Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 9, No. 1, March 2023, pp. 165-173
ISSN: 2338-3070, DOI: 10.26555/jiteki.v9i1.25861

Comparison of Support Vector Machine (SVM) and Random Forest Algorithm for Detection of Negative Content on Websites

Hermawan Syahputra, Aldiva Wibowo
Department of Computer Science, Universitas Negeri Medan, North Sumatera, Indonesia

ARTICLE INFO

Article history:

Received February 22, 2023
Revised March 19, 2023
Published March 21, 2023

Keywords:

Negative Content
Natural Language Processing
Machine Learning
Support Vector Machine
Random Forest

ABSTRACT

The amount of negative content circulating on the internet can damage people's morale so that social conflicts arise in society that threaten national sovereignty. Detecting negative content can help identify and prevent harmful events before they occur. This can lead to a safer and more positive online environment. Comparison of Support Vector Machine (SVM) and Random Forest (RF) Algorithm for Detection of Negative Content on Websites. The research contributions are 1) detect negative content on the internet with random forest and SVM, 2) comparing SVM and RF algorithms for detecting negative content on websites, 3) detection of negative content based on text focusing on the categories of fraud, gambling, pornography and Whitelist. The stages of this research are preparing a text content dataset on a website that has been labeled, preprocessing (duplicated data, text cleansing, case folding, stopward, tokenize, label encoding, data splitting, and determine the TF-IDF), finally performing the classification process with SVM and Random Forest. The dataset used in this study is a structured dataset in the form of text obtained from emails that have been registered on the TrustPositive website as negative content. Negative content includes fraud, pornography and gambling. The results show the accuracy of the SVM is 97%, Precision 90% and Recall 91%, while for Accuracy in Random Forest is 92%, Precision 71%, and Recall 86%. The value obtained is the result of testing using 526 website URLs. The test results show that the Support Vector Machine is better than the Random Forest in this study.

This work is licensed under a Creative Commons Attribution-Share Alike 4.0

Corresponding Author:

Hermawan Syahputra, Department of Computer Science, Universitas Negeri Medan, North Sumatera, Indonesia
Email: hsyahputra@unimed.ac.id

1. INTRODUCTION

The internet has become an integral part of modern life. The internet provides a lot of information that can be used by the global community. Information presented on the internet can be either positive information or negative information. For this reason, besides having a very large positive impact, it also has a negative impact. Negative impacts are caused due to bad data sources on the internet, such as fraudulent activity, gambling, and content that is not suitable for all people [1]. Negative elements on the internet make it difficult for users to determine a positive or negative website. Negative content can have a bad impact on users [1][2][3]. The government through Indonesian Ministry of Information and Communication has tried to overcome this by building a reporting system and DNS (Domain Name System) in the form of a Positive Trust Blacklist that can filter out websites that contain negative content such as fraud, gambling, and adult content. TrustPositive blocks websites that are reported and users can report websites with negative content through the complaints feature in it. TrustPositive is currently blocking many dangerous websites [4][5]. However, it is still ineffective because it is only based on DNS and public reports, namely from the public and related agencies which are adjusted to the dataset from the central server, before finally blocking the website in question.

This problem can be overcome by developing a negative content detection system on the internet and other social media. Negative content detection can be done using machine learning technology [6]. Artificial

intelligence is applied through machine learning, which enables systems to automatically learn from their surroundings and use that learning to improve their decision-making. Machine learning employs a variety of algorithms to iteratively analyze, explain, and enhance data, uncover patterns, and then take appropriate action [7]. The model created by machine learning may be descriptive or predictive. It is both descriptive and predictive in order to extract information from the data and create predictions about events in the future [8].

The negative content detection problem is a case of Text Classification. NLP is also known as computational linguistics, The field of NLP includes various techniques and methods for processing natural language, such as text analysis, language processing, and linguistic resource creation [9]. Methods used in NLP include text analysis, language processing, linguistic resource creation, and so on to help systems understand and manage language more effectively [10][11]. Text mining is a classifying concept in machine learning. The classification algorithm will group a set of text data into predetermined classes. By utilizing text mining, content circulating on the internet can be grouped into negative content classes and not negative content.

There are many studies that have been carried out to develop this negative content detection, including identification of negative content using machine learning [5], detection of negative content (hoax) on microblog data that contains covid-19 information [12], Deep Learning to detect inappropriate content in text [13], Identifying Harmful Text: Deep Learning from Public Comments and Emails [14], LSTM Neural Network for Adult Content Classification on Indonesian Tweets [15], Automated Pornographic and Gambling Website Identification Using a Decision Process Based on Visual and Textual Content [16], Use of Support Vector Machine and Naive Bayes to Classify Pornographic Material on Twitter [17].

One of the algorithms applied to text processing is the Vector Support Model for text to get an accuracy of 87.07% which indicates that the model is quite good at detecting negative content [18]. The random forest algorithm for sentiment analysis which produces quite good performance can correctly predict 810 out of a total of 1000 data and reaches 80.4% accuracy [19]. SVM classifier lebih baik dari beberapa classifier lainnya seperti Naïve Bayes [20], [21], Multinomial Naive Bayes (MNB) [22].

One of the most popular classifiers for classification and regression problems is random forest (RF). It is a desirable option for text classification due to its straightforward algorithm. Also, it has a substantial advantage over other machine learning models in that it can handle high-dimensional data and perform well with unbalanced datasets [23][24][25][26]. The RF is a prominent technique for handling imbalanced data and performs significantly better than other machine learning models due to its parallel architecture [27]. RF is preferred to the well-liked Decision Tree [28].

Based on the description above, it is interesting to develop a negative content detection system on websites based on text that focuses on the categories of fraud, gambling, and pornography using the Support Vector Machine (SVM) and Random Forest (RF). Then, the best performance comparison between the two methods is carried out. The research contributions are 1) detect negative content on the internet with random forest and SVM, 2) comparing SVM and RF algorithms for detecting negative content on websites, 3) detection of negative content based on text focusing on the categories of fraud, gambling, pornography and Whitelist.

2. METHODS

The stages in this study include preparing a text content dataset on a website that has been labeled.

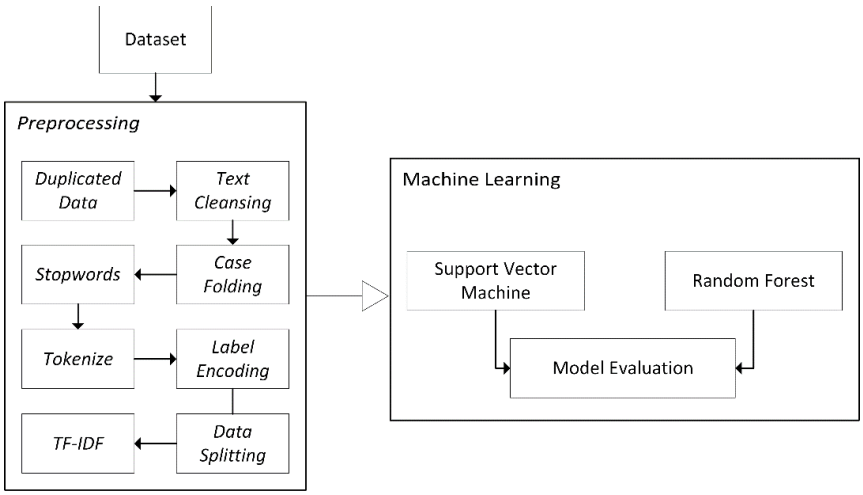

Fig. 1. Stages of The Negative Content Detection Process On Websites using SVM and RF

The next stage is to do pre-processing, furthermore, the data that has been pre-processed is divided into two parts, namely training data and testing data. The next stage is the classification process using the Support Vector Machine and Random Forest. The last evaluation was carried out on both models. The stages of the negative content detection process on websites with SVM and RF can be seen in Fig. 1.

2.1 Dataset

The dataset used in this study is a structured dataset in the form of text obtained from emails that have been registered on the TrustPositive website as negative content. In this case the website focuses on fraud, pornography, and gambling classes for negative content and one safe class labeled with the name whitelist as data containing safe text from negative content. By using a dataset of 2628 lines of data containing text from each website link, which is then divided into training data and test data. The dataset is arranged in a table with 2 columns, namely the text class category column and the contents of the text from the website that has been scraped (see Table 1).

Table 1. Dataset overview

	Category	Text
0	Gambling	Djarumsoccer adalah Situs Judi Online Terpercaya, Terbaik Di Indonesia yang menyediakan berbagai jenis permainan judi online secara lengkap seperti Sportsbok IBCbet, Live Kasino, Slot Games, Poker, Domino QQ, Ceme, Togel, Sabung Ayam, Fish World dan lainnya.
1	Gambling	Register | Deposit | Withdraw | Bonus | Hubungi Kami Copyright Â© 2010 Sports369, All Rights Reserved SPORTS369 Adalah Situs Judi Online Resmi Terbesar Terbaik di Indonesia yang menyediakan beragam jenis Games untuk Taruhan Bola Online, Casino Online, Mesin Slot, Poker, Togel, Tangkas dan sebagainya menggunakan Uang Asli.
2	Gambling	Cara Mendapatkan Banyak Referral Di Situs Judi Online â€“ Pada kesempatan kali ini saya akan memberikan kepada bettor beberapa cara Read More Tips Agar Bisa Merasakan Kesenangan Bermain DominoQQ â€“
3	Whitelist	Victoria Guardiola dibesarkan sebagai anak lakilaki dengan nama Victor. Dia mulai merasakan ada sesuatu yang "tidak beres" pada usia yang masih sangat muda. Di umur 17 tahun, Victor pun baru berani memutuskan untuk memulai perawatan hormonal dan mengganti namanya menjadi Victoria. Pada bulan Maret 2010 lalu, atau empat tahun setelah dia memulai perawatan, dia lalu menjalani menjalani operasi besar pertamanya, vaginoplasty.

2.2 Preprocessing

In data preprocessing, duplicated data, text cleansing, case folding, stopward, tokenize, label encoding, data splitting, and determine the TF-IDF will be carried out.

2.2.1. Duplicated Data

Duplicate data is data that exists more than one in the dataset. This will cause problems in model training and should be removed to maintain model quality.

2.2.2. Text Cleansing

Text cleansing Is the step to get rid of characters that are not needed from the text dataset. Characters such as numbers, non-ascii, and symbols are removed, leaving only words.

2.2.3. Case Folding

Case Folding is an important pre-processing process to be carried out on text datasets. By using Case Folding, all text is converted to the same lower case, so there is no difference in the values between uppercase and lowercase which can affect the prediction results.

2.2.4. Stopwords

Stopwords are words that are ignored in the machine learning process because they are too commonly used and need to be removed. Stopwords can be found in the existing stoplists for each language or can be added as needed.

2.2.5. Label Encoding

Label encoding is a preprocessing technique used to convert sequential number categorical data. As input labels in machine learning.

2.2.6. Data Splitting

Data splitting is the process of dividing the data into training data and test data with a specified ratio of 70:30. The data will be used to train and validate the model.

2.3 Term Frequency Inverse Document Frequency

A preprocessing technique called Term Frequency Inverse Document Frequency (TF-IDF) weighs words or groupings of words that are most frequently used in documents. TF gives weight to words in a document and IDF reduces the weight of words if they appear widely in various documents. To be machine understandable, we need to transform the data into vectors that represent each word used through pre-processing that measures the occurrence of the word in the document. This is a process known as the weighting process [30], [31]. Weighting using TF-IDF is the most common technique that is often applied.

To calculate TF-IDF use the following formula:

$$TF - IDF_{(t,d)} = TF_{(t,d)} \; x \; IDF_{(d)} \tag{1}$$

TF measures how often a term occurs in a document. Not all documents are the same length, so it's always possible for terms to appear more frequently in longer documents than in shorter ones.

$$TF_{(t,d)} = \frac{Frequency\ of\ term\ t, in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \tag{2}$$

IDF is used to measure the importance of terms in a document. All terms are considered equally important in calculating TF. However, certain terms such as "when", "that is", and "at". Although common, does not carry much meaning.

$$IDF_{(d)} = \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it} \tag{3}$$

2.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is an optimization-based learning technique that employs fictitious spaces represented as linear functions in high-dimensional data [32]. SVM is widely used in binary classification and regression problems [33]. Additionally, SVM is a family of strong classifications that have proved effective in a variety of natural language processing tasks [34]. SVMs, or support vector machines, are efficient machine learning techniques. SVM uses a classification technique to group data into two categories. When given labeled training data, SVM can be used to classify new data. SVM is optimal when using limited data. The main function of SVM is to classify data that can be separated linearly, and has better and more efficient performance than other methods such as ANN when dealing with large data [35]. SVM is designed to determine the best line in n-dimensional space to separate data into different groups, so that new data can be classified accurately in the future [36].

2.5 Random Forest (RF)

Random Forest (RF) is a machine learning method that combines multiple classification trees to classify data [37]. This algorithm starts with selecting many samples from the data which is called a bootstrap sample. In each sample, most of the observations from the original data will appear at least once. Each classification tree is formed for each sample using only a small portion of the total available variables. Each tree was fully grown and used to predict observations that were not included in the sample (out-of-bag observations). The predicted class for an observation is calculated based on the majority of the tree predictions for that observation, with the condition of the series determined randomly [38]. By growing individual trees to very deep levels (typically one observation per terminal node) and using the average of the tree predictions, Random Forests reduce the issue of high variation in predictions common to tree-based methods while maintaining the benefits of tree-based methods, such as the ability to capture complex interactions and keep bias low [39].

2.6 Evaluation

To be able to use or disseminate the model that has been made, the system needs to be tested and assessed for its performance first using data that is not used during training or testing data. This can include comparing with experts in the field, as well as choosing the right metrics to use in the comparisons. Many models can be made, but it is not certain which model is good and which is bad from the many models that have been built, which are able to provide the desired results in accordance with a given set of inputs. The evaluation process

is used to determine model performance because more than one model can provide the expected results and may be more suitable for certain needs. Evaluation of a classification model can be done using a classification report, which makes it easier to evaluate metrics such as recall, F1-score, and support [40]. A high accuracy score does not always guarantee the validity of the model, so it needs to be evaluated using other metrics such as: accuracy, recall, precision, F1-score obtained from the confusion matrix (see Fig. 2).

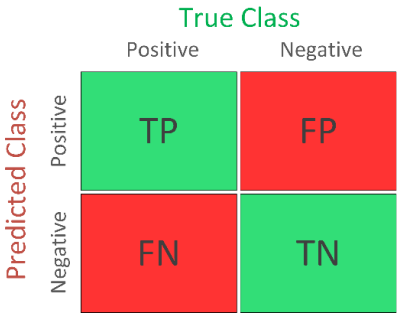

Fig. 2. Confusion Matrix

- True Positive (TP): the model correctly predicts the positive class.
- True Negative (TN): the model correctly predicts a negative class.
- False Positive (FP): the model incorrectly predicts the positive class.
- False Negative (FN): the model incorrectly predicts a negative class.

Accuracy is the ratio between the number of correct predictions and the total number of predictions. Precision is defined as the proportion of TP value with the number of TP and FP. Recall is defined as the proportion of TP value with the number of TP and FN. F1-score is the harmonic average of precision and memory. The closer the F1 score is to 1, the better the performance of the model. Accuracy, recall, precision, and F1-score values can be determined by:

- Accuracy $= \frac{TP+TN}{TP+TN+FP}$
- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$
- F1 score $= 2 \times \frac{(Recall\ x\ Precision)}{(Recall + Precision)}$

3. RESULTS AND DISCUSSION

3.1. Model Performance Results using Support Vector Machine (SVM)

Table 2 shows the results of the performance evaluation obtained by machine learning SVM which is trained using text data that has 4 classes to get an accuracy of 0.97. from the confusion matrix shown by this model has high accuracy based on a large number of TP (True Positive) so as to obtain high accuracy. In Table 3 it can be seen that the metric values for each class, the results of the SVM for each class show high scores in most every class. The results of all macro metrics for SVM can be seen in Table 4.

Table 2. SVM Confusion Matrix

		Predicted Value			
		Gambling	Pornography	Fraud	Whitelist
Actual Value	Gambling	172	5	1	2
	Pornography	5	124	5	0
	Fraud	2	2	15	0
	Whitelist	0	0	0	193
Accuracy			0.97		

Table 3. SVM Class metrics

Class	**TP**	**FN**	**FP**	**TN**	**Accuracy**	**Precision**	**Recall**	**F1**
Gambling	172	8	7	339	0.97	0.95	0.96	0.95
Pornography	124	10	7	385	0.96	0.94	0.92	0.93

Fraud	15	4	6	501	0.98	0.71	0.78	0.75
Whitelist	195	0	2	331	0.99	0.98	1	0.99

Table 4. Macro Metrics SVM

Metric	Score
Accuracy	0.979087452
Precision	0.902872011
Recall	0.917600594
F1	0.909727922

3.2. Model Performance Results using Random Forests

Table 5 shows the results obtained from the evaluation results of the Random Forest machine learning performance. The results obtained are not much different from the previous SVM model, this can be seen from the number of TP (True Positive) which is also quite high. However, it can be seen that in the gambling class there is a mismatch in the fraud class with a total of 22 and also in the pornography class there is a mismatch with a total of 30 fraud which can affect precision. The results of class metrics and all macro metrics for random forest can be seen in Table 6 and Table 7.

Table 5. Random Forest Confusion Matrix

		Predicted Value			
		Gambling	Pornography	Fraud	Whitelist
Actual Value	Gambling	153	2	22	3
	Pornography	2	102	30	0
	Fraud	1	0	18	0
	Whitelist	13	3	4	173
Accuracy			0.92		

Table 6. Random Forest Class Metrics

Class	TP	FN	FP	TN	Accuracy	Precision	Recall	F1
Gambling	153	27	16	330	0.91	0.90	0.85	0.87
Pornography	102	32	5	387	0.92	0.95	0.76	0.84
Fraud	18	1	56	451	0.89	0.24	0.94	0.38
Whitelist	173	20	3	330	0.95	0.98	0.89	0.93

Table 7. Macro Metrics Random Forest

Metric	Score
Accuracy	0.923954373
Precision	0.771198565
Recall	0.863733877
F1	0.762007503

3.3. Comparison of SVM and RF Models

Based on the test results at Table 2 and Table 5, SVM gets a higher score than the Random Forest. In this case the SVM model works quite well in classifying, while the number of misclassifications in the Random Forest model for two classes, namely the gambling and pornography classes is quite high and affects the precision of the model.

Fig. 3 shows a graphic comparison of performance metrics from SVM and RF using negative content datasets from websites. The results show that both algorithms achieve high scores in each evaluation matrix. The metric from SVM shows an accuracy of 97%, 90% precision, 91% recall and 90% F1 (see Table 4), and the metric results from RF show an accuracy of 92% accuracy, 77% precision, 86% recall and F176% . Therefore, in the detection of negative content based on text from this website it can be said that SVM is better than RF, where the accuracy, precision, recall and F1 values of SVM are all higher than RF.

In the study of text classification, SVM is better than other classifiers including RF. This is the result of the theoretical guarantee that SVM provides for the best classification performance. Regarding the parameters to be considered while choosing a text classification method, SVM is increasingly being observed to perform better than other classifiers in text mining applications such as text categorization and text filtering. [19][20][29][41]. In general, SVM often performs better for linearly separable problems with high-dimensional

data, whereas random forests are better suited for problems with complex interactions between features and classes. However, the most effective strategy will depend on the unique features of the problem and the data set used, hence it is often necessary to try both approaches and compare their effectiveness empirically [42].

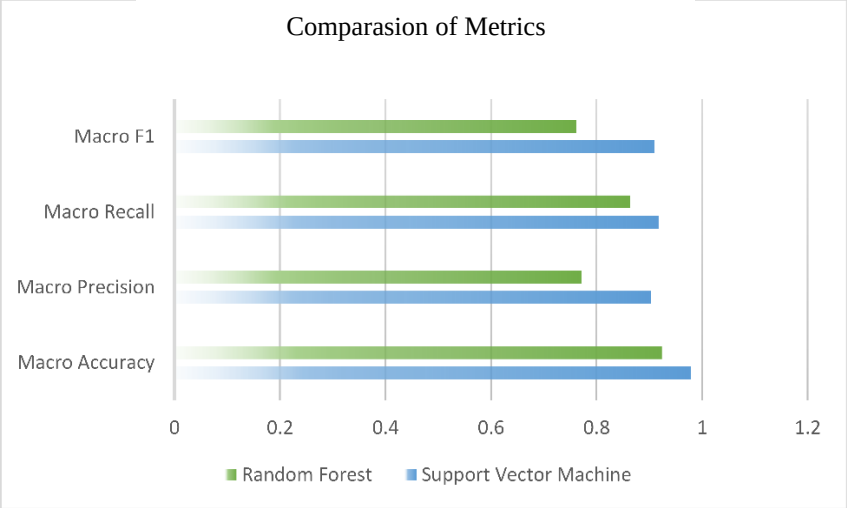

Fig. 3. Comparison of metrics based on the model used

4. CONCLUSION

Detection of negative content (with labels of fraud, pornography and gambling) on websites using SVM and Random Forest can work very well. The test results show that both SVM and Random Forest models can classify text with negative content from websites. The SVM algorithm is better than Random Forest in classifying negative content which can be seen from the results of the performance evaluation metrics obtained from SVM, namely 97% accuracy, 90% precision, 91% recall and 90% f1, while in Random Forest an accuracy score of 92%, 77% precision, 86% recall and 76% f1. Thus it can be concluded that SVM works better than Random Forest in detecting negative content on Indonesian language websites. In the future, we can add more classifications and feature extraction methods to this experimental investigation. This can be expanded to test a variety of regional and foreign languages.

REFERENCES

[1] A. J. Scheerder, A. J. A. M. van Deursen, and J. A. G. M. van Dijk, "Negative outcomes of Internet use: A qualitative analysis in the homes of families with different educational backgrounds," *Inf. Soc.*, vol. 35, no. 5, pp. 286–298, 2019, https://doi.org/10.1080/01972243.2019.1649774.

[2] P. R. Bazán *et al.*, "Can news with positive or negative content affect and a relaxation pause improve the emotional state of health care professionals? A randomized online experiment during COVID-19 pandemic," *Internet Interv.*, vol. 26, 2021, https://doi.org/10.1016/j.invent.2021.100441.

[3] S. Abdulmana and B. Saleh, "Coordinate negative content filtering and threat detection in Thailand on the Internet infrastructure," *2014 5th Int. Conf. Inf. Commun. Technol. Muslim World, ICT4M 2014*, pp. 1-5, 2014, https://doi.org/10.1109/ICT4M.2014.7020647.

[4] N. Ulfah, N. O. Irawan, P. D. Nurfadila, P. Y. Ristanti, and J. A. . Hammad, "Blocking pornography sites on the internet private and university access," *Bull. Soc. Informatics Theory Appl.*, vol. 3, no. 1, pp. 22–29, 2019, https://doi.org/10.31763/businta.v3i1.161.

[5] A. Amalia, D. Gunawan, M. S. Lydia, and Wesley, "The Identification of Negative Content in Websites by Using Machine Learning Approaches," *5th Int. Conf. Comput. Eng. Des. ICCED 2019*, pp. 1-6, 2019, https://doi.org/10.1109/ICCED46541.2019.9161105.

[6] K. Vo *et al.*, "Handling negative mentions on social media channels using deep learning," *J. Inf. Telecommun.*, vol. 3, no. 3, pp. 271–293, 2019, https://doi.org/10.1080/24751839.2019.1565652.

[7] H. Tatsat, S. Puri, and B. Lookabaugh, *Machine Learning and Data Science Blueprints for Finance*. O'Reilly Media, Inc., 2020, https://books.google.co.id/books?id=rZwAEAAAQBAJ.

[8] S. Dridi, V. Machine, D. Tree, R. Forest, and L. Regression, "Supervised Learning - A Systematic Literature Review," 2021, https://osf.io/tysr4.

[9] K. Bock and S. M. Garnsey, "Advances in Language Processing," *A Companion to Cogn. Sci.*, pp. 226–234, 2008, https://doi.org/10.1002/9781405164535.ch14.

[10] U. Arbieu, K. Helsper, M. Dadvar, T. Mueller, and A. Niamir, "Natural Language Processing as a tool to evaluate emotions in conservation conflicts," *Biol. Conserv.*, vol. 256, p. 109030, 2021,

https://doi.org/10.1016/j.biocon.2021.109030.
[11] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *npj Digit. Med.*, vol. 5, no. 1, pp. 1–13, 2022, https://doi.org/10.1038/s41746-022-00589-7.
[12] S. Lia, "Detection of Negative Content (Hoax) on Microblog Data that Contains Covid-19 Information," *J. Ilm. Indones.*, vol. 7, no. 8.5.2017, pp. 2003–2005, 2022, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders.
[13] H. Yenala, A. Jhanwar, M. K. Chinnakotla, and J. Goyal, "Deep learning for detecting inappropriate content in text," *Int. J. Data Sci. Anal.*, vol. 6, no. 4, pp. 273–286, 2018, https://doi.org/10.1007/s41060-017-0088-4.
[14] A. Baccouche, S. Ahmed, D. Sierra-Sosa, and A. Elmaghraby, "Malicious text identification: Deep learning from public comments and emails," *Inf.*, vol. 11, no. 6, 2020, https://doi.org/10.3390/info11060312.
[15] A. F. Hidayatullah, A. M. Hakim, and A. A. Sembada, "Adult content classification on Indonesian tweets using LSTM neural network," *2019 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2019*, pp. 235–240, 2019, https://doi.org/10.1109/ICACSIS47736.2019.8979982.
[16] Y. Chen, R. Zheng, A. Zhou, S. Liao, and L. Liu, "Automatic detection of pornographic and gambling websites based on visual and textual content using a decision mechanism," *Sensors (Switzerland)*, vol. 20, no. 14, pp. 1–21, 2020, https://doi.org/10.3390/s20143989.
[17] N. Izzah, I. Budi, and S. Louvan, "Classification of pornographic content on Twitter using support vector machine and Naive Bayes," *2018 4th Int. Conf. Comput. Technol. Appl. ICCTA 2018*, pp. 156–160, 2018, https://doi.org/10.1109/CATA.2018.8398674.
[18] T. B. Adji, Z. Abidin, and H. A. Nugroho, "System of negative Indonesian website detection using TF-IDF and Vector Space Model," *Proc. 2014 Int. Conf. Electr. Eng. Comput. Sci. ICEECS 2014*, pp. 174–178, 2014, https://doi.org/10.1109/ICEECS.2014.7045240.
[19] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Comput. Sci.*, vol. 127, pp. 511–520, 2018, https://doi.org/10.1016/j.procs.2018.01.150.
[20] N. Nurhasanah, D. E. Sumarly, J. Pratama, I. T. K. Heng, and E. Irwansyah, "Comparing SVM and Naïve Bayes Classifier for Fake News Detection," *Eng. Math. Comput. Sci. J.*, vol. 4, no. 3, pp. 103–107, 2022, https://doi.org/10.21512/emacsjournal.v4i3.8670.
[21] E. K. Andana, M. Othman, and R. Ibrahim, "Comparative analysis of text classification using naive bayes and support vector machine in detecting negative content in Indonesian twitter," in *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.3 S1, pp. 356–362, 2019, https://doi.org/10.30534/ijatcse/2019/6481.32019.
[22] M. G. Hussain, M. Rashidul Hasan, M. Rahman, J. Protim, and S. Al Hasan, "Detection of Bangla Fake News using MNB and SVM Classifier," *Proc. - 2020 Int. Conf. Comput. Electron. Commun. Eng. iCCECE 2020*, pp. 81–85, 2020, https://doi.org/10.1109/iCCECE49321.2020.9231167.
[23] M. Z. Islam, J. Liu, J. Li, L. Liu, and W. Kang, "A semantics aware random forest for text classification," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 1061–1070, 2019, https://doi.org/10.1145/3357384.3357891.
[24] C. Sun and B. Luo, "Analysis of English Writing Text Features Based on Random Forest and Logistic Regression Classification Algorithm," *Mob. Inf. Syst.*, 2022, https://doi.org/10.1155/2022/6306025.
[25] A. Bouaziz, C. Dartigues-Pallez, C. Da Costa Pereira, F. Precioso, and P. Lloret, "Short text classification using semantic random forest," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8646 LNCS, pp. 288–299, 2014, https://doi.org/10.1007/978-3-319-10160-6_26.
[26] B. P. O. Lovatti, M. H. C. Nascimento, Á. C. Neto, E. V. R. Castro, and P. R. Filgueiras, "Use of Random forest in the identification of important variables," *Microchem. J.*, vol. 145, pp. 1129–1134, 2019, https://doi.org/10.1016/j.microc.2018.12.028.
[27] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 2733–2742, 2022, https://doi.org/10.1016/j.jksuci.2022.03.012.
[28] P. Agarwal, A. P. H, and N. Bhatt, "News Text Classification Using Machine Learning Algorithms," *Elem. Educ. Online*, vol. 20, no. 02, pp. 2659–2665, 2021, https://doi.org/10.17051/ilkonline.2021.02.282.
[29] R. A. I. Berliana Putri Meliani, Oktariani Nurul Pratiwi, "Comparison of Support Vector Machine and Random Forest Algorithms in Sentiment Analysis on Covid-19 Vaccination on Twitter using Vader and Textblob Labelling," in *Proceedings of the International Conference on Applied Science and Technology on Social Science 2022*, vol. 10, no. 1, p. 1, 2022, https://doi.org/10.30595/juita.v10i1.12394.
[30] H. Syahputra, "Sentiment Analysis of Community Opinion on Online Store in Indonesia on Twitter using Support Vector Machine Algorithm (SVM)," *J. Phys. Conf. Ser.*, vol. 1819, no. 1, 2021, https://doi.org/10.1088/1742-6596/1819/1/012030.
[31] H. Syahputra, L. K. Basyar, and A. A. S. Tamba, "Setiment Analysis of Public Opinion on the Go-Jek Indonesia Through Twitter Using Algorithm Support Vector Machine," *J. Phys. Conf. Ser.*, vol. 1462, no. 1, 2020, https://doi.org/10.1088/1742-6596/1462/1/012063.
[32] L. Auria and R. A. Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis," *SSRN Electron. J.*, 2011, https://doi.org/10.2139/ssrn.1424949.
[33] R. Rodriguez-Pérez, M. Vogt, and J. Bajorath, "Support vector machine classification and regression prioritize

different structural features for binary compound activity and potency value prediction," *ACS Omega*, vol. 2, no. 10, pp. 6371–6379, 2017, https://doi.org/10.1021/acsomega.7b01079.
[34] F. Trevor Hastie Robert, Tibshirani Jerome, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. vol. 2, pp. 1-758, 2009, https://doi.org/10.1007/978-0-387-21606-5.
[35] S. V. N. Vishwanathan and M. N. Murty, "SSVM: A simple SVM algorithm," *Proc. Int. Jt. Conf. Neural Networks*, vol. 3, no. 1, pp. 2393–2398, 2002, https://doi.org/10.1109/IJCNN.2002.1007516.
[36] P. Chhajer, M. Shah, and A. Kshirsagar, "The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction," *Decis. Anal. J.*, vol. 2, no. June 2021, p. 100015, 2022, https://doi.org/10.1016/j.dajour.2021.100015.
[37] Y. Kong and T. Yu, "A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, 2018, https://doi.org/10.1038/s41598-018-34833-6.
[38] D. R. Cutler *et al.*, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007, https://doi.org/10.1890/07-0539.1.
[39] P. R. Cole Brokampa, Roman Jandarov, M.B. Rao, Grace LeMastersb, "Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches Cole," *Atmos Env.*, vol. 151, no. 1, pp. 1–11, 2017, https://doi.org/10.1016/j.atmosenv.2016.11.066.
[40] B. S. F. Astuti, N. A. Firdausanti, and S. W. Purnami, "Model Evaluation for Logistic Regression and Support Vector Machines in Diabetes Problem," *Inferensi*, vol. 1, no. 2, p. 77, 2018, https://doi.org/10.12962/j27213862.v1i2.6728.
[41] S. G. Kanakaraddi, A. K. Chikaraddi, K. C. Gull, and P. S. Hiremath, "Comparison Study of Sentiment Analysis of Tweets using Various Machine Learning Algorithms," *Proc. 5th Int. Conf. Inven. Comput. Technol. ICICT 2020*, pp. 287–292, 2020, https://doi.org/10.1109/ICICT48043.2020.9112546.
[42] A. Sabat-Tomala, E. Raczko, and B. Zagajewski, "Comparison of support vector machine and random forest algorithms for invasive and expansive species classification using airborne hyperspectral data," *Remote Sens.*, vol. 12, no. 3, 2020, https://doi.org/10.3390/rs12030516.

BIOGRAPHY OF AUTHORS

Hermawan Syahputra, He is a lecturer in Computer Science Department of Universitas Negeri Medan. He received the B.Sc (2003) in Department of Mathematics in North Sumatera University (USU), M.Sc (2009) in Department of Computer Science in Bogor University (IPB), Ph.D (2016) in Department of Computer Science from Gadjah Mada University (UGM). He works since 2003, a lecturer in the Department of Computer Science and Mathematic of Universitas Negeri Medan. His areas of expertise are image processing, pattern recognition, artificial intelligence and computer vision. He has 3 in H-index with Scopus, and 5 H-index with Google scholar. Email: hsyahputra@unimed.ac.id.

Aldiva Wibowo, He received the B.Sc (2023) in Department of Computer Science in Universitas Negeri Medan. His experience includes having been an Associate participant in the Data Scientist Training Held by the Directorate of Higher Education and the University of Indonesia (2021) and a participant in the BISA AI Academy program (2021). Email: aldivaawibowo@gmail.com.