Jurnal Media Teknik Elektro dan Komputer
Vol. 02, No.02 (July - December, 2025)
E-ISSN: 3063-2927

Implementation of SVM Algorithm in Online Gambling Comment Classification using RapidMiner

Priti Rindi Artika[1], Khafifah Dwi Meilianasari [2], Ali Ikhwan[3]

[1]Universitas Islam Negeri Sumatera Utara; pritirindyartika@gmail.com
[2]STIKOM Tunas Bangsa; kmeilianasari@gmail.com
[3]Universiti Malaysia Perlis; aliikhwan@studentmail.unimap.edu.my

ABSTRACT

The spread of negative comments containing elements of online gambling on digital platforms is increasingly affecting users. To address this issue, this study implemented a Support Vector Machine (SVM) algorithm to classify comments into two categories: those containing elements of online gambling and those without. The classification process was carried out using RapidMiner software, which allows data processing without the need for extensive coding. The dataset used was obtained from the Kaggle website and consisted of 8,442 comments. The data underwent preprocessing stages such as tokenization, normalization, and stopword removal. The SVM model was drilled and evaluated using cross-validation and evaluation metrics, with an accuracy of 97.91%, precision of 96.94%, recall of 99.81%, and an F1-score of 98.45%. The results showed that the SVM model achieved an accuracy of 97.91%, with high precision and recall across both classes. This demonstrates that the SVM algorithm is effective and efficient in automatically detecting comments containing elements of online gambling and is suitable for implementation as a content moderation system on digital platforms.

Keywords: *Text Classification, Online Gambling Comments, Support Vector Machine, RapidMiner*

Corresponding Author:
Priti Rindi Artika
Universitas Islam Negeri Sumatera Utara; pritirindyartika@gmail.com

1. INTRODUCTION

In today's digital age, the spread of negative content on the internet is increasing, one of which is comments containing elements of online gambling. Such comments not only undermine communication ethics but can also have a detrimental impact on users, particularly on social media platforms or public forums. Early detection of online gambling comments is crucial for maintaining the safety and comfort of digital spaces. One effective approach to addressing this issue is by using machine learning-based classification methods. Support Vector Machine (SVM) is one of the algorithms proven to be reliable in handling text classification problems, including detecting comments containing negative content such as gambling.

Previous research 1. By (Simanjuntak & Muhammad, 2025) entitled "The Application of Natural Language Processing in Online Gambling Sentiment Analysis on Twitter Social Media." This study explains that online gambling has become a popular activity among the public, especially in this digital era. However, many criticize this activity because of its negative aspects. This study achieved an accuracy rate of 76%, a precision rate of 58%, and a recall rate of 76%.

Previous research 2. By (Maulana & Yuliana, 2024) entitled "Analysis of Public Opinion Sentiment Related to Online Gambling Among Users of Application X Using the Naïve Bayes Algorithm and Support Vector Machine." This study explains that the analyzed data shows that the majority of public opinion regarding online gambling is neutral. These results provide valuable insights for policymakers in taking more appropriate steps in addressing the issue of online gambling.

Previous research 3. By (Simanjuntak & Muhammad, 2025) with the title "Detection of Online Gambling Promotional Text Using AI with a Combination of NLP and Deep Learning." This study explains that the results of this research indicate that AI technology can be effectively utilized as an automated tool in the process of moderating digital content, particularly to combat the spread of gambling content on online platforms such as YouTube. The results show that the transformer-based model (IndoBERT) has the highest accuracy compared to SVM and Naive Bayes.

Previous research 4. By (Pangestu & Harahap, 2024) with the title "Sentiment Analysis Related to Online Gambling on Instagram Using Naive Bayes" explains that from the analysis conducted on a dataset containing four comments, it was found that negative sentiment dominated with a proportion of 50%, while positive and neutral sentiment each only reached 25%. Negative comments generally contained social criticism related to issues such as corruption and poverty, indicating public concern about the social and economic impacts of online gambling.

Previous research 5. By (Simanjuntak & Muhammad, 2025) et al. With the title "Comparative Analysis of SVM and CNN Algorithms in Detecting Online Gambling Websites Based on Text Content." In this study, it is explained that with the increasing number of online gamblers in Indonesia, it is important to develop effective methods for identifying gambling content. The SVM model demonstrated an accuracy of 99%, with evaluation metrics of precision 1.00, recall 0.99, and F1-score 0.99. On the other hand, the CNN model achieved perfect accuracy of 100%, with precision, recall, and F1-score each at 1.00. These findings contribute significantly to the development of methods for detecting online gambling websites in Indonesia and open up opportunities for further research in this field.

Based on the above explanation, the researcher is interested in classifying comments related to online gambling because of the prevalence of online gambling users. It is interesting to analyze because online gambling is not just an online game but also has the potential to cause financial losses, addiction, and even family crises. According to reports from the Ministry of Communication and Information Technology and the Police, cases of online gambling in Indonesia have surged sharply in recent years, with thousands of new sites emerging each month. The impact extends to all segments of society, including teenagers and students, who are targeted by covert promotions through social media.

2. LITERATURE REVIEW

2.1 *Online Gambling*

Gambling is one form of social disease and qualifies as a crime. Gambling, in terms of definition, is a deliberate bet where one stakes a value or something considered valuable, fully aware of the risks and certain expectations associated with the outcomes of games, matches, races, and events whose

results are uncertain or not yet determined. The prevalence of gambling will damage the social system of society itself. Similarly, in Islam, gambling, gambling activities, and betting are considered sins or forbidden acts. Gambling is a temptation from the devil to disobey God's commands. Therefore, it is inherently evil and destructive (Jadidah et al., 2023).

2.2 *SVM (Support Vector Machine)*

Support Vector Machine is one of the classification methods that uses machine learning (supervised learning) to predict classes based on patterns from the training process created by Vladimir Vapnik. Several studies have explained that the Support Vector Machine (SVM) method is an efficient method (Ghifari et al., 2025).

2.3 *RapidManer*

RapidMiner is an application or software that functions as a learning tool in data mining science. The platform was developed by a company dedicated to all steps involving large amounts of data in commercial business, research, education, training, and learning. RapidMiner offers approximately 100 learning solutions for clustering, classification, and regression analysis. RapidMiner also supports around 22 file formats, such as .xls, .csv, and others (Sutoyo & Permana, 2025).

2.4 *Kaggle*

Kaggle is one of the most popular websites in the world for Data Science and Machine Learning, featuring over 6,000 datasets available for download in CSV format. Kaggle is extremely useful for those studying Data Science (Rahmat et al., 2023) (Fashakh et al., 2025).

3. METHODS

3.1 Dataset

Kaggle is one of the most popular websites in the world for Data Science and Machine Learning, featuring over 6,000 datasets that can be downloaded in CSV format. Kaggle is extremely useful for those studying Data Science (Rahmat et al., 2023).

3.2 Preprocessing

Preprocessing is a step to convert raw data into data or a format that is suitable for the next stage of analysis (Wulandari N. H., 2023) (Rambe et al., 2025) (Gibran et al., 2024). The following are the stages of preprocessing.

1. *Cleaning*

   Cleaning involves changing irrelevant characters in the text, such as changing letters to lowercase, removing symbols (read, numbers, URLs), and removing excess spaces.

   An example of changing letters to lowercase: “bro ALEXIS-â˜¯17â˜¯ lagi ngadain giveaway nih, buruan ikutan”

   Become: “bro alexis -â˜¯17â˜¯ lagi ngadain giveaway nih, buruan ikutan”

   Example of removing punctuation marks: “bro alexis -â˜¯17â˜¯ lagi ngadain giveaway nih, buruan ikutan”

   Become : “bro alexis a a lagi ngadain giveaway nih buruan ikutan”

   Example of removing extra spaces : “bro alexis a a lagi ngadain giveaway nih buruan ikutan”

   Become : “bro alexis a a lagi ngadain giveaway nih buruan ikutan”

2. Case Folding

Case folding is the process of converting all letters to lowercase for consistency.
For Example: "UNTUNGLAH YOUTUBE TIDAK BODOH DAN TUNDUK TERHADAP KOMINFO !!"
Become : "untunglah youtube tidak bodoh dan tunduk terhadap kominfo !!"

3. Tokenizing
Tokenizing is breaking down text into words (tokens) .
For Example : "mending jualan online cuan halal masa depan lebih cerah"
Become : "mending", "jualan", "online", "cuan", "halal", "masa", "depan", "lebih", "cerah"
4. Stopword Removal
Stopword Removal is the process of removing common words that do not add much meaning to the analysis. Such as words (in, which, and, to, more, this, that, etc.).
For Example : "harus segera ditonton sebelum video nya di take down"
Become : "Harus segera ditonton sebelum videonya take down"

3.3 Support Vector Machine (SVM)

The SVM method is a relatively new method for making predictions in regression or classification cases. SVM is also a set of related learning methods that analyze data and recognize patterns (Muhathir et al., 2021). Support Vector Machine was first introduced by Vapnik in 1992 as a harmonious series of leading concepts in the field of pattern recognition. SVM is very fast and effective for text classification problems. In geometric terms, a binary classification can be viewed as a hyperplane in feature space that separates points representing positive examples from those representing negative examples. This classification is selected during training as a unique hyperplane that separates known positive instances from negative instances. In SVM classification, it has an important advantage in its theoretical approach, which addresses the issue of overfitting, enabling it to perform well (Isnain et al., 2021) (Ma, 2025).

Data that has undergone processing is divided into training data and testing data. Training data is data used to train the model to recognize patterns from inputs and outputs. Testing data is used to measure the accuracy or performance of the model after training. In this study, data division was performed using the Split Data operator available in RapidMiner, with a ratio of 80% for training data and 20% for testing data. The dataset, totaling 8,442 records, was divided into 6,753 training data and 1,689 testing data. This division aims to ensure that the model has sufficient data to learn while also providing representative data to measure the model's generalization ability against new data that has never been seen before.

3.4 Flowchart

Flowchart explanation:

1. Start, the process begins. Indicates the start of the SVM pipeline.
2. Input Data, takes raw data to be analyzed in the form of Judi_online comments. At this stage, the data has not been processed or cleaned.
3. Set role, determines the role or division of data, which consists of training data and testing data in the form of 80% training data and 20% test data.
4. Data Preprocessing, performs cleaning, case folding, tokenizing, and stop removal.
5. Data Preprocessing Results, displaying or storing the processed data results in the form of numeric vectors ready to be input into the SVM model.

6. SVM Performance Testing, training the SVM model with training data, then testing it with testing data, measuring model performance using metrics (accuracy, precision, recall, F1-Score).
7. SVM Performance Test Results, the final stage after the SVM model has been trained and tested, producing the model's performance results.
8. Finish, the process is complete, the model results have been obtained and are ready to be reported or used.

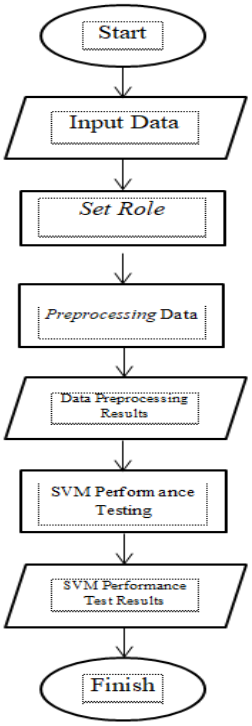


Figure 1. *Flowchart*

3.5 Evaluation

The evaluation of the SVM algorithm aims to measure the performance of the classification model after the model has been trained with data. This is important to determine whether the model is capable of classifying data accurately, efficiently, and without overfitting. At this stage, there are two parts that can determine the score of the model created. First, cross-validation is used to assess how well the model performs. Second, the model is tested by evaluating its accuracy, precision, recall, and F1-score. This involves comparing the number of true positives with the number of false positives within the classification class (Admojo & Sulistya, 2022) (Gupta & Rattan, 2023) (Romano & Conversano, 2025) (Mesanda & Sitompul, 2025) . For the evaluation process using a confusion matrix, precision, recall, and accuracy values will be obtained from the following formula :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$Precision = \frac{TP}{TP + FP} \times 100\%$$

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\%$$

Description :
TP (True Positive) = Number of positive data correctly predicted by the model.
TN (True Negative) = Number of negative data correctly predicted by the model.
FP (False Positive) = Number of negative data incorrectly predicted as positive by the model.
FN (False Negative) = Number of positive data incorrectly predicted as negative by the model.

4. RESULTS AND DISCUSSION

The classification process in this study was carried out in systematic stages: starting from reading raw data, organizing attribute roles, converting data to text format, preparing text through preprocessing, to training and evaluating the SVM model using Cross Validation.

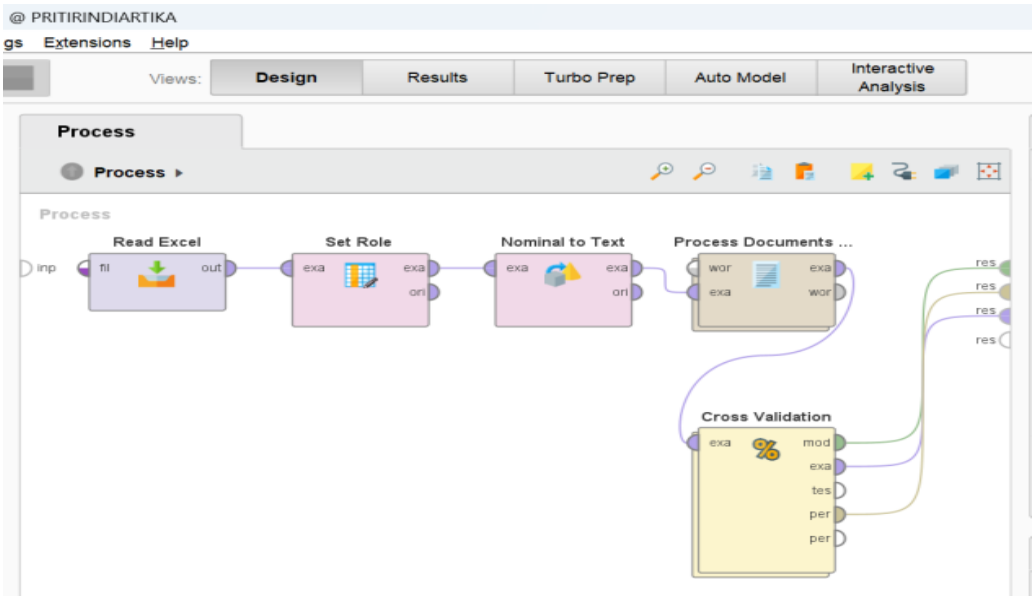


Figure 2. Workflow SVM

Figure 2. above shows the process of classifying online gambling comments using the Support Vector Machine (SVM) algorithm in the RapidMiner application. This process begins by reading data from an Excel file containing a collection of comments, followed by setting the role of the existing attributes—usually to designate the label column as the classification target.

After that, the data is processed through the "Nominal to Text" operator to convert the attribute format into text that can be further processed. The next step uses "Process Documents from Data," which prepares the text data before it is fed into the algorithm. At this stage, the comment text is usually broken down into tokens (words), cleaned of stopwords, and converted into numerical form using techniques such as TF-IDF so that it can be understood by the machine learning model.

Next, the main classification process takes place in the Cross Validation block. Within it, the data is divided into several parts (folds), where some will be used to train the model and the rest to test its performance. The model used is Support Vector Machine, which is known for its ability to handle high-dimensional data such as text. The evaluation results of this process, such as accuracy, precision, recall,

and F1-score, will appear after Cross Validation is run. However, these results are not yet visible in the image because the process has not been fully executed.

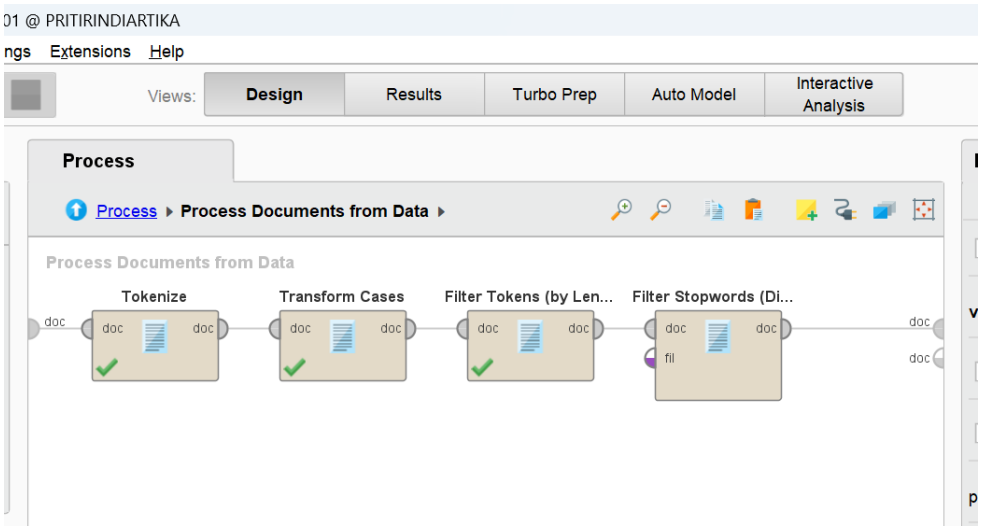


Figure 3. Preprocessing data

Next, process the text document. This process begins with the "Tokenize" operator, which breaks the text into small pieces called tokens, usually individual words. After the text is broken into tokens, the "Transform Cases" operator is used to convert all letters to lowercase. The purpose of this step is to ensure that words that are the same but have different capitalization, such as "Data" and "data," are recognized as the same word. The next step is to filter tokens based on their length using the "Filter Tokens (by Length)" operator. Here, tokens that are too short and considered irrelevant, such as one or two letters, will be removed from the data. Then, the process continues with the use of the "Filter Stopwords (Dictionary)" operator, which serves to remove common words that have no significant meaning in the analysis, such as "and," "which," or "is." These words are called stopwords and are usually determined based on a specific dictionary that corresponds to the language of the text being analyzed.

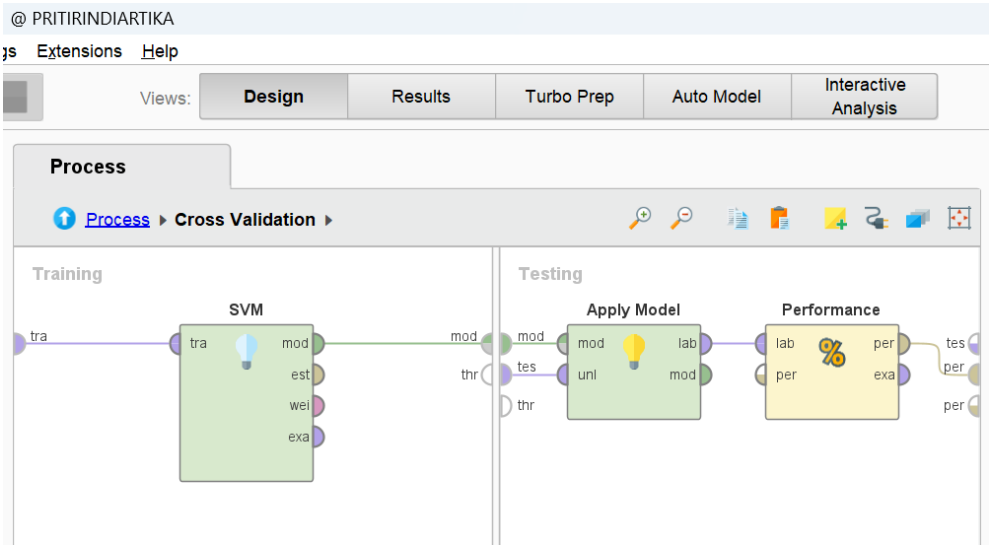
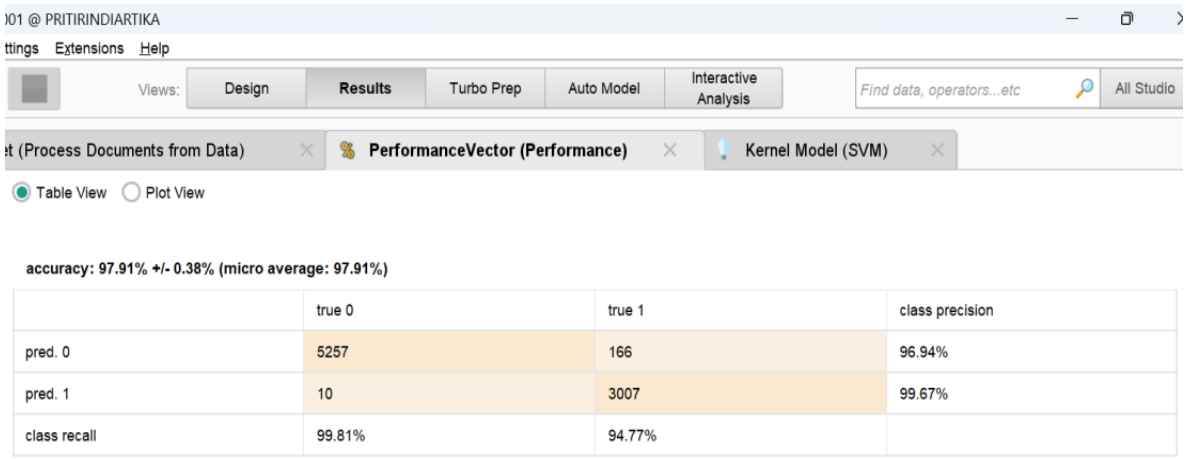


Figure 4. SVM Performance Testing

This process consists of two main parts, namely the Training section on the left and the Testing section on the right. In the training section, the algorithm used is SVM (Support Vector Machine). At this stage, the training data is entered into the SVM model to build a classification model based on the patterns found in the data. The trained model is then forwarded to the testing stage for evaluation. Moving on to the testing section, the model that has been created is applied to the test data using the Apply Model operator. This process generates predictions of labels or classifications for the test data based on the SVM model that has been created. The results of these predictions are then sent to the Performance operator, which is responsible for measuring how well the model performs. This measurement typically includes evaluation metrics such as accuracy, precision, recall, and F1-score.



	true 0	true 1	class precision
pred. 0	5257	166	96.94%
pred. 1	10	3007	99.67%
class recall	99.81%	94.77%	

Figure 5. Confusion Matrix

Figure 5. above shows a very high model accuracy value of 97.91% with a deviation of ±0.38%. This indicates that the model is able to classify data very well overall. After obtaining the confusion matrix, precision, recall, and F1-score are calculated to assess the accuracy, detection capability, and performance balance of the model in classification, thereby facilitating the identification of the model's strengths and weaknesses.

TP positif = 5257
FP Positive = 166
FN Positive = 10
TN Positive = 3007

Precission Positive = 5257/(5257+166) x 100% = 96,94%
Recall Positive =5257/(5257+10) x 100% = 99,81%
F1-Score Positive = 2x(96,94% x 100%)/(96,94% + 100%) = 98,45%

TP (Negative) = 3007
FP (Negative) = 10
FN (Negative) = 166
TN (Negative) = 5257

Precision Negative =3007/(3007+10) ×100%= 99,67%
Recall Negative =3007/(3007+166) ×100%=94,77%
FI-Score Negative =2×(99,67% × 94,77%)/(99,67% + 94,77%)=97,13

5. CONCLUSION

The conclusion of this study shows that the Support Vector Machine (SVM) algorithm applied through the RapidMiner platform is very effective in classifying comments containing elements of online gambling. By utilizing the dataset from Kaggle and through appropriate preprocessing stages such as tokenization, letter normalization, removal of irrelevant words (stopwords), and filtering based on token length, the text data was successfully prepared for model training. The training and testing process was conducted using cross-validation and data splitting into training and testing datasets, resulting in evaluation outcomes that are sufficiently accurate and reliable. The evaluation results show that the SVM model achieved a high accuracy rate of 97.91%, with precision of 96.94%, recall of 99.81%, and an F1-Score of 98.45%, which is also very high for both classes, whether the comments contain gambling elements or not. This indicates that the model is not only accurate but also balanced in identifying and distinguishing between the two types of comments. With these results, the SVM model has proven capable of automatically and efficiently detecting online gambling comments, making it highly potential for application in content moderation systems on digital platforms to maintain a healthy and safe communication space free from negative content.

REFERENCES

Admojo, F. T., & Sulistya, Y. I. (2022). Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin. *Indonesian Journal of Data and Science*, *3*(1), 1–8.

Fashakh, A. M., Çevik, M., Aydoğan, Ş. K., & Ibrahim, A. A. (2025). Detection cyberbullying using AI and sentiment analysis to examine psychological impacts on vulnerable groups. *Egyptian Informatics Journal*, *32*, 100856. https://doi.org/https://doi.org/10.1016/j.eij.2025.100856

Ghifari, A. G., Ananada, G. Y., & Purwandari, K. (2025). A Comparative Sentiment Analysis of Public Opinion on Indonesia's National Football Coach Using CRNN and SVM. *Procedia Computer Science*, *269*, 1485–1493. https://doi.org/https://doi.org/10.1016/j.procs.2025.09.090

Gibran, M. K., Rifki, M. I., Hasugian, A. H., Siahaan, A. T. A. A., Sahputra, A., & Ong, R. (2024). Sentiment Analysis of Platform X Users on Starlink Using Naive Bayes. *Instal: Jurnal Komputer*, *16*(03), 210–220. https://doi.org/10.54209/jurnalinstall.v16i03.240

Gupta, V., & Rattan, D. P. (2023). Improving Twitter Sentiment Analysis Efficiency with SVM-PSO Classification and EFWS Heuristic. *Procedia Computer Science*, *230*, 698–715. https://doi.org/https://doi.org/10.1016/j.procs.2023.12.125

Isnain, A. R., Sakti, A. I., Alita, D., & Marga, N. S. (2021). *Sentimen analisis publik terhadap kebijakan lockdown pemerintah Jakarta menggunakan algoritma SVM*. Jdmsi. https://doi.org/10.33365/jdmsi.v2i1.1021

Jadidah, I. T., Lestari, U. M., Fatiha, K. A. S., Riyani, R., & Wulandari, C. A. (2023). Analisis maraknya judi online di Masyarakat. *Jurnal Ilmu Sosial Dan Budaya Indonesia*, *1*(1), 20–27. https://doi.org/https://doi.org/10.61476/8xvgdb22

Ma, Y. (2025). Construction and Data Analysis of a New Media Content Popularity Prediction Model Based on Naive Bayes Algorithm. *Procedia Computer Science*, *261*, 294–302. https://doi.org/https://doi.org/10.1016/j.procs.2025.04.207

Maulana, A., & Yuliana, A. (2024). Analisis Sentimen Opini Publik Terkait Judi Online Pada Pengguna Aplikasi X Menggunakan Algoritma Naïve Bayes Dan Support Vector Mechine. *Jurnal Informatika*

Dan Teknik Elektro Terapan, *12*(3S1). https://doi.org//10.23960/jitet.v12i3S1.5187

Mesanda, Z., & Sitompul, B. A. (2025). Sentiment Analysis of Instagram Comments on Capital Relocation Using SVM and Random Forest. *Jurnal Metrokom : Media Teknik Elektro Dan Komputer*, *2*(1), 66–79. https://doi.org/10.65371/metrokom.v2i1.59

Muhathir, M., Santoso, M. H., & Larasati, D. A. (2021). Wayang image classification using SVM method and GLCM feature extraction. *Journal Of Informatics And Telecommunication Engineering*, *4*(2), 373–382. https://doi.org/10.31289/jite.v4i2.4524

Pangestu, A. D., & Harahap, L. S. (2024). Analisis Sentimen Terkait Judi Online di Media Sosial Instagram Menggunankan Naïve Bayes. *Indonesian Journal of Education and Development Research*, *3*(1), 556–561. https://doi.org/https://doi.org/10.57235/ijedr.v3i1.4798

Rahmat, A., Syafiih, M., & Faid, M. (2023). Implementasi Klasifikasi Potensi Penyakit Jantung Dengan Menggunakan Metode C4. 5 Berbasis Website (Studi Kasus Kaggle. Com). *INFOTECH Journal*, *9*(2), 393–400. https://doi.org/10.31949/infotech.v9i2.6295

Rambe, M. R. A., Zufria, I., & Rifki, M. I. (2025). Analisis Sentimen Masyarakat pada Platform Media Sosial X (Twitter) terhadap Pelantikan Kabinet Merah Putih Menggunakan Bernoulli Naïve Bayes. *DEVICE: JOURNAL OF INFORMATION SYSTEM, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY*, *6*(1), 83–102. https://doi.org/10.46576/device.v6i1.6360

Romano, M., & Conversano, C. (2025). Stairway to heaven: An emotional journey in Divina Commedia with threshold-based Naïve Bayes classifier. *Machine Learning with Applications*, *19*, 100613. https://doi.org/https://doi.org/10.1016/j.mlwa.2024.100613

Simanjuntak, N., & Muhammad, A. H. (2025). Analisis Perbandingan Algoritma SVM dan CNN dalam Mendeteksi Website Judi Online Berdasarkan Konten Teks. *Bulletin of Computer Science Research*, *5*(4), 361–371. https://doi.org//10.47065/bulletincsr.v5i4.586

Sutoyo, E., & Permana, M. C. (2025). Enhancing telemedicine service quality through sentiment analysis of user review dataset in Indonesia. *Data in Brief*, *61*, 111878. https://doi.org/https://doi.org/10.1016/j.dib.2025.111878