Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11430
e-ISSN : 2541-2019
p-ISSN : 2541-044X

SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter

Rahmat Syahputra[1)*], Gomal Juni Yanris[2)], Deci Irmayani[3)]
[1)2)3)]Universitas Labuhanbatu, Indonesia
[1)]rahmatsyahputrasilalahi13@gmail.com, [2)]gomaljuniyanris@gmail.com, [3)]deacyirmayani@gmail.com

Submitted : May 5, 2022 | **Accepted** : May 8, 2022 | **Published** : May 10, 2022

Abstract:. With the emergence of the Peduli Protect application, which is used by the government to monitor the spread of Covid-19 in Indonesia, it turns out to be reaping the pros and cons of public opinion on Twitter. From this phenomenon, a research was conducted by mapping the sentiment analysis of twitter users towards the Peduli Protect application. This study aims to compare two classification algorithms that are included in the supervised learning category. The two algorithms are Support Vector Machine (SVM) and Naïve Bayes. The two algorithms are implemented in analyzing the sentiment analysis of twitter user reviews on the Peduli Protect application. The dataset used in this research is tweets of twitter users with a total of 4,782 tweets. Then, compared to how much accuracy and processing time required of the two algorithms. The stages of the method in this research are: collecting data from user tweets with a crawling technique, preprocessing text, weighting words using the TF-IDF method, classification using the SVM and Naïve Bayes algorithm, k-fols cross validation test, and drawing conclusions. The results showed that the accuracy of the SMV algorithm with the k-fold test method was 86% and the split 8020 technique resulted in an accuracy of 79%. Meanwhile, the Naïve Bayes algorithm produces an accuracy of 85% with k-fold, and an accuracy of 80% with a split 8020. From these results it can be concluded that both algorithms have the same level of accuracy, only different in processing time, where Naïve Bayes algorithm is faster with time required 0.0094 seconds.

Keywords: Classification; Naïve Bayes; Sentiment Analysis; SVM; Twitter.

INTRODUCTION

In dealing with the spread of Covid-19, the Government of Indonesia created an application called Peduli Protect which is used to monitor the activities of the spread of Covid-19 in Indonesia (Mustopa et al., 2020). The presence of the application has generated various responses among the public, thus giving rise to public opinion pros and cons on Twitter social media.

Twitter has become the most popular social media among internet users which has 106 million users and 180 million monthly visitors, which continues to increase every day (Kristiyanti, Umam, Wahyudi, Amin, & Marlinda, 2018). By using data through posts on Twitter, various opinions on any issue can be classified using sentiment analysis into 2 types of opinions, namely, Positive and Negative (Goyal, 2021). Sentiment analysis is needed to find out an in-depth assessment of an object. By using information from Twitter social media, sentiment analysis can be carried out on the Peduli Protect Application reviews (Illia, Eugenia, & Rutba, 2021).

Support Vector Machine (SVM) is a classification algorithm that has a good level of accuracy compared to other algorithms because it is able to define the hyperplane well. While the Naïve Bayes algorithm is a machine learning classification algorithm with probability reasoning that is not inferior to other algorithms. The Naïve Bayes algorithm also has the advantage that it uses less training data and has a good level of accuracy (Fitriana, Utami, & Al Fatta, 2021).

Research related to sentiment analysis using a classification algorithm has been carried out by a number of previous researchers. Sentiment analysis using a combination of SVM and Naïve Bayes Classifier methods using a dataset of 2,378 has resulted in an accuracy rate of 81.61% on SVM and an accuracy of 67.29% on Nave Bayes (Tuhuteru & Iriani, 2018). Furthermore, the sentiment analysis of the Covid-19 vaccine opinion on Twitter social media using a dataset taken from Twitter as many as 1000 records showed that the SVM method was slightly better than the Naïve Bayes method with an accuracy rate of 90.47% for the SVM algorithm, while the Nave Bayes method by 88.64% (Fitriana et al., 2021). In another study related to sentiment analysis, it also proved that the

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11430
e-ISSN : 2541-2019
p-ISSN : 2541-044X

SVM algorithm was better than Naïve Bayes (Siswanto, Wibawa, Gata, Gata, & Kusumawardhani, 2018) (Rahat, Kahir, & Masum, 2019).

Based on the description of the SVM and Naïve Bayes methods in previous studies. So this study aims to compare the SVM and Naïve Bayes algorithms in terms of sentiment analysis of Twitter users towards the Peduli Protect application. As for the formulation of the problem in this study, how much accuracy and processing time is required for the SVM and Naïve Bayes algorithms in analyzing twitter users' sentiment towards the Peduli Protect application.

LITERATURE REVIEW

Sentiment Analysis is the task of seeking written opinions about a particular entity. The decision-making process of people is influenced by opinions formed, by the general public, by the thoughts of leaders. Sentiment analysis sources are obtained from social media, Twitter, Facebook, Blogs and other user forums. Sentiment analysis is usually used to take user feedback from a product or an organization's performance (Yaakub, Latiffi, & Safra, 2019). Sentiment analysis is a term that has the same meaning as Text Mining, which aims to find words that can represent what is in the document so that an analysis of the relationship between documents can be carried out (Rolliawati, Khalid, & Rozas, 2020). Sentiment analysis can map opinions on a review into three opinion classifications, namely: positive, neutral, or negative automatically (State, Muhardi, & Putri, 2020).

Twitter is a social media with microblogging type as an interaction service. Twitter is one of the most popular social media services in the world with 200 million active users and more than 10.6 billion tweets that have been generated (Ibrahim, Abdillah, Wicaksono, & Adriani, 2015). Twitter was created as a platform for exchanging experiences and sharing anything among its users without any barriers. By using Twitter, users will find it easier to follow trends as well as information and news from all over the world. In addition, Twitter also helps users to always be connected with the people closest to them (Hadna, Santosa, & Winarno, 2016).

Support Vector Machine (SVM) is a supervised learning method that can analyze data and recognize patterns, and is used for classification and regression analysis. The working principle of this method is to find the most optimal separator space from a dataset in different classes. This classification is done by looking for a hyperplane or dividing line that separates one class from another (Kurniawan et al., 2019).

Naïve Bayes Classifier (NBC) is a classification algorithm which is rooted in Bayes' theorem. Naïve Bayes Classifier works very well compared to other classifier models such as Decision Tree or Neural Network. The advantage of using this method is that this method only requires small training data to determine the parameters needed in the classification process (Iskandar & Nataliani, 2021).

METHOD

This study compares the accuracy and processing time of the Support Vector Machine and Naïve Bayes methods in conducting sentiment analysis on user reviews of the Peduli Lingdungi application on Twitter. This study uses data originating from user tweets on Twitter with the keyword PeduliLindungi in the period 15 September 2021 to 22 September 2021. The number of datasets taken is 4,782 datasets. The stages of the research carried out are illustrated in Figure 1.

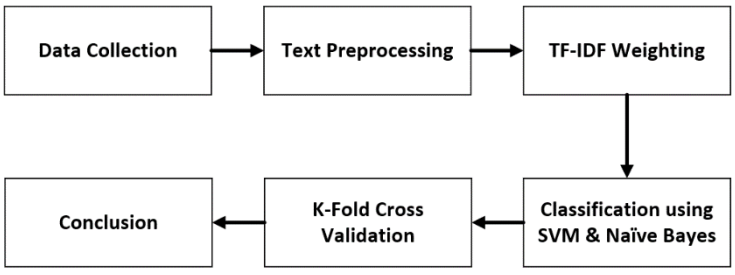

Fig 1. Research Framework

Data Collection

The dataset is collected from tweets of twitter users with the keyword PeduliLindungi using a crawling technique. In this data collection, twint tools are used to retrieve data from Twitter. Then the data is labeled manually so that it produces a positive label and a negative label.

Text Preprocessing

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11430
e-ISSN : 2541-2019
p-ISSN : 2541-044X

At this stage there are four processes carried out. The first process is Case Folding, in this process all letters contained in the data will be converted into lowercase letters and eliminate characters other than letters. The second process is tokenizing, in this process cutting or separating each word contained in the dataset is carried out. The third process is Stopword removal, in this process all words that have no meaning such as the words 'and', 'di', 'by' are removed so that only meaningful words are left. The fourth process is Stemming, in this process all words will be converted into basic word forms by removing or eliminating the affixes contained in the word.

TF-IDF Weighting

At this stage, the value of each word is given a weight using the Terms Frequency-Inverse Document Frequency (TF-IDF) algorithm. This process is done to count how many occurrences of words in the dataset.

Classification using SVM & Naïve Bayes

At this stage the SVM and Naïve Bayes algorithms are implemented. In this stage the machine is taught to recognize existing data patterns and then the preprocessed data that has been given the TF-IDF weight is classified into two classes, namely, Positive and Negative.

K-Fold Cross Validation

Using K-Fold Cross Validation, the text document will be divided into 10 sections. Then the experiment was carried out 10 times (k = 10), the document classification and each trial the data will be randomized first before finally being included in the fold. This process is done to avoid grouping documents from one particular category in a fold. At this stage, 80% of training data is used, and 20% of testing data is used.

Conclusion

At this stage, a comparison of the accuracy and time processing results obtained from the Naive Bayes algorithm and SVM is carried out. And take the results and conclusions in this study.

RESULT

Based on the research stages that have been described previously, this study aims to compare the accuracy and processing time of the SVM and Naïve Bayes algorithms by testing using K-Fold Cross Validation. The following are the results and discussion of the tests that have been carried out.

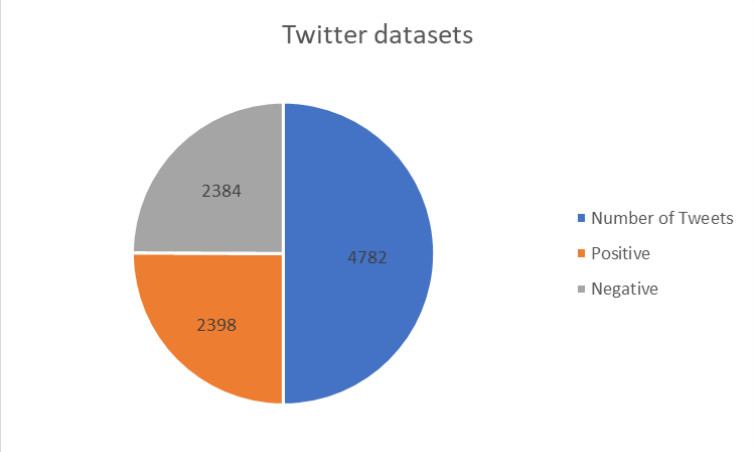

Fig 2. Twitter Opinion Dataset

Figure 1 shows that there are 4,782 user tweets taken from Twitter using a crawling technique with the keyword PeduliLindungi. From the dataset, data labeling was carried out, resulting in 2,398 positive labels, and 2,384 negative labels.

After labeling the text, the next step is to preprocess the text using case folding, tokenizing, stopwords removal and stemming techniques. The results of preprocessing the Twitter user's tweet dataset are shown in Table 1.

Table 1. Text Preprocessing Results

Label	Tweets	Case Folding	Tokenizing	Stopwords removal	Stemming

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11430
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Positive	kabar gembira, ancol taman impian sudah dibuka ???? tapi, jangan lengah! main ke ancol harus senang selamat bareng-bareng dengan tetap patuhi protokol kesehatan selama berkegiatan, siapkan e-tiket, serta aplikasi pedulilindungi ya, sob! yuk cari tahu cara masuknya lewat video ini https://t.co/m62aanyx7h	kabar gembira ancol taman impian sudah dibuka tapi jangan lengah main ke ancol harus senang selamat barengbareng dengan tetap patuhi protokol kesehatan selama berkegiatan siapkan etiket serta aplikasi pedulilindungi ya sob yuk cari tahu cara masuknya lewat video ini	['kabar', 'gembira', 'ancol', 'taman', 'impian', 'sudah', 'dibuka', 'tapi', 'jangan', 'lengah', 'main', 'ke', 'ancol', 'harus', 'senang', 'selamat', 'barengbareng', 'dengan', 'tetap', 'patuhi', 'protokol', 'kesehatan', 'selama', 'berkegiatan', 'siapkan', 'etiket', 'serta', 'aplikasi', 'pedulilindungi', 'ya', 'sob', 'yuk', 'cari', 'tahu', 'cara', 'masuknya', 'lewat', 'video', 'ini']	['kabar', 'gembira', 'ancol', 'taman', 'impian', 'dibuka', 'lengah', 'main', 'ancol', 'senang', 'selamat', 'barengbareng', 'patuhi', 'protokol', 'kesehatan', 'berkegiatan', 'siapkan', 'etiket', 'aplikasi', 'pedulilindungi', 'sob', 'cari', 'masuknya', 'video']	kabar gembira ancol taman impi buka lengah main ancol senang selamat barengbare ng patuh protokol sehat giat siap etiket aplikasi pedulilindungi sob cari masuk video
Negative	staisun cikarang gajelassss bangettt, masa suruh cek peduli lindungi sama suhu aja harus turun dulu kebawah yg bisa ngabisin waktu 5-7 menit???? (turun tangga pula), gajelas bgtt kenapa ga diatas aja dah cek ek begituan	staisun cikarang gajelassss bangettt masa suruh cek peduli lindungi sama suhu aja harus turun dulu kebawah yg bisa ngabisin waktu menit turun tangga pula gajelas bgtt kenapa ga diatas aja dah cek ek begituan	['staisun', 'cikarang', 'gajelassss', 'bangettt', 'masa', 'suruh', 'cek', 'peduli', 'lindungi', 'sama', 'suhu', 'aja', 'harus', 'turun', 'dulu', 'kebawah', 'yg', 'bisa', 'ngabisin', 'waktu', 'menit', 'turun', 'tangga', 'pula', 'gajelas', 'bgtt', 'kenapa', 'ga', 'diatas', 'aja', 'dah', 'cek', 'ek', 'begituan']	['staisun', 'cikarang', 'gajelassss', 'bangettt', 'suruh', 'cek', 'peduli', 'lindungi', 'suhu', 'turun', 'kebawah', 'ngabisin', 'menit', 'turun', 'tangga', 'gajelas', 'bgtt', 'diatas', 'cek', 'ek', 'begituan']	staisun cikarang gajelassss banget suruh cek peduli lindung suhu turun bawah ngabisin menit turun tangga enggak jelas banget atas cek ek begitu

After going through the testing process, the results of the confusion matrix (precision, recall, f1-score, and accuracy) of the two algorithms are obtained. The comparison results from the confusion matrix are shown in Figure 3 below.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11430
e-ISSN : 2541-2019
p-ISSN : 2541-044X

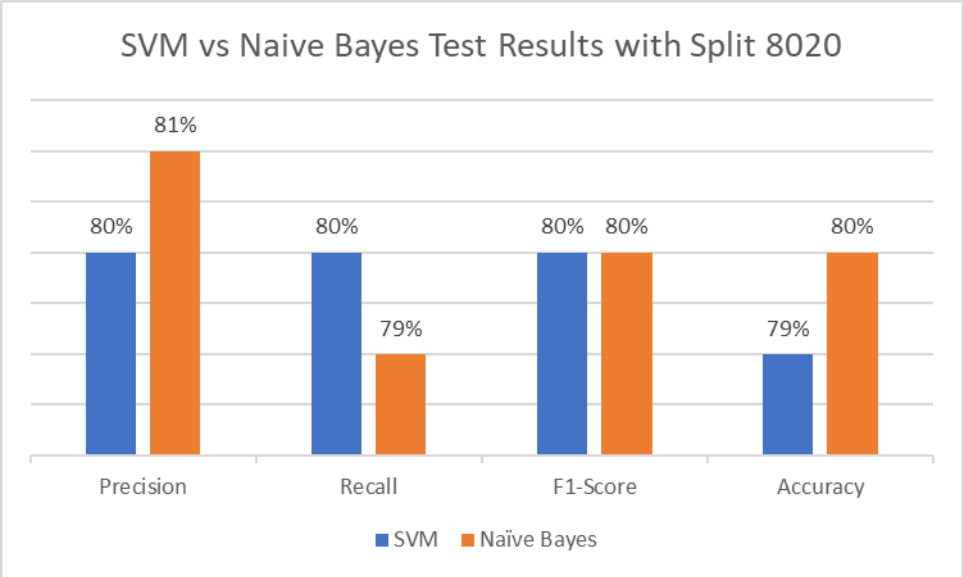

Fig 3. Confusion Matrix SVM vs Naive Bayes Classification with Split 8020

From the test results using training data of 80% and test data of 20% (split 8020) shown in Figure 3, the results show that, the SVM algorithm obtains 80% precision, 80% recall, 80% F1-score, and The resulting accuracy rate is 79%. While the Naïve Bayes algorithm, obtained 81% precision, 79% recall, 80% F1-score, and 80% accuracy.

Next, the SVM and Naïve Bayes algorithms were tested by applying K-Fold Cross Validation which used 80% training data and 20% test data. The results of the K-Fold Cross Validation test of the two algorithms are shown in Figures 4 and 5.

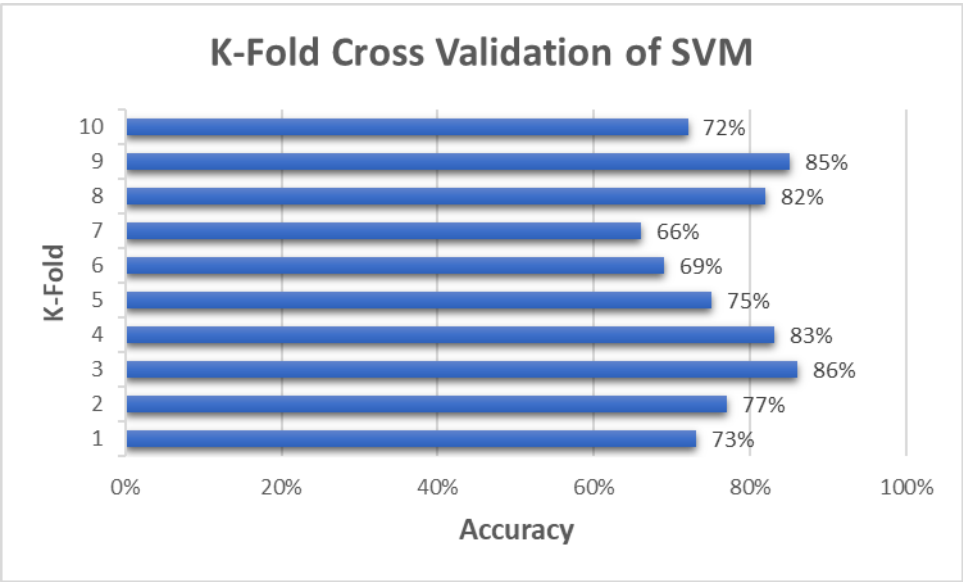

Fig 4. K-Fold Cross Validation Test Results SVM Algorithm

Figure 4 shows that the highest level of accuracy in the SVM algorithm is at the K-Fold = 3 value, with an accuracy value of 86%, where the overall average accuracy is 76.80%.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 2, April 2022
DOI : https://doi.org/10.33395/sinkron.v7i2.11430
e-ISSN : 2541-2019
p-ISSN : 2541-044X

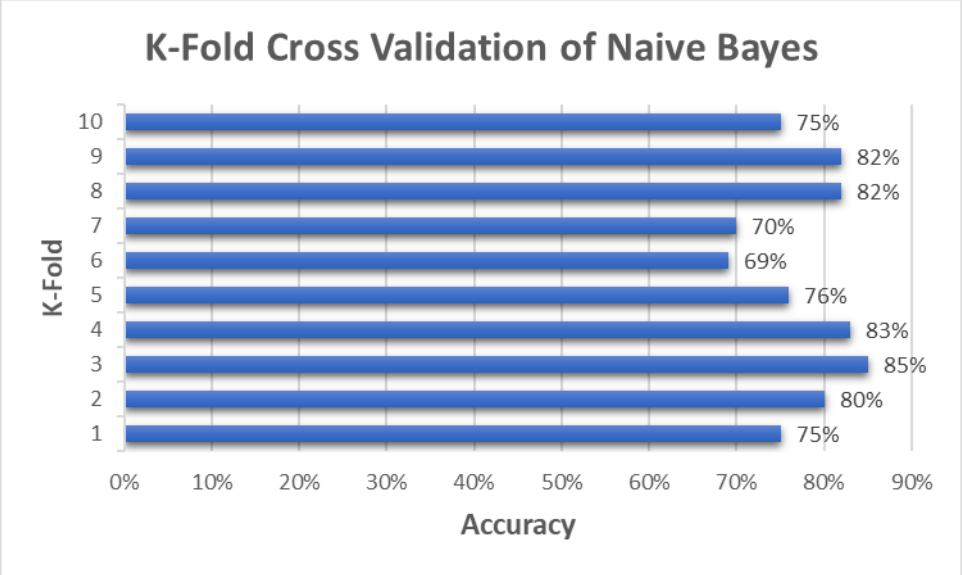

Fig 5. K-Fold Cross Validation Test Results Naïve Bayes Algorithm

Figure 5 shows that the highest level of accuracy in the Naïve Bayes algorithm is at the value of K-Fold = 3, with an accuracy value of 85%, where the overall average accuracy is 77.70%.

Furthermore, the processing time of the SVM and Naïve Bayes algorithms was tested. Processing time includes the time required for training time and prediction time. The processing time test results of the two algorithms are shown in Figure 6.

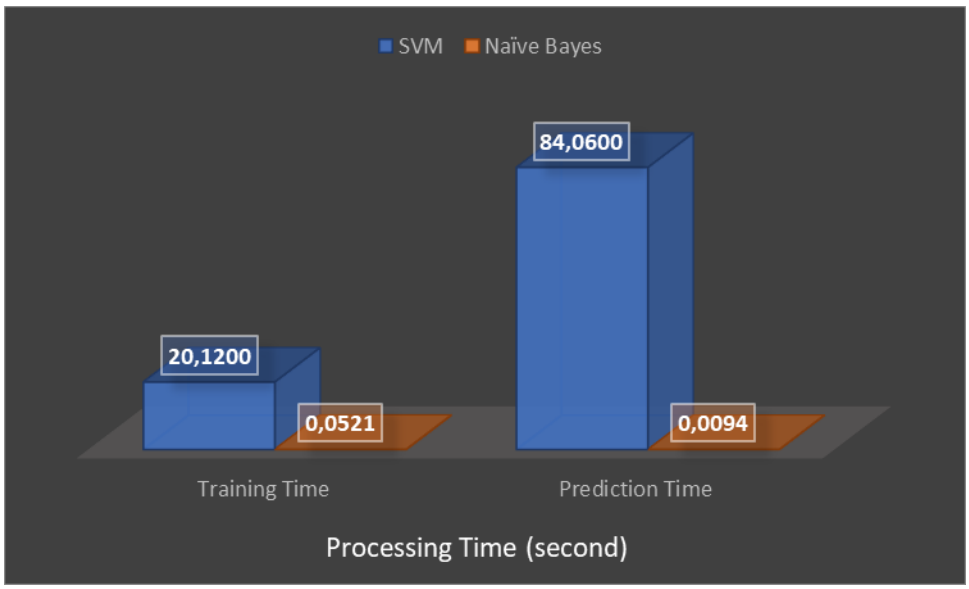

Fig 6. SVM vs Naïve Bayes Processing Time Comparison

From Figure 6, it can be seen that the time required to train data on the SVM algorithm is 20.12 seconds, the Naïve Bayes algorithm takes 0.0521 seconds. While the time needed to make predictions on the SVM algorithm is 84.06 seconds, for the Naïve Bayes algorithm it takes 0.0094 seconds to predict.

DISCUSSIONS

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

After getting the results from each test on the SVM and Naïve Bayes algorithm using the Split 8020 test method and K-Fold Cross Validation, the results are as shown in Table 2 below.

Table 3. Comparison of SVM vs Naïve Bayes Algorithm Results

Algorithms	Accuracy		Processing Time (second)
	K-Fold = 10	Split 8020	
SVM	86%	79%	84.06
Naïve Bayes	85%	80%	0.0094

Based on the results of the comparison of the SVM and Nave Bayes algorithms in Table 3 above, it is found that the accuracy level with the K-Fold Cross Validation test on the Naïve Bayes Algorithm gets an accuracy of 85%, while the SVM algorithm gets better accuracy with an accuracy rate of 86%. While the level of accuracy generated by testing the split 8020 technique, the Naïve Bayes algorithm has a better accuracy rate of 80% than the SVM algorithm which has an accuracy of 79%.

Meanwhile, based on processing time, the SVM algorithm is longer than the Naïve Bayes algorithm with a processing time of 84.06 seconds. This happens because the way the SVM algorithm works is more complex by using a kernel that aims to find a hyperplane, causing the process time to tend to be longer. This is different from the way the Naïve Bayes algorithm works, namely by calculating the probability of one class for each existing attribute group and determining the most optimal probability and using training data which tends to be less, resulting in a shorter processing time of 0.0094 seconds.

CONCLUSION

This research has succeeded in conducting a sentiment analysis on user reviews of the Cares Protect application on Twitter. From the results of the research that has been carried out, it is concluded that the accuracy level of the SVM algorithm is 86% with k-fold cross validation, the accuracy is 79% with the 8020 split technique. While the accuracy level of the Naïve Bayes algorithm is 85% with k- fold cross validation, 80% accuracy with the split 8020 technique. Based on the test results, the processing time required for the SVM algorithm is 84.06 seconds, while the Naïve Bayes algorithm is 0.0094 seconds. From these results, it is known that the SVM algorithm has the same level of accuracy as the Naïve Bayes algorithm, which differs only in terms of processing time, namely, the Naïve Bayes algorithm has a better processing time than the SVM algorithm.

REFERENCES

Fitriana, F., Utami, E., & Al Fatta, H. (2021). Analisis Sentimen Opini Terhadap Vaksin Covid - 19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes. *Jurnal Komtika (Komputasi Dan Informatika)*, *5*(1), 19–25. https://doi.org/10.31603/komtika.v5i1.5185

Goyal, G. (2021). Twitter Sentiment Analysis- A NLP Use-Case for Beginners. Retrieved April 25, 2022, from Analytics Vidhya website: https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/#:~:text=Sentiment analysis refers to identifying,about a variety of topics.

Hadna, N. M., Santosa, P., & Winarno, W. (2016). *Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen di Twitter*.

Ibrahim, M., Abdillah, O., Wicaksono, A. F., & Adriani, M. (2015). Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 1348–1353. https://doi.org/10.1109/ICDMW.2015.113

Illia, F., Eugenia, M. P., & Rutba, S. A. (2021). Sentiment Analysis on PeduliLindungi Application Using TextBlob and VADER Library. *Proceedings of The International Conference on Data Science and Official Statistics*, *1*(1), 278–288. https://doi.org/10.34123/icdsos.v2021i1.236

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. https://doi.org/10.29207/resti.v5i6.3588

Kristiyanti, D. A., Umam, A. H., Wahyudi, M., Amin, R., & Marlinda, L. (2018). Comparison of SVM & Naive Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Bases on Public Opinion on Twitter. *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 1–6. https://doi.org/10.1109/CITSM.2018.8674352

Kurniawan, S., Gata, W., Puspitawati, D. A., Nurmalasari, Tabrani, M., & Novel, K. (2019). Perbandingan Metode Klasifikasi Analisis Sentimen Tokoh Politik Pada Komentar Media Berita Online. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 176–183. https://doi.org/10.29207/resti.v3i2.935

Mustopa, A., Hermanto, Anna, Pratama, E. B., Hendini, A., & Risdiansyah, D. (2020). Analysis of User Reviews for the PeduliLindungi Application on Google Play Using the Support Vector Machine and Naive Bayes

*name of corresponding author

Algorithm Based on Particle Swarm Optimization. *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 1–7. https://doi.org/10.1109/ICIC50835.2020.9288655

Negara, A. B. P., Muhardi, H., & Putri, I. M. (2020). Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(3), 599. https://doi.org/10.25126/jtiik.2020711947

Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019). Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 266–270. https://doi.org/10.1109/SMART46866.2019.9117512

Rolliawati, D., Khalid, K., & Rozas, I. S. (2020). Teknologi Opinion Mining untuk Mendukung Strategic Planning. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(2), 293. https://doi.org/10.25126/jtiik.2020721685

Siswanto, Wibawa, Y. P., Gata, W., Gata, G., & Kusumawardhani, N. (2018). Classification Analysis of MotoGP Comments on Media Social Twitter Using Algorithm Support Vector Machine and Naive Bayes. *2018 International Conference on Applied Information Technology and Innovation (ICAITI)*, 96–101. https://doi.org/10.1109/ICAITI.2018.8686751

Tuhuteru, H., & Iriani, A. (2018). Analisis Sentimen Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode Support Vector Machine dan Naive Bayes Classifier. *Jurnal Informatika: Jurnal Pengembangan IT*, *3*, 394–401. https://doi.org/10.30591/jpit.v3i3.977

Yaakub, M. R., Latiffi, M. I. A., & Safra, L. (2019). A Review on Sentiment Analysis Techniques and Applications. *IOP Conference Series: Materials Science and Engineering*, *551*, 12070. https://doi.org/10.1088/1757-899X/551/1/012070

*name of corresponding author