Analyzing Indonesian Football Sentiment Towards PSSI Performance Using Support Vector Machines

Faturrahman Hakim
Faculty of computer science and informatics,
University Amikom,
Yogyakarta, Indonesia
email:
faturrahman.hakim@students.amikom.ac.id

Yuli Astuti[1]
Departement of Informatic Mangement,
University Amikom,
Yogyakarta, Indonesia
email: yuli@amikom.ac.id

Football is a popular and widely engaged sport in Indonesia, attracting individuals across various age groups, including teenagers, adults, and children. The Indonesian Football Association (PSSI), established on April 19, 1930, originally named the All-Indonesian Football Association, is the governing body responsible for managing and overseeing football activities in the country. Despite its long history, PSSI has faced significant criticism for its perceived lack of professionalism in handling and managing Indonesian football. This discontent was notably amplified in the wake of the cancellation of the U-20 World Cup, leading to a surge of negative sentiments on social media platforms, particularly Twitter. This study aims to analyze public opinion regarding PSSI's performance. Public opinion, which emerges in response to various events, tends to be diverse due to the differing perspectives of individuals. The research focuses on assessing the balance between positive and negative sentiments towards PSSI's performance. By employing a comprehensive approach to sentiment analysis, including stages such as data preprocessing, labeling, modeling, and evaluation, this study provides a detailed examination of public sentiment. The methodology involves the application of the Support Vector Machine (SVM) algorithm across four tests with different data splits and the use of the SMOTE technique to address class imbalance. The findings reveal that the fourth test yielded the most effective results in sentiment classification, achieving an accuracy of 70.75%, precision of 67.16%, recall of 68.18%, and an F1 score of 67.66%

***KeyWords**: PSSI, Twitter, U-20 National Team, Sentiment Analysis, Support Vector Machine*

This Article was:
submited: 05-06-24
accepted: 14-06-24
publish on: 20-07-24

How to Cite:
F. Hakim, et al, "Analyzing Indonesian Football Sentiment Towards PSSI Performance Using Support Vector Machines", Journal of Intelligent Software Systems, Vol.3, No.1, 2024, pp.26–30, 10.26798/jiss.v3i1.1330

1 Introduction

Football is a familiar sport that is very popular among Indonesian people, both among teenagers, parents and children. Football is a team game, each team consists of eleven people, and one of them is a goalkeeper. Football is very embedded in Indonesian society, the public's concern and expectations for Indonesian footballers are very high. The All Indonesian Football Association or the abbreviation PSSI, is an organization responsible for managing football in Indonesia. PSSI was founded on April 19 1930 with the initial name of the All Indonesian Football Association, which is now chaired by Erick Thohir in February 2023.

Nowadays, the rapid development of technology makes the flow of information about Indonesian football easier to receive and share. Currently, mass media has experienced a sharp increase from conventional media such as radio, television and print media until now it has entered the era of new media or what is usually called digital media. In digital media the internet plays a very important role. Only with a smartphone can we get information quickly and flexibly, anytime and anywhere. Actual and factual information is now a priority for society. People can easily access news that is going viral through the use of social media. One example is FIFA canceling Indonesia's hosting of the U-20 World Cup, said new PSSI chairman Erick Thohir. Don't forget the Kanjuruhan Malang tragedy which resulted in many deaths. This news became a trending topic on almost all social media, including Instagram and Twitter.

Many people feel disappointed with PSSI's performance which is considered less professional in handling and managing Indonesian football. Apart from that, prospective young people who dream of playing in the U-20 World Cup feel that their hopes have been dashed, even though the opportunity seems to be right before their eyes. PSSI also received a lot of criticism from netizens, so it became a trending topic on Twitter with the topic "Cancellation of the U-20 World Cup." Many people argue that the cancellation of the U-20 World Cup in Indonesia had various impacts, such as the refusal of the Israeli national team to play in Indonesia on the grounds of supporting Palestine. Apart from that, there is also the view that this is related to the political interests of high-ranking state officials. A news report from Kompas TV also noted that the national team coach, Shin Tae Yong, was very disappointed with FIFA's decision to cancel the U-20 event in Indonesia. Coach Shin Tae Yong even said that his team had been preparing for more than 3 years. All of this is very unfortunate because it results in the players losing their hopes and dreams [1].

Regarding this discussion, researchers interested in this research found unstable conditions regarding PSSI's performance. Using the Support Vector Machine (SVM) algorithm method, the author

[1]Corresponding Author.

conducted research on Twitter Sentiment Analysis to assess public opinion regarding PSSI's performance. The result of this research is to provide information to PSSI, PT LIB (New Indonesia League), and Broadcasting to be able to find out the comparison of the number of positive and negative comments. From these results, we can evaluate the policies that must be implemented so that this incident does not happen again.

2 Research Method

2.1 Sentiment Analysis. Sentiment analysis is the process of collecting, processing and interpreting data to determine public sentiment or opinion towards a particular topic or entity. Sentiment analysis can be carried out through data obtained from social media, online forums, or other data sources. Sentiment analysis is a classification task that places text into positive or negative aspects [2].

2.2 Text Mining. Text mining, or what is often referred to as text mining or text data mining, is the process of extracting valuable information from unstructured text data. It involves analyzing texts to identify patterns, trends, entities, or relationships that exist in those texts. Text mining techniques are often used to extract knowledge from documents, articles, messages, or other text sources. In the Indonesian context, text mining can also be referred to as "text mining" or "text analysis" [3].

2.3 Prepossessing . Prepossessing is an important step in text analysis and text mining. The main goal of prepossessing is to clean and convert unstructured text into a more structured form, so that it can be further processed with data analysis or machine learning algorithms.

2.4 Classification . Classification is a technique in data mining and machine learning that is used to group or separate data into categories or classes based on certain characteristics or attributes. Classification techniques take data samples that have been labeled with a class (classification) and use patterns found in the data to predict the class of data that has not been labeled. This is a form of supervised learning where an algorithm learns from training data to make predictions about new data [4].

2.5 Support Vector Machine. Support Vector Machine (SVM) is a machine learning algorithm used in classification and regression. SVM works by finding the best hyperplane that can separate two classes of data with a maximum margin. In the context of sentiment analysis, SVM can be used to classify text or sentiment data into two main classes, namely positive and negative, based on the characteristics extracted from the text. Support Vector Machine (SVM) with a linear kernel is usually used for classification tasks. The linear SVM formula describes the relationship between features and class decisions. Following is the general formula (1).

Decision Function:

$$f(x) = sign(w \cdot x + b) \tag{1}$$

There is also a hyperplane equation that divides the feature space into two parts, representing two different classes. The following is the formula for the hyperplane equation

$$w \cdot x + b = 0 \tag{2}$$

Then there is a support function measuring how far a data point is from the separating hyperplane. Points that have a support function smaller than 1 are within the margin and can be considered as a support vector. The following is the formula for the function

$$Support\ (x) = \frac{|w \cdot x + b|}{||w||} \tag{3}$$

SVM also involves choosing the parameter C as a trade-off factor between achieving the maximum margin and reducing decision boundary violations. Optimizing parameters and handling imbalanced data are additional considerations in using SVM.

2.6 Python. Python programming language is a high-level programming language known for its easy-to-understand syntax and is used for developing various types of applications such as web development, data analysis, artificial intelligence, desktop software development, and many more. Python was designed with a focus on code readability, so it is suitable for both beginners and experienced developers [5].

2.7 Confusion matrix. Confusion matrix is a method that is generally used to calculate the level of accuracy in data mining. The confusion matrix contains information about the classifications that are correctly predicted by a classification system. There are three parameters to be calculated, namely accuracy, recall, and precision [6].

2.8 TF-IDF. TF-IDF (Term Frequency-Inverse Document Frequency) is an algorithm method used to give weight to words in text. This method is used to analyze text and determine how important a word is in the context of a particular document, In more detail:

TF (Term Frequency) measures how often a word appears in a particular document. The TF value will be high if the word appears many times in the document. Meanwhile, IDF (Inverse Document Frequency) is the inverse value of the number of documents containing that word in the collection of documents being analyzed. Words that appear in few documents will have a high IDF value, while words that are common and appear in many documents will have a low IDF value.

The word weight in TF-IDF is calculated by multiplying the TF value by the IDF value. The results of this multiplication produce a weight that shows how important a word in a particular document is in the context of the entire collection of documents [6]. The following is the general formula for calculating TF-IDF (4) and (5).

$$\text{IDF}(t) = \log\left(\frac{N}{\text{df}(t)}\right) \tag{4}$$

$$\text{TF.IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{5}$$

3 Result and Discussion

3.1 Research Flow. The research methodology used in research regarding Support Vector Machines for Analysis of Indonesian Football Sentiment Regarding PSSI Performance can be divided into four stages, namely data collection, data prepossessing, modeling, and evaluation or testing.

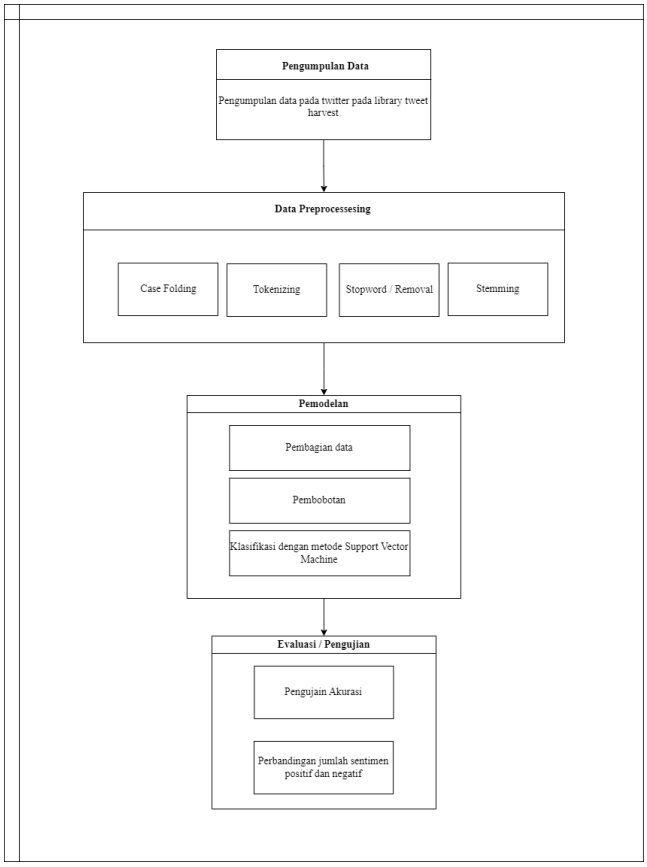

Fig. 1 Research Flow Diagram

3.2 Research Flow. The data used in this research is tweets data containing the keywords "PSSI" and "U-20 National Team". The data collection stage in this research is by using the tweet harvest library. Tweets data was taken from 13 August 2023 to 15 November 2023. The library used for data collection was tweet harvest. Produces approximately 1600 data. The results of data collection can be seen in Figure 2 After the data is collected, it continues with prepossessing the data. This study utilized Support Vector Machine (SVM) to classify the sentiment of tweets related to PSSI's performance. Overall, the SVM model achieved an accuracy of 70.75%, with a precision of 67.16%, recall of 68.18%, and an F1 score of 67.66%. These metrics indicate the model's effectiveness in identifying sentiment, although there is still room for improvement, particularly in enhancing precision and recall.

	id_str	full_text	username	tweet_url
0	1,71153E+18	Malu woy! @PSSI	gutheutsub	https://twitter.com/gutheutsub/status/17115299...
1	1,71153E+18	Kompetisi tingkat kampung bisa menghasilkan se...	YuliSugiyanto06	https://twitter.com/YuliSugiyanto06/status/171...
2	1,71152E+18	@Rheza_pradita dari persiapan dan ujicoba aja ...	semi_sjw	https://twitter.com/semi_sjw/status/1711523319...
3	1,71152E+18	@nikosuke_1 pssi!	0107ai_	https://twitter.com/0107ai_/status/17115191454...
4	1,71152E+18	Ketua PSSI, Erick Thohir, membantah narasi yan...	bolacomID	https://twitter.com/bolacomID/status/171151689...
...	...	...	...	...
1617	1,69063E+18	Eks kiper andalan Shin Tae-yong di Timnas U-20...	BolaSportcom	https://twitter.com/BolaSportcom/status/169063...
1618	1,69063E+18	Eks kiper andalan Shin Tae-yong di Timnas U-20...	tribunSUPERBALL	https://twitter.com/tribunSUPERBALL/status/169...
1619	1,69059E+18	Thomas Doll sudah menyiapkan kiper pengganti u...	BolaSportcom	https://twitter.com/BolaSportcom/status/169059...
1620	1,69059E+18	Thomas Doll sudah menyiapkan kiper pengganti u...	tribunSUPERBALL	https://twitter.com/tribunSUPERBALL/status/169...
1621	1,69055E+18	@Hudakeyy Lonceng kyk bola, mgkn karna tolak t...	PaberSitorus1	https://twitter.com/PaberSitorus1/status/16905...

1622 rows × 12 columns

Fig. 2 Research Datasets

3.3 Labeling. The data labeling process was carried out manually using Microsoft Excel. Each tweet data is given a positive or negative label. From the initial data collection process of 1622, after going through the text prepossessing stage, the amount of data was reduced to 1466. As a result of 1466 data labels, 713 positive labels and 753 negative labels were produced. The results of data labeling can be seen in Figure 3 below.

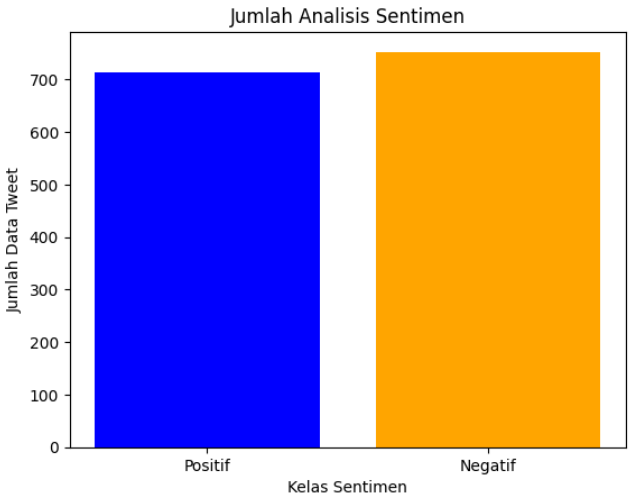

Fig. 3 Dataset Labeling Number Display

3.4 Modeling. The data division process is carried out to divide the dataset into training data and test data. The comparison of training data and test data in this research is 70% training data and 30% test data. Based on the number of training data: 1026 and the number of test data: 440, the researcher added a data division of 80%: 20% with the number of training data: 1172 and test data: 294. The library used to divide the data was the Sklearn library.

TF-IDF (Term Frequency-Inverse Document Frequency) weighting is a method used in text processing and data modeling to evaluate how important a word in a document is in a larger collection of documents. TF-IDF weighting helps evaluate words that appear in a document in a way that takes into account how often they appear in a particular document and how unique they are across a collection of documents.

3.5 Testing models. Testing is carried out to assess the level of accuracy of the classification model based on test data. This testing process involves comparing the classification results of the test data with the classification results carried out manually by the author. The method used in this test is confusion matrix. In the evaluation using the confusion matrix, the accuracy value, precision value and recall value are measured.

3.6 Comparison Across Tests. To gain deeper insights into how data distribution and parameter settings affect model performance, we conducted four different tests with varying data splits and the application of SMOTE techniques to address data imbalance.

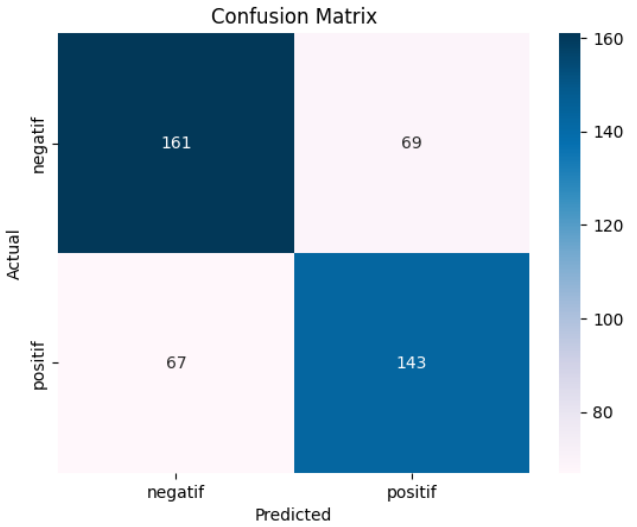

Fig. 4 First test of the confusion matrix

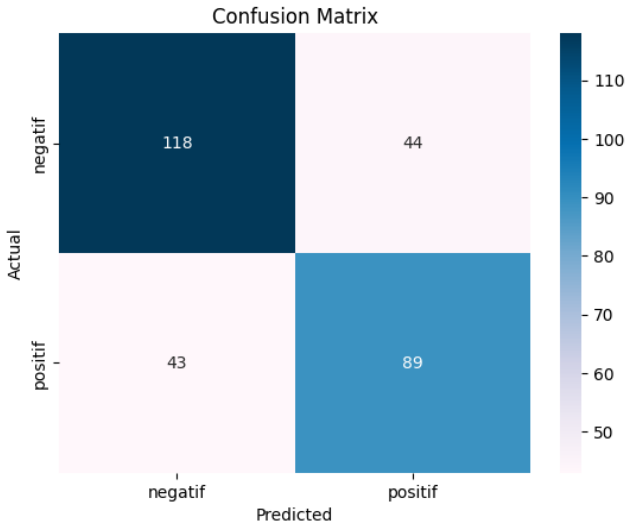

Fig. 6 Third test of confusion matrix

Test 1. In the first test, the data was split with a proportion of 70% for training and 30% for testing. This test resulted in an accuracy of 68.20%, precision of 65.00%, recall of 66.25%, and an F1 score of 65.60%.

Test 3. In the third test, the proportion of training data was increased to 80%, with 20% of the data used for testing. The results from this configuration showed a slight decrease in all metrics: accuracy of 68.00%, precision of 64.50%, recall of 65.75%, and an F1 score of 65.10%.

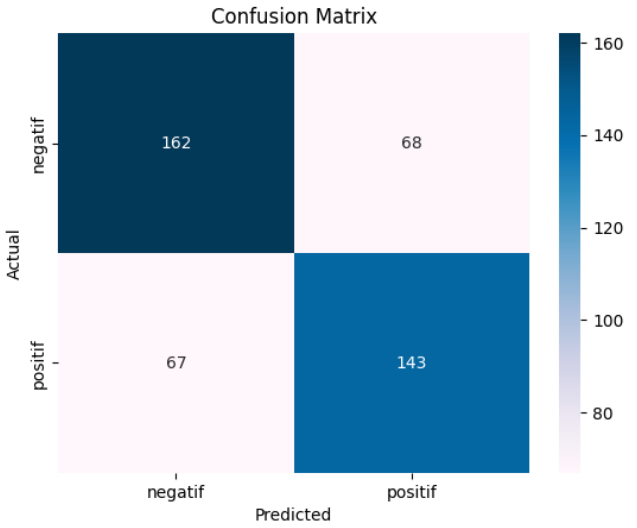

Fig. 5 The second test uses a smote confusion matrix

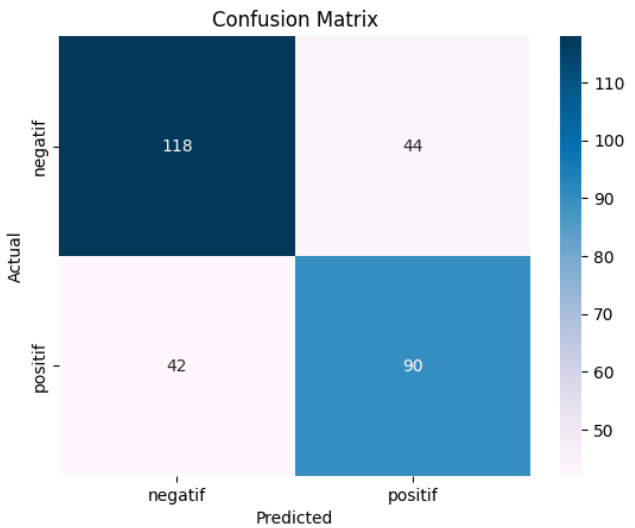

Fig. 7 The fourth test uses a smote confusion matrix

Test 2. For the second test, we applied the SMOTE technique to the training data to address class imbalance. The data split remained the same, but with adjusted class distribution. This improved the accuracy to 69.50%, precision to 66.50%, recall to 67.75%, and the F1 score to 67.10%.

Test 4. In the fourth test, SMOTE was also applied with a data split of 80% for training and 20% for testing. This resulted in significant improvements across all metrics: accuracy of 70.75%, precision of 67.16%, recall of 68.18%, and an F1 score of 67.66%.

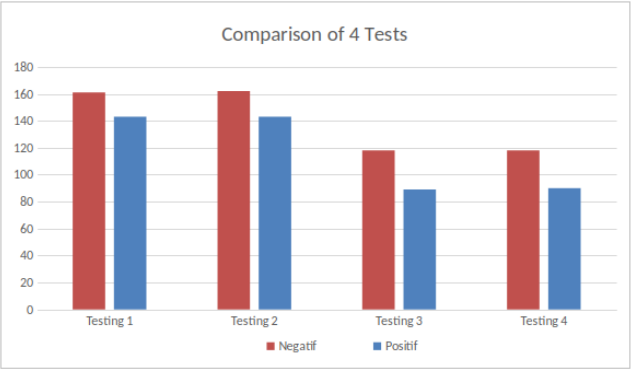

Fig. 8 The fourth test uses a smote confusion matrix

4 Conclusion

Based on the results of this research, evaluating sentiment regarding PSSI's performance and the cancellation of the U-20 National Team World Cup in 1622 data after going through a series of stages, including text preprocessing, labeling, weighting, and implementation of the Support Vector Machine (SVM) method. The results show that from the total data, 713 comments were categorized as positive, while 753 comments were categorized as negative. This research uses the Support Vector Machine algorithm method for four tests with different data divisions and also applies the SMOTE technique. Based on the performance results obtained, it was found that in the fourth test it was superior in classifying sentiment, with accuracy results of 70.75%, precision 67.16%, recall 68.18%, and f1-score 67.66%. This test performance evaluation provides insight into the effectiveness of SVM in classifying sentiment related to PSSI performance. Based on the performance results obtained, it was found that in the fourth test it was superior in carrying out sentiment classification.

References

[1] Rahayuu, S., 2023, "Shin Tae-yong Sakit Hati Piala Dunia U20 Batal Digelar di Indonesia," Accessed: Mar. 30, 2023, https://bola.kompas.com/read/2023/03/30/22000068/shin-tae-yong-sakit-hati-piala-dunia-u20-batal-digelar-di-indonesia

[2] Hendrastuty, N., Isnain, A. R., and Rahmadhani, A. Y., 2021, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," http://situs.com

[3] Siregar, D., Ladayya, F., Albaqi, N. Z., and Wardana, B. M., 2023, "Penerapan Metode Support Vector Machines (SVM) dan Metode Naïve Bayes Classifier (NBC) dalam Analisis Sentimen Publik terhadap Konsep Child-free di Media Sosial Twitter," Jurnal Statistika dan Aplikasinya, **7**(1).

[4] Wibawa, M. G. A. P. M. F. A. A. P., 2018, "Metode-metode Klasifikasi," *Pros. Semin. Ilmu Komput. dan Teknol. Inf.*, 1st ed., Vol. 3.

[5] Melinda, R. N., Ningrum, L. M., Suryabrata, I. B., Bayu, G. S., Dwipa, A., and Sukoco, T. P., 2021, "Program Perhitungan RAB Pekerjaan Struktur Baja (WF BEAM) Menggunakan Bahasa Python," TIERS Information Technology Journal, **2**(1), pp. 31–38.

[6] Salam, R. R., Jamil, M. F., and Ibrahim, Y., 2023, "Analisis Sentimen Terhadap Bantuan Langsung Tunai (BLT) Bahan Bakar Minyak (BBM) Menggunakan Support Vector Machine," .

[7] Witanti, A., Wates-Jogjakarta, B. Y. J. R., Sedayu, K., Bantul, K., and Yogyakartalamat, D. I., 2022, "ANALISIS SENTIMEN MASYARAKAT TERHADAP VAKSINASI COVID-19 PADA MEDIA SOSIAL TWITTER MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)," Jurnal Sistem Informasi dan Informatika (Simika), **5**, pp. 2622–6901.

[8] Wicaksono, A. S., 2019, "ANALISIS SENTIMEN SEPAKBOLA INDONESIA MENGGUNAKAN SUPPORT VECTOR MACHINE," .

[9] Asshiddiqi, M. F. and Lhaksmana, K. M., 2020, "Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI," .

[10] Prajamukti, J. M. M. S. R., 2021, "KLASIFIKASI DAN ANALISIS SENTIMEN PADA DATA," .

[11] Gifari, O. I., Adha, M., Hendrawan, I. R., Freddy, F., and Durrand, S., 2022, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," JIFOTECH (JOURNAL OF INFORMATION TECHNOLOGY), **2**(1).

[12] Tineges, R., Triayudi, A., and Sholihati, I. D., 2020, "Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM)," JURNAL MEDIA INFORMATIKA BUDIDARMA, **4**(3), p. 650.

[13] Zai, C. and Komputer, T., 2022, "IMPLEMENTASI DATA MINING SEBAGAI PENGOLAHAN DATA," .

[14] Pakpahan, S. R. M., 2019, "Analisis Sentimen Tentang Opini Performa Klub Sepak Bola Pada Dokumen Twitter Menggunakan Support Vector Machine Dengan Perbaikan Kata Tidak Baku," http://j-ptiik.ub.ac.id

[15] Sulaiman, J. K., di Indonesia Herdianti Darwis, P. A., Wanaspati, N., Anraeni, S., and Abstrak, I. A., 2023, "Support Vector Machine untuk Analisis Sentimen Masyarakat Terhadap Penggunaan Antibiotik di Indonesia," .