ISSN 2087-3336 (Print) | 2721-4729 (Online)

TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika
Vol. 11, No. 1, January 2024, hlm. 1-7
http://jurnal.sttmcileungsi.ac.id/index.php/tekno
DOI: 10.37373

Opinion mining toward work from office policies on post-pandemic covid-19 by using supervised learning

Tri Hadi Wicaksono[1*], Imam Yuadi[1,2], Ira Puspitasari[1,3]

[1*] Master of Human Resource Development, Postgraduate School, Airlangga University, Indonesia
[2] Information and Library Science, Faculty of Social and Political Sciences, Airlangga University, Indonesia
[3] Information Systems, Faculty of Science and Technology, Airlangga University, Indonesia

[1*, 2] Jl. Airlangga No. 4-6, City of Surabaya 60286, East Java, Indonesia
[3] Jl. Dr. Ir. H. Soekarno, City of Surabaya 60115, East Java, Indonesia

[*] ✉tri.hadi.wicaksono-2022@pasca.unair.ac.id

Submitted: 29/03/2023 Revised: 26/04/2023 Accepted: 27/04/2023

ABSTRACT

Post-pandemic COVID-19, many companies have re-implemented work from office policies for their employees. However, the policy has been controversial among social media activists, especially in Twitter. The sentiment arose according to them, during the pandemic COVID-19 they believe that working from home has many advantages over working in an office. The emergence of 'work from office' sentiments is an interesting target for opinion-mining research. Opinion mining or sentiment analysis is a general research area of data mining that helps to explore and analyze existing views and opinions to obtain useful information. The analysis process involves the use of machine learning with several supporting algorithms. This study used four classification algorithm models of supervised learning, including naive Bayes, support vector machines, k-nearest neighbors, and random forests. The selection of those algorithms also aims to find out which model produced a good performance for the results. The performance results of each model were evaluated by the confusion matrix and the k-nearest neighbor algorithm model with an accuracy value of 96.62% was found to give the best results and to be the most used model in the classification process. On the other hand, the algorithm model that obtains the lowest accuracy is a random forest with 72.08%.

Keywords: WFO, opinion mining, supervised learning, machine learning, post-pandemic

1. INTRODUCTION

The company's post-pandemic work from office (WFO) policy was deemed ineffective, created and added to urban congestion, contributed to added pollution, and caused workers to become stressed, which has an impact less productive work performance, sparking the creation of a petition asking for work from home (WFH) to be re-implemented. According to a YouGov study conducted for the BBC, over 70% of the 1,684 respondents believe that workers won't ever return to the workplace once the pandemic is ended **[1]**. Additionally, according to a McKinsey survey, about 52% of the 5,043 full-time employees at the company prefer post-pandemic work flexibility [2]. This can improve organizational performance, cut expenses, and make the most of talent no matter where employees are situated. This is consistent with previous results showing WFH has an impact on boosting job satisfaction and work system flexibility [3], efficiency, and productivity [4], when working in a pandemic, feeling secure [5], heaving fewer conflicts at work, and in the home during the pandemic [6].

Following the pandemic, policies shaped public opinion, which in turn influenced social media activists' positive, negative, and neutral feelings. Sentiment analysis is the process of examining arising to learn more about or gain a general sense of the subjects being discussed. Because this analysis can be used by businesses, governments, and researchers to explore and analyze public sentiment or views, gain business insights, and improve decision-making, it is a popular research area in data mining [7]. Sentiment analysis, sometimes referred to as opinion mining, is a method of using natural language processing (NLP) to comprehend, extract, and analyze textual data to acquire information that contains sentiments in an opinion [8]. By categorizing opinion attitudes, this study can provide helpful information for organizations, including firms and agencies [9]. Utilizing text mining and machine learning, the analysis process looks at text data to find sentiments and subjective information [10].

Previous studies on opinion mining or sentiment analysis connected to work from home activities have been conducted, and the findings suggest that the support vector machine algorithm model produced an F1 score of 83.36% [11]. However, sentiment analysis research contrasting work from home and work from the office has not yet been conducted. Meanwhile, many researchers have evaluated machine learning techniques using sentiment analysis to compare the efficacy, advantages, and drawbacks of each mechanism [12]. The results of the study demonstrate that supervised learning methods produce results with more accuracy than unsupervised learning methods. On the other hand, researchers discovered that the naive Bayesian classifier is best suited for small data sets since as the data amount increases, its accuracy drastically declines. Support vector machines, on the other hand, are efficient with any size of data collection. Future study will concentrate on investigating deep learning techniques for sentimental classification.

The researcher is interested in undertaking a study to learn more about opinion mining on this subject in light of the context that has been given. This study is expected to be able to offer both theoretical and practical scope. Regarding theoretical scope, this study employs Twitter data with WFO keywords and a machine learning technique with supervised learning algorithms for WFO policy sentiment analysis. Additionally, it is anticipated that this research will be able to address the results of earlier studies by improving the precision of sentiment analysis, classification, and detection. Regarding the evaluation of WFO policies put in place after the pandemic, this analytical sentiment can be applied in a variety of industries including office businesses, services, and academia. The study suggests using machine learning with supervised learning in scenarios resembling those in other studies so that it can be modified and enhanced to obtain a higher degree of accuracy, particularly in challenging contexts when performing text analysis.

2. LITERATURE REVIEW

The supervised learning algorithm models used in this work, naïve Bayes (NB), support vector machine (SVM), k-nearest neighbors (KNN), and random forest (RF), were examined in the literature.

a) Naive Bayes

One of the simplest probabilistic classification techniques based on the Bayes theorem, Naive Bayes is a popular classification technique and one of the top 10 data mining algorithms [13]. To make calculations simpler, Naive Bayes assumes that the impact on attribute values for specific classes is independent of the impact on other attribute values [14].

b) Support vector machine

A guided learning technique for classification is the support vector machine. Both linear and non-linear classification can employ SVM. In essence, non-linear SVM solves the linear SVM issue by doing kernel operations in high-dimensional feature space [15].

c) K-nearest neighbors

An algorithm called K-nearest neighbors uses learning data (train data set), which is drawn from the set of 'k' closest neighbors (nearest neighbors), to categorize data. where k represents the quantity of closest neighbors. The majority of the proximity of the closest neighbors is used to classify the results [16].

d) Random forest

As a development of the decision tree approach, random forest divides each characteristic into trees that are randomly chosen from a subset of attributes after each decision tree is trained using a sample [17]. This algorithm model has several advantages that can improve accuracy, where missing data is found and reject outliers. In addition, random forest is also used as an efficient data store. On the other hand, this algorithm also has a feature selection process that can use the best features to improve the performance of the classification model. Of course, with a choice of features that can work effectively with large data with complex parameters.

3. METHOD

The design flow used for this study methodology is as follows,

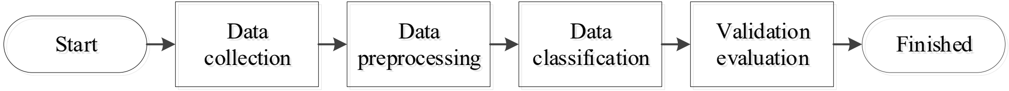


Figure 1. Research design flow

As shown in **Figure 1**, the process begins with gathering the necessary data, followed by data preparation, data classification using an algorithm model, and assessment of the algorithm model's performance using a confusion matrix.

a) Data collection

The text was initially gathered via secondary Twitter data in the early phases. The data used has anything to do with the word "WFO.".

b) Data preprocessing

This stage involves numerous text processing operations, including tokenization, case transformation, stopword filtering, stemming, and token weighting.

c) Data classification

RapidMiner Studio, a widely used data mining analytical tool that supports a several machine learning algorithms, is used in the classification process [18]. This program, which is supported by 100 learning schemes for clustering, classification, and regression analysis, can be utilized for predictive analysis [19]. Furthermore, an analysis was carried out with the RapidMiner Studio application in making an appropriate and automatic classification program.

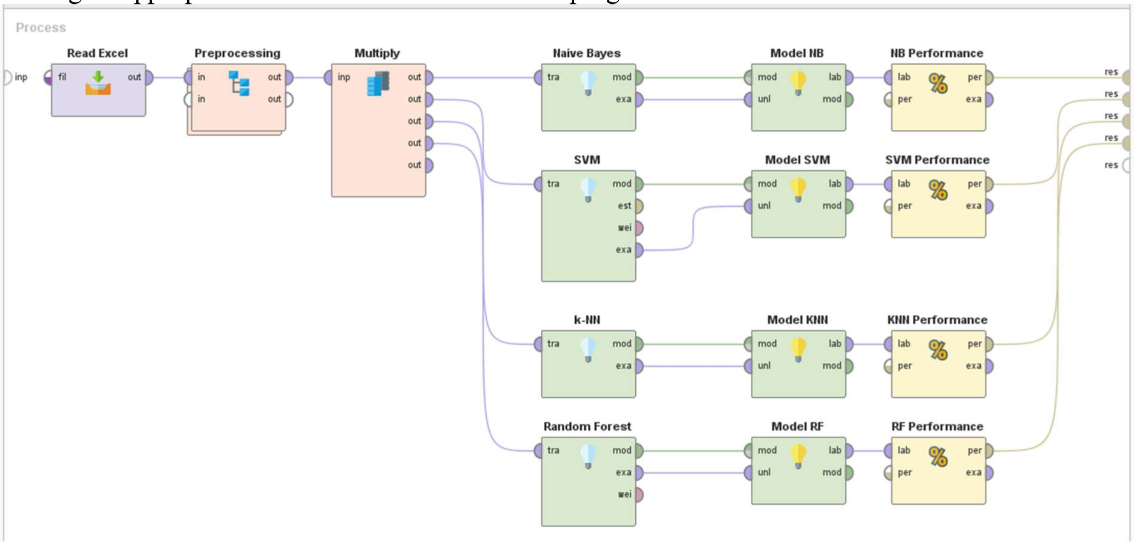


Figure 2. Rapidminer Studio's classification procedure.

Figure 2 is a representation of the categorization method used by RapidMiner Studio, as is well known. The method starts with entering data that was gathered during the data collecting stage, and it then moves on to text mining utilizing the naive Bayes algorithm model, support vector machine, k-nearest neighbors, and random forest from the outcomes of preprocessing. The performance results of the algorithm model then show that a text classification is formed into many categories.

d) Validation evaluation

The confusion matrix, a technique for gauging the algorithm model's level of accuracy throughout the classification phase, is used for the model's evaluation and validation at this point [20], and a table that can show how well the categorization model performs [21].

Table 1. Classification of two classes with confusion

		Predictions	
		Positive	**Negative**
Actual	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

The classification of prediction data derived from models based on real or accurate data is shown in **Table 1**. True positive (abbreviated as TP) will result in a positive prediction model because it is based on a large amount of real data in a positive class, while true negative (abbreviated as TN) will result in a negative prediction model because it relates to a large amount of real data in the negative class [22]. A false positive, or FP, on the other hand, will produce a positive prediction model potential but a negative conclusion regarding the actual data being in the negative class. In contrast, a false negative (abbreviated as FN) will result in a negative prediction model linked to positive classification mistakes since a large portion of the real data is in the positive class [23]. The solution of equation 1 can be used to formulate a model's accuracy level.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} \times 100\% \tag{1}$$

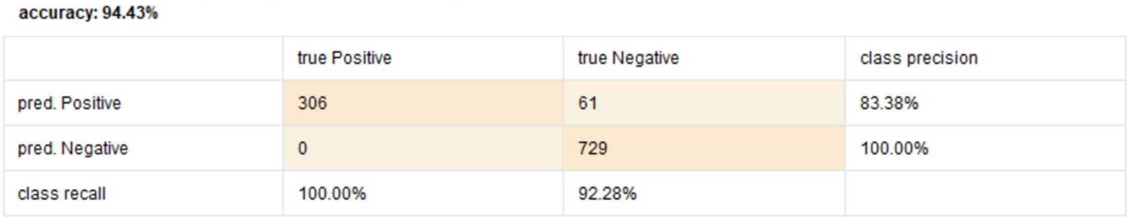
Where:

- *True Positive* (TP)
- *False Positive* (FS)
- *False Negative* (FN)
- *True Negative* (TN)

4. RESULTS AND DISCUSSION

In evaluating the model using the confusion matrix and the performance results of the algorithm model are obtained.

a) The Naive Bayes algorithm model's performance



accuracy: 94.43%

	true Positive	true Negative	class precision
pred. Positive	306	61	83.38%
pred. Negative	0	729	100.00%
class recall	100.00%	92.28%	

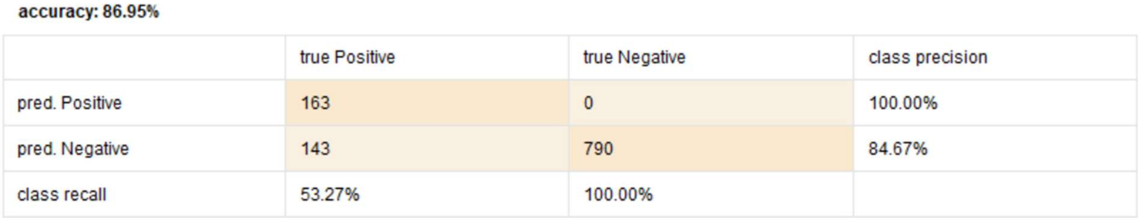
Figure 3. Naive Bayes with a confusion matrix

According to **Figure 3**, 306 of the data are classified as true positive, 729 as true negative, 61 as false negative, and none as false positive. The outcome is that the naive Bayes algorithm model performs with a result of 94.43%. The outcome of manually calculating the accuracy value is.

$$Accuracy = \frac{306 + 729}{306 + 729 + 0 + 61} \times 100\%$$

$$= \frac{1.035}{1.096} \times 100\% \;=\; 94{,}43\%$$

b) The effectiveness of the model of the support vector machine algorithm.



accuracy: 86.95%

	true Positive	true Negative	class precision
pred. Positive	163	0	100.00%
pred. Negative	143	790	84.67%
class recall	53.27%	100.00%	

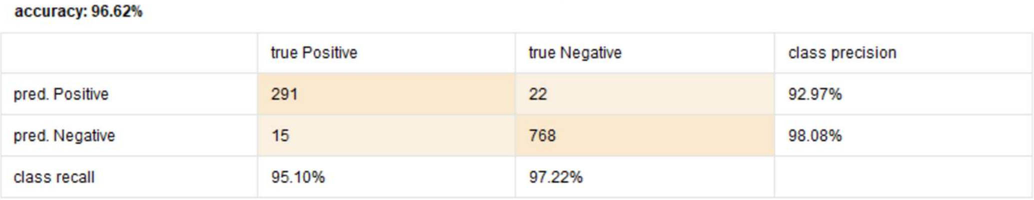
Figure 4. Support vector machine for the confusion matrix

According to **Figure 4**, 163 data are categorized as true positive, 790 as true negative, 143 as false positive, and no data are categorized as false negative. This results in a performance result of 86.95% for the support vector machine algorithm model. if the accuracy value was determined by hand.

$$Accuracy = \frac{163 + 790}{163 + 790 + 143 + 0} \times 100\%$$

$$= \frac{953}{1.096} \times 100\% \;=\; 86{,}95\%$$

c) The performance of the k-nearest neighbors algorithm model



accuracy: 96.62%

	true Positive	true Negative	class precision
pred. Positive	291	22	92.97%
pred. Negative	15	768	98.08%
class recall	95.10%	97.22%	

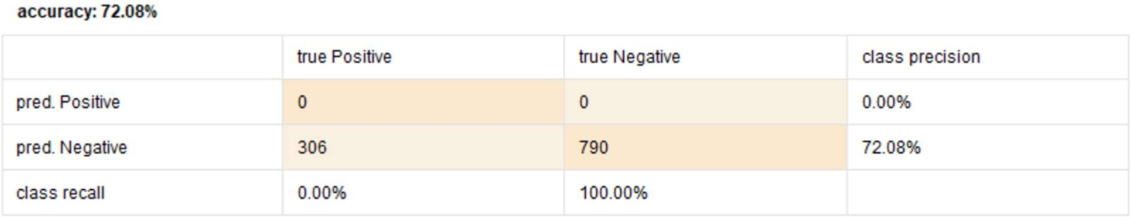
Figure 5. K-nearest neighbors confusion matrix

By **Figure 5**, 291 data are categorized as true positives, 768 data as true negatives, 15 data as false positives, and 22 data as false negatives. As a consequence, 96.62% is the performance score for the k-nearest neighbors method model. The outcome of manually calculating the accuracy value is.

$$Accuracy = \frac{291 + 768}{291 + 768 + 15 + 22} \times 100\%$$

$$= \frac{1.059}{1.096} \times 100\% \;=\; 96{,}62\%$$

d) The effectiveness of a model using the random forest algorithm



accuracy: 72.08%

	true Positive	true Negative	class precision
pred. Positive	0	0	0.00%
pred. Negative	306	790	72.08%
class recall	0.00%	100.00%	

Figure 6. Matrix random forest with confusion

The model's effectiveness reveals a random forest accuracy of 72.08%. **Figure 6** shows that no data are classified as real positives or false positives and that 790 data are classed as genuine negatives, 306 data as false negatives, and no data as true positives. Thus, a performance result of 72.08% is displayed for the random forest algorithm model. The outcome of manually calculating the accuracy value is.

$$Accuracy = \frac{0 + 790}{0 + 790 + 0 + 306} \times 100\%$$

$$= \frac{790}{1.096} \times 100\% \;=\; 72{,}08\%$$

The k-nearest neighbors algorithm model produced a high accuracy value of 96.62% and became an algorithm with greater performance than the other models, according to the performance findings of the four algorithm models examined in the confusion matrix. This is connected to the k-nearest neighbors technique, which produced 291 real data points for the positive class and 768 real data points for the negative class. The term "accuracy value" also describes the degree to which the model correctly predicts the positive and negative outcomes about the total amount of data.

5. CONCLUSION

The k-nearest neighbors algorithm has the best accuracy value of 96.62% when compared to other algorithms in the sentiment analysis of work from office policies. This study combines machine learning with a variety of algorithm models from supervised learning. The results of this study also address earlier research on improving detection, classification, and sentiment analysis precision. Unfortunately, more research is necessary to fully understand the limitations of this finding because of the small amount of processed data. An F1 score of 83.36% was obtained with the support vector machine algorithm model, according to earlier research on opinion mining or sentiment analysis connected to work from home activities. Unfortunately, there hasn't been any sentiment analysis research contrasting working from home and working in an office up until now. Additionally, sentiment analysis is a well-known research area in data mining because it offers businesses, governments, and researchers a useful tool for exploring and analyzing public sentiment or views, gaining business insights, and improving decision-making. The execution time at which the method is applied determines whether the trend in sentiment analysis is rising or dropping. Therefore, it is anticipated that future research will be able to assess the model using its running time.

REFERENCES

[1] R. Jones, L., and Wearn, "'Most Workers Do Not Expect Full-Time Office Return, Survey Says.,'" 2021. https://www.bbc.com/news/business-58559179.

[2] A. Alexander, A. De Smet, M. Langstaff, and D. Ravid, "What employees are saying about the future of remote work," *McKinsey Co.*, no. April, 2021, [Online]. Available: https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/what-employees-are-saying-about-the-future-of-remote-work.

[3] G. Kaufman and H. Taniguchi, "Working from Home and Changes in Work Characteristics during COVID-19," *Socius*, vol. 7, 2021, doi: 10.1177/23780231211052784.

[4] P. Choudhury, C. Foroughi, and B. Larson, "Work-from-anywhere: The productivity effects of geographic flexibility," *Strateg. Manag. J.*, vol. 42, no. 4, 2021, doi: 10.1002/smj.3251.

[5] A. D. Dubey and S. Tripathi, "Analysing the sentiments towards work-from-home experience during COVID-19 pandemic," *Journal of Innovation Management*, vol. 8, no. 1. 2020, doi: 10.24840/2183-0606_008.001_0003.

[6] S. Schieman, P. J. Badawy, M. A. Milkie, and A. Bierman, "Work-Life Conflict During the COVID-19 Pandemic," *Socius*, vol. 7, 2021, doi: 10.1177/2378023120982856.

[7] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, 2021, doi: 10.1016/j.knosys.2021.107134.

[8] A. Nurzahputra and M. A. Muslim, "Analisis Sentimen pada Opini Mahasiswa Menggunakan Natural Language Processing," *Semin. Nas. Ilmu Komput.*, no. Snik, 2016.

[9] P. A. Permatasari, L. Linawati, and L. Jasa, "Survei Tentang Analisis Sentimen Pada Media Sosial," *Maj. Ilm. Teknol. Elektro*, vol. 20, no. 2, 2021, doi: 10.24843/mite.2021.v20i02.p01.

[10] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 2, no. 5, 2022, doi: 10.30865/klik.v2i5.362.

[11] Elly Pusporani, Siti Qomariyah, and Irhamah, "Klasifikasi Pasien Penderita Penyakit Liver," vol. 2, no. March, 2019.

[12] S. C. Agrawal, S. Singh, and S. Gupta, "Evaluation of Machine Learning Techniques in Sentimental Analysis," 2021, doi: 10.1109/ISCON52037.2021.9702430.

[13] A. I. Lubis, U. Erdiansyah, and R. Siregar, "Komparasi Akurasi pada Naive Bayes dan Random Forest dalam Klasifikasi Penyakit Liver," *J. Comput. Eng. Syst. Sci.*, vol. 7, no. 1, pp. 81–89, 2022, doi: DOI: https://doi.org/10.24114/cess.v7i1.28888.

[14] H. Nalatissifa, W. Gata, S. Diantika, and K. Nisa, "Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, 2021, doi: 10.32493/informatika.v5i4.7575.

[15] F. Handayani, "Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network Dalam Prediksi Penyakit Jantung," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 3, 2021, doi: 10.26418/jp.v7i3.48053.

[16] K. Kristiawan and A. Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 1, 2021, doi: 10.28932/jutisi.v7i1.3182.

[17] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, “Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah,” *E-Bisnis J. Ilm. Ekon. dan Bisnis*, vol. 13, no. 2, 2020, doi: 10.51903/e-bisnis.v13i2.247.

[18] Sheena Angra and Sachin Ahuja, “Analysis of Student’s Data using Rapid Miner,” *J. Today’s Ideas - Tomorrow’s Technol.*, vol. 4, no. 2, 2016, doi: 10.15415/jotitt.2016.42007.

[19] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, “YALE: Rapid prototyping for complex data mining tasks,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, vol. 2006.

[20] I. P. Rahayu, A. Fauzi, and J. Indra, “Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Naive Bayes Dan Support Vector Machine,” vol. 4, pp. 296–301, 2022, doi: 10.30865/json.v4i2.5381.

[21] R. Bold, H. Al-Khateeb, and N. Ersotelos, “Reducing False Negatives in Ransomware Detection: A Critical Evaluation of Machine Learning Algorithms,” *Appl. Sci.*, vol. 12, no. 24, 2022, doi: 10.3390/app122412941.

[22] Fatihah Rahmadayana and Yuliant Sibaroni, “Sentiment Analysis of Work from Home Activity using SVM with Randomized Search Optimization,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 5, 2021, doi: 10.29207/resti.v5i5.3457.

[23] I. W. Saputro and B. W. Sari, “Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa,” *Creat. Inf. Technol. J.*, vol. 6, no. 1, 2020, doi: 10.24076/citec.2019v6i1.178.