Performance Analysis of Naïve Bayes and Support Vector Machine in Sentiment Analysis of Clash of Champions Event by Ruangguru on X/Twitter Media

Leni Muslimah[1], Dr. Ravi Ahmad Salim[2]

Department of Information Systems Management, Gunadarma University, Depok, Indonesia, 16424

Abstract—This research aims to analyze user sentiment towards the Ruangguru Application on the X/Twitter social media platform following the Launch of "Clash of Champions," a web reality series produced by Ruangguru. The dataset used in this study comprises 1,559 tweets Collected over 68 days using the keyword "Clash of Champions" through Crawling data techniques on the Google Colab research platform. After Text preprocessing, 1,469 structured tweet data were obtained and Classified using the Naïve Bayes and Support Vector Machine (SVM) Methods. The classification results using Naïve Bayes showed an Accuracy of 91.35%, 87.56% recall for positive data. Meanwhile, the SVM method yielded an accuracy of 92.24, 100% recall for positive Data. Based on accuracy values, the SVM method outperformed Naïve Bayes in analyzing sentiment. Positive sentiment was found to be more Dominant in user responses to the Ruangguru application after the Launch of "Clash of Champions". This sentiment analysis provides Valuable contributions to the development of digital marketing Strategies in improving Ruangguru's services. It offers insights into Public perception of the event, which can serve as evaluation material To enhance service quality.

Keywords—Sentiment Analysis, Naïve Bayes, Support Vector Machine, Twitter, Ruangguru.

I. INTRODUCTION

The increasing development of information technology requires learning methods to follow technological developments. The learning methods used should be able to utilize various media to improve the quality of learning outcomes. One of the learning media is to use e-learning or electronic-based learning using computers or computer-based. There are several e-learning applications in Indonesia, from paid to unpaid, one of which is the Ruangguru application [1].

Ruangguru has partnered with 32 provincial governments and 326 cities and regions in Indonesia to develop technology-based learning services, including virtual classroom services, online testing platforms, educational video sharing, marketplaces, courses and other educational programs available Through Ruangguru website and app. [2]

To improve the quality of its services, Ruangguru often holds various events and competitions. One event that is currently attracting attention is the Clash of Champions. This event aims to give awards to outstanding students and motivate other users to improve their academic achievements through the Ruangguru platform.

Despite having a positive purpose, this event also garnered various responses from the public, which were conveyed through various social media platforms, including X (formerly known as Twitter). X/Twitter is one of the platforms widely used by users to express their opinions. Data obtained in April 2024 showed that Indonesia was ranked 4th as the largest X user [3].

Sentiment analysis is used to find valuable information needed from unstructured data, so it is hoped that this study can determine the sentiment of Twitter users towards the Ruangguru application. With this approach, companies can gather insights from customer responses or reviews about their products or services, which can then be used to improve their business strategies [4].

Several researchers have previously conducted research to measure sentiment analysis with various algorithms, including the Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Decision Tree, K-Medoid, and Backpropagation Neural Network (BNN) algorithms. Naïve Bayes and Support Vector Machine (SVM) are two methods commonly used in sentiment analysis. Naïve Bayes is a simple yet effective probabilistic method for text classification, while SVM is a powerful classification method and is often used in various machine learning applications. By combining these two methods, sentiment analysis can be done more accurately and comprehensively [5].

Based on the description above, the researcher proposes a study of Clash of Champions Sentiment Analysis of the Ruangguru Application Using Naïve Bayes and Support Vector Machine in Media X/Twitter.

II. METHOD

A. Research Object

Ruangguru Clash of Champions is an Indonesian reality web series produced by Ruangguru, a startup in the field of education. This event presents a competition that tests the intelligence of students from various leading universities, both from Indonesia and abroad. In the Ruangguru application, it provides an opportunity for students to study privately. Students can take pictures of questions that are considered difficult, upload them, and communicate with teachers via messages or calls online using a smartphone or laptop [6].

B. Data Collection Methodology

The data collection method in this research is through literature studies and data crawling.

C. Data Analysis and Interpretation of Results

The data analysis stages used in this study are the SEMMA (Sample, Explore, Modify, Model, Assess) method.

D. Research Area

Based on relevant research, research conducted by Aliyah et al. (2020a) is the main reference in this study. The study has similarities in research objects, application of the SEMMA method, use of the Naïve Bayes and Support Vector Machine algorithms, and similar data sources and labeling processes [7]. With the comparison between these two methods, the research results become more accurate because the method with the highest level of accuracy is selected.

III. RESULT AND DISSCUSSION

A. Sample

This stage collects data; data is collected through the data-crawling process. In this data crawling, researchers use the Python programming language using tweet-harvest to retrieve tweet data stored in xlsx file format.
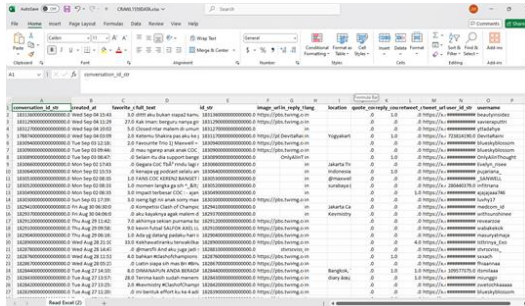
Fig. 1. Crawling Data Results after saving in .xlsx format

The results of the successful Twitter data crawling collected amounted to 1559 data points. After changing the crawling data results into Excel format, the next step is to process the text to reduce interference so that it facilitates the next steps.

B. Explore

Before proceeding to the text processing stage, we need to select the data attributes that have been obtained from data crawling using the full_text attribute.

Fig. 2. Attribute Selection Results (Explore)

Figure 2. is the result of the full_text attribute selection that will be used in this sentiment analysis process. With 1559 data points and 1 regular attribute.

C. Modify

1. Cleaning

The cleaning stage is carried out by removing attributes that are not needed and do not have important meaning, such as hashtags, mentions, retweets, excess spaces (whitespace), links, and other symbols.

Fig. 3. Results Before Cleaning Process

Figure 3. is the result of the dataset before the cleaning process. In this cleaning stage, the author also deletes duplicate tweet data. Deleting unnecessary symbols, hashtags, mentions, and urls so that the data is ready to continue to the next process.

Fig. 4. Results After Cleaning Process

Figure 4 shows the results after the data cleaning process. It can be seen that symbols, mentions, hashtags, and URLs have been successfully removed, making it easier for the data to proceed to the next stage.

2. Tokenizing

Tokenizing is the process of separating words in a text document into independent units that stand alone. This process is done using the Process Documents from Data operator, which uses TF-IDF vectorization. The result of this process successfully breaks the entire text into 7492 regular word attributes.

164

Word	Attribute Name	Total Occurences ↓	Document Occurences
ClashofChampions	ClashofChampions	852	849
ini	ini	319	284
di	di	314	269
clashofchampions	clashofchampions	282	280
aku	aku	263	195
yang	yang	262	208
coc	coc	250	219
yg	yg	246	200
dan	dan	245	211
ClashOfChampion	ClashOfChampion	223	223
ruangguru	ruangguru	202	184
ga	ga	169	141
bisa	bisa	162	144
ya	ya	157	144
D	D	152	100
sama	sama	151	135
banget	banget	150	129
ada	ada	148	128

Fig. 5. Tokenization Results

3. Transform Case

Transform Case is a process of standardizing letters in which only Latin letters from 'a' to 'z' are accepted and converted to all lowercase.

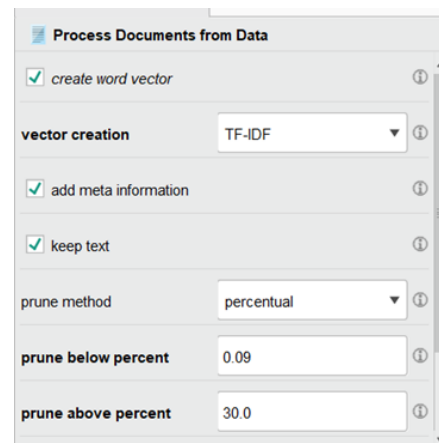| Word	Attribute Name	Total Occure... ↓	Document Occurences
clashofchampions	clashofchampions	1255	1250
coc	coc	448	363
ini	ini	365	323
di	di	344	293
aku	aku	313	232
yang	yang	285	226
clashofchampion	clashofchampion	277	277
ruangguru	ruangguru	271	230
yg	yg	263	209
dan	dan	255	217
sandy	sandy	224	190
maxwell	maxwell	208	169
grabcoc	grabcoc	195	98
ga	ga	191	158
bisa	bisa	184	163
ada	ada	179	152
ya	ya	175	161
banget	banget	171	150

Fig. 6. Transform Case Results

4. Stopwords Removal

The stopwords removal stage aims to remove words that are often used but have no influence on the sentiment of a sentence. In this study, stopword removal was carried out using an Indonesian stopword dictionary obtained from the corpus on Kaggle.

After going through the processing stages as above, at this stage, the frequency of data occurrence is evaluated to ensure that there is no data with a frequency that is too low or too high, which can affect the performance of the method used. The pruning process is carried out by removing words whose frequency is below 0.09% or above 30%. As a result, the number of data in the dataset changes from 1,559 crawling data to 1,469 data.

Word	Attribute Name	Total Occure... ↓	Document Occure...
clashofchampions	clashofchampions	1255	1250
coc	coc	448	363
clashofchampion	clashofchampion	277	277
ruangguru	ruangguru	271	230
yg	yg	263	209
sandy	sandy	224	190
maxwell	maxwell	208	169
grabcoc	grabcoc	195	98
ga	ga	191	158
ya	ya	175	161
banget	banget	171	150
axelots	axelots	166	162
d	d	155	103
timsekop	timsekop	154	150
kevin	kevin	150	124
axel	axel	142	121
sandyourheart	sandyourheart	134	134
shakira	shakira	128	107

Fig. 7. Stopwords Results

Process Documents from Data

- ☑ create word vector
- **vector creation** — TF-IDF ▼
- ☑ add meta information
- ☑ keep text
- **prune method** — percentual ▼
- **prune below percent** — 0.09
- **prune above percent** — 30.0

Fig. 8. Parameter Process Documents from Data

5. Filter

Filter is the stage of deleting words that have letters that are too short and letters that are too long. In this study, filtering is set with a minimum of 4 letters and a maximum of 25 letters.

Word	Attribute Name	Total Occurences ↓	Document Occurences
clashofchampions	clashofchampions	1255	1250
clashofchampion	clashofchampion	277	277
ruangguru	ruangguru	271	230
sandy	sandy	224	190
maxwell	maxwell	208	169
grabcoc	grabcoc	195	98
banget	banget	171	150
axelots	axelots	166	162
timsekop	timsekop	154	150
kevin	kevin	150	124
axel	axel	142	121
sandyourheart	sandyourheart	134	134
shakira	shakira	128	107
finalchaptercoc	finalchaptercoc	124	124
udah	udah	109	95
thankyoucoc	thankyoucoc	107	107
grabid	grabid	98	98
champions	champions	96	88

Fig. 9. Filter Results

Fig. 10. Data Preprocessing Results

In Figure 10. it can be seen that the data obtained after processing the data is 1469 data from 1559 crawling data. This data is the dataset that will be used in the next stage.

After data processing is complete, the word cloud feature can be used to illustrate the most frequently occurring words in the dataset.

Row No.	word	in documents	total
1	clashofchampion	277	277
2	ruangguru	230	271
3	sandy	190	224
4	maxwell	169	208
5	grabcoc	98	195
6	banget	150	171
7	axelots	162	166
8	timsekop	150	154
9	kevin	124	150
10	axel	121	142
11	sandyourheart	134	134
12	shakira	107	128
13	finalchaptercoc	124	124
14	udah	95	109
15	thankyoucoc	107	107
16	grabid	98	98
17	champions	88	96
18	nonton	86	93

Fig. 11. Frequently Appearing Words

Figure 11 displays the sequence of words that frequently appear in the dataset, using the total attribute sorting operator, it will display a word cloud visualization as in Figure 12.

Fig. 12. Word Cloud dataset Clash of Champions

D. ModeL

The next stage involves machine learning modeling using

the Naïve Bayes and Support Vector Machine algorithms, as well as the Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction method. In this study, the data (tweets) that have gone through the text pre-processing stage are divided into two parts: training data and testing data. After that, feature extraction is carried out from the selected tweets. Training data is used to predict sentiment labels on tweets in the test data [7]. Of the total 1469 preprocessed data, 635 data are used as training data, while the remaining 834 data are used as testing data.

To train the machine learning algorithm, manual labeling is done on the training data. According to research conducted by Prasetyo & Hidayatullah (2020), this manual labeling helps machine learning to recognize sentence patterns that will later be used in labeling automation [8].

Fig. 13. Test Results using the Naïve Bayes method

Figure 13 shows the test results using the Naïve Bayes method which provides sentiment predictions along with the level of confidence for each prediction.

Fig. 14. Hasil Uji using method Support Vector Machine

Figure 14 is the test result using the Support Vector Machine method which provides sentiment predictions along with the level of confidence for each prediction. The values range from 0 to 1. The closer to 1, the higher the confidence of the model.

E. Assess

1. Naïve Bayes
In the first trial stage using the Naïve Bayes classification

with the Cross Validation method and the K-fold value as a parameter, at this stage the K-fold value used is K 10 with accuracy results. The cross-validation process using the Naïve Bayes algorithm includes the training and testing stages, including model application and performance evaluation (performance-Naïve Bayes).

This process consists of two main stages: training and testing. Through the cross-validation process involving these two stages, the Naïve Bayes algorithm for Clash of Champions produces the following levels of accuracy:

Fig. 15. Confusion Matrix Results with Naïve Bayes

The AUC (Area Under Curve) distribution of this test can be seen as follows:

Fig. 16. Naïve Bayes classification AUC graph

From the graph above, it can be seen that the AUC figure of 0.876 shows that the Naive Bayes model used has a fairly good level of performance in separating positive and negative sentiments. So that this model is able to distinguish positive and negative sentiments as much as 87.6% of the time. The greater the AUC value (closer to 1), the better the model is at classifying positive and negative classes.
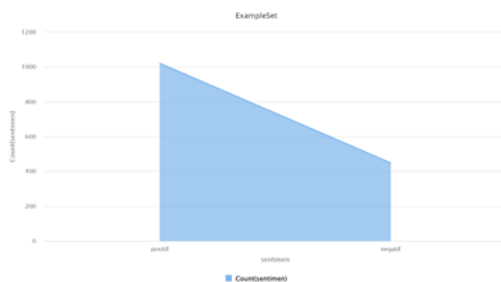
Fig. 17. Naïve Bayes Area Graph

Figure 17 shows that there were more positive tweets, namely 1021 tweets, compared to negative tweets, which were

only 448 tweets.

2. Support Vector Machine

In the next test, namely by using the Support Vector Machine algorithm model and using the same K-fold value as the previous test, namely 10, the test using cross-validation produced a different accuracy value from the previous algorithm but not small enough when compared to the Naïve Bayes algorithm. The cross-validation process using the Support Vector Machine (SVM) algorithm includes the training and testing stages, including model application (apply model) and performance evaluation (performance-SVM).

This process consists of two main stages: training and testing. By conducting cross-validation involving these two stages, the SVM algorithm for Clash of Champions produces the following accuracy:

Fig. 18. Test results based on the Support Vector Machine algorithm with K-Fold

The AUC (Area Under Curve) distribution of this test can be seen as follows:
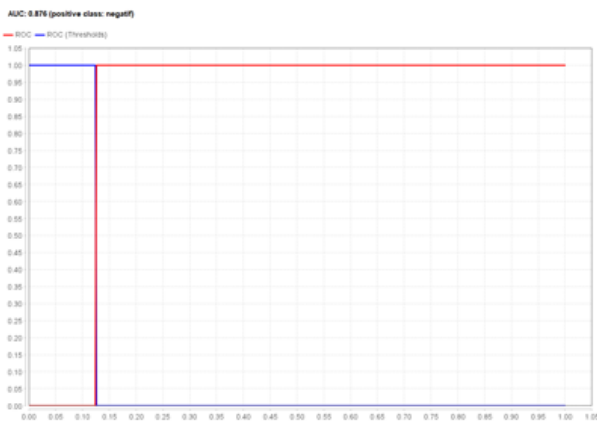

Fig. 19. Support Vector Machine classification AUC graph

From the graph above, it can be seen that the AUC figure of 0.977 shows that the SVM model used has a good level of performance in separating positive and negative sentiments. So this model is able to distinguish positive and negative sentiments as much as 97.7% of the time.
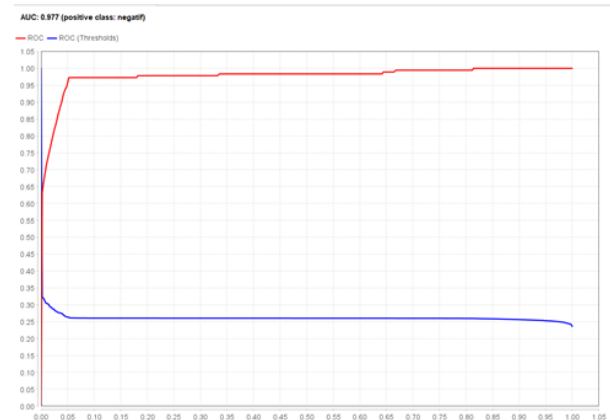

Fig. 20. Graph Area Support Vector Machine

Figure 20 shows that there were more positive tweets, namely 1286 tweets, compared to negative tweets, namely only 183 tweets.

F. Interpretation of Results

Based on the classification model testing process using the Naive Bayes and Support Vector Machine methods, the accuracy results of the two methods are obtained in Table 1.

TABEL 1. Performance of Naïve Bayes and Support Vector Machine

	Naive Bayes				Support Vector Machine				
Accuracy	(+) Precision	(-) Precision	Recall (+)	Recall (-)	Accuracy	(+) Precision	(-) Precision	Recall (+)	Recall (-)
91.35%	100%	77.91%	87.56%	100%	92.24%	91.86%	100%	100%	37.70%

This stage collects data, data is collected through the data Crawling process. In this data Crawling, researchers use the Python programming language using tweet-harvest to retrieve tweets data stored in xlsx file format.

The performance obtained from Naive Bayes is as follows:
1. *Accuracy:* useful for seeing the closeness of the system's prediction value and the human prediction produced, where the closeness produced from the two predictions is 91.35% in the system using the naive bayes algorithm.
2. *Positive Precision:* useful for seeing the comparison of true positive data and true positive predictions produced by the system, in naive bayes it produces a positive precision value of 100% in the system using the naive bayes algorithm.
3. *Negative Precision:* useful for seeing the comparison of true negative data and true negative predictions produced by the system, in naive bayes it produces a negative precision value of 77.91% in the system using the naive bayes algorithm.
4. *Positive Recall:* useful for measuring the model's ability to predict positive data, in naive bayes it produces a positive recall value of 87.56% in the system using the naive bayes algorithm.
5. *Negative Recall:* useful for measuring the model's ability to predict negative data, in naive bayes it produces a value of 100% in the system using the naive bayes algorithm.

The performance obtained from the Support Vector Machine is as follows:
1. *Accuracy:* useful for seeing the closeness of the system's prediction value and the human prediction produced, where the closeness produced from both predictions is 92.24% in the system using the Support Vector Machine algorithm.
2. *Positive Precision:* useful for seeing the comparison of true positive data and true positive predictions produced by the system, in naive bayes it produces a positive precision value of 91.86% in the system using the Support Vector Machine algorithm.
3. *Negative Precision:* useful for seeing the comparison of true negative data and true negative predictions produced by the system, in naive bayes it produces a negative precision value of 100% in the system using the Support Vector Machine algorithm.

4. *Positive Recall:* useful for measuring the model's ability to predict positive data, in naive bayes it produces a positive recall value of 100% in the system using the Support Vector Machine algorithm.
5. *Negative Recall:* useful for measuring the model's ability to predict negative data, in naive bayes it produces a value of 37.70% in the system using the Support Vector Machine algorithm.

Research (Sujadi, 2022) on sentiment analysis with 1625 data sets using Naïve Bayes and Support Vector Machine produced a Naïve Bayes accuracy of 78.3% and a Support Vector Machine accuracy of 81.6% [9]. Research by Asro et al. (2024) with 1800 data sets produced a Naïve Bayes accuracy of 72.19% and a Support Vector Machine accuracy of 76.33% [10]. Research (Pamungkas & Kharisudin, 2021) on sentiment analysis of 10000 data sets using the Naïve Bayes method produced an accuracy of 79.2% and a support vector machine accuracy of 90.01%. [11]. Research (Hidayat & Sugiyono, 2023) on sentiment analysis of 482 datasets using the Naïve Bayes method produced an accuracy of 96.14% and a support vector machine accuracy of 94.80% [12]. Research (Alizah et al., 2020) sentiment analysis of 15494 datasets using the Naïve Bayes method produced 81% accuracy and 87% support vector machine accuracy [7].

TABEL 2. Performance Comparison Results

Researcher	Methods	
	NB	SVM
(Sujadi, 2022)	78.3%	81.6%
(Asro et al., 2024)	72.19%	76.33%
(Pamungkas & Kharisudin, 2021)	79.2%	90.01%
(Hidayat & Sugiyono, 2023)	96.14%	94.80%
(Alizah et al., 2020)	81%	87%
Penulis	91.35%	92.24%

Table 2 shows that the SVM results have a superior value compared to Naïve Bayes, so the results of the classification comparison used are SVM. The SVM algorithm successfully predicted the sentiment category in this study. The analysis of tweet sentiment related to Clash of Champions with the SVM method is more dominated by positive sentiment than negative sentiment. The SVM classification produced 1286 positive sentiments and 183 negative sentiments.

IV. CONCLUSION

Based on the results of the classification discussion, the following conclusions can be drawn from this study:
1. Classification with Naïve Bayes produces 1021 positive sentiments and 448 negative sentiments, while Support Vector Machine produces 1286 positive sentiments and 183 negative sentiments.
2. Based on the results of the analysis, the interpretation of the study states that the Support Vector Machine method has better performance.
3. Based on the results of the analysis and the results of several classification model tests above, several conclusions can be obtained, including that reviews or comments on X/Twitter media aimed at Clash of Champions on the Ruangguru application, when viewed

from the table of preprocessing dataset results, show that the number of positive comments is more than the number of negative comments or reviews from X users. And indicates that X/Twitter users like the Clash of Champions event more.

ACKNOWLEDGMENT

REFERENCES

[1] H. Wahyono, "Pemanfaatan Teknologi Informasi dalam Penilaian Hasil Belajar pada Generasi Milenial di Era Revolusi Industri 4.0," *Proceeding Biol. Educ.*, vol. 3, no. 1, pp. 192–201, 2019, doi: 10.21009/pbe.3-1.23.

[2] T. Listyorini and A. Widodo, "Perancangan Mobile Learning Mata Kuliah Sistem Operasi Berbasis Android," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 3, no. 1, p. 25, 2013, doi: 10.24176/simet.v3i1.85.

[3] Statistica. (2024). *Leading countries based on number of X (formerly Twitter) users as of April 2024*. Statista Research Department.

[4] Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, *2*, 100003. https://doi.org/10.1016/j.nlp.2022.100003

[5] Chen, Y., Xie, H., & Wang, Y. (2020). Sentiment Analysis of Social Media on Business: Applications and Theories. *EEE Transactions on Computational Social Systems*, *7*(4), 923–934.

[6] Silvia, D., & Setiawati, I. (2019). Ruangguru sebagai Platform Edukasi Onlin. *Jurnal Ilmiah Pendidikan*, *12*(1), 45–55.

[7] Alizah, M. D., Nugroho, A., Radiyah, U., & Gata, W. (2020). Sentimen Analisis Terkait Lockdown pada Sosial Media Twitter. *Indonesian Journal on Software Engineering (IJSE)*, *6*(2), 223–229. https://doi.org/10.31294/ijse.v6i2.8991

[8] Prasetyo, D. B., & Hidayatullah, A. F. (2020). Identifikasi Dual Sentimen Terhadap Ulasan Objek Wisata di Daerah Istimewa Yogyakarta. *Automata*, *1*(1), 1–5.

[9] Sujadi, H. (2022). Analisis Sentimen Pengguna Media Sosial Twitter Terhadap Wabah Covid-19 Dengan Metode Naive Bayes Classifier Dan Support Vector Machine. *INFOTECH Journal*, *8*(1), 22–27. https://doi.org/10.31949/infotech.v8i1.1883

[10] Asro, Azizah, N., & Sudaryono. (2024, June 9). Public Sentiment Analysis on the 2024 Presidential Election Using Naive Bayes Classifier (NBC) and Support Vector Machine (SVM) On Social Media Data. *PROSIDING SEMINAR INTERNASIONAL*.

[11] Pamungkas, F. S., & Kharisudin, I. (2021b). Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter. *PRISMA, Prosiding Seminar Nasional Matematika*.

[12] Hidayat, F. N., & Sugiyono, S. (2023b). Analisis Sentimen Masyarakat Terhadap Perekrutan Pppk Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine. *Jurnal Sains Dan Teknologi*, *5*(2), 665–672. https://doi.org/10.55338/saintek.v5i2.1359