[1]st Proceeding of International Conference on Science and Technology UISU (ICST)
"Contribution of Engineering Disciplines to Achieving Sustainable Development Goals Research Realizing the Vision of Golden Indonesia 2045"
ISSN: 3063-7929 (Online)
DOI: https://doi.org/10.30743/icst

SVM Optimization with Kernel Function for Sentiment Analysis on Social Media twitter (X) in AFC U23 Asian Cup Case Study

Hardi Prasetya[1*], Zakarias Situmorang[2], Rika Rosnelly[3]
[1]Master of Computer Science, Faculty of Engineering and Computer Science, Universitas Potensi Utama, Medan. Indonesia
[2]Department of Computer Science, Faculty of Information Systems, Universitas Katolik Santo Thomas, Medan. Indonesia
[3]Department of Computer Science, Faculty of Engineering and Computer Science, Universitas Potensi Utama, Medan. Indonesia
[*]Corresponding author: hardi.prasetya@gmail.com

Abstract - This study aims to optimize the performance of the Support Vector Machine (SVM) in sentiment analysis on social media by using various kernel functions, namely linear, polynomial, and Radial Basis Function (RBF). The case study taken was a conversation related to the AFC Asian Cup U-23 taken from social media platforms. The data used in this study included three classes of sentiment: positive, neutral, and negative. The experimental results show that the linear kernel achieves the highest accuracy of 93.55% with an F1-score of 0.9296. The RBF kernel shows almost comparable performance with an accuracy of 90.05% and an F1-score of 0.8820. In contrast, the polynomial kernel showed lower performance with an accuracy of 80.65% and an F1-score of 0.7346. The results of the analysis using the confusion matrix show that linear kernels and RBF are more effective in classifying neutral and positive sentiment than polynomial kernels. This study confirms that the right selection of kernels in SVM greatly affects the accuracy and effectiveness of sentiment analysis. Linear kernels and RBFs have proven superior in handling complex sentiment analysis datasets, such as those related to the AFC U-23 Asian Cup. These findings can be used as a basis for further development in sentiment analysis applications across various domains.

Keywords : Sentiment Analysis, SVM, Linier Kernel, Polynomial Kernel, RBF Kernel, Social.

1. INTRODUCTION

In today's digital age, social media has become a major platform for people to express their opinions and feelings about various events, including major sporting events such as the AFC Asian Cup U-23. The diversity of opinions expressed on this platform creates fertile ground for sentiment analysis, which plays an important role in understanding public perception. Sentiment analysis utilizes Natural Language Processing (NLP) techniques and machine learning to classify texts based on the sentiments expressed, i.e., positive, negative, or neutral. Sentiment analysis has become an increasingly important topic in today's digital age, where unstructured text data, such as reviews, comments, and social media, is increasingly accessible and abundant. Sentiment analysis is a branch of *text mining* and *Natural Language Processing* (NLP) that focuses on identifying and classifying opinions, emotions, and attitudes expressed in texts that can be used to extract and classify opinions, emotions, and attitudes from texts originating from social media platforms, product reviews, news articles, and other data sources. Sentiment analysis has become an increasingly important topic in today's digital age, where unstructured text data, such as reviews, comments, and social media, is increasingly accessible and abundant. Sentiment analysis is a branch of *text mining* and *Natural Language Processing* (NLP) that focuses on identifying and classifying opinions, emotions, and attitudes expressed in texts that can be used to extract and classify opinions, emotions, and attitudes from texts originating from social media platforms, product reviews, news articles, and other data sources. In conducting sentiment analysis, *text mining* and *natural language processing* (NLP) techniques play an important role. Through *text mining*, unstructured text data can be processed and extracted to identify relevant patterns, trends, and insights. Meanwhile, NLP can be used to understand and analyze the natural language used by people in providing their responses and comments. In conducting sentiment analysis, *machine learning* is becoming an increasingly popular approach. *Machine learning* algorithms, both *supervised learning*, *unsupervised learning*, and *reinforcement learning*, can be used to classify and predict the sentiment contained in text. Some of the algorithms that are often used in sentiment analysis include *Support Vector Machine (SVM), Naive Bayes, and Logistic Regression.* Support Vector Machine (SVM) is one of the most effective machine learning algorithms for classification tasks, including sentiment analysis. SVM performance is greatly influenced by the selection of kernel functions, which can transform input data into higher feature spaces. Common kernel functions used in SVM include linear, polynomial, and Radial Basis Function (RBF) kernels. Each kernel has its own characteristics and advantages, which can affect the results of sentiment analysis. This study aims to optimize SVM's performance in sentiment analysis on social media related to the AFC Asian Cup U-23 by comparing the effectiveness of three different kernel functions: linear, polynomial, and RBF. This study is expected to provide insight into the selection of the most effective kernel functions in the context of sentiment analysis at major sporting events, as well as contribute to the development of more sophisticated and accurate analysis methods.

Previous research on the application of *the Support Vector Machine (SVM) algorithm* has been conducted by Saputra, A. (2023). Comparison of the Naïve Bayes Classifier and Support Vector Machine methods for Twitter user sentiment analysis regarding the 2022 FIFA World Cup. From these results, it can be concluded that the performance of the SVM method is better than the NBC method in finding Accuracy with an average of 85%. Aulia,

ISSN: 3063-7929 (Online)
DOI: https://doi.org/10.30743/icst

et al. (2021), Comparison of Kernel Support Vector Machine (SVM) in the Application of Covid-19 Vaccination Sentiment Analysis. The research data is to find the best kernels among linear, sigmoid, polynomial, and RBF kernels. Dewi, et al. (2023). Comparison of the implementation of the Smote method on the Support Vector Machine (SVM) algorithm in the analysis of public opinion sentiment about Mixue. The Support Vector Machine (SVM) algorithm based on SMOTE with the evaluation results can be concluded that SMOTE has an effect on increasing accuracy and precision values, but there is a decrease in recall and F1-Score values. Fitriyah, et al. (2020). Gojek sentiment analysis on Twitter social media with a Support Vector Machine (SVM) classification. The kernels used are linear kernels and RBF kernels. Nasution, et al. (2019). Comparison of accuracy and runtime of K-NN and SVM algorithms in twitter sentiment analysis. The results of the accuracy calculation show that the Support Vector Machine method is superior with a value of 89.70% without K-Fold Cross Validation and 88.76% with K-Fold Cross Validation. Meanwhile, in the calculation of the process time, the K-Nearest Neighbor method is superior with a process time of 0.0160s without K-Fold Cross Validation and 0.1505s with K-Fold Cross Validation. This research will use *machine learning* algorithms, specifically *the Support Vector Machine* (SVM), to classify public sentiment towards the AFC Asian Cup U23. SVM is one of the *supervised learning* algorithms that is very effective in classifying data with high accuracy. With the growing amount of textual data related to football competitions, such as comments on social media, news articles, and discussion forums, sentiment analysis can be a very useful tool for understanding how people respond to and give their assessment of the AFC Asian Cup U23. By understanding the important role of kernel function selection in SVM, this study will explore the extent to which these kernels can improve the performance of the model in classifying public sentiment, and how these results can be applied in practical scenarios to monitor and analyze public perceptions of major events.

2. METHOD

This research method adopts the Knowledge Discovery in Databases (KDD) approach, which is a systematic methodology to obtain knowledge from relevant, useful, and understandable data. The KDD process involves several interconnected stages, as shown in Figure 1.

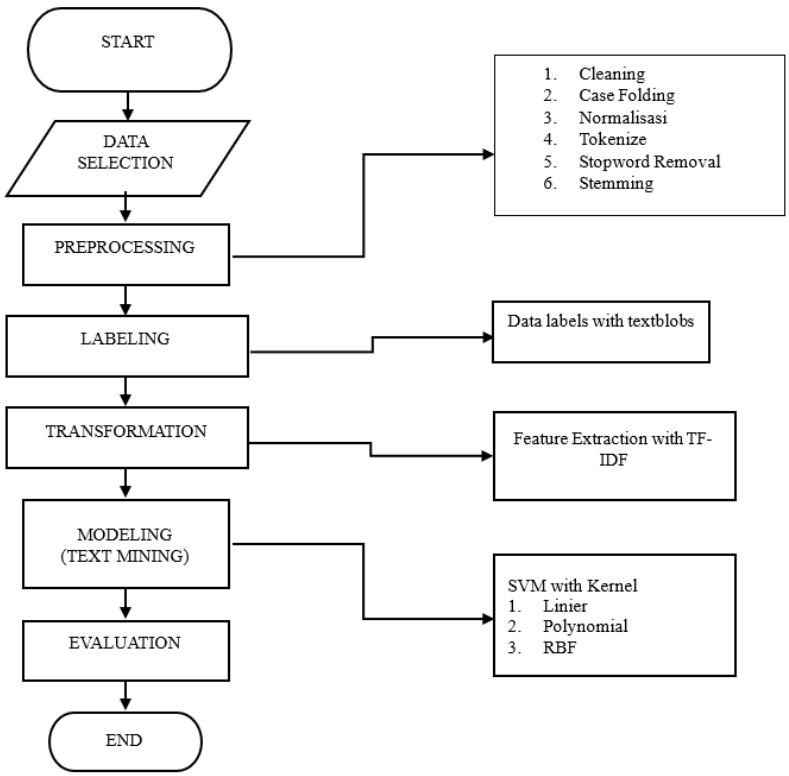

Figure 1. Research Stages

2.1 Data Selection

ISSN: 3063-7929 (Online)
DOI: https://doi.org/10.30743/icst

Data selection is the initial stage to collect the data needed for the research. The data used in this study comes from the Twitter platform, collecting tweet data on Twitter social media by crawling data. The time for data collection is during the 2024 AFC Asian Cup U23 Cup. The keyword used to collect data is the AFC Asian Cup U23.

2.2 Preprocessing

Preprocessing is an important step in preparing data for sentiment analysis. In this study, we used six main steps in preprocessing tweet data. Here is an explanation of each of these steps:

a. Cleaning

The cleaning process aims to remove irrelevant elements from the text of the tweet. This includes the removal:

1. Punctuation, such as periods, commas, and other symbols.
2. The URL contained in the tweet, as it does not contribute to sentiment analysis.
3. Emoticons and special characters that don't provide additional information.

	full_text	cleaning
0	Kapan Arhan Pratama Debut dengan Suwon Fc?? #A...	Kapan Arhan Pratama Debut dengan Suwon Fc AFCU...
1	Setelah afcu23 kemaren lagi seneng ngikutin an...	Setelah afcu kemaren lagi seneng ngikutin anak...
2	BOCORAN POLA GACOR IBETSLOT HARI INI LINK DAFT...	BOCORAN POLA GACOR IBETSLOT HARI INI LINK DAFT...
3	IBETSLOT SITUS SLOT GACOR MUDAH MENANG SAAT IN...	IBETSLOT SITUS SLOT GACOR MUDAH MENANG SAAT IN...
4	GAME GACOR IBETSLOT HARI INI 22 MEI 2024 LINK ...	GAME GACOR IBETSLOT HARI INI MEI LINK DAFTAR...
...	...	...

Figure 2. Results of the Data Cleaning Process

b. Case Folding

Case folding is the process of converting all letters in text to lowercase. It is important to ensure that the same words, regardless of capitalization, are recognized as the same entity. For example, "Cup" and "Cup" will be considered identical after this process.

cleaning	case_folding
Kapan Arhan Pratama Debut dengan Suwon Fc AFCU...	kapan arhan pratama debut dengan suwon fc afcu...
Setelah afcu kemaren lagi seneng ngikutin anak...	setelah afcu kemaren lagi seneng ngikutin anak...
BOCORAN POLA GACOR IBETSLOT HARI INI LINK DAFT...	bocoran pola gacor ibetslot hari ini link daft...
IBETSLOT SITUS SLOT GACOR MUDAH MENANG SAAT IN...	ibetslot situs slot gacor mudah menang saat in...
GAME GACOR IBETSLOT HARI INI MEI LINK DAFTAR...	game gacor ibetslot hari ini mei link daftar...
...	...

Figure 3. Case Folding Process Results

c. Normalization

Normalization aims to simplify the form of words. This process includes:

1. Remove variations of words that have similar meanings, such as unnecessary suffixes or prefixes.
2. Changing words that have different forms to their basic forms to maintain consistency in analysis.

ISSN: 3063-7929 (Online)
DOI: https://doi.org/10.30743/icst

case_folding	normalized
kapan arhan pratama debut dengan suwon fc afcu...	kapan arhan pratama debut dengan suwon fc afc_...
setelah afcu kemaren lagi seneng ngikutin anak...	setelah afc_u kemaren lagi seneng ngikutin ana...
bocoran pola gacor ibetslot hari ini link daft...	bocoran pola gacor ibetslot hari ini link daft...
ibetslot situs slot gacor mudah menang saat in...	ibetslot situs slot gacor mudah menang saat in...
game gacor ibetslot hari ini mei link daftar...	game gacor ibetslot hari ini mei link daftar r...
...	...

Figure 4. Results of the Normalization Process

d. Tokenize

Tokenization is the process of breaking down text into smaller units, usually in the form of words or phrases. By tokenizing, we can analyze each word in the tweet separately. For example, the sentence "Indonesia U23 national team wins!" will be broken down into tokens: ["National team", "U23", "Indonesia", "win"].

normalized	tokenize
kapan arhan pratama debut dengan suwon fc afc_...	[kapan, arhan, pratama, debut, dengan, suwon, ...
setelah afc_u kemaren lagi seneng ngikutin ana...	[setelah, afc_u, kemaren, lagi, seneng, ngikut...
bocoran pola gacor ibetslot hari ini link daft...	[bocoran, pola, gacor, ibetslot, hari, ini, li...
ibetslot situs slot gacor mudah menang saat in...	[ibetslot, situs, slot, gacor, mudah, menang, ...
game gacor ibetslot hari ini mei link daftar r...	[game, gacor, ibetslot, hari, ini, mei, link, ...

Figure 5. Results of the Tokenization process

e. Stopword Removal

Stopword removal is a step to remove common words that do not give significant meaning in the context of analysis, such as "and", "in", "to", and other connecting words. By removing the stopwords, we can focus on the more important and relevant words for sentiment analysis.

tokenize	stopword removal
[kapan, arhan, pratama, debut, dengan, suwon, ...	[arhan, pratama, debut, suwon, fc, afc_u, arha...
[setelah, afc_u, kemaren, lagi, seneng, ngikut...	[afc_u, kemaren, seneng, ngikutin, anaknya, co...
[bocoran, pola, gacor, ibetslot, hari, ini, li...	[bocoran, pola, gacor, ibetslot, link, daftar,...
[ibetslot, situs, slot, gacor, mudah, menang, ...	[ibetslot, situs, slot, gacor, mudah, menang, ...
[game, gacor, ibetslot, hari, ini, mei, link, ...	[game, gacor, ibetslot, mei, link, daftar, rtp...
...	...

Figure 6. Results of the stopword removal process

ISSN: 3063-7929 (Online)
DOI: https://doi.org/10.30743/icst

f. Stemming

Stemming is the process of reducing a word to its basic form. For example, the words "play", "play", and "play" will be changed to "play". This process helps to bring together variations of words that have the same meaning, thereby increasing the accuracy in sentiment analysis.

stopword removal	stemming_data
[arhan, pratama, debut, suwon, fc, afcu, arhan...	arhan pratama debut suwon fc afcu arhanpratama...
[afcu, kemaren, seneng, ngikutin, anaknya, coa...	afcu kemaren neng ngikutin anak coach nopa ana...
[bocoran, pola, gacor, ibetslot, link, daftar,...	bocor pola gacor ibetslot link daftar rtpibets...
[ibetslot, situs, slot, gacor, mudah, menang, ...	ibetslot situs slot gacor mudah menang link da...
[game, gacor, ibetslot, mei, link, daftar, rtp...	game gacor ibetslot mei link daftar rtpibetslo...
...	...

Figure 7. Results of the stemming process

2.3 Labeling

Labeling is an important step in sentiment analysis that aims to assign labels or categories to text data based on the sentiment it contains. In this study, we used the **TextBlob** library to perform a labeling process on previously processed tweets. TextBlob is a simple and effective Python library for natural language processing (NLP). One of its main features is its ability to analyze sentiment from text. Using a TextBlob, we can determine whether the sentiment in a tweet is positive, negative, or neutral.

2.4 Transformation

Transformation The selection feature is the process of converting categorical data into numerical data. This time it will be used using TF-IDF, the frequency requirement of the inverse frequency document (TF-IDF) is a weighting process where the word will be extracted into a value form file.

2.5 Modelling

The data from the transformation is then modeled using the SVM algorithm. The following modeling process is also preprocessing and transformation. using the Python programming language. A total of 3719 data will be divided into training data and test data, classified into Positive, Negative, and Neutral sentiments. In this study, 3 modeling process scenarios will be used using three different kernels, namely polynomial, linear, and RBF kernels.

2.6 Evaluation

The evaluation stage is the last stage in the KDD process, the evaluation is carried out to check the results of the model that has been built contrary to existing facts or not. The results of the research that have been carried out need to be tested to determine the accuracy of the research results that have been carried out. The data testing method in this study will calculate the value of accuracy, precision, recall and f-measure.

3. RESULTS AND DISCUSSION

3.1 Data description

The total data collected through social media Twitter is 3719 in Indonesian. The data used for preprocessing has a classification comparison as shown in figure 8.

ISSN: 3063-7929 (Online)
DOI: https://doi.org/10.30743/icst

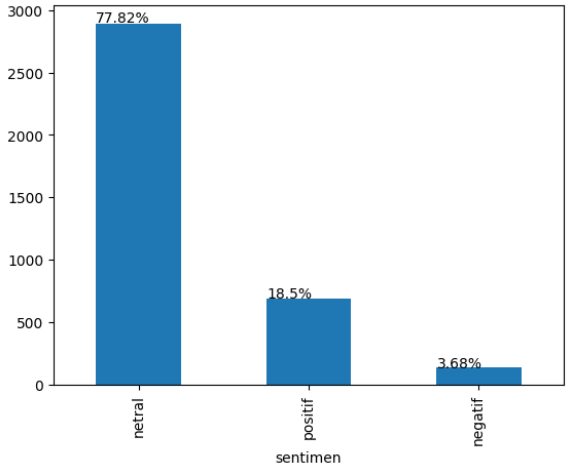

Figure 8. Comparison chart of sentiment numbers

3.2 Discussion

The data that has been collected is then separated into training data and test data. The data will then be classified into three sentiments, namely negative, positive and neutral using the SVM Algorithm. Because the data is non-linear, it is necessary to use the karnel function. This study will try to compare the results of three SVM kernels, namely polynomial, linear, and RBF kernels. Then the test calculates the accuracy value of the sentiment classification. The result is as shown in Table 1.

Table 1. Accuracy Results

Linear Kernel		**Polynomial Kernel**		**RBF Kernel**	
Accuracy:	0.9355	**Accuracy**:	0.8065	**Accuracy**:	0.9005
Precision:		**Precision**:		**Precision**:	
Negatif:	0.62	Negatif:	0.00	Negatif:	0.50
Netral:	0.93	Netral:	0.80	Netral:	0.89
Positif:	0.98	Positif:	1.00	Positif:	1.00
Recall:		**Recall**:		**Recall**:	
Negatif:	0.35	Negatif:	0.00	Negatif:	0.04
Netral:	1.00	Netral:	1.00	Netral:	1.00
Positif:	0.75	Positif:	0.10	Positif:	0.61
F1-score:		**F1-score**:		**F1-score**:	
Negatif:	0.44	Negatif:	0.00	Negatif:	0.08
Netral:	0.97	Netral:	0.89	Netral:	0.94
Positif:	0.85	Positif:	0.18	Positif:	0.76

4. CONLUSION

This study succeeded in optimizing the performance of the Support Vector Machine (SVM) in sentiment analysis on social media related to the AFC U-23 Asian Cup by comparing three kernel functions: linear, polynomial, and Radial Basis Function (RBF). The experimental results show that the linear kernel provides the highest accuracy of 93.55% with an F1-score of 0.9296, followed by the RBF kernel with an accuracy of 90.05% and an F1-score of 0.8820. In contrast, polynomial kernels show lower performance with an accuracy of 80.65% and an F1-score of 0.7346.

Analysis using a confusion matrix indicates that linear and RBF kernels are more effective in classifying neutral and positive sentiment than polynomial kernels, which show poor results especially in the negative class. These findings confirm the importance of proper kernel selection in SVM, which significantly affects the accuracy and effectiveness of sentiment analysis.

As such, this research provides valuable insights for the development of more sophisticated and accurate sentiment analysis methods across various domains, as well as highlighting the potential of SVMs in understanding public perception of major events such as the AFC U-23 Asian Cup. The results of this study are expected to be the basis for further research in the application of sentiment analysis on social media.

ISSN: 3063-7929 (Online)
DOI: https://doi.org/10.30743/icst

REFERENCES

[1] Abdurrahman, G. (2023). Klasifikasi Kanker Payudara Menggunakan Algoritma SVM dengan Kernel RBF, Linier, dan Sigmoid. JUSTIFY: Jurnal Sistem Informasi Ibrahimy, 2(1), 74-80.

[2] Ansori, Y., & Holle, K. F. H. (2022). Perbandingan metode machine learning dalam analisis sentimen Twitter. JUSTIN (Jurnal Sistem dan Teknologi Informasi), 10(4), 429-434.

[3] Arsi, P., & Waluyo, R. (2021). Analisis sentimen wacana pemindahan ibu kota Indonesia menggunakan algoritma Support Vector Machine (SVM). J. Teknol. Inf. dan Ilmu Komput, 8(1), 147.

[4] Aulia, T. M. P., Arifin, N., & Mayasari, R. (2021). Perbandingan Kernel Support Vector Machine (Svm) Dalam Penerapan Analisis Sentimen Vaksinisasi Covid-19. SINTECH (Science and Information Technology) Journal, 4(2), 139-145.

[5] Darwis, D., Pratiwi, E. S., & Pasaribu, A. F. O. (2020). Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. Jurnal Ilmiah Edutic: Pendidikan dan Informatika, 7(1), 1-11.

[6] Dewi, T. A., & Mailoa, E. (2023). Perbandingan Implementasi Metode Smote Pada Algoritma Support Vector Machine (Svm) Dalam Analisis Sentimen Opini Masyarakat Tentang Mixue. Jurnal Indonesia Manajemen Informatika dan Komunikasi, 4(3), 849-855.

[7] Fitriyah, N., Warsito, B., & Di Asih, I. M. (2020). Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (SVM). Jurnal Gaussian, 9(3), 376-390.

[8] Hartono, F., & Novitasari, D. (2019). Overfitting and Underfitting in Machine Learning. Prosiding Konferensi Ilmiah Nasional Teknologi Informasi, 9, 23-32.

[9] Husada, H. C., & Paramita, A. S. (2021). Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM). Teknika, 10(1), 18-26.

[10] Idris, I. S. K., Mustofa, Y. A., & Salihi, I. A. (2023). Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM). Jambura Journal of Electrical and Electronics Engineering, 5(1), 32-35.

[11] Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed. draft). Pearson.

[12] Kelvin, K., Banjarnahor, J., Nababan, M. N., & Sinurat, S. H. (2022). Analisis perbandingan sentimen Corona Virus Disease-2019 (Covid19) pada Twitter Menggunakan Metode Logistic Regression Dan Support Vector Machine (SVM). Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA), 5(2), 47-52.

[13] Noviana, R., & Rasal, I. (2023). Penerapan Algoritma Naive Bayes Dan Svm Untuk Analisis Sentimen Boy Band Bts Pada Media Sosial Twitter. Jurnal Teknik dan Science, 2(2), 51-60.

[14] Omnicore. (2023). Twitter by the Numbers: Stats, Demographics & Fun Facts. https://www.omnicoreagency.com/twitter-statistics/

[15] Rahardjo, D., & Simatupang, N. (2021). Algorithms in Machine Learning. Jurnal Teknologi Informasi, 18(2), 67-82.

[16] Saputra, A., & Wibowo, B. (2022). Data and Features in Machine Learning. Jurnal Ilmiah Teknologi Informasi, 12(3), 45-57.

[17] Tineges, R., Triayudi, A., & Sholihati, I. D. (2020). Analisis sentimen terhadap layanan indihome berdasarkan twitter dengan metode klasifikasi support vector machine (SVM). Jurnal Media Informatika Budidarma, 4(3), 650-658.

[18] Utami, D. S., & Erfina, A. (2021, September). Analisis Sentimen Pinjaman Online di Twitter Menggunakan Algoritma Support Vector Machine (SVM). In Prosiding Seminar Nasional Sistem Informasi dan Manajemen Informatika Universitas Nusa Putra (Vol. 1, pp. 299-305).

[19] Wijaya, L., & Suryadi, H. (2020). Supervised and Unsupervised Learning. Buku Teks Machine Learning, Penerbit Informatika, Bandung.