JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 5, No. 1. December 2022 DOI: https://doi.org/10.34288/jri.v5i1.189
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Support Vector Classification with Hyperparameters for Analysis of Public Sentiment on Data Security in Indonesia

Siti Ernawati [-1*)], Risa Wati [-2], Nuzuliarini Nuris [-3]

Sistem Informasi
Universitas Nusa Mandiri
Jakarta, Indonesia
www.nusamandiri.ac.id
[1*)]siti.ste@nusamandiri.ac.id

Sistem Informasi
Universitas Bina Sarana Informatika
Jakarta, Indonesia
www.bsi.ac.id
[2]risawati.rwx@bsi.ac.id, [3]nuzuliarini.nzn@bsi.ac.id

(*) Corresponding Author

Abstract

The development of Information Technology makes increasing use of the internet. This raises the vulnerability of data security. Cyber attacks in Indonesia caused many tweets on social media Twitter. Some are positive, and some are negative. The problem of this study is to determine the public sentiment towards data security in Indonesia, while the purpose of this study is how the response or evaluation of the government of Indonesia to the many perceptions of people who lack confidence in data security in Indonesia. Data obtained from twitter with as much as 706 data was processed using python with a percentage of 10% test data and 90% training data. Weighting is done using TF-IDF, and then the data is processed using the Support Vector Machine algorithm using the SVC (Support Vector Classification) library. Support Vector Classification with RBF kernel classifies Text well to obtain AUC value with good classification category. Utilizing one of the hyperparameter techniques, which is a grid search technique that can compare the accuracy of test results. The test results using SVC with RBF kernel obtained an accuracy value of 0.87, Precision of 0.82, recall of 0.94, and F1_Score of 0.87. This study is expected to be used by decision-makers related to public confidence in data security in Indonesia.

Keywords: Data Security; Grid Search ; Hyperparameter; Support Vector Classification

Abstrak

Berkembangnya teknologi Informasi membuat meningkatnya penggunaan internet. Hal ini menimbulkan rentannya keamaan data. Serangan siber di Indonesia menimbulkan banyaknya cuitan pada media social twitter, ada yang beropini positif dan ada yang beropini negative. Permasalahan dalam penelitian adalah untuk mengetahui sentimen masyarakat terhadap keamanan data di Indonesia, sedangkan tujuan dari penelitian ini adalah bagaimana tanggapan atau evaluasi pemerintah Indonesia terhadap banyaknya persepsi masyarakat yang kurang percaya terhadap keamanan data di Indonesia. Data diperoleh dari twitter dengan jumlah data sebanyak 706 data diolah menggunakan python dengan prosentase 10% data test dan 90% data training. Dilakukan pembobotan menggunakan TF-IDF selanjutnya data diolah menggunakan algoritma Support Vector Machine dengan memanfaatkan library SVC (Support Vector Classification). Support Vector Classification dengan kernel RBF mengklasifikasikan teks dengan baik memperoleh nilai AUC dengan kategori good classification. Memanfaatkan salah satu teknik hyperparameter yaitu teknik grid search yang dapat membandingkan keakuratan hasil uji. Hasil uji menggunakan SVC dengan kernel RBF didapatkan nilai akurasi sebesar 0.87, Precision 0.82, recall 0.94 dan F1_Score 0.87. Penelitian ini diharapkan dapat dijadikan pengambil keputusan terkait kepercayaan masyarakat terhadap keamanan data di Indonesia.

Kata kunci: Grid Search; Hyperparameter; Keamanan Data; Support Vector Classification

INTRODUCTION

Information technology is growing rapidly, causing an impact on many community activities. Utilizing information technology resulted in the vulnerability of personal data security. Every year internet users in Indonesia increase, based on the results of a survey by the Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) in the 2021-2022 period reaching 210.03 million users (Bayu, 2022). Increased use of the internet makes the risk of hacking even higher. Until April 2022, Badan Siber dan Sandi Negara (BSSN) recorded cyber attacks in Indonesia, reaching 100 million cases (Andreya, 2022). The issue of personal data protection is critical. The theft of personal data such as names, addresses, emails, telephone numbers, bank account numbers, and medical history in the thousands to millions can threaten national security (Wisnubroto, 2021). Regulation of personal data protection is essential due to the emergence of various problems as the use of personal data in information technology-based transactions increases (Rumlus & Hartadi, 2020). Kemenkominfo revealed that Indonesia was the country that experienced the most cyber attacks (Widyanuratikah, 2018).

In the last few years, social media, one of which is Twitter, has gotten much attention, and users can express their opinions. Users generate information that shows the user's views on a particular topic, which is very useful for analyzing public opinion (Naz et al., 2018). The number of cyber attacks that occurred in Indonesia made people have given opinions on social media such as Twitter. Various tweets appeared with both positive opinions and negative opinions. Therefore, researchers will analyze public perceptions of data security in Indonesia. The purpose of this study is to determine the value of performance metrics generated by SVC using the hyperparameter grid search technique and is expected to be helpful for the government of Indonesia in evaluating the perception of people who lack confidence in data security in Indonesia.

The proposed model will analyze the sentiment of public perception of data security in Indonesia using the SVM algorithm with SVC library and Grid-Search techniques. With two or more classes, SVM is commonly used in classification problems (Hsu, Chang, & Lin, 2010). The SVM algorithm will find the best parameters (Ahmad, Aftab, Bashir, Hameed, et al., 2018) and compare the accuracy of the results obtained and then choose the best parameters. The most recommended Kernel in SVM is RBF, and this Kernel nonlinearly maps samples into higher dimensional space (Hsu et al., 2008). The parameters to be searched are Gamma and C. To evaluate the performance of the proposed model using a confusion matrix by calculating the performance metrics of accuracy, Precision, recall, and F1-Score, also using the ROC curve (Receiver Operator Characteristic) and AUC value (Area Under Curve).

The first related research compares Twitter's sentiment classification algorithm to Tokopedia's data leak incident. Conducted observations of three different classifiers, it was concluded that of the total 494 tweets analyzed, Support Vector Machine is the classifier with the best performance, resulting in the highest f1-score of 0.503583 (Wibowo et al., 2021). The following research is the application of SVM in various research fields such as text categorization, protein fold, remote homology detection, Image classification, Bioinformatics, Hand-written character recognition, Face detection, Generalized predictive control, and many more. Many researchers have shown that in classification techniques, SVM is better than other algorithms (Cervantes et al., 2019).

Research on Gopay user sentiment analysis using Lexicon Based and Support Vector Machine method shows SVM classification method by comparing the Kernel is quite good, for linear Kernel get an accuracy value of 89.17% while the polynomial Kernel of 84.38% (Mahendrajaya et al., 2019). Application of the Support Vector Machine method for sentiment analysis of indihome services based on tweets (Tineges et al., 2020) resulted in an accuracy of 87% with accuracy between the predicted results with the actual data (Precision) of 86%, the success rate of the system in predicting a data (recall) of 95%, while for the average comparison value of Precision and recall (f1-score) of 90%. Research conducted on sentiment analysis of gojek on social media using the SVM method (Fitriyah et al., 2020) has the best overall accuracy rate of 79.19%. The accuracy is obtained from modelling using RBF kernel with Cost=1000 and γ=0,00026.

RESEARCH METHODS

The stages of research conducted are:

1. Identification of Problems and Objects of Research

The problem of this study is to determine people's sentiment about data security in Indonesia, while the purpose of this study is how the response or evaluation of the government of Indonesia to the many perceptions of people who lack confidence in

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i1.189
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

data security in Indonesia. The object of this study is the opinions written by people on social media. This opinion continues to increase over time, becomes a problem, and can be used by researchers to determine public sentiment by analyzing public opinion.

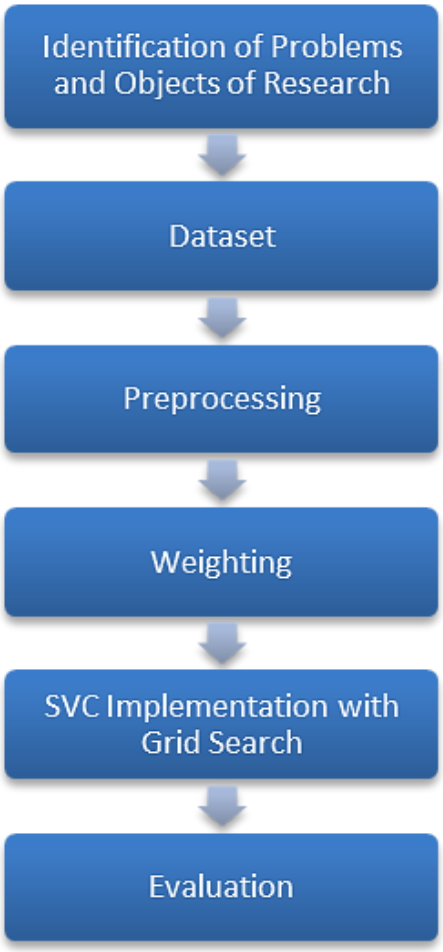

Figure 1. Stages of Research

2. Data set

Using public Data is data collected from Twitter, namely data on public opinion about data security in Indonesia. The data used as much as 706, labelled positive and negative and then processed using python. The percentage of test data and training data is 10% test data and 90% training data.

3. Prepossessing

Preprocessing converts raw data into data that uses for subsequent data processing. Several stages of preprocessing conducted in this study are:

a. Case Folding converts each word in the dataset into lowercase (Fikri et al., 2020).
b. Cleaning Text removes unused characters such as double spaces, hashtags, links, Retweets, mentions, and punctuation (Rahmawati & Sukmasetya, 2022).
c. Tokenization is the process used to divide Text into single words (Unigrams) or consecutive combinations of words (n-grams) (Chiny et al., 2021).
d. Stopword removes words that have no meaning, such as the, in, and on (Fikri et al., 2020).
e. Stemming is changing a word into a base word by removing affixes (Rahmawati & Sukmasetya, 2022).

4. Weighting

TF-IDF (Term Frequency-Inverse Document Frequency) is an algorithm for assigning weight to Text (Fikri et al., 2020). TF (Term Frequency) is the frequency of occurrence of a term in the document in question, while the IDF is the relationship between the availability of a term in all documents (Mahendrajaya et al., 2019).

TF-IDF is used to find out how often words appear in a document, the following formula in TF-IDF weighting (Liu & Yang, 2012).

$$a_{i,j} = tf_{i,j} * log\left(\frac{N}{n_j}\right) \quad \text{.....(1)}$$

$tf_{i,j}$ Is term frequency of term j in document i. N represents the total number of documents in the dataset. n_j is the number of the emergence of documents in term i.

5. Support Vector Classification Implementation with Grid Search

Support vector machine is an important and fundamental technique in machine learning (Yin & Li, 2019). SVM is one of the widely used machine-learning techniques to detect the polarity of Text (Ahmad, Aftab, Bashir, & Hameed, 2018). SVM can predict both classification and regression. In implementing the performance model, researchers will use the sklearn SVC library, a Support Vector Classification, to implement the libsvm library (Chang & Lin, 2011). How SVM works first, SVM looks for a support vector in each class. The support vector is a sample from each class with the closest distance to other class samples. After the support vector obtains, SVM then calculates the margin. We can think of margin as the path that separates two classes. Margin is created based on the support vector where the support vector works as the edge of the road, or often we know as the shoulder of the road. SVM seeks the largest margin or widest path to separate the two classes.

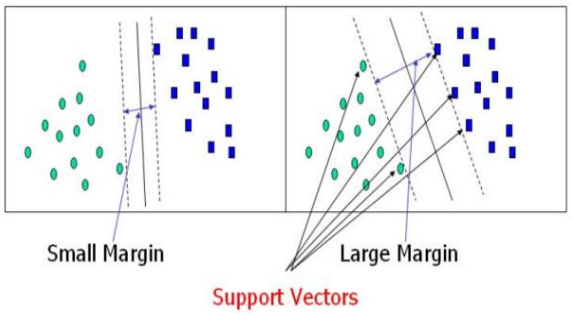

Figure 2. SVM Working Technique (Ahmad, Aftab, Bashir, & Hameed, 2018)

Based on Figure 2, SVM will select the margin on the right because the 'path 'or margin on the right image is wider than the' path ' on the left. It is why images placed off to the left are said to have a "big margin," while images placed off to the right have a "small margin."

6. Evaluation

An essential step in this evaluation is to measure the performance of the proposed model. This evaluation is used as a consideration to choose the best model. One technique to measure the proposed model's performance is the fusion matrix. The performance metrics calculated in the confusion matrix are accuracy, Precision, recall, and F1-Score. In addition to the confusion matrix, the researcher used the ROC curve (Receiver Operator Characteristic) and AUC value (Area Under Curve) in this evaluation stage.

This evaluation phase will explain whether proven Support Vector Classification is a good classification model for sentiment analysis of data security in Indonesia.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (2)$$

$$precision = \frac{tp}{tp+fp} \quad (3)$$

$$recall = \frac{tp}{tp+fn} \quad (4)$$

$$F1 - Score = 2 * (recall * precision)/(recall + prescision) \quad (5)$$

RESULTS AND DISCUSSION

1. Classification Of Models

This study collected data from social media Twitter in the form of public opinion about data security in Indonesia. The data used as much as 706 data, and then the data was separated manually into positive and negative opinions. With several positive opinions as much as 379 and negative opinions as much as 328 data. Data sharing between test data and training data. Comparison of data made 10% test data and 90% training data. Based on Figure 3, the negative words in the opinion include data, private, cyber, bocor, and bjorka.

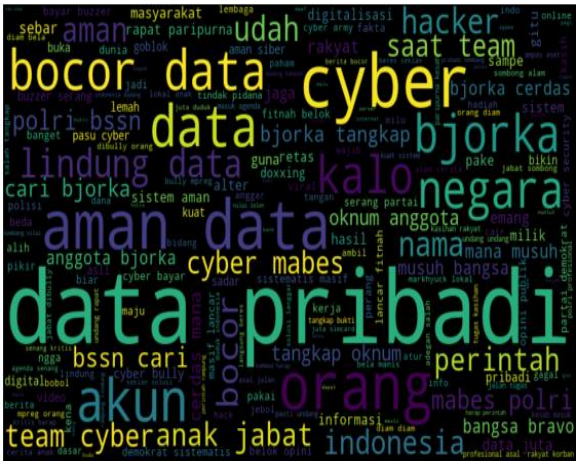

Figure 3. Word Cloud Negatif

In Figure 4, the positive words that appear are data, pribadi, aman, lindung, and undang-undang. Positive and negative Wordcloud shows the words used in the sentiment, the more words that appear, the larger the size of the word in the image.

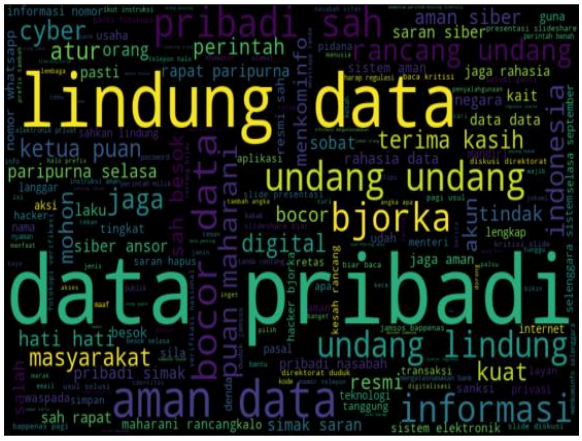

Figure 4. Word Cloud Positif

After the data becomes positive and negative opinions, preprocessing the data removes data noise so that the data is used for further data processing. Here is the form of data before and after preprocessing.

Table 1. Preprocessing Result

Tweet	Preprocessing Result
#kuliahasyik #digitalactivism #Bjorka bjorka sosok hero indonesia? Bjorka menunjukan bahwa betapa lemahnya sistem perlindungan dan keamanan data indonesia,seharusnya pemerintah kita sadar dengan kebocoran sistem pemerintah dari dulu sampek sekarang.	bjorka sosok hero indonesia bjorka tunjuk betapa lemah sistem lindung aman data indonesia perintah sadar bocor sistem perintah sampek
Di tengah isu privasi dan kebocoran data warga yang gagal diamankan negara, @IndiHome memperbarui kebijakan. Pengguna harus merelakan datanya untuk ???afiliasi Telekom dan tujuan-tujuan lain.??? Di poin 8, Indihome tak bertanggung jawab jika data bocor karena ???gangguan keamanan.??? https://t.co/zMsX5O5L09	privasi bocor data warga gagal aman negara baru bijak guna rela data afiliasi telekom tuju tuju poin indihome tanggung data bocor ganggu aman

2. Experiment Using the SVC model with Grid Search

Statistical approaches such as machine learning and deep learning work best when using numerical data. Therefore, opinion data is a collection of words converted into numbers or numeric. This process is often called the weighting process. Some techniques often used in weighting include Bag of Words, N-grams, Word2Vec, and TF-IDF. This study uses TF-IDF (Term Frequency-Inverse Document Frequency) for weighting in sentiment analysis (Chiny et al., 2021). TF-IDF is used to find out how often words appear in a document, the following formula in TF-IDF weighting (Liu & Yang, 2012).

After the weighting process, the SVC library implements on the Support Vector Machine by building a model and utilizing several kernels. There are four kernels, linear, RBF, poly, and sigmoid. One hyperparameter technique will be used based on the four kernels: grid search. One technique to optimize the accuracy value is to use grid search (Yan et al., 2022), a grid search technique to determine the result of a combination of parameters that are best for the performance of the proposed model (K et al., 2019). The parameters tested can be seen in table 2.

Table 2. Parameters to be Tested

Parameter	**Description**
kernel	linear, RBF, poly, sigmoid
C	1, 100, 1000, 10000
gamma	0.01, 0.1, 1, 10, 100

Table 3. SVC Model Accuracy Results with Grid Search

Kernel	**C**	**Gamma**	**Accuracy Results**
Linear	1	0.01	Precision : [0.94 0.79] Accuracy: 0.86 F1_Score: 0.86 Recall: 0.94
RBF	**1**	**1**	**Precision : [0.94 0.82]** **Accuracy: 0.87** **F1_Score: 0.87** **Recall: 0.94**
Poly	1	1	Precision : [0.87 0.85] Accuracy: 0.86 F1_Score: 0.85 Recall: 0.76
Sigmoid	100	0.01	Precision : [0.94 0.79] Accuracy: 0.86 F1_Score: 0.86 Recall: 0.94

Table 3 shows the results of the SVC model calculation with grid search. From the test results can be seen that the best parameter is C= 1 and Gamma = 1, which produces a value of negative Precision = 0.94, positive Precision = 0.82, Accuracy = 0.87, F1_Score = 0.87 and recall = 0.94

3. Evaluation

To evaluate the performance of the proposed model using performance metrics of accuracy, Precision, recall, and F1-Score. Performance metrics are calculated based on the value of the true positive (tp), false positive (FP), true negative (TN), and false negative (fn) class set. At the same time, the Precision of the correct prediction is positive compared to the overall predicted outcome.

Figure 5. Confusion Matrix obtained a True negative value of 43.66%, False positive of 9.86%, False negative of 2.82%, and True Positive of 43.66%.

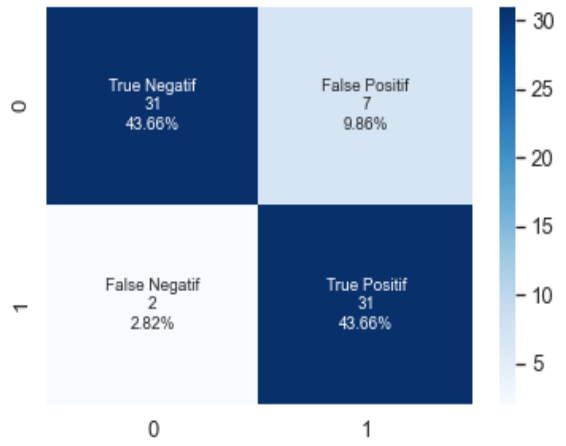

Figure 5. Confusion Matrix

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} = \\ = \frac{(43.66+43.66)}{(43.66+9.86+2.82+43.66)} = 0.87$$

$$precision = \frac{tp}{tp+fp} = \frac{43.66}{43.66+9.86} = 0.82$$

$$recall = \frac{tp}{tp+fn} = \frac{43.66}{43.66+2.82} = 0.94 \quad (8)$$

$$F1 - Score = 2 * (recall * precision)/(recall + prescision) = 2 * (0.94 * 0.82)/(0.94 + 0.82) = 0.87$$

Figure 6 illustrates the ROC curve. The ROC curve shows accuracy and visually compares classifications. The ROC curve expresses the confusion matrix. ROC is a two-dimensional graph with false positives as horizontal lines and true positives as vertical lines. This study obtained an AUC value of 0.89, and the value entered into the category of good classification.

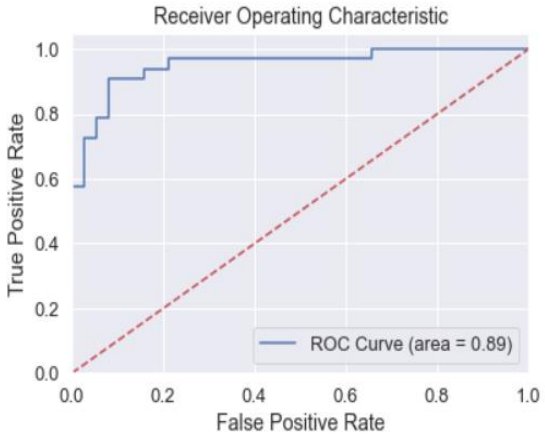

Figure 6. ROC Curve

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the test results for sentiment analysis of Indonesian public confidence in data security, SVC with RBF kernel can classify Text well, as evidenced by the AUC value of 0.89. The grid search technique can compare the accuracy of the test results and then choose the best parameters. The test results using SVC with RBF kernel obtained an accuracy value of 0.87, Precision of 0.82, recall of 0.94, and F1_Score of 0.87. This research Model can be applied by agencies in decision-making related to public confidence in data security so that the government can evaluate the number of perceptions of people who lack confidence in data security in Indonesia.

Suggestion

Further research can focus on comparing techniques or other classification methods that can create a new model by combining many classification and selection methods to produce more accurate values.

REFERENCES

Ahmad, M., Aftab, S., Bashir, M. S., & Hameed, N. (2018). Sentiment Analysis using SVM : A Systematic Literature Review. *(IJACSA)*

International Journal of Advanced Computer Science and Applications, *9*(2), 182–188. https://doi.org/10.14569/IJACSA.2018.090226

Ahmad, M., Aftab, S., Bashir, M. S., Hameed, N., Ali, I., & Nawaz, Z. (2018). SVM optimization for sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *9*(4), 393–398. https://doi.org/10.14569/IJACSA.2018.090455

Andreya, E. (2022). *Antisipasi Bersama Tingkatkan Sistem dan Cegah Serangan Siber*. Aptika.Kominfo.Go.Id. https://aptika.kominfo.go.id/2022/09/antisipasi-bersama-tingkatkan-sistem-dan-cegah-serangan-siber/

Bayu, D. (2022). *APJII: Pengguna Internet Indonesia Tembus 210 Juta pada 2022*. DataIndonesia.Id. https://dataindonesia.id/digital/detail/apjii-pengguna-internet-indonesia-tembus-210-juta-pada-2022

Cervantes, J., Garcia-lamont, F., Rodríguez-mazahua, L., & Lopez, A. (2019). Neurocomputing A comprehensive survey on support vector machine classification : Applications, challenges and trends. *Neurocomputing*, *xxxx*. https://doi.org/10.1016/j.neucom.2019.10.118

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–40. https://doi.org/10.1145/1961189.1961199

Chiny, M., Chihab, M., Chihab, Y., & Bencharef, O. (2021). LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model. *International Journal of Advanced Computer Science and Applications*, *12*(7), 265–275. https://doi.org/10.14569/IJACSA.2021.0120730

Fikri, M. I., Sabrila, T. S., & Azhar, Y. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. *Smatika Jurnal*, *10*(02), 71–76. https://doi.org/10.32664/smatika.v10i02.455

Fitriyah, N., Warsito, B., & Maruddani, D. A. I. (2020). Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (SVM). *Jurnal Gaussian*, *9*(3), 376–390. https://doi.org/10.14710/j.gauss.v9i3.28932

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2008). A Practical Guide to Support Vector Classification. *BJU International*, *101*(1), 1396–1400. http://www.csie.ntu.edu.tw/%7B~%7Dcjlin/papers/guide/guide.pdf

K, R. G. S., Verma, A. K., & Radhika, S. (2019). K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries. *2019 5th International Conference for Convergence in Technology (I2CT)*, 9–13. https://doi.org/10.1109/I2CT45611.2019.9033691

Liu, M., & Yang, J. (2012). An improvement of TFIDF weighting in text categorization. *2012 International Conference on Computer Technology and Science (ICCTS 2012)*, *47*(Iccts), 44–47. https://doi.org/10.7763/IPCSIT.2012.V47.9

Mahendrajaya, R., Buntoro, G. A., & Setyawan, M. B. (2019). Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based Dan Support Vector Machine. *Komputek*, *3*(2), 52–63. https://doi.org/10.24269/jkt.v3i2.270

Naz, S., Sharan, A., & Malik, N. (2018). Sentiment Classification on Twitter Data Using Support Vector Machine. *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 676–679. https://doi.org/10.1109/WI.2018.00-13

Rahmawati, C., & Sukmasetya, P. (2022). Sentimen Analisis Opini Masyarakat Terhadap Kebijakan Kominfo atas Pemblokiran Situs non-PSE pada Media Sosial Twitter. *JURIKOM (Jurnal Riset Komputer)*, *9*(5), 1393–1400. https://doi.org/10.30865/jurikom.v9i5.4950

Rumlus, M. H., & Hartadi, H. (2020). Kebijakan Penanggulangan Pencurian Data Pribadi dalam Media Elektronik. *Jurnal HAM*, *11*(2), 285–299. https://doi.org/10.30641/ham.2020.11.285-299

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. https://doi.org/10.30865/mib.v4i3.2181

Wibowo, N. I., Maulana, T. A., Muhammad, H., & Rakhmawati, N. A. (2021). Perbandingan Algoritma Klasifikasi Sentimen Twitter Terhadap Insiden Kebocoran Data Tokopedia. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(2), 120–129. https://doi.org/10.14421/jiska.2021.6.2.120-129

Widyanuratikah, I. (2018, October 8). Indonesia

Negara Ketiga Paling Sering Terkena Serangan Siber. *Republika.*, Nasional. https://www.republika.co.id/berita/pg9slu354/indonesia-negara-ketiga-paling-sering-terkena-serangan-siber

Wisnubroto, K. (2021). *Memastikan Data Pribadi Aman*. Indonesia.Go.Id. https://www.indonesia.go.id/kategori/editorial/3272/memastikan-data-pribadi-aman

Yan, T., Shen, S.-L., Zhou, A., & Chen, X. (2022). Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(4), 1292–1303. https://doi.org/10.1016/j.jrmge.2022.03.002

Yin, J., & Li, Q. (2019). A semismooth Newton method for support vector classification and regression. *Computational Optimization and Applications*, *73*(2), 477–508. https://doi.org/10.1007/s10589-019-00075-z