Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12214
e-ISSN : 2541-2019
p-ISSN : 2541-044X

IMPLEMENTATION OF K-NEAREST NEIGHBOR ALGORITHM WITH SMOTE FOR HOTEL REVIEWS SENTIMENT ANALYSIS

Firman Gazali Mahmud[1)*], Teguh Iman Hermanto[2)], Imam Maruf Nugroho[3)]
[1,2,3)]Sekolah Tinggi Teknologi Wastukancana, Purwakarta, Indonesia
[1)]firmangazali85@wastukancana.ac.id, [2)]teguhiman@wastukancana.ac.id, [3)]imam.ma@wastukancana.ac.id

Submitted : Feb 18, 2023 | **Accepted** : Mar 3, 2023 | **Published** : Apr 1, 2023

Abstract: Indonesia has considerable tourism development potential, this phenomenon is in accordance with the number of foreign tourist visits to Indonesia from January to September 2022 recorded by Badan Pusat Statistik many as 2,397,181 visitors. This research focuses on super-priority destinations in Labuan Bajo, East Nusa Tenggara, based on the government's plan that the focus of developing this destination is to increase hotel development to meet the need for an additional 2,000 hotel rooms. Thus, the available hotel rooms are still limited. Then for need to choose a hotel based on the November 2021 survey by the Populix website, 76% of 1,012 respondents chose to book hotels online with the majority using the Traveloka website. However, making decisions in choosing hotels using the reviews feature in the Traveloka website still raises various problems, such as biased information and even the rating values given do not match the reviews submitted. So that users to know what becomes the perception of positive and negative ratings, it is necessary to do in-depth research on satisfaction factors to find out positive and negative sentiments of hotel visitors. This study uses the k-nearest neighbor algorithm with SMOTE on the research objects of the three most popular hotels in Labuan Bajo. Data testing uses a value of k = 3 so that it produces an accuracy value of 87.71% - 93.47% with a maximum error tolerance of 12.29%. In addition, the performance of accuracy results is validated by the appropriate AUC value, namely the good classification category.

Keywords: Hotel Review, K-Nearest Neighbor, Sentiment Analysis, SMOTE, Traveloka

INTRODUCTION

Indonesia has considerable tourism development potential, based on the phenomenon of the Badan Pusat Statistik, this is in line with the number of foreign tourists visiting Indonesia from January to September 2022 of 2,397,181 visitors. This research focuses on tourist destinations in Labuan Bajo in NTT with the research object being the three most popular hotels selected based on the most reviews on the Traveloka website using the web scraping method. Web scraping is the process of extracting content in the form of data or information from a website. Web scraping is used because the required data is not available in RSS or API. In addition to extracting content, data and information, this technique is also used to automate the data retrieval process or called a robot (Adila, 2022). In addition, how to develop web scraping techniques, namely the maker of the scraping script must first learn the HTML document from a website that will be enclosed in information to use HTML tags (Sahria, 2020).

Based on the government's plan to develop tourist destinations to pursue the construction of an additional 2,000 hotel rooms, currently, Labuan Bajo is reportedly experiencing a shortage of hotels. Tourist visitors do not have much information to choose a satisfactory hotel so to get the best hotel choices to use references from online hotel booking sites. Survey results on the Populix website in November 2021 stated that 76% of 1,012 respondents chose online hotel bookings and the majority used Traveloka. Reviews given by hotel visitors are more credible and important than information obtained from advertisements (Lo & Yao, 2019). However, currently, the review

*Firman Gazali Mahmud

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12214
e-ISSN : 2541-2019
p-ISSN : 2541-044X

feature on the Traveloka website only displays general information, so the information obtained is still considered biased. It was found that the rating value given was not following the review submitted. Previous research stated the need to classify sentiments in reviews from previous hotel visitors to determine the factors of visitor satisfaction during their stay at the hotel (Chory et al., 2018). Sentiment analysis is a type of analysis that involves extracting sentiment to find attitudes and emotions associated with the text or data supplied (Patel et al., 2023).

Based on the problems above, it can be seen that Traveloka users have difficulty getting clear information from the currently available review features and hotels have difficulty getting clear feedback to evaluate their performance, so in-depth research is needed on the satisfaction and perception factors of positive and negative sentiment reviews from visitors. hotel. Previous research by (Damarta et al., 2021) classifies sentiment from Twitter using the k-nearest neighbor algorithm classification method without resampling unbalanced datasets, but produces an accuracy value of 87.41%. So this study intends to carry out tests using the classification method with the k-nearest neighbor algorithm for different research subjects and process flows. (Safitri, 2020) concluded that the k-nearest neighbor is a supervised learning algorithm where the results of new instances are classified based on the majority of the k-nearest neighbor categories. This research is expected to help solve the problem of increasing hotel service factors, especially in super-priority destinations that need serious attention. In addition to anticipating the occurrence of dissatisfaction problems from Traveloka users.

(Tsujii et al., 2015) conducted a study to extract hidden information from comments on HotelClub.com, Agoda.com, Booking.com, and Expedia.com sites using text mining techniques to evaluate expressions in comments based on positive and negative labels. Apart from sentiment, there are also several categories tested, namely aspects of location, rooms, hotels and services which are the most discussed topics by tourists. Text mining is an effective and efficient approach that has been applied in a wide range of research interests. With the purpose of organizing and understanding documents, text mining disclosed hidden semantic patterns in a corpus (Karami et al., 2020) This study is very likely to provide biased test results in the majority class because the percentage of class composition in the training data is not balanced. Unbalanced data sets make it difficult for machine learning to predict classifications and so minority classes tend to be misclassified. The research entitled The Implementation Of Genetic Algorithm In SMOTE (Synthetic Minority Oversampling Technique) For Handling Imbalance Dataset Problem (Tallo & Musdholifah, 2018) states that the solution to the dataset imbalance problem is to use the SMOTE method. (Cahyaningtyas et al., 2021) concluded that the Synthetic Minority Oversampling Technique (SMOTE) is a technique used to overcome the class imbalance problem (CIP). SMOTE works by modifying unbalanced datasets by creating new synthetic data from minority classes to increase the performance of the classification method.

METHOD

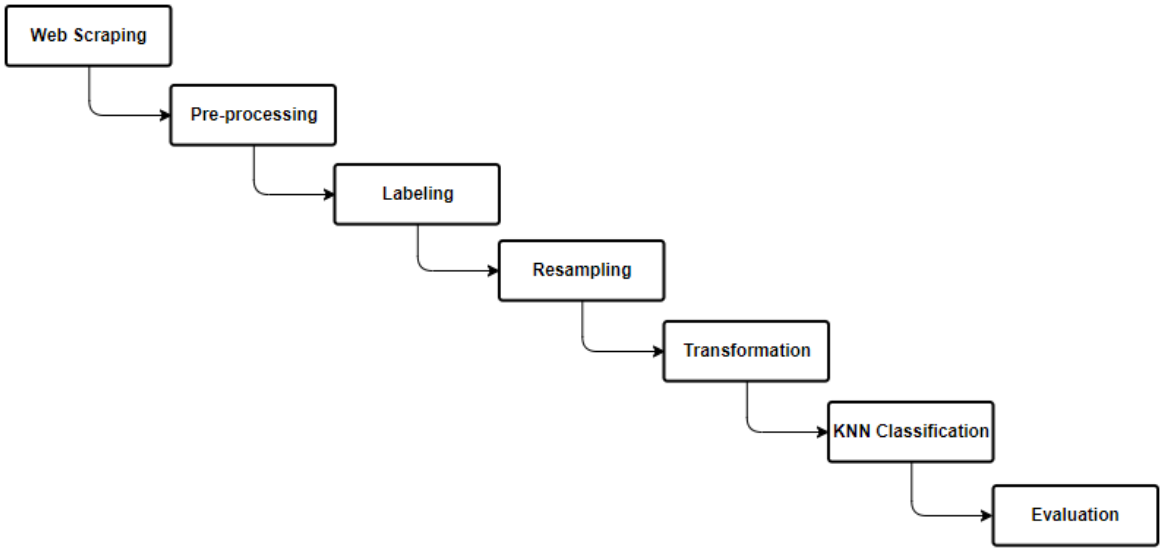

Fig. 1 Stage of research using k-nearest neighbor

*Firman Gazali Mahmud

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12214
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Web Scraping

The first stage, namely the data collection process, was carried out from November 9, 2022, until November 12, 2022, using web scraping on the Traveloka website using the Web Scraper extension from Google Chrome for the three most popular hotels in Labuan Bajo. The review data collected is the result of posts from 2015 until 2022. Namely Laprima Hotel with a total of 1,318 data reviews, Sylvia Hotel & Resort Komodo with a total of 918 data reviews, and Hotel Bintang Flores with a total of 693 data reviews. The order of the fourth most popular hotel and so on is not used as a sample object because the amount of review data is almost close to the amount of data on the third most popular hotel, it's feared that the similarity in the number of review data will not produce calculations that have significantly different characteristics between one object and another sample object.

Pre-processing

The second stage is cleaning the dataset by removing unnecessary attributes and characters and then eliminating missing values and duplicate data, to minimize the effect of noise on the dataset.

Labeling

The data set review will be labeled manually. Determining the classification of positive and negative sentiment labels refers to previous research from (Setiawan et al., 2022) which is determined based on the rating value that has been given. If the rating value is ≥ 8.5 then the review sentiment is declared positive, whereas if the rating value is ≤ 6.5 then the review sentiment is declared negative. In the rating range of 6.6 to 8.4, manual labeling will be carried out based on sentences that represent positive or negative sentiments.

Resampling SMOTE

After being labeled. The dataset will be split into 2 parts, 70% used for training data and 30% for testing data. Initial labeling on data training resulted in a total of 282 positive and 71 negative classes of Laprima Hotel sentiment, 286 positive and 56 negative classes of Sylvia Hotel & Resort Komodo, and 121 positive and 20 negative classes of Bintang Flores Hotel. The results of the labeling class cause an imbalanced class, namely the condition of classes from a data set where one class has a very large amount of data compared to other classes. The large difference in the amount of data between classes can cause the classification model to often fail to predict and cause bias. Over-sampling in SMOTE takes instances of the minority class and then finds the k-nearest neighbors of each instance, then generates synthetic instances instead of replicating the minority class instances, thereby avoiding the problem of excessive overfitting (Haryanto et al., 2022).

Transformation

The next stage is transformcase, tokenizing, stopword, and stem on rapidminer. The algorithm used to perform feature extraction in this study is TF-IDF, Bag of Words, and improvement TF-IDF. The three methods will be used to perform feature extraction on each review that has been collected where later the extracted features will be used for classification where feature extraction will convert text data into word features and weights for classification (Dharma & Saragih, 2022). The TF-IDF calculation results will also be used to calculate visitor satisfaction aspects.

K-NN Classification

The k-nearest neighbor algorithm is divided into two phases, namely learning (training) and classification (testing). In the training phase, this algorithm only stores feature vectors and classifies learning data. In the testing phase, the same features are calculated for the data to be tested. The distance from this new vector to all learning data vectors is calculated, and the k-nearest neighbors are taken (Angreni et al., 2018). Calculation of similarity and distance for each attribute uses the cosine similarity formula. In using the k-nearest neighbor algorithm, constraints are found in its application, namely determining the value of k. To find out the best value of k, it is necessary to try using odd values.

Evaluation

The evaluation stage is the last stage which aims to measure the results of the classification test on data testing. Test evaluation uses the confusion matrix for classification tests using SMOTE and without SMOTE so that it can produce accuracy, precision, and recall values. Furthermore, the calculation of the AUC (Area Under Curve) value is carried out which aims to measure the performance of the results of the Confusion Matrix.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x\ 100\% \qquad (1)$$

*Firman Gazali Mahmud

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12214
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$Precision = \frac{TP}{TP+FP} x\ 100\% \quad (2)$$

$$Recall/Sensitivity = \frac{TP}{TP+FN} x\ 100\% \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} x\ 100\% \quad (4)$$

$$FPR = 1 - Specificity \quad (5)$$

$$AUC = \frac{1+Sensitivity-FPR}{2} \quad (6)$$

RESULT

In the case of this study, the percentage of unbalanced data training classes showed as much as 83.10% for the positive class and 16.90% for the negative class. Because class balance has an important role in determining classification, resampling is needed using the SMOTE technique. This process adds data to the minority class by duplicating the minority class data randomly to produce the same amount of data as the majority class.

Table 1. Resampling data training

Hotel Name	Before SMOTE		After SMOTE	
	Positive	Negative	Positive	Negative
Laprima Hotel	282	71	282	282
Sylvia Hotel & Resort Komodo	286	56	286	286
Bintang Flores Hotel	121	20	121	121

The following is a model implemented using the k-nearest neighbor algorithm. This model will solve 70% of the data used as training data in conducting learning so that it can find its correlation, while the remaining 30% of the data is used as test data to test the classification results. Calculation of similarity between data and k-nearest neighbors uses the Data to Similarity operator.

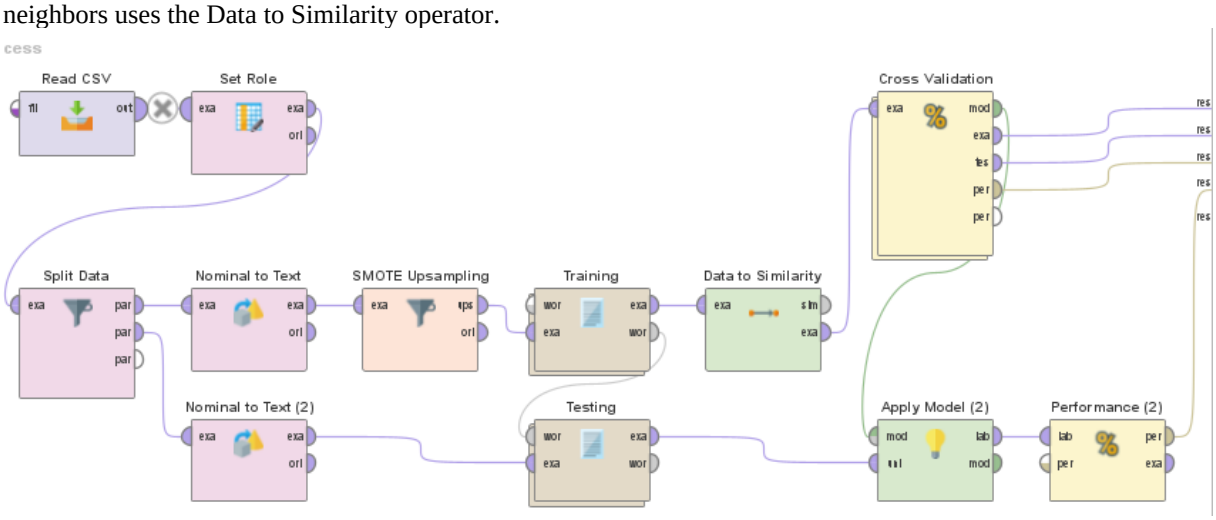

Fig. 2 K-nearest neighbor model on rapidminer

In using the k-nearest neighbor algorithm, problems were found during its application, namely determining the value of k, to find out the best k value, trials using odd values were needed. An even value is not used because if the number of class data in the k calculation is balanced it has the potential to cause ambiguity in classification, besides that the value 1 is not used because determining the number of nearest neighbors requires a comparison, then starting with a value of k = 3 using SMOTE on one of the data objects, namely Sylvia Hotel & Resort Komodo for testing. Based on the results in Table 2 in determining the value of k, in this case, it can be concluded that the higher the value of k, the lower the accuracy value. Thus, the value used is k = 3.

*Firman Gazali Mahmud

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12214
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2. Determine the value of k

k	Accuracy	Precision	Recall
3	90,19%	89,86%	91,03%
5	84,61%	86,32%	82,49%
7	83,04%	85,06%	80,28%
9	83,03%	84,76%	80,69%
11	80,81%	83,82%	76,34%

Furthermore, the results of the classification test using the k-nearest neighbor will be evaluated using the confusion matrix. Generates an accuracy value, namely the degree of closeness between the predicted value and the actual value, the precision value as the degree of accuracy between the information requested by the user and the answer given by the system, and the recall value as the success of the system level in retrieving information. The following is a comparison of the evaluation of the confusion matrix if resampling is not done without SMOTE and if resampling is done with SMOTE.

Table 3. Confusion matrix result without SMOTE

No	Hotel Name	Accuracy	Precision	Recall
1	Laprima Hotel	79,85%	60,00%	16,25%
2	Sylvia Hotel & Resort Komodo	80,70%	61,29%	31,19%
3	Bintang Flores Hotel	88,33%	50,00%	14,29%

Table 4. Confusion matrix result with SMOTE

No	Hotel Name	Accuracy	Precision	Recall
1	Laprima Hotel	87,71%	87,58%	88,44%
2	Sylvia Hotel & Resort Komodo	90,19%	89,86%	91,03%
3	Bintang Flores Hotel	93,47%	92,68%	95,06%

The next evaluation process calculates the AUC value which aims to show the accuracy of the resulting confusion matrix performance. The AUC value ranges from 0 to 1, the classification is said to be good if the AUC value is higher.

Table 5. AUC value result

No	Hotel Name	AUC Value	Meaning
1	Laprima Hotel	0,944	Excellent Classification
2	Sylvia Hotel & Resort Komodo	0,934	Excellent Classification
3	Bintang Flores Hotel	0,821	Good Classification

After the performance evaluation shows good results, it can be done to calculate the visitor satisfaction factor which is divided into 5 main factors. Namely the factors of facilities, service, restaurant, cleanliness, and location. To be able to determine which attributes are included in one of the factors, it is necessary to calculate the TF summary. In determining the factor value, the positive and negative sentiment classes need to be separated and the percentage value calculated. The results of factor calculations from TF can be seen in Table 6.

Table 6. Determining the attribute satisfaction factors

Hotel Name	Positive Sentiment Class			Negative Sentiment Class		
	Attribute	TF Sum	Factor	Attribute	TF Sum	Factor
Laprima Hotel	pemandangan	183	Facility	kamar	48	Facility
Laprima Hotel	kamar	149	Facility	makanan	37	Restaurant

*Firman Gazali Mahmud

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12214
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Hotel Name	Positive Sentiment Class			Negative Sentiment Class		
	Attribute	TF Sum	Factor	Attribute	TF Sum	Factor
Laprima Hotel	makanan	124	Restaurant	kotor	17	Cleanliness
Laprima Hotel	bersih	106	Cleanliness	staff	12	Services
Laprima Hotel	staff	46	Services	pantai	11	Facility
Laprima Hotel	pantai	41	Facility	wifi	11	Facility
Laprima Hotel	lokasi	38	Location	toilet	11	Facility
Laprima Hotel	pelayanan	36	Services	menu	9	Restaurant
Laprima Hotel	kolam	34	Facility	pelayanan	8	Services
Laprima Hotel	dermaga	26	Facility	kolam	8	Facility
Laprima Hotel	balkon	26	Facility	pemandangan	7	Facility
Laprima Hotel	menu	17	Restaurant	rusak	7	Facility
Laprima Hotel				lalat	7	Cleanliness
Sylvia Hotel & Resort	pantai	121	Facility	kamar	14	Facility
Sylvia Hotel & Resort	makanan	115	Restaurant	jauh	10	Location
Sylvia Hotel & Resort	kamar	112	Facility	staff	7	Services
Sylvia Hotel & Resort	pemandangan	112	Facility	pantai	6	Facility
Sylvia Hotel & Resort	bersih	83	Cleanliness	sarang	5	Cleanliness
Sylvia Hotel & Resort	kolam	38	Facility	lebah	4	Cleanliness
Sylvia Hotel & Resort	pelayanan	37	Services	restoran	4	Restaurant
Sylvia Hotel & Resort	menu	20	Restaurant	lokasi	4	Location
Sylvia Hotel & Resort	restoran	17	Restaurant	sampah	3	Cleanliness
Sylvia Hotel & Resort				kotor	3	Cleanliness
Sylvia Hotel & Resort				lalat	3	Cleanliness
Sylvia Hotel & Resort				pemandangan	2	Facility
Sylvia Hotel & Resort				tawon	2	Cleanliness
Sylvia Hotel & Resort				jendela	2	Facility
Sylvia Hotel & Resort				kebersihan	2	Cleanliness
Bintang Flores Hotel	kamar	71	Facility	kamar	14	Facility
Bintang Flores Hotel	bersih	66	Cleanliness	jauh	10	Location
Bintang Flores Hotel	makanan	50	Restaurant	staff	7	Services
Bintang Flores Hotel	kolam	43	Facility	pantai	6	Facility
Bintang Flores Hotel	pemandangan	43	Facility	sarang	5	Cleanliness
Bintang Flores Hotel	pantai	38	Facility	lebah	4	Cleanliness
Bintang Flores Hotel	staff	31	Services	restoran	4	Restaurant
Bintang Flores Hotel	pelayanan	26	Services	lokasi	4	Location
Bintang Flores Hotel				sampah	3	Cleanliness
Bintang Flores Hotel				kotor	3	Cleanliness
Bintang Flores Hotel				lalat	3	Cleanliness
Bintang Flores Hotel				pemandangan	2	Facility
Bintang Flores Hotel				tawon	2	Cleanliness

*Firman Gazali Mahmud

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12214
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Hotel Name	Positive Sentiment Class			Negative Sentiment Class		
	Attribute	TF Sum	Factor	Attribute	TF Sum	Factor
Bintang Flores Hotel				jendela	2	Facility
Bintang Flores Hotel				kebersihan	2	Cleanliness

The calculation results from Table 6 can be simplified using a radar chart to make it easier to read the value of the visitor satisfaction factor.

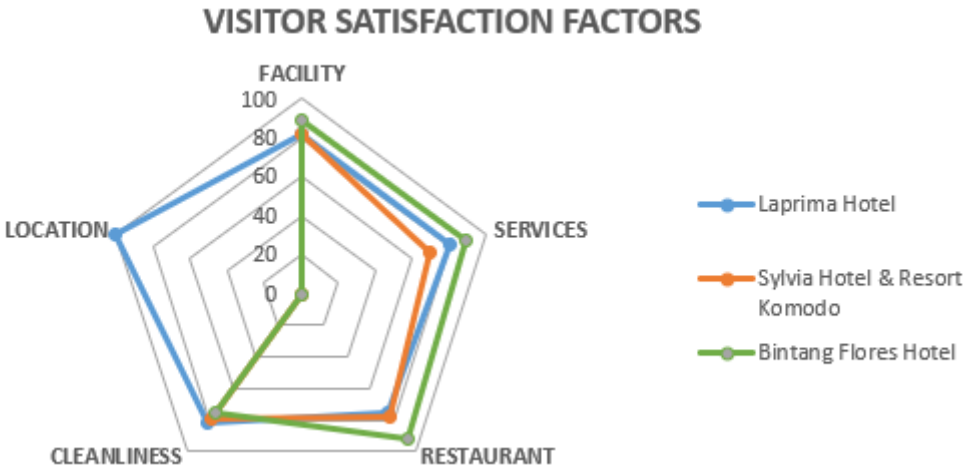

Fig. 3 Visitor satisfaction factors result

From the results of the sentiment classification that has passed the pre-processing and transformation stages of the three most popular hotel in Labuan Bajo. Resulting in 397 positive and 107 negative sentiments for Laprima Hotel, then 321 positive and 88 negative sentiments for Sylvia Hotel & Resort Komodo, and 177 positive and 24 negative sentiments for Bintang Flores Hotel.

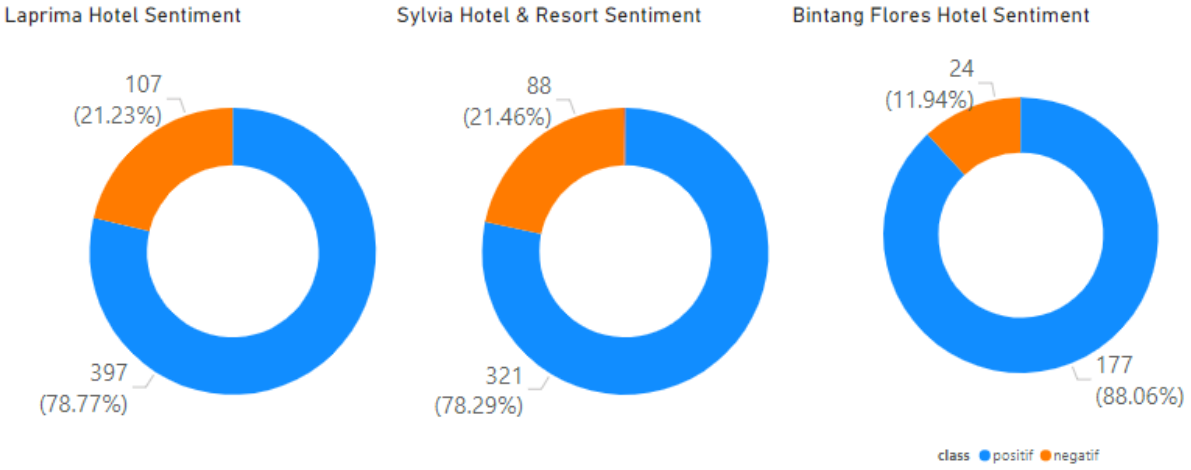

Fig. 4 Sentiment percentage result on the hotel

DISCUSSIONS

Based on the test results that have been obtained, resampling the training data using the SMOTE technique produces better accuracy performance, which is around 87.71% - 93.47% with a maximum error tolerance of 12.29%. Meanwhile, if you do not use the SMOTE technique, it will produce an accuracy of around 79.85% - 88.33%. To increase the effectiveness of testing in improving accuracy performance, it is suggested that further research apply other data optimization techniques such as Particle Swarm Optimization (PSO).

*Firman Gazali Mahmud

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i2.12214 p-ISSN : 2541-044X

CONCLUSION

The application of the k-nearest neighbor algorithm with SMOTE for reviews sentiment analysis of the most popular hotels in Labuan Bajo provides benefits that not only provide accurate sentiment analysis around 87.71% - 93.47%. But it also generates visitor satisfaction factors based on TF-IDF calculations which consist of facilities, service, restaurant, cleanliness, and location to provide further information about the contents of Traveloka hotel reviews feature. The data collection resulted in 1,318 reviews for Laprima Hotel, then 918 reviews for Sylvia Hotel & Resort Komodo, and 693 reviews for Bintang Flores Hotel. After carrying out a cleaning process and filtering data that only use Indonesian, it resulted in 504 reviews for Laprima Hotel, then 409 reviews for Sylvia Hotel & Resort Komodo, and 201 reviews for Bintang Flores Hotel. In addition, the percentage of sentiment for the three most popular hotels resulted in 78.77% positive sentiment for Laprima Hotel, then 78.29% positive sentiment for Sylvia Hotel & Resort Komodo, and 88.06% positive sentiment for Bintang Flores Hotel. Before resampling with smote, it produces larger positive sentiment data. However, this data cannot be accepted because it does not take into account negative sentiment and tends to produce biased data.

REFERENCES

Adila, N. (2022). Implementation of Web Scraping for Journal Data Collection on the SINTA Website. *SinkrOn: Jurnal Dan Penelitian Teknik Informatika*, *7*(4), 2478–2485. https://doi.org/10.33395/sinkron.v7i4.11576

Angreni, I. A., Adisasmita, S. A., Ramli, M. I., & Hamid, S. (2018). Pengaruh Nilai K Pada Metode K-Nearest Neighbor (KNN) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan. *Rekayasa Sipil*, *7*(2), 63. https://doi.org/10.22441/jrs.2018.v07.i2.01

Cahyaningtyas, C., Nataliani, Y., & Widiasari, I. R. (2021). Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE. *Jurnal Teknologi Informasi*, *18*(2), 173–184. https://doi.org/10.24246/aiti.v18i2.173-184

Chory, R. N., Nasrun, M., & Setianingsih, C. (2018). Sentiment Analysis On User Satisfaction Of Mobile Data Services Using Support Vector Machine (SVM) Algorithm. *2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS)*, 194–200. https://doi.org/10.1109/IOTAIS.2018.8600884

Damarta, R., Hidayat, A., & Abdullah, A. S. (2021). The application of k-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality. *Journal of Physics: Conference Series*, *1722*(1), 012002. https://doi.org/10.1088/1742-6596/1722/1/012002

Dharma, A. S., & Saragih, Y. G. R. (2022). Comparison of Feature Extraction Methods on Sentiment Analysis in Hotel Reviews. *Sinkron*, *7*(4), 2349–2354. https://doi.org/10.33395/sinkron.v7i4.11706

Haryanto, E. M. O. N., Estetikha, A. K. A., & Setiawan, R. A. (2022). Implementasi Smote Untuk Mengatasi Imbalanced Data Pada Sentimen Analisis Sentimen Hotel Di Nusa Tenggara Barat Dengan Menggunakan Algoritma Svm. *Jurnal Informasi Interaktif*, *7*(1), 16–20. http://e-journal.janabadra.ac.id/index.php/informasiinteraktif/article/view/1615

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and Research: A Systematic Literature Review through Text Mining. *IEEE Access*, *8*, 67698–67717. https://doi.org/10.1109/ACCESS.2020.2983656

Lo, A. S., & Yao, S. S. (2019). What makes hotel online reviews credible? *International Journal of Contemporary Hospitality Management*, *31*(1), 41–60. https://doi.org/10.1108/IJCHM-10-2017-0671

Patel, A., Oza, P., & Agrawal, S. (2023). Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model. *ScienceDirect Procedia Computer Science*, *218*, 2459–2467. https://doi.org/10.1016/j.procs.2023.01.221

Safitri, R. N. (2020). Analisis Sentimen Review Pelanggan Hotel Menggunakan Metode K-Nearest Neightbor (K-NN). In *Unirvesitas Dinamika* (Vol. 2507, Issue 1). https://doi.org/10.1016/j.solener.2019.02.027%0A

Sahria, Y. (2020). Implementasi Teknik Web Scraping pada Jurnal SINTA Untuk Analisis Topik Penelitian Kesehatan Indonesia. *URECOL (Unversity Research Colloqium)*, 297–306. http://repository.urecol.org/index.php/proceeding/article/view/1079

Setiawan, R. A., Estetikha, A. K. A., Nurharyanto, E. M. O., Asmara, Y., & Wahyudi, A. (2022). Analisis Sentimen Hotel di Nusa Tenggara Barat Menggunakan Algoritma SVM. *Jurnal Informasi Interaktif*, *7*(1), 149–155. https://papersmai.mercubuana-yogya.ac.id/index.php/smai/article/view/98

Tallo, T. E., & Musdholifah, A. (2018). The Implementation of Genetic Algorithm in Smote (Synthetic Minority Oversampling Technique) for Handling Imbalanced Dataset Problem. *Proceedings - 2018 4th International Conference on Science and Technology, ICST 2018*, *1*, 1–4. https://doi.org/10.1109/ICSTC.2018.8528591

Tsujii, K., Tsuda, K., & Takahashi, M. (2015). Towards Extracting the Hotel Evaluations from the Comments by the Foreign Tourists with Text Mining. *2015 IIAI 4th International Congress on Advanced Applied Informatics*, 46–49. https://doi.org/10.1109/IIAI-AAI.2015.172

*Firman Gazali Mahmud