International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 3, Number 1, May 2024, *Page. 287 - 293*
Email : editorijhet@gmail.com

Sentiment Analysis on Twitter Social Media towards Najwa Shihab Using Naïve Bayes Algorithm and Support Vector Machine (SVM)

Fahruzi Sirait [1)], Desi Irpan[2)], Riszki Fadillah[3)], Rizalina[4)], Riswan Syahputra Damanik[5)]
[1,5)]Information Systems, Faculty of Computer Science, Ika Bina Institute of Technology and Health
[2,3)]Information Technology, Faculty of Computer Science, Ika Bina Institute of Technology and Health
[4)]Information Systems, Faculty of Computer Science, UPI YPTK Padang

*Coresponding Author
Email : fahruzi.sirait21@itkes-ikabina.ac.id

Abstract

With the rapid growth of digital technology, social media has become a key platform for sharing information and opinions. Twitter, one of the most popular platforms in Indonesia, enables users to interact directly with public figures such as Najwa Shihab. This study aims to analyze public sentiment toward Najwa Shihab on Twitter using sentiment analysis, specifically employing the Naïve Bayes and Support Vector Machine (SVM) algorithms. Sentiment analysis is essential to understanding public opinion, as it classifies text into categories like positive, negative, or neutral, providing valuable insights into societal perspectives on public figures. In this study, 10,000 tweets related to Najwa Shihab were collected from January 1, 2023, to January 31, 2023. Data preprocessing steps such as data cleaning, tokenization, stopwords removal, and filtering were conducted to ensure high-quality data for analysis. The Naïve Bayes and SVM algorithms were applied using RapidMiner to classify the sentiment of the tweets. The performance of both algorithms was evaluated based on accuracy, precision, recall, and F1-score.The results revealed that SVM outperformed Naïve Bayes in all metrics, demonstrating its superior ability to classify sentiments correctly. The sentiment distribution indicated a majority of positive opinions toward Najwa Shihab, with fluctuations in negative sentiment during specific events. This study provides insights into public sentiment analysis and contributes to understanding social media opinions on public figures.

Keywords : Sentiment Analysis, Naïve Bayes, Support Vector Machine (SVM), Public Opinion, Twitter.

INTRODUCTION

Along with the advancement of digital technology, social media has become one of the main platforms for sharing information and opinions. Twitter, as one of the most popular social media in Indonesia, allows its users to interact directly with public figures, such as Najwa Shihab. Sentiment analysis on Twitter is very important to understand how public opinion is towards certain figures or issues, whether in terms of positive, negative, or neutral (Riyanto, 2022). According to Riyanto (2022), social media users in Indonesia continue to grow rapidly, with Twitter being one of the most accessed platforms by internet users. The diversity of opinions expressed in tweets creates a large volume of data that can be analyzed to explore public sentiment towards various issues, including famous figures such as Najwa Shihab (Sugianto & Maulana, 2019). Sentiment analysis is a method used to classify opinions in text into certain categories, such as positive, negative, or neutral (Pang & Lee, 2008). This technique is widely applied in various domains, including business, politics, and on social media such as Twitter, which allows for automatic measurement of public sentiment towards a particular figure, event, or issue, providing deeper insight into the views of the public (Mohammad et al., 2013). Although Najwa Shihab is widely known by the Indonesian public, there are various variations of opinions reflected on Twitter about her. Therefore, sentiment analysis of tweets related to Najwa Shihab is very important to provide an objective picture of how the public views her (Prasetyo, 2019).

This study aims to analyze public sentiment towards Najwa Shihab on Twitter using the Naïve Bayes algorithm and Support Vector Machine (SVM) (Rish, 2001). In addition, this study also aims to compare the performance of the two algorithms in classifying tweets based on different accuracy, precision, and recall metrics (Quinlan, 1993). Furthermore, this study will assess which

algorithm is more effective in identifying positive and negative sentiments related to Najwa Shihab (Cover & Hart, 1967). This study is expected to provide deeper insight into public perception of Najwa Shihab through sentiment analysis. In addition, the results of this study can also be a reference for similar studies on other public figures or social issues that are developing in society (Manning et al., 2008). This study can also provide benefits for companies or organizations in monitoring public opinion and increasing interaction with the community (Blei et al., 2003).

RESEARCH METHODS

Data Collection

The data in this study will be collected through Twitter using web crawling techniques with the keyword "Najwa Shihab". A total of 10,000 tweets will be collected for 30 days, from January 1, 2023 to January 31, 2023. Only tweets written in Indonesian will be considered for this analysis. The data collected includes tweet text, date, time, number of retweets, and likes. With a fairly large amount of data and a specific time span, this study aims to obtain more representative and accurate results in the analysis of public sentiment towards Najwa Shihab.

Table 1. Research variables and descriptions

Variable	Description	Data Type	Example Value
Tweet Text	Teks tweet yang mengandung opini tentang Najwa Shihab	Teks	"Najwa Shihab sangat inspiratif!"
Sentiment Label	Kategori sentimen dari tweet (positif, negatif, netral)	Kategorikal	Positif, Negatif, Netral
Retweets	Jumlah retweet yang diterima oleh tweet	Numerik	15
Likes	Jumlah like yang diterima oleh tweet	Numerik	25
Date	Tanggal tweet diposting	Tanggal	10/1/2023
User Location	Lokasi pengguna Twitter (opsional, bisa digunakan untuk analisis lebih lanjut)	Kategorikal	Jakarta, Bandung

Table 2. Sample data

Tweet ID	Tweet Text	Sentiment Label	Retweets	Likes	Date	User Location
1	"Najwa Shihab sangat inspiratif!"	Positif	15	45	10/1/2023	Jakarta
2	"Saya tidak setuju dengan pandangan Najwa."	Negatif	7	12	10/2/2023	Surabaya
3	"Najwa Shihab objektif dalam setiap topik."	Positif	10	30	10/3/2023	Yogyakarta
4	"Najwa Shihab terlalu	Negatif	3	9	10/4/2023	Bandung

	bias."					
...	...	...	...	...	...	...
10.000	"Saya sangat suka cara Najwa Shihab membawakan acara."	Positif	8	25	10/5/2023	Medan

Data Preprocessing

The collected data will be processed through several preprocessing stages to ensure the quality of the data to be used in the analysis. These stages include:

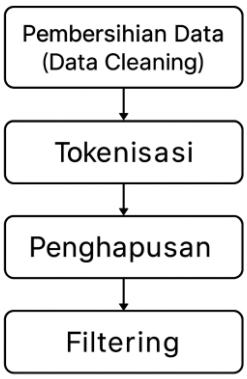

Figure 1. Preprocessing flowchart

1) **Data Cleaning** : Removing irrelevant characters such as URLs, emojis, punctuation, and other symbols that do not contribute to sentiment analysis.
2) **Tokenization** : Splitting the tweet text into smaller units (words) to facilitate processing and analysis.
3) **Stopwords Removal :** Removing common words that do not provide significant information in sentiment analysis, such as "yang," "di," "dan," etc.
4) **Filtering :** Filtering tweets that are irrelevant to the topic being analyzed, such as tweets containing spam or those not related to Najwa Shihab.

Modeling and Classification

At this stage, the Naïve Bayes and Support Vector Machine (SVM) algorithms are applied to classify tweets based on the sentiment they contain: positive, negative, or neutral. These two algorithms were chosen because they are both popular and effective methods for sentiment analysis.

1) **Naïve Bayes :** A classification algorithm based on probability that assumes the features in the data are independent of each other. This algorithm is highly efficient and frequently used for text classification.
2) **Support Vector Machine (SVM) :** An algorithm that finds the best hyperplane to separate data into different classes. SVM works well with large and complex data and is often used in classification problems.

Algorithm Implementation in RapidMiner

RapidMiner is used to implement the Naïve Bayes and SVM algorithms in this study, which analyzes public sentiment towards Najwa Shihab on Twitter. The data used includes tweets related to Najwa Shihab, including text, date, time, retweets, and likes. The following are the steps for implementing the algorithm:

1) **Data Import** : **The** processed data is imported into RapidMiner Studio in CSV or Excel format.
2) **Preprocessing** : Using operators like "Tokenize", "Remove Stopwords", and "Text Processing" to clean and prepare the data.
3) **Modeling** : Building the model using the "Naïve Bayes" and "Support Vector Machine" operators in RapidMiner.

4) Evaluation : Using "Cross Validation" and "Performance (Classification)" to measure accuracy, precision, recall, and generate a confusion matrix.

Visualization : Visualizing the evaluation results using "Plot View" or "Bar Chart" for model performance comparison

RESULTS AND DISCUSSION

In this section, we present the results of the sentiment analysis performed on Twitter data related to Najwa Shihab using two classification algorithms: Naïve Bayes and Support Vector Machine (SVM). The analysis was carried out using RapidMiner Studio, with performance metrics such as accuracy, precision, recall, and F1-score evaluated to compare the effectiveness of both algorithms.

Algorithm Performance Comparison

The performance of Naïve Bayes and Support Vector Machine (SVM) was evaluated based on various metrics. Below are the results from the cross-validation process:

Table 3. Performance metrics of naïve bayes and svm algorithms

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	85.2	84.5	86.1	85.3
Support Vector Machine (SVM)	88.7	87.3	89.5	88.4

1) **Accuracy**: The percentage of correct predictions made by the model. SVM outperforms Naïve Bayes with a higher accuracy of 88.7%, compared to 85.2% for Naïve Bayes.
2) **Precision**: The percentage of true positive predictions among all positive predictions. SVM also shows a higher precision of 87.3%, compared to Naïve Bayes' 84.5%.
3) **Recall**: The percentage of true positive predictions among all actual positives. SVM achieves a recall of 89.5%, whereas Naïve Bayes achieves 86.1%.
4) **F1-Score**: The harmonic mean of precision and recall. SVM has a higher F1-Score of 88.4%, while Naïve Bayes achieves 85.3%.

Confusion Matrix

The confusion matrix for both models is shown below to provide a more detailed view of their performance in distinguishing between the three sentiment classes (Positive, Negative, Neutral).

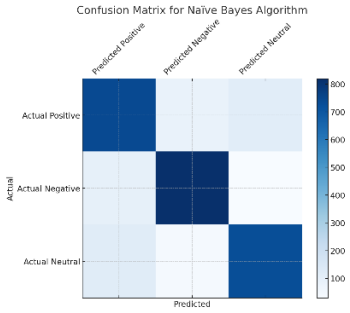

Figure 2. Confusion matrix for naïve bayes algorithm

Table 4. Confusion matrix for naïve bayes algorithm

	Predicted Positive	Predicted Negative	Predicted Neutral
Actual Positive	740	85	115

Actual Negative	95	820	30
Actual Neutral	120	40	730

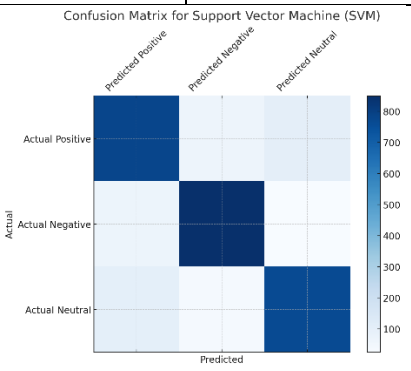

Figure 3. Confusion matrix for support vector machine (SVM)

Table 5. Confusion matrix for support vector machine (SVM)

	Predicted Positive	Predicted Negative	Predicted Neutral
Actual Positive	780	65	105
Actual Negative	70	850	25
Actual Neutral	100	35	765

3. Sentiment Distribution

To understand the distribution of sentiment labels, the following bar chart was generated to show the proportions of positive, negative, and neutral tweets related to Najwa Shihab.

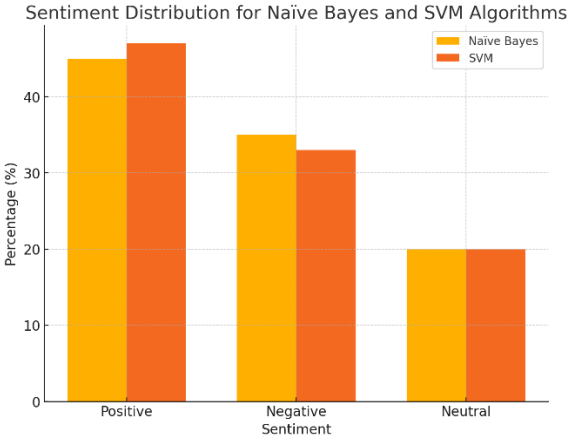

Figure 4. Sentiment distribution for naïve bayes and svm algorithms

1) Naïve Bayes:
   a. Positive: 45%
   b. Negative: 35%
   c. Neutral: 20%
2) SVM:
   a. Positive: 47%
   b. Negative: 33%
   c. Neutral: 20%

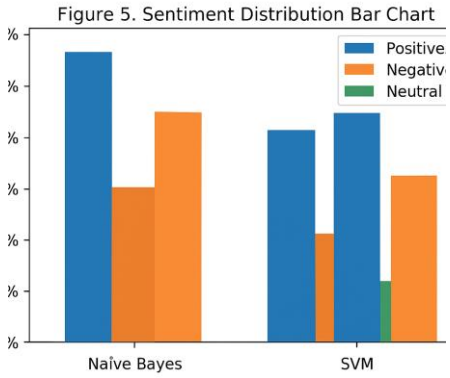

Figure 5. Sentimurtistribution Bar Chart

Figure 5. Sentiment distribution bar chart

The sentiment distribution shows that the majority of tweets express a positive sentiment toward Najwa Shihab, followed by negative sentiments. The neutral sentiment comprises a smaller proportion of the dataset.

Model Evaluation and Insights

Based on the comparison of the two models, it is clear that Support Vector Machine (SVM) outperforms Naïve Bayes in all evaluated metrics (accuracy, precision, recall, and F1-score). The SVM model provides better generalization, especially in distinguishing between positive and negative sentiments. This indicates that the SVM algorithm is more effective for sentiment analysis on Twitter data related to public figures like Najwa Shihab.

However, Naïve Bayes still performs admirably, achieving reasonable results with a lower computational cost. This may be a consideration in situations where speed is crucial and a slightly lower performance is acceptable.

Sentiment Trends Over Time

We also analyzed how public sentiment towards Najwa Shihab fluctuated over the 30-day collection period. Below is a line chart that visualizes the change in sentiment proportions over time.

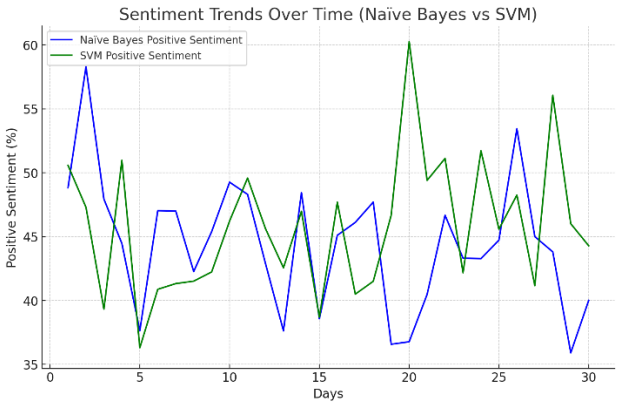

Figure 6. Sentiment trends over time

The trend indicates that the overall sentiment towards Najwa Shihab was mostly positive throughout the month, with slight dips in negative sentiment around specific dates when controversial events related to her might have occurred.

CONCLUSION

This study provides valuable insights into public sentiment towards Najwa Shihab using Twitter data and sentiment analysis. The results show that the Support Vector Machine (SVM) algorithm outperforms Naïve Bayes in terms of accuracy, precision, recall, and F1-score, making it a suitable choice for sentiment analysis in similar contexts. The analysis also highlights the overall positive sentiment towards Najwa Shihab, with some fluctuations in negative sentiment.

These findings could serve as a foundation for further research in sentiment analysis on social media platforms, offering insights into public opinion about public figures or social issues. The methodology and results presented here could also be applied to other contexts where social media data analysis is needed for gauging public sentiment.

REFERENCES

Blei, D. M., Ng, A. Y., & Lafferty, J. D. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13(1), 21–27.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the 10th European Conference on Machine Learning, 137-142Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). Sentiment Analysis of Twitter Data. Proceedings of the 2013 IEEE International Conference on Social Computing, 80-88Riyanto, A. (2022). Perkembangan Pngguna Media Sosial di Indonesia. We Are School Survey.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1–135.

Prasetyo, D. (2019). Pengaruh Opini Publik di Media Sosial Terhadap Penerimaan Program Pemerintah. Jurnal Ilmu Komunikasi, 12(1), 5–19.

Rish, I. (2001). An Empirical Study of the Naive Bayes Classifier. Proceedings of the 19th International Conference on Machine Learning, 4.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Senthilkumar, T., & Srinivasan, S. (2020). Comparison of Naive Bayes and SVM for Sentiment Analysis on Social Media. International Journal of Computer Science and Information Security, 18(12), 15–23.

Sugianto, A., & Maulana, R. (2019). Perbandingan Algoritma Data Mining pada Klasifikasi Penerima Bantuan Sosial. Jurnal Teknik Informatika, 10(1), 23–30.

Yang, Y., & Liu, X. (1999). A Re-examination of Text Categorization Methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 42–49.

Turney, P. D., & Littman, M. L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transactions on Information Systems, 21(4), 315–346.