


Comparative Study of Support Vector Machine and Naive Bayes for Sentiment Analysis on Lecturer Performance

Debora Chrisinta^{*1}, Justin Eduardo Simarmata²

¹Information Technology Study Program, University of Timor, Kefamenanu, 85613, Indonesia

²Mathematics Education Study Program, University of Timor, Kefamenanu, 85613, Indonesia

*Corresponding Author: deborachrisinta@unimor.ac.id

ARTICLE INFO

Article history:

Received: 01 January 2023

Revised: 02 February 2023

Accepted: 29 March 2023

Available online: 30 March 2023

E-ISSN: 2656-1514

P-ISSN: -

How to cite:

Chrisinta, D., Simarmata, J.E., "Comparative Study of Support Vector Machine and Naive Bayes for Sentiment Analysis on Lecturer Performance", Journal of Research in Mathematics Trends and Technology, vol. V5, no. 1, Mar. 2023, doi: 10.32734/jormtt.v5i1.15864

ABSTRACT

This study addresses the challenge of sentiment analysis within the Information Technology study program at Universitas Timor, aiming to compare the performance of Support Vector Machines (SVM) and Naive Bayes (NB) through 100 iterations. The dataset, comprising 21 instances of negative sentiment and 18 instances of positive sentiment, is analyzed using both methods, with accuracy and Area Under the ROC Curve (AUC) serving as key metrics. The sample size consists of 39 instances, and the results indicate significant variability in both accuracy and AUC, emphasizing the sensitivity of the models to dataset characteristics and random initialization. On average, SVM outperforms NB, with an accuracy of 0.5846 compared to 0.5075 and an AUC of 0.5916 compared to 0.4607.

Keyword: Support Vector Machines, Naive Bayes, Sentiment Analysis



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International.

<http://doi.org/10.32734/jormtt.v5i1.15864>

1. Introduction

Sentiment analysis, also known as opinion mining, is a computational technique designed to discern the emotional tone expressed in written text. The primary objective is to determine whether the sentiment conveyed is positive, negative, or neutral [1]. This process is crucial in understanding public opinions, customer feedback, and social media interactions. The workflow involves collecting textual data, preprocessing it by cleaning and organizing the text, and extracting relevant features for analysis [2]. Sentiment analysis employs various approaches, including rule-based methods, where predefined rules dictate sentiment based on specific words or patterns, and machine learning techniques that leverage algorithms trained on labeled datasets [3]. These algorithms learn to associate features with sentiments, allowing them to classify new data accurately. The applications of sentiment analysis are widespread, ranging from customer sentiment in product reviews to monitoring social media for brand reputation. As technology evolves, sentiment analysis continues to play a vital role in extracting valuable insights from vast amounts of textual information in the

digital landscape [4], [5], [6]. Sentiment analysis employs various classification models to discern the emotional tone expressed in text. Naive Bayes, a probabilistic model, assumes word independence, while Support Vector Machines construct a hyperplane for classification [7], [8], [9]. Logistic Regression predicts probabilities using a logistic function, and Random Forests use an ensemble of decision trees. Gradient Boosting methods sequentially build models to improve accuracy. Neural Networks, including deep learning architectures, learn intricate patterns, while Transformer models, like BERT and GPT, capture contextual information. The choice depends on factors such as dataset size, language complexity, and interpretability preferences. Pre-trained models, especially in deep learning and transformers, offer high-performance outcomes when fine-tuned for specific sentiment analysis tasks [10]. In this research, the focus lies on sentiment analysis using Support Vector Machines (SVM) and Naive Bayes (NB) methods. Both SVM and NB have seen extensive use in recent sentiment analysis research. The emphasis is on the effectiveness of SVM in high-dimensional spaces, its adeptness at handling non-linear relationships in data, and its suitability for feature-rich datasets. Conversely, NB, a probabilistic model assuming feature independence, is recognized for its computational efficiency and demonstrated effectiveness, particularly in smaller datasets [11]. The study aims to compare the performance of SVM and NB on the specific sentiment analysis dataset, evaluating metrics such as accuracy and AUC (Area Under the ROC Curve). Base on Huang proved that AUC is, in general, a better measure (defined precisely) than accuracy [12]. This comparison seeks to provide insights into the strengths and limitations of each method, contributing valuable knowledge to the field of sentiment analysis.

In the context of prior research, Rana & Singh demonstrated that SVM achieved the highest accuracy in sentiment analysis [13], in contrast to Lawal et al, who favored the Naïve Bayes classifier [14]. Rahat et al. subsequently supported SVM's superiority over Naïve Bayes [15]. Given these conflicting results, this study seeks to apply sentiment analysis for evaluating lecturer performance in the Technology Information department of Universitas Timor, contributing to the ongoing exploration of the most effective machine learning approach in this specific domain.

2. Methodology

The methodology stage of the study involves outlining the systematic approach and procedures for conducting sentiment analysis on lecturer performance in the Technology Information department of Universitas Timor. The general outline of the methodology stage as below:

1. Collect a representative dataset of comments or reviews related to lecturer performance from various student evaluations. The instrument used to collect responses from a questionnaire.
2. Clean and preprocess the collected data, including tasks such as text normalization, removal of irrelevant information, handling of missing data, and tokenization.
3. Divide the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance.
4. Applying SVM formulation, the optimization problem involves finding the optimal weight vector (ω) and bias term (b) such that the decision boundary effectively separates the classes. The constraints and the

summation term in the objective function are defined for each of the m data points:

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m \max(0, 1 - y_i(\omega \cdot x_i + b)) \quad (1)$$

where x_i is feature vector for the i th data point, y_i is Class label for the i th data point (1 for positive class, -1 for negative class), ω is the weight vector, b is the bias term, C is the regularization parameter, and m is the number of data points in the dataset.

- Applying NB formulation with calculate the posterior probability of each sentiment class given the observed features (words) from Bayes' theorem:

$$p(\text{Class} = c | \text{Words}) \propto P(\text{Class} = c) \times \prod_{i=1}^n P(\text{Word}_i | \text{Class} = c) \quad (2)$$

where c is represents the sentiment class, \propto is represents proportionality and Word_i is represents the i -th word in the document

- Train the SVM and NB model using the prepared training dataset.
- Use the trained SVM and NB model to predict sentiment labels on the testing dataset.
- Performing step 4 (Model Training) in a loop for 100 iterations allows to observe the variability in the performance metrics across different training runs.
- Evaluate the performance of the SVM and NB model on the testing dataset using metrics such as accuracy and AUC.
- Select the better-performing model based on the looped, based on average accuracy and AUC for both SVM and Naive Bayes.

3. Result and Discussion

The domain of sentiment analysis within the Information Technology study program at Universitas Timor, the provided counts of 21 for negative sentiment and 18 for positive sentiment represent the numerical distribution of sentiments within the analyzed dataset. These figures signify that, based on the content derived from the Information Technology study program, there are 21 instances identified as having a negative sentiment and 18 instances characterized as expressing a positive sentiment (Figure 1).

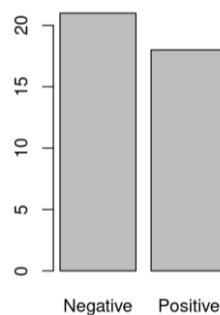


Figure 1. Data Distribution on Sentiment Analysis

Given the specific characteristics of dataset, which revolves around sentiment labels derived from the Information Technology study program at Universitas Timor, it is advisable to employ both Support Vector

Machines (SVM) and Naive Bayes (NB) methods for sentiment analysis. These machine learning algorithms are widely used in the field and are known for their effectiveness in uncovering patterns and relationships within textual data. Support Vector Machines (SVM) are particularly suitable for scenarios where the feature space is complex and high-dimensional. Their ability to capture non-linear relationships makes them well-suited for sentiment analysis tasks, where the nuances of language may require a more sophisticated approach. On the other hand, Naive Bayes (NB) is a probabilistic model that assumes feature independence. Despite its simplicity, NB is computationally efficient and has shown effectiveness, especially in situations with smaller datasets. Its quick training times and robustness in cases where feature independence assumptions align with the data characteristics make it a viable choice for sentiment analysis. By applying both SVM and NB methods, such that enable a comparative analysis of their performance on specific sentiment analysis dataset. Metrics such as accuracy and the AUC can be evaluated to discern which model aligns better with the inherent characteristics of data and provides more accurate sentiment predictions. To ensure a robust evaluation, proper preprocessing of the data, division into training and testing sets, and fine-tuning of model parameters are crucial steps. Through this comparative analysis, can make an informed decision about whether SVM or NB is better suited for sentiment analysis within the unique context of the Information Technology study program at Universitas Timor.

The sentiment analysis task has been repeated or looped 100 times. Each iteration might involve different subsets of the data, different training/test splits, or variations in the model parameters to account for randomness or variability in the results. The sentiment analysis has been conducted using two different methods or algorithms. These could be Support Vector Machines (SVM) and Naive Bayes (NB), as previously discussed. The results, likely in the form of plots or graphs, are presented in two separate figures. One figure is dedicated to displaying the accuracy trends over the 100 iterations for both methods, and the other figure illustrates the AUC trends over the same iterations. These figures provide a visual representation of how the performance of the two methods evolves across multiple loops. The statement indicates a rigorous evaluation of sentiment analysis models through repeated iterations, with a focus on accuracy and AUC as the performance metrics, and the results are visually presented in two separate figures for comprehensive analysis and comparison.

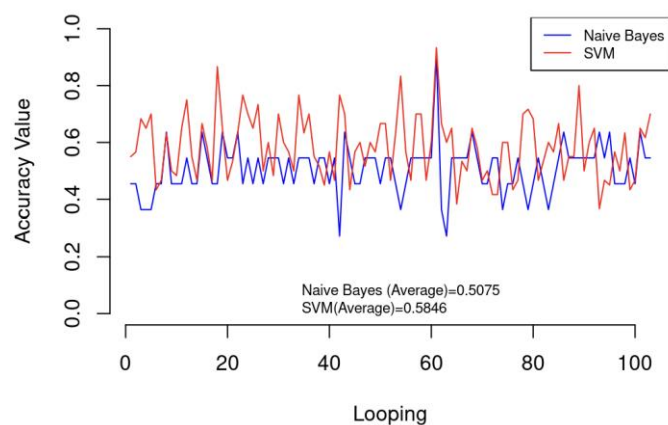


Figure2. The Accuracy Trends Over the 100 Iterations

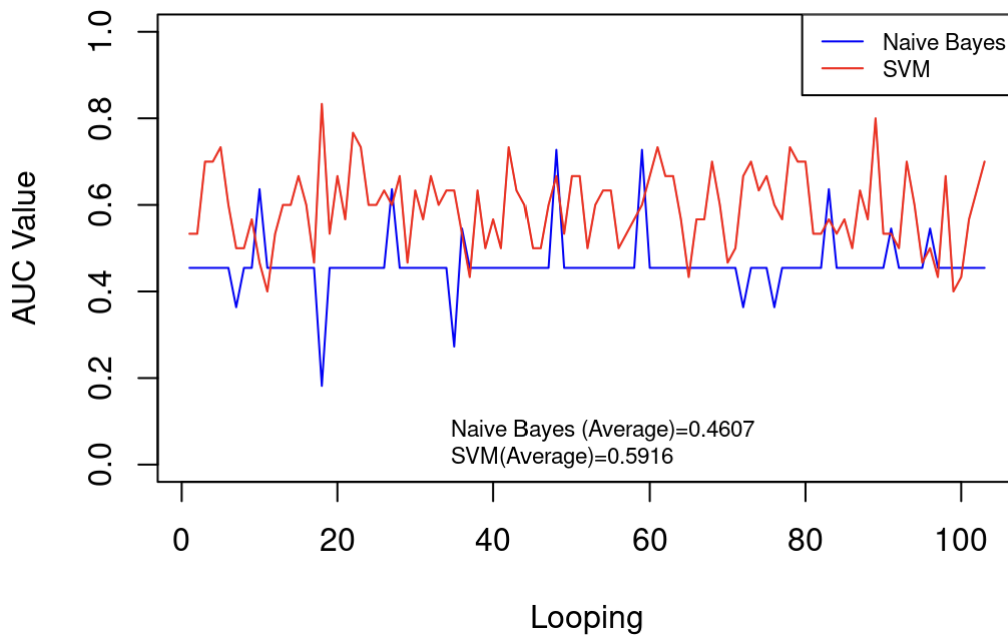


Figure3. The AUC Trends Over the 100 Iterations

The results from sentiment analysis experiments employing NB and SVM models are presented in the provided table, spanning multiple iterations (1 to 100). The accuracy values for both NB and SVM exhibit considerable variation, ranging from approximately 0.36 to 0.91 for NB and 0.38 to 0.93 for SVM. Similarly, the AUC values, representing the models' discrimination ability, fluctuate between approximately 0.18 and 0.73 for NB and 0.40 and 0.80 for SVM. The observed variability in performance suggests that the models' effectiveness is influenced by factors such as dataset characteristics or random initialization during training. To determine the superior model, consideration of average performance metrics or specific criteria, such as consistently higher AUC values, may be employed. The average performance metrics across multiple iterations indicate that, on average, the Support Vector Machines (SVM) method outperforms Naive Bayes (NB) in the context of sentiment analysis. The average accuracy for SVM is 0.5846, compared to 0.5075 for NB. Additionally, the average Area Under the ROC Curve (AUC) for SVM is 0.5916, while NB has an average AUC of 0.4607. These results suggest that SVM exhibits a more consistent and superior performance in distinguishing between positive and negative sentiments in the analyzed dataset. However, it is crucial to consider the specific requirements of the sentiment analysis task and the associated limitations of each method when making a final decision. The observed differences in average performance metrics provide valuable insights for selecting the most suitable method for the given application.

4. Conclusion and Future Research

the sentiment analysis within the Information Technology study program at Universitas Timor, conducted through 100 iterations using SVM and NB reveals a dataset with 21 instances of negative sentiment and 18 instances of positive sentiment. Both SVM and NB were chosen for their effectiveness in handling textual data patterns, with SVM exhibiting superior performance on average. Visualizations of accuracy and AUC trends over iterations illustrate considerable variability, emphasizing sensitivity to dataset characteristics and random

initialization. Average accuracy metrics of 0.5846 for SVM and 0.5075 for NB, along with average AUC values of 0.5916 for SVM and 0.4607 for NB, suggest SVM's consistent and superior ability to distinguish between positive and negative sentiments.

References

- [1] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research*, vol. 50, pp. 723–762, 2014, doi: 10.1613/jair.4272.
- [2] A. Kathuria, A. Gupta, and R. K. Singla, "A review of tools and techniques for preprocessing of textual data," in *Computational Methods and Data Engineering: Proceedings of ICMDE Volume 1*, 2021, pp. 407–422. doi: 10.1007/978-981-15-6876-3_31.
- [3] B. Pham *et al.*, "Text mining to support abstract screening for knowledge syntheses: a semi-automated workflow," *Syst Rev*, vol. 10, no. 1, p. 156, 2021, doi: 10.1186/s13643-021-01700-x.
- [4] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, p. 1, 2020, doi: 10.3390/bdcc4010001.
- [5] S. Yadav, A. Kaushik, M. Sharma, and S. Sharma, "Disruptive technologies in smart farming: an expanded view with sentiment analysis," *AgriEngineering*, vol. 4, no. 2, pp. 424–460, 2022, doi: 10.3390/agriengineering4020029.
- [6] P. Sánchez-Núñez, M. J. Cobo, C. De Las Heras-Pedrosa, J. I. Peláez, and E. Herrera-Viedma, "Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis," in *IEEE Access* 8, 2020, pp. 134563–134576. doi: 10.1109/ACCESS.2020.3009482.
- [7] K. K. Kiilu, "Sentiment Classification for Hate Tweet Detection in Kenya on Twitter Data Using Naïve Bayes Algorithm," Doctoral dissertation, JKUAT-COETEC, 2021.
- [8] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl Based Syst*, vol. 226, p. 107134, Mar. 2021, doi: 10.1016/j.knosys.2021.107134.
- [9] D. Chrisinta and J. E. Simarmata, "Analisis Sentimen Penilaian Masyarakat Terhadap Pejabat Publik Menggunakan Algoritma Naïve Bayes Classifier," *Komputika: Jurnal Sistem Komputer*, vol. 12, no. 1, pp. 93–101, 2023, doi: 10.34010/KOMPUTIKA.V12I1.9638.
- [10] S. Singh and A. Mahmood, "The NLP cookbook: modern recipes for transformer based deep learning architectures," in *IEEE Access* 9, 2021, pp. 68675–68702. doi: 10.1109/ACCESS.2021.3077350.
- [11] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, 2021, doi: 10.1007/s00500-020-05297-6.
- [12] J. Huang, J. Lu, and C. X. Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy," in *Third IEEE International Conference on Data Mining*, 2003, pp. 553–556. doi: 10.1109/ICDM.2003.1250975.

- [13] S. Rana and A. Singh, “Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques,” in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 106–111. doi: 10.1109/NGCT.2016.7877399.
- [14] M. A. Lawal, R. A. Shaikh, and S. R. Hassan, “Security analysis of network anomalies mitigation schemes in IoT networks,” in *IEEE Access*, 8, 2023, pp. 43355–43374. doi: 10.1109/ACCESS.2020.2976624.
- [15] A. M. Rahat, A. Kahir, and A. K. M. Masum, “Comparison of Naive Bayes and SVM Algorithm based on sentiment analysis using review dataset,” in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, pp. 266–270. doi: 10.1109/SMART46866.2019.9117512.