confusion matrix of each algorithm can be seen the following tables:

Table 3. Confusion Matrix of Logistic Regression

	precision	**recall**	**F1-score**	**Support**
0	0.91	0.59	0.71	620
1	0.93	0.99	0.96	3479
Accuracy			0.93	4099
Macro avg	0.92	0.79	0.84	4099
Weighted avg	0.93	0.93	0.92	4099

Table 4. Confusion Matrix of Support Vector Machine

	precision	**recall**	**F1-score**	**Support**
0	0.87	0.69	0.77	620
1	0.95	0.98	0.96	3479
Accuracy			0.94	4099
Macro avg	0.91	0.84	0.87	4099
Weighted avg	0.94	0.94	0.93	4099

Table 5. Confusion Matrix of Naïve-Bayes

	precision	**recall**	**F1-score**	**Support**
0	0.89	0.49	0.63	620
1	0.92	0.99	0.95	3479
Accuracy			0.91	4099
Macro avg	0.90	0.74	0.79	4099
Weighted avg	0.91	0.91	0.90	4099

Based on **Table 3, Table 4, and Table 5** it can be seen the confusion matrix of 3 models using Logistic Regression, Support Vector Machine, and Naïve-Bayes algorithm. Logistic Regression made 91% correct prediction in negative class, and 93% correct prediction in positive class, making it 92% in average. This model had 59% recall in negative class and 99% recall in positive class. The F1-score of this model was 71% for negative class and 96% for positive class, making the average was 84%. The overall accuracy of this model was 93%

Compared to Logistic Regression, Support Vector Machine had lower precision in negative class with 87%, but higher precision in positive class with 95%. This model had a better average recall than Logistic Regression with 84% average recall score. Support Vector Machine also had a better F1-score with an average of 87%. The overall accuracy of this model was 94%

For the third algorithm, Naïve-Bayes had the worst performance out of all three. The average score of precision of this model was 90%, 74% for the average recall, and 79% for the average F1-score. This model has 91% overall accuracy.

From the comparison above, it reveals differences in the 3 algorithms. The choice of the best algorithm for doing sentiment analysis toward hotel reviews goes to Support Vector Machine, which performed better than Logistic Regression and Naïve-Bayes. With the average score 91% precision, 84% recall, 87% F1-score, and 94% overall accuracy, Support Vector Machine performs better than other two in doing sentiment analysis toward hotel reviews.

D. Conclusion

In this study, using 20,491 reviews of hotel data from Kaggle site, which was crawled from Trip Advisor. The data then went through some processes such as preprocessing, doing feature selection using TF-IDF, then splitting the data into 80:20 ratio. Now the data is ready to train the sentiment analysis model using Support Vector Machine, Logistic Regression, and Naïve-Bayes algorithm. After the models have been trained and tested, it can be concluded that Support Vector Machine performed better than Logistic Regression, and Naïve-Bayes in doing sentiment analysis toward hotel review data. The result of Support Vector Machine is precision 91%, recall 84%, F1-score 87%, and the accuracy is 94%.

E. Acknowledgment

The researcher would like to thank Yogyakarta University of Technology who has helped the researcher doing this research and giving advice during the research.

F. References

[1] Y. J. Kim and H. S. Kim, "The Impact of Hotel Customer Experience on Customer Satisfaction through Online Reviews," *Sustainability (Switzerland)*, vol. 14, no. 2, Jan. 2022, doi: 10.3390/su14020848.

[2] H. M. Gonçalves, G. M. Silva, and T. G. Martins, "Motivations for posting online reviews in the hotel industry," *Psychol Mark*, vol. 35, no. 11, pp. 807–817, Nov. 2018, doi: 10.1002/mar.21136.

[3] J. Wen, Z. Lin, X. Liu, S. H. Xiao, and Y. Li, "The Interaction Effects of Online Reviews, Brand, and Price on Consumer Hotel Booking Decision Making," *J Travel Res*, vol. 60, no. 4, pp. 846–859, Apr. 2021, doi: 10.1177/0047287520912330.

[4] O. A. El-Said, "Impact of online reviews on hotel booking intention: The moderating role of brand image, star category, and price," *Tour Manag Perspect*, vol. 33, Jan. 2020, doi: 10.1016/j.tmp.2019.100604.

[5] J. Khatib Sulaiman Dalam No, R. Afrinanda, W. Tawa Bagus, and L. Efrizoni, "Comparison of Machine Learning Algorithm Models in Bitcoin Price Sentiment Analysis," *Indonesian Journal of Computer Science Attribution*, vol. 12, no. 2, pp. 2023–502.

[6] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," in *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, IEEE, May 2018, pp. 1–6. doi: 10.1109/ICIRD.2018.8376299.

[7] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler : Twitter."

[8] K. Zahoor, N. Z. Bawany, and S. Hamid, "Sentiment analysis and classification of restaurant reviews using machine learning," in *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ACIT50332.2020.9300098.

[9] Larxel, "Trip Advisor Hotel Reviews," Kaggle. Accessed: Sep. 20, 2023. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews

[10] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Comput Math Organ Theory*, vol. 25, no. 3, pp. 319–335, Sep. 2019, doi: 10.1007/s10588-018-9266-8.

[11] Jouf University and Institute of Electrical and Electronics Engineers, *2019 International Conference on Computer and Information Sciences (ICCIS) : Jouf University - Aljouf - kingdom of Saudi Arabia, 03-04 April 2019.*

[12] K. K. Agustiningsih, E. Utami, and M. A. Alsyaibani, "Sentiment Analysis of COVID-19 Vaccines in Indonesia on Twitter Using Pre-Trained and Self-Training Word Embeddings," *Jurnal Ilmu Komputer dan Informasi*, vol. 15, no. 1, pp. 39–46, Feb. 2022, doi: 10.21609/jiki.v15i1.1044.

[13] M. Khairul Anam, M. Bambang Firdaus, T. Arita Fitri, and W. Agustin, "Analisis Pilkada Medan pada Sosial Media Menggunakan Analisis Sentimen dan Social Network Analyisis," *Indonesian Journal of Computer Science Attribution*, vol. 11, no. 1, pp. 2022–101.

[14] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 2, p. 112, Oct. 2020, doi: 10.20473/jisebi.6.2.112-122.