Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

Sentiment Analysis of Cigarette Use Based on Opinions from X Using Naive Bayes and SVM

Tundo [1], Ratih Eldina [2*], Kiki Setiawan [3], Raisah Fajri [4]

[1,2*,3,4] Program Studi Teknik Informatika, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, Indonesia.

Corresponding Email: ratiheldina1234@gmail.com [2*]

Article History:
Submitted May 20, 2024; *Revised* June 25, 2024*;* Accepted July 15, 2024*; Published* September 20, 2024.

Abstrak

Penelitian ini menggunakan teknik klasifikasi Naive Bayes dan Support Vector Machine (SVM) untuk menganalisis sikap terhadap konsumsi rokok berdasarkan opini pengguna Twitter. Twitter, sebagai salah satu platform media sosial paling populer, berfungsi sebagai sumber yang sangat baik untuk mengukur sentimen publik tentang berbagai isu, termasuk merokok, yang disebut di sini sebagai "X." Beragamnya opini menimbulkan tantangan untuk klasifikasi sentimen yang akurat. Studi ini mengevaluasi efektivitas algoritma Naive Bayes dan SVM dalam mengkategorikan sentimen sebagai positif, negatif, atau netral. Data dikumpulkan melalui web scraping, dan langkah-langkah praproses seperti pembersihan teks, tokenisasi, dan stemming diterapkan. Kinerja klasifikasi dinilai menggunakan metrik seperti akurasi, presisi, recall, dan skor F1. Hasilnya menunjukkan bahwa SVM mengungguli Naive Bayes dalam analisis sentimen terkait penggunaan rokok. Temuan ini memberikan wawasan baru tentang opini publik dan bertujuan untuk membantu pembuat kebijakan dalam mengembangkan strategi pengendalian tembakau yang efektif.

Kata Kunci: Rokok; Naive Bayes; Analisis Sentimen; SVM.

Abstract

The research employs Naive Bayes and Support Vector Machine (SVM) classification techniques to analyze attitudes toward cigarette consumption based on Twitter user opinions. Twitter, being one of the most popular social media platforms, serves as an excellent source for gauging public sentiment on various issues, including cigarette smoking, referred to here as "X." The diverse array of opinions poses a challenge for accurate sentiment classification. This study evaluates the effectiveness of the Naive Bayes and SVM algorithms in categorizing sentiment as positive, negative, or neutral. Data is collected through web scraping, and preprocessing steps such as text cleaning, tokenization, and stemming are implemented. The performance of the classification is assessed using metrics like accuracy, precision, recall, and F1-score. The results indicate that SVM outperforms Naive Bayes in sentiment analysis related to cigarette use. These findings provide new insights into public opinion and aim to assist policymakers in developing effective tobacco control strategies.

Keyword: Cigarettes; Naive Bayes; Sentiment Analysis; SVM.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

1. Introduction

Smoking in public places is a common occurrence worldwide. This behavior has detrimental effects, impacting not only the health of active smokers but also increasing the risk for those nearby who are indirectly exposed to cigarette smoke. According to data from the World Health Organization (WHO), more than 1.2 million people die annually from passive smoke exposure. In the digital age, public opinion on indiscriminate smoking is reflected across various social media platforms. Statistical data indicates that over 3.8 billion people globally are active social media users, making these platforms valuable for understanding public views on health issues like indiscriminate smoking. However, deciphering and analyzing the widespread opinions on social media, particularly on platform X (formerly known as Twitter), is a challenging task. The complexity and sheer volume of data necessitate tools that can effectively analyze and interpret the sentiments expressed. Sentiment analysis methods such as Naive Bayes and Support Vector Machine (SVM) are employed to tackle this challenge. Naive Bayes is a widely used classification method known for its good accuracy. It does not require complex modeling or statistical testing and classifies data based on simple probabilities, assuming that the explanatory variables are independent. One of the benefits of the Naive Bayes algorithm is its lower error rate and faster processing on large datasets. The classification process in Naive Bayes is divided into two stages: the learning/training stage and the testing/classification stage.

During the learning stage, part of the known class data is used to create an estimation model, and in the testing stage, the remaining data is tested. Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for regression and classification. The concept of Structural Risk Minimization (SRM) underpins SVM, which aims to process data into a hyperplane that classifies the input space into two classes. The SVM theory starts with the collection of linearly separable cases by a hyperplane, categorized by their class. SVM begins with a two-class classification problem, requiring a set of positive and negative training samples Previous studies, such as Santoso (2020), "Spam Email Classification Using Naive Bayes Algorithm," successfully classified spam emails with 95% accuracy. Additionally, Lestari's (2021) research, "Product Sentiment Analysis with Support Vector Machine," showed that SVM effectively identifies customer sentiments with a 92% accuracy rate. Furthermore, Styawati's (2021) study, "Public Sentiment Analysis Towards the Pre-Employment Card Program on Twitter Using the Support Vector Machine Method," demonstrated that SVM achieved a 98% accuracy rate in classifying opinions on the pre-employment card program. This study, unlike previous research, uses Naive Bayes and Support Vector Machine (SVM) algorithms to analyze sentiments regarding cigarette use on the social media platform X (formerly known as Twitter). This approach introduces a new area of study in the field. The research aims to present sentiment analysis results based on user opinions on X, categorizing these opinions as neutral, negative, or positive.

2. Research Methods

This research was conducted through several interconnected stages, each designed to achieve the desired research outcomes.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E - ISSN: 2723-7079 | P - ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

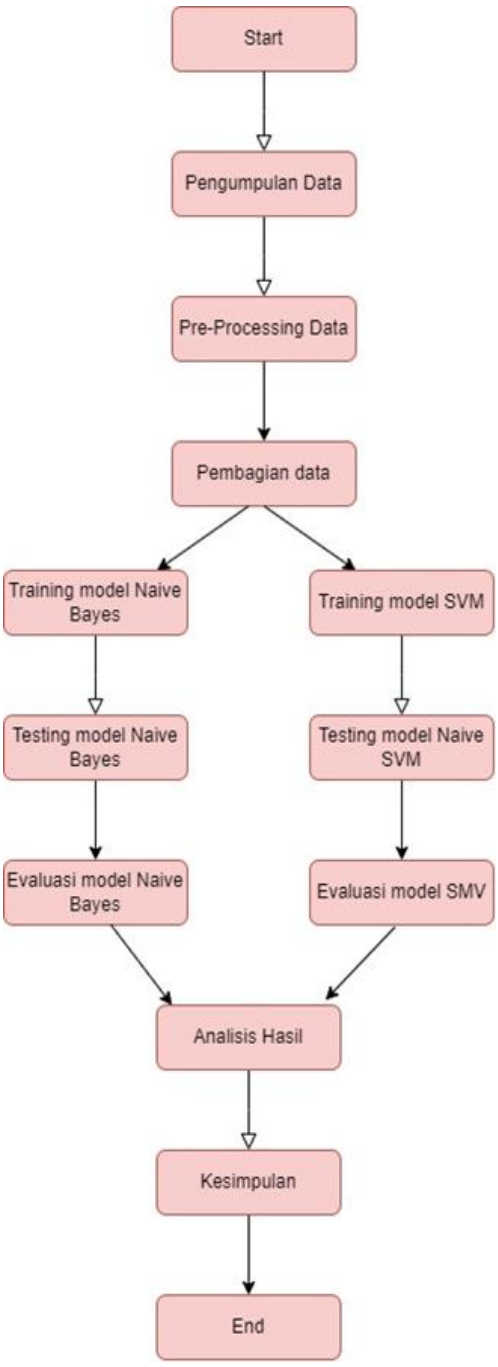

Figure 1. Research Process Model

The study commenced with data collection, which is a critical initial step in analyzing social media users' sentiments regarding cigarette use. Data was gathered from the platform X (formerly known as Twitter) using the keyword "rokok" (cigarette). Data collection was performed using the Python programming language, implemented in Google Colab, allowing for efficient and effective data processing. The data collection period spanned from 2019 to 2024, providing a comprehensive timeframe to capture relevant opinions. The tweet-harvest package was utilized to access data from the X API, requiring essential credentials such as access tokens, secret access tokens, and Twitter API keys. This process resulted in approximately one thousand tweets containing the word "rokok," which

E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

were subsequently saved in a .csv file for further analysis. Following data collection, the next step was data preprocessing. Data preprocessing is a crucial phase that aims to clean and format raw data so that it can be used for subsequent analysis. Data collected from social media often contains unnecessary elements such as usernames, hashtags, and URLs. Therefore, the first step in data preprocessing involved removing these elements to ensure that only relevant data was used. Next, case folding was performed to convert all text to lowercase, ensuring consistency across the dataset. The data was then tokenized, meaning that the text was broken down into individual words or tokens. This step was followed by stemming, which reduces words to their base forms and eliminates common meaningless words, known as stop words. The preprocessed data was then saved again in a .csv file for use in the subsequent research stages.

The next phase involved splitting the data into two portions: training data and testing data. This data splitting is essential to ensure that the trained model can be tested with previously unseen data, allowing for an objective evaluation of its performance. In this research, 20% of the data was used for training the model, while the remaining 80% was reserved for testing. Each tweet in the dataset was labeled according to the sentiment it conveyed, categorized as positive, negative, or neutral. This labeling process is crucial as it provides the model with the necessary context to understand the sentiment within the data. After data splitting, the next step was the implementation of the Naive Bayes model. Naive Bayes is a classification algorithm that operates on the probability of each class, assuming that each attribute is independent of the others. The Naive Bayes model was trained using the training dataset. Once the model was trained, the testing data was used to predict the sentiment of each tweet. The model's performance was then evaluated by analyzing the predicted outcomes.

In addition to Naive Bayes, the research also implemented the Support Vector Machine (SVM) model. SVM is a machine learning algorithm used for classification by dividing data into classes using a hyperplane function. The hyperplane serves as a boundary that separates data into different classes based on their features. In this study, an SVM model with a linear kernel was employed to classify text from the downloaded dataset. After training the model, the prediction values were calculated and evaluated using the testing data to assess the model's performance. To determine the effectiveness and performance of both models, an evaluation was conducted using a confusion matrix. The confusion matrix displays a comparison between the actual classification results and the system's predictions, allowing for the calculation of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. From these values, evaluation metrics such as accuracy, precision, recall, and F1-score can be calculated. This evaluation is critical in determining how well the models perform in accurately classifying sentiment. After the evaluation was completed, the results of the Naive Bayes and SVM models were compared to identify the most effective algorithm for categorizing users' opinions on cigarette use on X. This comparison was based on the previously calculated evaluation metrics. The analysis provided insights into the strengths and weaknesses of each model in handling sentiment data from social media.

3. Results and Discussion

3.1 Results

1) Scraping

For this study, data was gathered from Twitter through the use of scraping tools, with a particular focus on tweets that included the term "rokok" (cigarette). The Python programming language was used to collect a total of 1000 tweets. After that, the collected data was placed in a document in a.csv format for later examination. The outcome of this scraping procedure is displayed in Table 1.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

Table 1. Scraping data snippet

Comment	Location
Worldly deviations begin to feel the hustle and bustle of increasingly deviant human interactions. Truly cigarettes can be calming.	Bandung
tasting the world which is said to be very exciting for the soul but not with cigarette smoke	Bandug
I'm sick from cigarette smoke	Jakarta
smoke is also dangerous #passivesmoking	Bekasi

2) Preparation

During the preprocessing stage, the data that was scraped and saved in CSV format is reprocessed. This makes unstructured information more structured by removing elements that aren't needed for sentiment analysis. Table 2 presents the results of the text preparation.

Table 2. Preprocessing text snippet

Text preprocessing	Scraping
Save the world began to feel human really cigarette	Worldly deviations begin to feel the hustle and bustle of increasingly deviant human interactions. Truly cigarettes can be calming.
taste the world of words very passionate soul but no cigarette smoke	Taste the world which is said to really excite the soul but not with cigarette smoke
cigarette smoke sickness	I'm sick from cigarette smoke
smoke is also dangerous	smoke is also dangerous #passivesmoking

3) Sectioning

At this point, the previously processed twitter data is subjected to manual labeling. Each tweet is assigned to a certain class via this labeling. There are three classes in use: neutral, negative, and positive. "1" denotes the positive class, "-1" the negative class, and "0" the neutral class. Details on the labeled text preprocessing outcomes are shown in Table 3.

Table 3. Snippet of data labeling

Text preprocessing	Label
Save the world began to feel human really cigarette	-1
taste the world of words very passionate soul but no cigarette smoke	1
cigarette smoke sickness	-1
smoke is also dangerous	-1

4) Classification with Naive Bayes Algorithm

An 80:20 split between training and testing data is used for sentiment analysis during the classification stage. Approximately 1,000 tweets' worth of preprocessed and manually labeled data are used in this trial. The outcomes of this classification are displayed in the following figure:

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

```
for i, tweet in enumerate(data_tweet):
    analysis = TextBlob(tweet)
    polaritas += analysis.polarity

    if analysis.sentiment.polarity > 0.0:
        total_positif += 1
        status.append('Positif')
    elif analysis.sentiment.polarity == 0.0:
        total_netral += 1
        status.append('Netral')
    else:
        total_negatif += 1
        status.append('Negatif')

    total += 1

print(f'Hasil Analisis Data:\nPositif = {total_positif}\nNetral = {total_netral}\nNegatif = {total_negatif}')
print(f'\nTotal Data : {total}')
```

```
Hasil Analisis Data:
Positif = 416
Netral = 135
Negatif = 454
```

Figure 2. Naive Bayes polarity results

The distribution of polarities in the used data is displayed in Figure 2 above. Every class is represented by these polarities: positive, negative, and neutral. There are 135 data points for the neutral class, 416 data points for the positive class, and 454 data points for the negative class.

```
# Calculate accuracy, recall, and precision
accuracy = (cm[0][0] + cm[1][1]) / len(dataset)
recall = cm[0][0] / (cm[0][0] + cm[0][1])
precision = cm[0][0] / (cm[0][0] + cm[1][0])

# Print results
print("Accuracy:", accuracy)
print("Recall:", recall)
print("Precision:", precision)
```

```
[[454   0   0]
 [135   0   0]
 [416   0   0]]
Accuracy: 0.45174129353233833
Recall: 1.0
Precision: 0.7707979626485568
```

Figure 3. Confusion matrix calculation

5) Classification with Support Vector Machine (SVM) Algorithm
Sentiment analysis is done during the classification phase using the Support Vector Machine (SVM) technique. The dataset is divided between training and testing data in an 80:20 ratio following preprocessing and manual labeling. The training data is then used to train the SVM model. Next, the model's performance is assessed using the test data. The classification results are displayed in the image below.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi
https://journal.stmiki.ac.id
E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

	precision	recall	f1-score	support
Negatif	0.17	0.12	0.14	16
Netral	0.33	0.68	0.45	19
Positif	0.00	0.00	0.00	16
accuracy			0.29	51
macro avg	0.17	0.27	0.20	51
weighted avg	0.18	0.29	0.21	51

Figure 4. SVM calculation results

The SVM model is trained using the training data following the phases of preprocessing and human labeling. The model's performance is then assessed by testing it using the testing data. The classification results are displayed in the image below.

```
# Menampilkan confusion matrix sebagai diagram (heatmap)
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', xt
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title('Confusion Matrix')
plt.show()

 Output:
Akurasi Test:  0.6666666666666666
```

Figure 5. SVM accuracy calculation

The data used's classification outcomes and confusion matrix are displayed in Figure 5. A 66.6% accuracy percentage is indicated by the classification findings.

3.2 Discussion

The results of this research provide crucial insights into the comparative effectiveness of the Naive Bayes and Support Vector Machine (SVM) algorithms in the context of sentiment analysis, specifically regarding cigarette use as expressed on social media. The findings demonstrate that while both algorithms are competent in processing sentiment data, SVM consistently outperforms Naive Bayes, particularly in accuracy and robustness. This is aligned with previous research by Alita and Shodiqin (2023), which showed that SVM achieved superior performance in sentiment analysis of COVID-19 vaccine discussions compared to Naive Bayes, with higher accuracy and precision rates. The Naive Bayes algorithm, known for its simplicity and computational efficiency, performed adequately in this study but lagged behind SVM in terms of accuracy. This outcome resonates with the findings of Normawati and Prayogi (2021), who implemented Naive Bayes for sentiment analysis on Twitter and noted that while the algorithm is effective for smaller datasets, it may struggle with more complex or nuanced sentiments.

The probabilistic nature of Naive Bayes, which assumes independence among features, can sometimes lead to oversimplification, especially in text data where word dependencies often play a critical role in sentiment detection. On the other hand, the SVM algorithm, which operates on the principle of maximizing the margin between classes, showed better performance in this study. This result is consistent with several other studies, including research by Husen *et al.* (2023) and Tineges *et al.* (2020), where SVM demonstrated superior classification accuracy in analyzing public sentiments

E - ISSN: 2723-7079 | P - ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

on various issues, including banking and telecommunication services. The strength of SVM lies in its ability to handle high-dimensional data and its effectiveness in text classification tasks, as it finds an optimal hyperplane that best separates the classes, making it particularly adept at managing the intricacies of social media text data. The data preprocessing stage, which included text cleaning, tokenization, and stemming, played a significant role in ensuring the quality and reliability of the sentiment analysis. Proper preprocessing is vital in sentiment analysis, as highlighted by Darwis *et al.* (2021) in their application of Naive Bayes for analyzing sentiment in BMKG's Twitter data. They emphasized the importance of preprocessing steps to enhance the model's ability to accurately classify sentiments. In this study, preprocessing ensured that irrelevant elements such as special characters and stop words were removed, thus refining the input data for better model performance.

Despite the advantages observed with SVM, it is important to acknowledge the limitations encountered, particularly in manual labeling, which could introduce bias. Although this is a common challenge in sentiment analysis, as noted by Rahayu *et al.* (2022) in their study on Spotify sentiment analysis, future research should consider alternative approaches such as semi-supervised learning or crowdsourcing to improve labeling consistency. Additionally, the focus on tweets containing the keyword "rokok" may not fully capture the broader spectrum of public opinion on smoking. Expanding the dataset to include related keywords or phrases could provide a more comprehensive understanding of public sentiment. The implications of these findings are significant, particularly for public health policy and tobacco control strategies. Accurate sentiment analysis can provide policymakers with real-time insights into public opinion, enabling more targeted and effective interventions. For example, Hendrastuty *et al.* (2021) demonstrated how sentiment analysis using SVM could inform government policies by monitoring public reactions to the pre-employment card program on Twitter . Similarly, the insights gained from this study could be utilized to shape public health campaigns and regulations related to smoking, ensuring they are more aligned with public sentiment and, therefore, more likely to be effective.

4. Conclusion

It is clear from the results and the above talks that sentiment analysis on passive smoking was done using the Naïve Bayes and SVM algorithms with data taken from Twitter user tweets. Results of the sentiment analysis on about a thousand tweets revealed that SVM performs better in this situation, with an accuracy of 66% compared to 45.15% for Naïve Bayes. It is intended that this study will serve as a foundation for future research and offer fresh perspectives on the risks that passive smoking poses to the general public.

5. Acknowledgments

Praise be to Allah for His endless mercy, guidance, and blessings, which have enabled me to complete this research successfully. With all humility, I extend my deepest gratitude to Allah for His immeasurable help. I also wish to express my sincere thanks to my parents, who have consistently provided unwavering support, prayers, and unconditional love. Without their moral and material support, this research would not have been possible. They are my source of inspiration and strength, and for that, I am deeply grateful. May Allah always bestow His grace and blessings upon us all. Aamiin

E-ISSN: 2723-7079 | P-ISSN: 2776-8074
Vol. 5 No. 3 (2024) | September
OPEN ACCESS
https://doi.org/10.35870/jimik.v5i3.947

6. References

Guia, M., Silva, R. R., & Bernardino, J. (2019). Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis. *KDIR*, *1*, 525-531.

Husen, R. A., Astuti, R., Marlia, L., Rahmaddeni, R., & Efrizoni, L. (2023). Analisis Sentimen Opini Publik pada Twitter Terhadap Bank BSI Menggunakan Algoritma Machine Learning: Sentiment Analysis of Public Opinion on Twitter Toward BSI Bank Using Machine Learning Algorithms. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 211-218. DOI: https://doi.org/10.57152/malcom.v3i2.901.

Iskandar, J. W., & Nataliani, Y. (2021). Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(6), 1120-1126.

Millennianita, F., Athiyah, U., & Muhammad, A. W. (2024). Comparison of Naïve Bayes Classifier and Support Vector Machine Methods for Sentiment Classification of Responses to Bullying Cases on Twitter. *Journal of Mechatronics and Artificial Intelligence*, *1*(1), 11-26.

Normawati, D., & Prayogi, S. A. (2021). Implementation of Naive Bayes Classifier and Confusion Matrix in Text-Based Sentiment Analysis on Twitter. *J-SAKTI (Jurnal Sains Komput. Dan Inform.*, *5*(2), 697-711.

Oktavia, D., Ramadahan, Y. R., & Minarto, M. (2023). Analisis Sentimen Terhadap Penerapan Sistem E-Tilang Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM). *KLIK: Kajian Ilmiah Informatika dan Komputer*, *4*(1), 407-417.

Petiwi, M. I., Triayudi, A., & Sholihati, I. D. (2022). Analisis Sentimen Gofood Berdasarkan Twitter Menggunakan Metode Naïve Bayes dan Support Vector Machine. *Jurnal Media Informatika Budidarma*, *6*(1), 542-550. DOI: http://dx.doi.org/10.30865/mib.v6i1.3530.

Putri, D. D., Nama, G. F., & Sulistiono, W. E. (2022). Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Informatika dan Teknik Elektro Terapan*, *10*(1). DOI: https://doi.org/10.30865/klik.v4i1.1040.

Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019, November). 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)* (pp. 266-270). IEEE. DOI: 10.1109/SMART46866.2019.9117512.

Rahayu, A. S., Fauzi, A., & Rahmat, R. (2022). Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Spotify. *Jurnal Sistem Komputer dan Informatika (JSON)*, *4*(2), 349-354. DOI: http://dx.doi.org/10.30865/json.v4i2.5398.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 616-623).

Tineges, R., Triayudi, A., & Sholihati, I. D. (2020). Analisis sentimen terhadap layanan indihome berdasarkan twitter dengan metode klasifikasi support vector machine (SVM). *Jurnal Media Informatika Budidarma*, *4*(3), 650-658. DOI: http://dx.doi.org/10.30865/mib.v4i3.2181.