Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 6 (2024) 819 - 827 e-ISSN: 2580-0760

Data Clustering for Sentiment Classification with Naïve Bayes and Support Vector Machine

Bayu Yanuargi[1]*, Ema Utami[2], Kusrini[3], Arli Aditya Parikesit[4]
[1,2,3] Department of Magister of Informatics Engineering, Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia
[4]Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences
[1]bayu.yanuargi@students.amikom.ac.id, [2]ema.u@amikom.ac.id, [3]kusrini@amikom.ac.id, [4]arli.parikesit@i3l.ac.id

Abstract

Visitor reviews play a crucial role in determining the success of a business, particularly those offering hospitality and services, such as hotels. The growth of internet technology has made it easier for guests to share their experiences, which can influence potential customers. Google Maps is one of the platforms used for giving and searching reviews This research uses data crawled from Google Maps Review using the playwright library. However, the large volume of reviews can make analysis and topic-based categorization—such as service quality, hotel location, and operational hours—challenging. To address this, DBSCAN is used to cluster reviews based on these topics. Clustering helps improve sentiment classification, making it more targeted and allowing a comparison of two machine learning algorithms: Naïve Bayes and Support Vector Machine (SVM). Naïve Bayes achieved higher accuracy (0.87) in the operational hours cluster, while SVM scored 0.78. However, SVM showed improved accuracy in the location (0.89) and service (0.88) clusters, with Naïve Bayes maintaining a stable 0.86 accuracy in both. Both models demonstrated an average training time of less than one second, excluding preprocessing.

Keywords: sentiment analysis; hotel; clustering; naïve bayes; support vector machine

How to Cite: B. Yanuargi, Ema Utami, Kusrini, and A. A. Parikesit, "Data Clustering for Sentiment Classification with Naïve Bayes and Support Vector Machine", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 6, pp. 819 - 827, Dec. 2024.
DOI: https://doi.org/10.29207/resti.v8i6.6139

1. Introduction

The most effective way today to understand individual preferences and dislikes regarding situations, events, and conditions is through community detection. The most challenging part is identifying valuable outcomes after detection, which involves categorizing different viewpoints within social networks. Therefore, community detection is highly useful for gaining insights into people's perspectives. [1]. In recent years, multi-density data clustering has become a focal point of research. As a widely used clustering algorithm, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm can effectively handle noisy data and has the ability to discover clusters of any shape, making it highly versatile in various applications [2].

The DBSCAN algorithm can identify more complex data variations and effectively separate points that are not part of the main cluster (outliers). DBSCAN produced a higher DBI score of 1.39 compared to K-Means, which had a score of 0.39. The DBI score is used to measure the quality of clustering, with lower scores indicating better separation between groups [3]. In a different study comparing the accuracy performance of the K-Means and DBScan algorithms in clustering product reviews, both algorithms were chosen due to their distinct clustering methods. K-Means employs a centroid-based approach, while DBScan uses a density-based approach. The study found that DBScan achieved an accuracy of 99.80%, which was higher than K-Means' accuracy of 99.50%. This demonstrates that DBScan was more effective in clustering product reviews in terms of accuracy, highlighting its ability to handle data differently from K-Means, particularly in dealing with complex data structures [4].

The development of internet technology provides an opportunity for tourists to plan their trips by searching and obtaining information from the internet, which brings major changes to the tourism industry, especially for hotels as a key industry in the tourism industry that

Received: 27-10-2024 | Accepted: 22-12-2024 | Published Online: 29-12-2024

must find ways to compete and read the diverse needs of customers that greatly affect the sustainability of their business [5]. The ease of access in the internet world provides the widest possible access for customers to provide reviews about the quality of service and their experience while staying or visiting the hotel as a whole. Customer reviews on online social media have become an important instrument for evaluating customer satisfaction with hotel services which is an important factor in influencing new customers, consumer loyalty, and the financial success of hotel companies [6].

One of the models or approaches that is quite reliable in text classification is the Bayes Hypothesis, which acknowledges that the emergence of several other features does not depend on the presence of certain elements in a class so that it can be used to solve multi-class prediction problems [7]. The classification of community tweet data using the N-Gram feature extraction of trigram type in the Naive Bayes algorithm has been shown to perform fairly well, with a total accuracy of 81%, a precision value of 78%, a recall of 91%, and a f1-Score of 84%. The Naive Bayes and Trigram algorithm settings, which are 84%, 84% for Precision, 86% for Recall, and 85% for f1-Score, produced the best results [8].

Negative or positive sentiments from application users can be input and evaluation for managers to maintain user loyalty, where the classification process can be started by preprocessing data starting from case folding, removing stop words, tokenization, stemming to TF-IDF, and the results of the preprocessing are then used as data to perform classification using Naïve Bayes [9]. Comparison of Naive Bayes classification and other algorithms such as XGBoost in classifying data into two classes, namely positive and negative, obtained Classification results using XGBoost proven to be able to classify unbalanced data better than Naïve Bayes where the combination of Word2vec + XGBoost produces a higher F1 score of 0.941, followed by TF-IDF + XGBoost with an F1 score of 0.940 and meanwhile, Naïve Bayes has an F1 score of 0.915 with TF-IDF and 0.900 with word2vec [10].

People may now voice their ideas on social media, and the process of mining social media user sentiment data will be highly beneficial. Support Vector Machine (SVM) outperforms the Naïve Bayes algorithm in terms of accuracy, precision, and recall, with values of 90.47%, 90.23%, and 90.78%, respectively, when used for sentiment analysis on the Covid-19 vaccine. The Naïve Bayes algorithm performs better, with values of 88.64%, 87.32%, and 88.13%, with differences in accuracy of 1.83%, precision of 2.91%, and recall of 2.65%. On the other hand, the Naïve Bayes method performs better in terms of processing time—it takes 8.1 seconds as opposed to 11 seconds for SVM. Sentiment analysis results for Naïve Bayes indicate that neutral sentiment is 8.76%, negative sentiment is 42.92%, and positive sentiment is 48.32%. In contrast, SVM showed 10.56% neutral sentiment, 41.28% negative sentiment, and 48.16% positive sentiment [11].

Support Vector Machines (SVM) have become a popular and effective method for classifying sentiment. SVM efficiently divides sentiment classes and achieves high sentiment prediction accuracy by locating the ideal hyperplane. One reason SVM is a popular choice for sentiment analysis jobs is its capacity to handle high-dimensional feature fields [12]. Research on how to use feature selection—particularly chi-square—to improve the accuracy of the Naive Bayes and Support Vector Machine algorithms in classifying Instagram comments reveals that the Multinomial Naive Bayes (MNB) algorithm outperforms the Support Vector Machine (SVM) algorithm, which achieves an accuracy of 82.31% without feature selection and 90% with feature selection. The MNB algorithm achieves an accuracy of 83.85% without feature selection and 90.77% with feature selection. [13].

Based on the background presented, this paper aims to classify hotel sentiment with three types of review data, namely reviews of location accuracy, reviews related to operations and reviews related to service satisfaction. In order to obtain these three types of data, the dataset will be clustered using the DBSCAN approach, so that three data with different amounts will be obtained.

Using the three types of review data from the DBSCAN clustering results, sentiment analysis will then be carried out on each data using the Naïve Bayes and Support Vector Machine algorithms. The experiment is expected to provide new knowledge related to the performance of the two models on different amounts of data, in this case six experimental results will be obtained, where Naïve Bayes with three experimental results and SVM with three experimental results.

The data utilized in this study were visitor reviews from various hotels associated with the Indonesian Hotel and Restaurant Association (PHRI) in the Sleman region. The findings from this research are intended to provide valuable insights to PHRI regarding the feedback and responses from tourists or customers who stayed at these affiliated hotels. This input is expected to assist PHRI in better understanding customer experiences and improving the quality of services offered by its member hotels, ultimately enhancing overall guest satisfaction and fostering better relationships between hotels and their customers.

2. Research Methods

In line with the research objectives, the author carried out the study following the steps outlined in Figure 1. The diagram provides a detailed representation of the process used to achieve the research goals. The purpose of clustering hotel review data is to categorize reviews based on key issues in the hotel industry. These differences in the number of reviews will be analyzed to determine whether the review count has an impact on

the accuracy and speed of classification. The study aims to evaluate how variations in review quantity influence classification performance and processing efficiency within the context of sentiment analysis and hotel-related feedback Each step was carefully executed to ensure the study's accuracy and effectiveness in addressing the intended objectives, as shown in Figure 1.

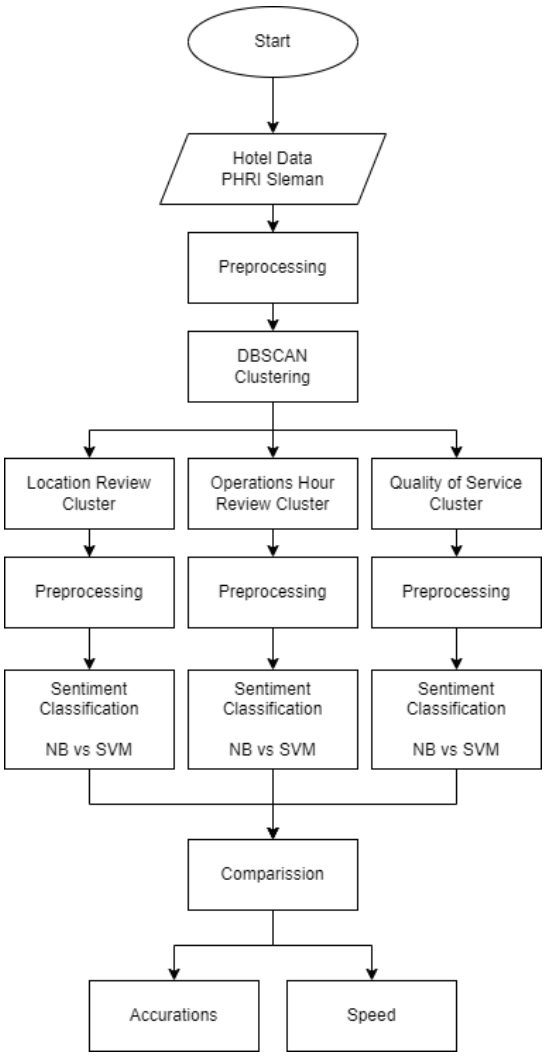

Figure 1. General Research Flow

2.1 Visitor Review Data Acquisition

This paper uses two types of data. The first dataset includes hotels that are members of PHRI Sleman, and the second consists of visitor reviews from those hotels. The data collection process is divided into two stages. The first stage involves gathering hotel data from the website https://phriyogyakarta.com/sleman/. The second stage involves collecting visitor reviews based on the hotel name and address. The steps involved in this process are illustrated in Figure 2.

The Hotel name and address from PHRI website used as keyword for the crawling activities on the google maps. This crawling process aims to get the visitor review data based on the hotel name and address. The combination of name and address from PHRI for crawling activities is important to ensure the crawling can get the correct hotel based on its address.

In the acquisition process, as explained in Figure 2 above, the Playwright library is used, which is a browser automation tool that can efficiently perform web scraping or data crawling from websites. In the context of crawling or web scraping, Playwright allows interaction with web pages as if it were a human user, handling dynamic JavaScript, pop-ups, page scrolling, and more.

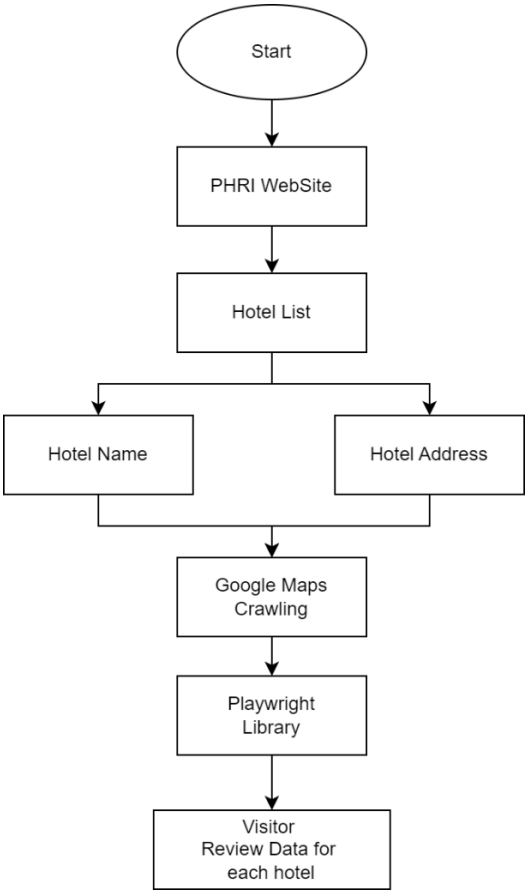

Figure 2. Data Acquisition Flow

The use of the Playwright library aims to avoid the need for APIs in the data acquisition process, thus reducing costs while still obtaining the desired results. The acquisition process resulted in data from a total of 36 hotels and the number of visitor reviews for each hotel, with a combined total of 53,000 reviews, as shown in Table 1.

Table 1. Research Data

No	Hotel Name	Review Numbers
1	Allstay Ecotel Yogyakarta	1556
2	Atrium Premiere Hotel Jogja	1026
3	Cakra Kusuma Hotel Yogyakarta	1603
4	Crystal Lotus Hotel Yogyakarta	2526
5	GRAMM HOTEL by Ambarrukmo	2210
6	Grand Keisha Yogyakarta	1859
7	Grand Serela Yogyakarta	1868
8	Grand Tjokro Yogyakarta	1711
9	Griya Persada Convention & Resort	1468
10	Hotel FortunaGrande Seturan	1555

No	Hotel Name	Review Numbers
11	Hyatt Regency Yogyakarta	3119
12	Ibis Yogyakarta Adi Sucipto	1737
13	Indoluxe Hotel Jogjakarta	2728
14	INNSiDE by Meliá Yogyakarta	1642
15	Lafayette Boutique Hotel	1291
16	LPP Garden Hotel	1113
17	Merapi Merbabu Hotels	2000
18	Platinum Adisucipto Hotel	2118
19	Prima SR Hotel & Convention	2026
20	Royal Ambarrukmo Yogyakarta	2195
21	Sahid Raya Hotel & Convention	2828
22	Satoria Hotel Yogyakarta	2372
23	Sheraton Mustika Resort & Spa	2558
24	Student Park Hotel	797
25	The Atrium Hotel and Resort	1719
26	The Jayakarta Hotel & Spa	2153
27	The Rich Jogja Hotel	2171
28	The Westlake Resort Jogja	2000

Each review data acquired in the next step undergoes a labeling process based on the ratings provided by visitors. Ratings of 1 and 2 are labeled as negative, a rating of 3 is considered neutral, and ratings of 4 and 5 are labeled as positive as shown in Table 2.

Table 2. Data Label Based on Rating

Rating/ Label	Cluster	Review Sample
Negative (1) 1547	Services	Soo disappointed with this hotel. The hotel is in bad repair and is in need of upgrading in a big way. The staff is also not good or helpful and their communication is very bad.
Negative (1) 1547	Location	The location of the hotel is on a small street, the check-in process took a long time
Negative (1) 1547	Operational Hour	Can only check in at 03.30 in the afternoon. even though it was already 2 o'clock.
Negative (2) 1057	Services	The service is lazy, like it doesn't intend to work. When asked, the answer is also like being lazy.
Negative (2) 1057	Location	The location is too crowded
Negative (2) 1057	Operational Hour	Checked in at 2 but until half past 4 the room wasn't ready yet
Neutral (3) 3796	Services	The meeting room is okay, but the sound echoes several times. Get a snack menu from the hotel. It feels 50:50.
Neutral (3) 3796	Location	Hotel area is not far away from Yogyakarta train station is about 15-20 minutes to the hotel
Neutral (3) 3796	Operational Hour	Actually, it was really okay
Positive (4) 11797	Services	The food is good and the service is also good
Positive (4) 11797	Location	The location of the hotel is really in the middle of the city
Positive (4) 11797	Operational Hour	My experience is that I have to wait for the elevator for up to 30 minutes because it is always full indicates that the hotel is full and the service is good …
Positive (5) 35752	Services	The hotel is good with an affordable price. The breakfast is varied and delicious
Positive (5) 35752	Location	Good 10 minutes from the airport and a good varied breakfast! …
Positive (5) 35752	Operational Hour	I was very grateful because I was allowed to check in at 12 noon

2.2 DBSCAN Cauterization

DBSCAN is an algorithm used to dynamically form clusters with the help of an epsilon value. The number of clusters generated depends on the epsilon value specified in the algorithm. Comparing results from different algorithms helps determine which one is the most effective for stock price prediction. Developing DBSCAN from scratch allows for flexibility in modifying the algorithm to fit specific needs, and we can also incorporate values such as centroids into the model. This adaptability ensures that the algorithm can be tailored to suit various data patterns and improve the accuracy of predictions based on clustering results [14].

The clustering process involves several steps, as illustrated in the flowchart in Figure 3. Initially, hotel review data is collected, followed by preprocessing, which includes various stages such as case folding, stop word removal, tokenization, and stemming. After the preprocessing steps are completed, the data is vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This ensures that the reviews are represented in a numerical format, ready for further analysis and clustering.

The selection of three clusters ensures that the data is effectively grouped into meaningful segments, improving the accuracy of the overall clustering process. The optimization of the DBSCAN algorithm not only prevents poor clustering results but also enhances the performance of the model in identifying patterns within the dataset. Consequently, this approach allows for better insights into the data and more reliable conclusions from the analysis. The data clustered to be three cluster as: Data Cluster based on the Hotel Location; Data Cluster based on the operations Hour; and Data Cluster based on the quality of the services

DBSCAN is a clustering algorithm that groups data points by evaluating the density of their distances. Its primary strength lies in its capacity to detect and manage outliers or noise within a dataset. Unlike other clustering methods, DBSCAN does not require a predetermined number of clusters and can form clusters of varying shapes and sizes. This flexibility makes it particularly useful for datasets with irregular patterns. By identifying points that don't fit within a dense region, DBSCAN effectively separates noise from meaningful clusters, making it ideal for more complex and noisy data structures.

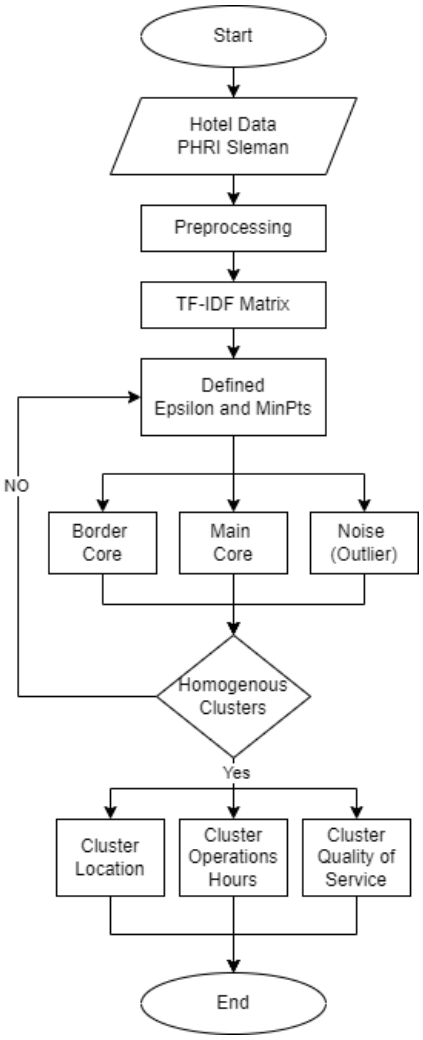

Figure 3. Hotel Data Clustering using DBSCAN

The algorithm achieves this by calculating distances between data points using a specific metric. In the case of DBSCAN, the most commonly used method for measuring these distances is the Euclidean Distance formula. This formula helps determine how close or far data points are from one another, which in turn helps to form clusters and distinguish outliers effectively as shown in Equation 1 [15].

$$Distance = \sqrt{(x - xp)^2 + (y - yp)^2} \quad (1)$$

With x and y as the coordinates of the target point, and Xp and Yp representing the coordinates of the reference point on the axis.

By using the DBSCAN clustering approach, the amount of data in each cluster is obtained in Figure 4, which we can see that the Operational hours have the fewers data and the quality of services have the biggest data. The quality-of-service data have a biggest number of reviews because hotel as the hospitality industry demanded to provide a excelent services by their customers. Comparing with another clusters like location that mostly already accurates based on google maps, the review is only about the distances with another places like shopping mall or transportation facilities, and the fewer review for operational hours because of the hotel mostly open 24 hours and the review mostly was about the check in or check out time that not as promised.

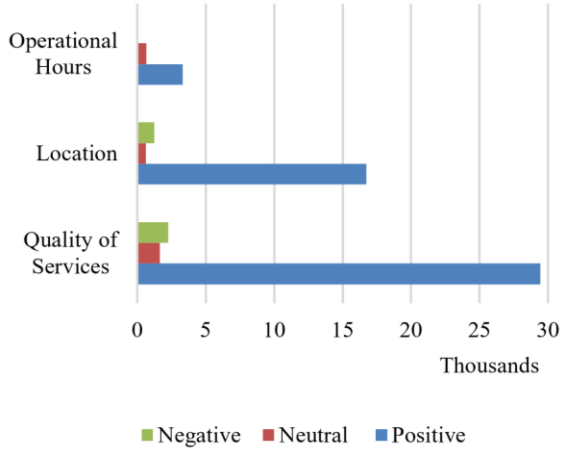

Figure 4. Data Cluster from DBSCAN

2.3. *Sentiment Analysis*

Social media sentiment mining for specific targets is a critical concern for decision-makers across various sectors, such as services, politics, entertainment, and manufacturing. This growing interest has led to a strong emphasis on Sentiment Analysis. Many studies have gone beyond simply extracting sentiment and have taken it a step further by diving deeper into subjective text. The goal of these advanced analyses is to uncover potential motivations behind the sentiments that are identified. By doing this, researchers aim to gain a more comprehensive understanding of the underlying factors that drive public opinion or customer feedback. This deeper insight can help businesses and organizations better tailor their strategies and responses, whether for marketing, public relations, or product development. In essence, sentiment mining serves as a powerful tool, not only to capture the mood of a target audience but also to interpret the reasons that fuel their opinions [16].

This paper presents an experiment comparing the performance of two algorithms, Naïve Bayes (NB) and Support Vector Machine (SVM). The aim is to determine which model performs better using different training data sets. These training data sets are organized based on three distinct clusters that were previously created, as illustrated in Figure 5.

By evaluating the models with data segmented into these clusters, the study seeks to gain insights into the strengths and weaknesses of each algorithm in handling different aspects of the data. The experiment is structured to analyze the efficiency of each model within the context of the clusters, ultimately revealing how they respond to varying characteristics within the training data. This process provides a comprehensive understanding of each algorithm's capabilities in classifying data from different clusters, helping to

identify the more suitable model for future applications in similar data-driven projects.

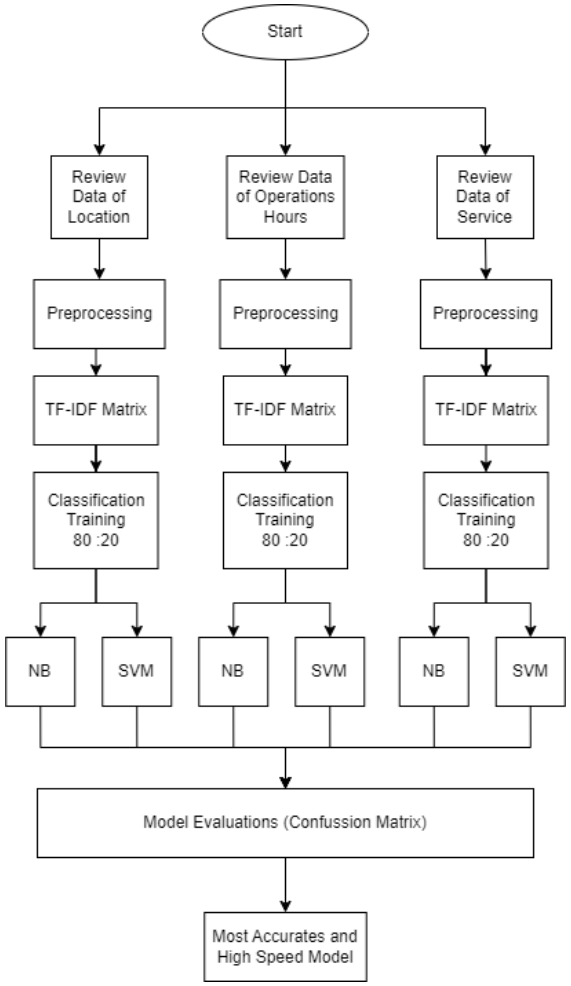

Figure 5. Classification Flow

Based on the diagram in Figure 4, it is evident that sentiment classification is conducted using three different types of data on two separate models. The data is split into 80% for training and 20% for validation. The algorithm employed in this paper follows the Naive Bayes theorem [17]as shown in Equation 3.

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad (3)$$

In the context of probability theory, X represents evidence, while H denotes a hypothesis. P(H|X) refers to the probability that hypothesis H is true given the evidence X, also known as the posterior probability of H, conditioned on X. Similarly, P(X|H) is the probability that the evidence X is true given hypothesis H, which is the posterior probability of X, conditioned on H. P(H) is the prior probability of the hypothesis H, and P(X) is the prior probability of the evidence X [17].

Support Vector Machine (SVM) uses a training data set in the format (Xi, yi), where Xi is a tuple and yi is a class label for i=1...N. Here, Xi belongs to R^d and yi belongs to {−1, 1}. The formula for SVM is presented in Equation 4 [18].

$$f(xi) = \{\geq 0, yi = +1 < 0, yi = -1\} \quad (4)$$

The formation of the hyperplane occurs is explained in Equation 5 [18]:

$$W.X + b = 0 \quad (5)$$

W is a weight vector consisting of w1, w2, w3, ..., wn, where n represents the number of attributes. b is a scalar, also referred to as bias. X represents the training dataset or the set of training tuples.

2.4. *Confusion Matrix*

The model evaluation process in this study will be conducted using the confusion matrix method. The use of a confusion matrix is particularly valuable for measuring how well the classification performs. Figure 6 illustrates a confusion matrix for a multi-class classification with three categories: Positive, Neutral, and Negative. This approach provides a clear visual representation of the model's accuracy in categorizing data into the defined classes. By assessing true positives, false positives, true negatives, and false negatives, the confusion matrix helps evaluate the overall performance and accuracy of the classification model.

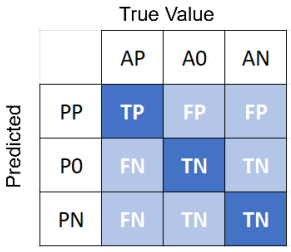

Figure 6. Confusion Matrix

Based on Figure 6, AP represents Actual Positive, A0 is Actual Neutral, and AN is Actual Negative, while PP stands for Predict Positive, P0 for Predict Neutral, and PN for Predict Negative. Using the information from TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative), accuracy, precision, and recall are calculated using Equations 6-8. These metrics help evaluate the performance of the model, offering insight into its ability to correctly predict positive, neutral, and negative outcomes, as well as minimizing false predictions in each category.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP+TF} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

3. Results and Discussions

This paper aims to compare two machine learning models, Naïve Bayes and Support Vector Machine, using various data clusters. The composition of the training and test data used in the comparison is presented in Table 3.

Table 3. Training and Test Data Composition on each Cluster

Cluster	Total Data	Training	Test
Services	33341	26673	6668
Location	18603	14883	3720
Operational Hours	4409	3528	881

The clustering process was performed using the DBSCAN method, dividing the data into three clusters based on the provided keywords. The time required for this clustering process is shown in Table 4.

Table 4. Speed or Timelog of DBSCAN Clustering

Cluster	Translating & Preprocessing (Hour)	TF-IDF Processing (Hour)	DBSCAN (Hour)
Services	5 : 57 : 41	0 : 0 : 22	0 : 0 : 13
Location	5 : 57 : 41	0 : 0 : 21	0 : 0 : 16
Operational Hours	5 : 57 : 41	0 : 0 : 19	0 : 0 : 32

The translation and preprocessing steps for all three clusters were performed once, and the resulting data was used to run DBSCAN for each cluster. As shown in Table 4, the operational hours cluster had the longest clustering time at 41 seconds, while the fastest was the Service cluster at 35 seconds.

Using the clustered data, model training was conducted for each cluster using the Naïve Bayes and Support Vector Machine algorithms, producing six classification results with varying accuracies. Figure 7 shows the confusion matrix related to the validation results for the operational hours cluster training.

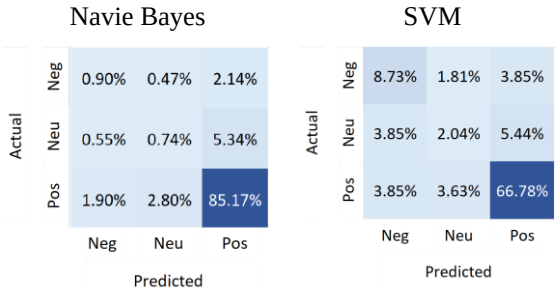

Figure 7 Confussion Matrix for Operational Hours Cluster

On the Operational hours cluster, the confusion matrix compares Naive Bayes (NB) and Support Vector Machine (SVM) performance in sentiment classification. NB demonstrates a higher accuracy in predicting positive sentiments (85.17%) but slightly struggles with neutral and negative predictions. SVM performs better in identifying negative sentiments (8.73%) but has lower accuracy in positive predictions (66.78%). NB's precision for positives is higher, while SVM excels at negatives. The overall accuracies of the Operational Hour cluster, can be seen on Table 5.

Table 5. Training Result for Operational Hours

Index	NB	SVM
Accuracy	0.87	0.78
Precision	0.85	0.76
Recall	0.87	0.78
F1 Score	0.86	0.77

Using location cluster, the training conducted using 18603 data, with 80% training data and 20% testing data, with result can be seen on Figure 8.

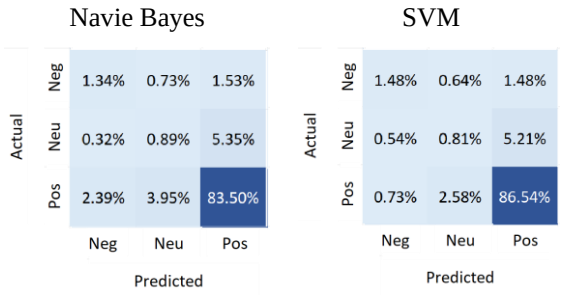

Figure 8. Confussion Matrix for Location Cluster

The confusion matrix on Figure 8 shows that SVM outperforms Naive Bayes in predicting positive sentiments (86.54% vs. 83.50%). However, Naive Bayes achieves better accuracy in classifying neutral sentiments, making both models useful in different scenarios and the overall accuracies of the Operational Hour cluster, can be seen on Table 6.

Table 6. Training Result for Location Cluster

Index	NB	SVM
Accuracy	0.86	0.89
Precision	0.85	0.87
Recall	0.86	0.89
F1 Score	0.85	0.88

The biggest data cluster used by the services cluster, since the biggest visitor review is about the quality of services, using 33341 data on below figure 9 we can see the result.

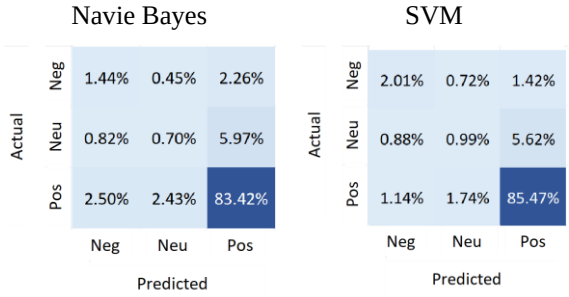

Figure 9 Confusion Matrix for Service Cluster

Based on Figure 9 the confusion matrix compares Naive Bayes (NB) and SVM performance in sentiment classification. NB performs slightly better in predicting positive sentiments (83.42%), while SVM excels with an 85.47% accuracy for positives. SVM is more effective at identifying neutral sentiments (0.99%) and negatives (2.01%). NB's neutral classification is weaker, but both models handle positive predictions reasonably well. The accuracies of service cluster can be seen on Table 7.

Based on the experiments conducted across the three clusters, it was found that Naïve Bayes performed better in terms of accuracy for the operational hours cluster. However, Support Vector Machine (SVM) showed improved accuracy, surpassing Naïve Bayes in the

location and service clusters, which had larger training datasets compared to the operational hours cluster. This improvement in SVM's performance is likely due to the increased amount of data available in these clusters. The results, which highlight the superior accuracy of SVM in these cases, can be observed in Figure 10.

Table 7. Training Result for Location Cluster

Index	NB	SVM
Accuracy	0.86	0.88
Precision	0.83	0.86
Recall	0.86	0.88
F1 Score	0.84	0.87

These findings suggest that while Naïve Bayes is effective for smaller datasets, it can be seen on the Operational hour accuracies 0.87 higher than other Naïve bayes accuracies, SVM tends to excel with larger training sets, particularly in clusters with more comprehensive data such as location and service.

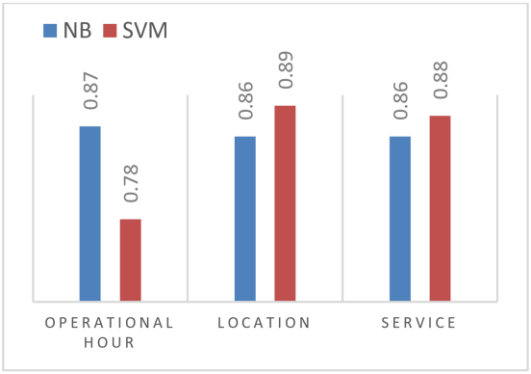

Figure 10 Accuracies Comparison On Each Cluster

Based on Figure 10, SVM has the lowest accuracy during cluster operational hours, this can be correlated with the amount of training and testing data during cluster operational hours which is quite low compared to cluster locations and services.

SVM is a complex machine learning model so that the use of more training data will produce optimal performance. It can be proven that SVM is superior to NB in clusters with larger data amounts, such as "Location," but its performance decreases in clusters with small data such as "Operational Hours." How the amount of data affects the accuracy of SVM can be seen in Figure 11.

Figure 11 SVM Accuracies Comparing with Data Training

One of the parameters used to compare algorithms is by analyzing the training speed, which helps in calculating the computational load, as shown in Figure 11.

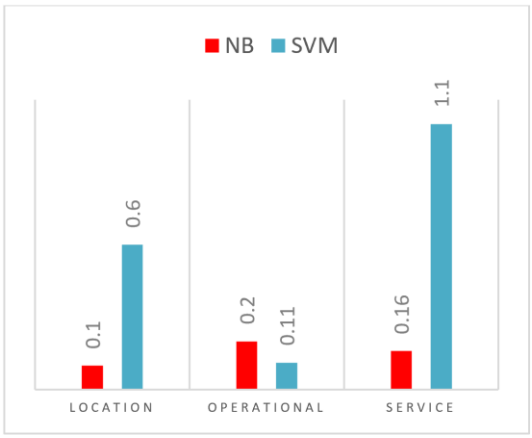

Figure 12 Speed Comparison On Each Cluster

Referring to Figure 12, the training speed, excluding preprocessing, is relatively fast, with times under 1 second. The fastest time was observed in the service cluster, where the Support Vector Machine model completed training in 1.1 seconds.

The comparison of accuracy between this study and other studies can be seen in Table 8. The results of this study are still within the accuracy range of previous studies, although not as high as the previous studies.

Table 8. Comparison with the previous research

Referensi	Research Object	NB	SVM
[19]	Opensea Apps Sentiment Analysis	89%	91%
[20]	Lazada Apps Sentiment Analysis	83%	88%
[20]	Tokopedia Apps Sentiment Analysis	85%	86%
[21]	Fake News Detection	89%	96%
[22]	Cyberbulying with Features Selection	91%	90%
[22]	Cyberbulying without Features Selection	84%	82%
This Research	Operational Hour Sentiment Analysis	87%	78%
This Research	Location Sentiment Analysis	86%	89%
This Research	Service Sentiment Analysis	86%	88%

4. Conclusions

In this study, we conducted sentiment analysis using an initial approach of clustering data based on specific categories, namely service quality, location, and operational hours. This categorization was done using the DBSCAN algorithm to obtain varying amounts of data, allowing for a comparison of classification models based on data volume. The data used were hotel reviews sourced from the website of the Indonesian Hotel and Restaurant Association (PHRI) in Sleman Regency, with a total of 53,000 reviews collected from Google Maps reviews. Using DBSCAN, it was found that the service quality cluster contained 33,341 reviews, the

location cluster had 18,603, and the operational hours cluster had 3,961. The experiments yielded varying accuracies with small gaps. In the operational hours cluster, Naïve Bayes achieved an accuracy of 0.87, while SVM only reached 0.78. However, in the location cluster, SVM improved to 0.89, compared to Naïve Bayes at 0.86. Similarly, in the service quality cluster, SVM had an accuracy of 0.88, and Naïve Bayes scored 0.86. The improved accuracy of SVM is likely influenced by the larger data volume used for training, particularly in the service quality cluster, which provided more data for SVM's training process. This also impacted training speed, with SVM taking 1.1 seconds in the service cluster, while the other clusters remained under 1 second. A limitation of this study is the use of ratings for data labeling, which may introduce some bias, though not significant. Future studies should aim to use more accurate and appropriate labels.

References

[1] M. Khatoon and W. A. Banu, "Unsupervised algorithms comparison in the perspective of community detection from social networks," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 391–395. doi: 10.1109/ICIRCA51532.2021.9544555.

[2] L. Ma, "An improved and heuristic-based iterative DBSCAN clustering algorithm," in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2021, pp. 2709–2714. doi: 10.1109/IAEAC50856.2021.9390918.

[3] N. P. Sutramiani, I. M. T. Arthana, P. F. Lampung, S. Aurelia, M. Fauzi, and I. W. A. S. Darma, "The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 13–24, 2024, doi: 10.20473/jisebi.10.1.13-24.

[4] F. Andriyani and Y. Puspitarani, "Performance Comparison of K-Means and DBScan Algorithms for Text Clustering Product Reviews," *SinkrOn*, vol. 7, no. 3, pp. 944–949, Jul. 2022, doi: 10.33395/sinkron.v7i3.11569.

[5] Y. Lu, Y. Huang, H. Yu, and Y. Lan, "Research on consumer service quality based on hotel online reviews," in *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, 2022, pp. 836–840. doi: 10.1109/AEMCSE55572.2022.00168.

[6] H. Adiningtyas and H. Millanyani, "Analysis of Customer Satisfaction Levels in Five-Star Hotels Based on Online Customer Reviews," in *2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)*, 2024, pp. 167–174. doi: 10.1109/ICoSEIT60086.2024.10497518.

[7] A. Abraham *et al.*, "Naïve Bayes Approach for Word Sense Disambiguation System with a Focus on Parts-of-Speech Ambiguity Resolution," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3453912.

[8] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 1, p. 55, Jan. 2021, doi: 10.22146/ijccs.60718.

[9] B. Yanuargi, "Analisis sentimen terhadap aplikasi Bukalapak sebelum dan sesudah IPO menggunakan algoritma Naïve Bayes," 2022, doi: 10.36802/jnanaloka.v3-no1-17-25.

[10] I. R. Hendrawan, E. Utami, and A. D. Hartanto, "Comparison of Naïve Bayes Algorithm and XGBoost on Local Product Review Text Classification," *Edumatic: Jurnal Pendidikan Informatika*, vol. 6, no. 1, pp. 143–149, Jun. 2022, doi: 10.29408/edumatic.v6i1.5613.

[11] F. Fitriana, E. Utami, and H. Al Fatta, "Analisis Sentimen Opini Terhadap Vaksin Covid - 19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 5, no. 1, pp. 19–25, Jul. 2021, doi: 10.31603/komtika.v5i1.5185.

[12] C. K. Wang, "Sentiment Analysis Using Support Vector Machines, Neural Networks, and Random Forests," 2023, pp. 23–34. doi: 10.2991/978-94-6463-300-9_4.

[13] S. Riadi, E. Utami, and A. Yaqin, "Comparison of NB and SVM in Sentiment Analysis of Cyberbullying using Feature Selection," *sinkron*, vol. 8, no. 4, pp. 2414–2424, Oct. 2023, doi: 10.33395/sinkron.v8i4.12629.

[14] L. S. Parvatha, D. Naga Veera Tarun, M. Yeswanth, and Jonnalagadda. S. Kiran, "Stock Market Prediction Using Sentiment Analysis and Incremental Clustering Approaches," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2023, pp. 888–893. doi: 10.1109/ICACCS57279.2023.10112768.

[15] M. Ula, Tsania Asha Fadilah Daulay, Richki Hardi, Sujacka Retno, Angga Pratama, and Ilham Sahputra, "Density Based Spatial Clustering of Applications and Spatial Pattern Analysis In Mapping the Distribution of ISPA Disease in Bireuen Regency," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 3, pp. 733–742, Jun. 2023, doi: 10.29207/resti.v7i3.4936.

[16] F. Alattar and K. Shaalan, "A Survey on Opinion Reason Mining and Interpreting Sentiment Variations," *IEEE Access*, vol. 9, pp. 39636–39655, 2021, doi: 10.1109/ACCESS.2021.3063921.

[17] F. Alghifari and D. Juardi, "Fauzan Alghifari Penerapan Data Mining Pada Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," 2021.

[18] Y. Kustiyahningsih and Y. Permana, "Penggunaan Latent Dirichlet Allocation (LDA) dan Support-Vector Machine (SVM) Untuk Menganalisis Sentimen Berdasarkan Aspek Dalam Ulasan Aplikasi EdLink," *Teknika*, vol. 13, no. 1, pp. 127–136, Mar. 2024, doi: 10.34148/teknika.v13i1.746.

[19] S. Riadi, E. Utami, and A. Yaqin, "Comparison of NB and SVM in Sentiment Analysis of Cyberbullying using Feature Selection," *sinkron*, vol. 8, no. 4, pp. 2414–2424, Oct. 2023, doi: 10.33395/sinkron.v8i4.12629.

[20] I. Kurniawan *et al.*, "Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 10, no. 1, 2023, [Online]. Available: http://jurnal.mdp.ac.id

[21] N. Nurhasanah, D. E. Sumarly, J. Pratama, I. T. K. Heng, and E. Irwansyah, "Comparing SVM and Naïve Bayes Classifier for Fake News Detection," *Engineering, MAthematics and Computer Science (EMACS) Journal*, vol. 4, no. 3, pp. 103–107, Sep. 2022, doi: 10.21512/emacsjournal.v4i3.8670.

[22] S. Riadi, E. Utami, and A. Yaqin, "Comparison of NB and SVM in Sentiment Analysis of Cyberbullying using Feature Selection," *sinkron*, vol. 8, no. 4, pp. 2414–2424, Oct. 2023, doi: 10.33395/sinkron.v8i4.12629.

