Journal of information and
communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 23(2): 94-100, Jun. 2025 Regular paper

Machine Learning-Based Fake Account Detection System: Instagram Case Study

Yulia[1], Hendy Gunawan[1], Gregorius Satia Budhi[1]*, and Kartika Gunadi Kartawidjaja[1]

[1]Informatics Department, Petra Christian University, Surabaya 60236, Indonesia

Abstract

People often create fake social media accounts to express themselves anonymously. However, these fake accounts can harm the reputation of individuals and businesses, resulting in fewer genuine likes and followers. Instagram, a top-rated social media platform often used for business and political engagement, suffers from the negative impacts of these accounts. This highlights the urgent need for a dependable system to identify whether Instagram accounts are genuine. This study investigated several machine learning models for developing a fake account detection system. Single models, such as support vector machines, naïve Bayes, logistic regression, multilayer perceptron, and ensemble models based on bootstrap aggregating techniques and boosting, were trained and tested. The training and testing processes were conducted using a 10-fold cross-validation to prevent overfitting. The test results indicated that the adaptive and gradient boosting models achieved the best accuracy and an F1 score of more than 92%, with precision surpassing 93%.

Index Terms: Fake account detection, Machine learning, Single and ensemble models, Social media

I. INTRODUCTION

Many individuals endeavor to increase their follower count for various reasons, such as seeking fame or earning trust from others based on a large follower count [1]. Consequently, individuals create fake accounts to inflate their follower counts and use platforms for malicious activities, such as fraud and cyberbullying [2,3]. Furthermore, individuals create fake accounts to express themselves, exploit social media, and engage in other online activities without revealing their true identities to others [4].

Fake accounts pose problems for business owners who use influencers to promote their products. The influencers are paid using endorsements. The total number of influencer followers determines the endorsement process. It is crucial to recognize that this number can be artificially inflated by up to 78% by using fictitious followers (fake accounts). Such manipulation distorts the influencer's genuine value and influence, resulting in business owners potentially overpaying for their endorsements [5]. The creation of fake accounts under false identities can be detrimental to the reputations of individuals and businesses, leading to a decrease in genuine likes and followers [1].

Instagram is one of the most active social media platforms worldwide [2,5]. It is used to share images and creative work for communication [1]. Over time, Instagram's role in social media has evolved. In addition to being a communication medium, Instagram is used for business and political purposes. Many celebrities have recently created Instagram accounts to develop their businesses and fan bases [6]. All types of fake accounts adversely affect social media benefits. This underscores the critical need for a reliable system to detect whether an Instagram account is fake. Real accounts are those in which the account owners utilize their real identity to make them

Received 19 November 2024, **Revised** 16 March 2025, **Accepted** 1 April 2025
***Corresponding Author** Gregorius Satia Budhi (E-mail: greg@petra.ac.id)
Petra Christian University, Informatics Department, Siwalankerto 103-144, Surabaya, East Java, 60236, Indonesia

Open Access https://doi.org/10.56977/jicce.2025.23.2.94 print ISSN: 2234-8255 online ISSN: 2234-8883

easily recognizable. This includes full names, short names, biographies, and profile pictures [7]. Such a system could provide comfort and security to Instagram users through social media interactions, particularly on Instagram.

A previous study by Albayati and Altamimi in 2019 [8] aimed to address this issue by utilizing data mining techniques to detect fake profiles on Facebook. The study proposed three supervised learning algorithms (k-nearest neighbor (K-NN) [9], support vector machines (SVMs) [10], and decision tree (DT)) [11] and two unsupervised learning algorithms, k-means [12] and k-medoids [13]. The study reported that the DT, SVMs, and K-NN with k = 3 achieved accuracies of 97.76%, 95.72%, and 91.45%, respectively. The unsupervised learning algorithms, k-means and k-medoids, achieved accuracies of 67.31% and 67.01%, respectively. In this study, the supervised learning algorithms outperformed the unsupervised learning algorithms. This study did not utilize cross-validation (CV).

In 2020, Sheikhi [1] conducted a study to identify the most efficient method for detecting fake accounts on Instagram. Various algorithms were employed in this study, including the Hoeffding tree [14], random forest (RF) [15], SVMs, naïve Bayes (NB) [16], multilayer perceptron (MLP) [17], and bagging predictors (BP) [18]. The study reported that BP achieved the highest accuracy of 98.45%, followed by RF, NB, and SVMs with accuracies of 97.2, 94.58, and 68.68%, respectively. This experiment was conducted using a 10-fold CV. The results of the study indicated that the BP method can accurately detect fake accounts.

In 2020, Purba et al. [5], conducted a study on classifying fake Instagram users. This study aimed to classify fake users using supervised learning algorithms such as RF, MLP, logistic regression (LR) [19], NB, and DT. Experiments were conducted using 2-classes (fake or authentic users) and 4-classes (authentic users, active-fake users, inactive-fake users, and spammers) classifications. In the 2-class classification, the RF achieved the highest accuracy of 90.1%. In the 4-class classification, the RF achieved the highest accuracy of 91.8%. This experiment was conducted using a 10-fold CV.

Based on the outlined problem background, the research questions were as follows: Q1: To identify which machine learning algorithms are most suitable for detecting fake accounts on Instagram, as measured by accuracy, precision, recall, and F1-score; Q2: What are the criteria for determining fake accounts on Instagram? An additional question was as follows: Q3: What if machine learning experiences overfitting?

This study examined the efficacy of single-model machine learning (ML), such as SVMs, NB, LR, and MLP, and ensemble models based on bootstrap aggregation techniques (RF and BP) and boosting techniques, such as Adaptive Boosting (AB) and Gradient Boosting (GB) to identify the most suitable model for detecting fake accounts on Instagram. The performance of each algorithm was evaluated using metrics such as accuracy, recall, precision, and F1-score. We evaluated the performance of the ML models using k-fold CV. This analysis aimed to ascertain the stable performance of each tested algorithm.

II. SYSTEM MODEL AND METHODS

A. Comparison Framework to Identify the Best Model

We designed a comparison framework to identify the best Instagram fake account detection model. Our previous studies

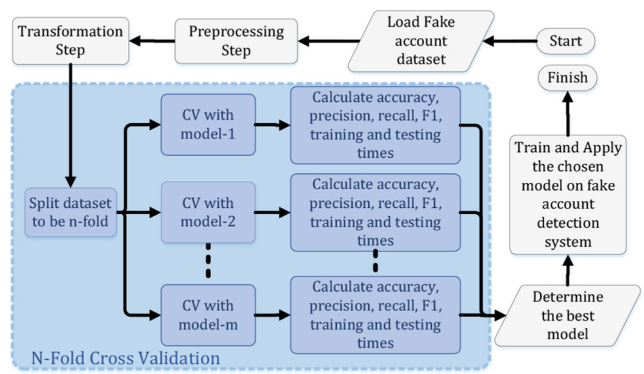

Fig. 1. Design of comparison framework.

inspired this framework [20,21]. The design is illustrated in Fig. 1.

1) Dataset

The dataset used in this study was obtained from the website kaggle.com, which was created by Bakhshandeh[1]. The dataset comprised 696 records, with 348 records labelled as fake or spammer accounts and 348 records labelled as genuine accounts, and consisted of 12 attributes. Although small, this well-balanced and carefully labeled dataset was ideal for our purposes. Furthermore, it exceeds the 10-times rule, which recommends at least 10 examples for each feature in every class [22].

2) Preprocessing Step

Preprocessing is a crucial step in data analysis and ML as it prepares raw data for further processing and analysis. In this study, the dataset underwent several preprocessing steps to ensure its quality and suitability for training the ML models. The preprocessing steps are as follows.

1) https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts

(1) All missing values were identified, removed, or imputed using appropriate methods to ensure completeness, integrity, and data quality [23].
(2) A new attribute was added to the dataset to represent the ratio of followers to followings. According to articles on social media [24,25], this ratio indicates accounts' level of engagement. A higher ratio indicates that the accounts are of better quality.
(3) The numerical data in the dataset were grouped into categories. This grouping simplified the analysis and made it easier to understand for specific ML algorithms [23].
(4) The final step was to change the dataset from comma-separated values (CSV) to a dataframe format. This change makes it easier to analyze data using tools and libraries designed for data processing [26].

3) Transformation Step

Data transformation is crucial for modifying data by altering their formats within a dataset. This step ensures that the data are suitable for subsequent classification processes. The data transformations performed at this stage can be categorized into form and value transformations. The detailed data transformation process is outlined as follows.

(1) Value Transformation: This process involves modifying the dataset to add new data attributes derived from calculations based on existing attributes. The value transformation performed in this study included adding the followers/followings ratio attributes. This transformation aims to determine how frequently an account follows other accounts and is followed in return. A higher value ratio indicates a higher quality account. This ratio is expressed in (1):

$$\text{Followers\_Followings\_Ratio} = \frac{\text{num of followers}}{\text{num of followings}} \quad (1)$$

(2) Form Transformation: This process involves modifying numerical and categorical attributes. This simplifies the data analysis process and facilitates a better understanding of the various ML algorithms [23]. The applications of each attribute are listed in Table 1.

4) Training and Testing of the Model

The initial stage of this process involves dividing the dataset into training and testing sets. Subsequently, the training data were processed using four methods. After training the data, an ML model was obtained from the trained data. Subsequently, the model was tested using the testing data, and its performance was evaluated using a confusion matrix with metrics such as accuracy, recall, precision, and F1-score. A 10-fold CV was performed. After all the models were trained, overfitting tests were performed on each method. To detect fake accounts, we investigated four single ML models, the SVMs linear kernel [27], NB [28], LR [19], and MLP [29,30], and four ensemble models, BP [18], RF [15], AB [31], and GB [32].

Table 1. Data transform from numerical to categorical

No.	Attribute	Rule	Categorical Value
1	Description Length (DL)	DL < 50 50 ≤ DL < 100 DL ≥ 100	Low Middle High
2	Username Length (UL)	UL < 0.3 0.3 ≤ UL < 0.6 UL ≥ 0.6	Low Middle High
3	Fullname Length (FL)	FL < 0.3 0.3 ≤ FL < 0.6 FL ≥ 0.6	Low Middle High
4	Fullname Words (FW)	FW < 4 4 ≤ FW < 8 FW ≥ 8	Low Middle High
5	Post (P)	P < 50 50 ≤ P < 100 P ≥ 100	Low Middle High
6	Followers (Fle)	Fle < 300 Fle ≥ 300	Low High
7	Following (Fli)	Fli < 500 Fli ≥ 500	Low High
8	Fle–Fli Ratio (FFR)	FFR < 0.5 0.5 ≤ FFR < 1 1 ≤ FFR < 2 2 ≤ FFR < 10 FFR ≥ 10	Very Bad Bad Normal Good Very Good

B. Design System

The best model found using the analysis presented in Section 2 was then trained using all the records in the dataset and applied to the fake account detection system. The design of this system is illustrated in Fig. 2.

The proposed system is straightforward. First, the user can insert a suspected username. The system runs a scrapping module to gather the metadata of the suspected username from Instagram. The metadata are then transformed into features, as listed in Table 1. Subsequently, the system detects whether the account is fake. The system shows the account details if the suspected account is the genuine account. If detected as fake, it runs a warning in addition to showing the account details. Although this design is technically feasible, individuals interested in its implementation should consider the equipment that will be used. It must be sufficiently robust to handle large amounts of data. Additionally, it is essential to be mindful of Instagram users' privacy.

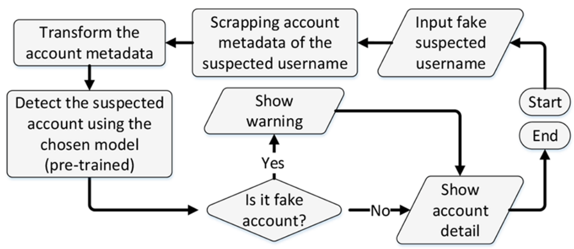

Fig. 2. Design of Fake Account Detection System.

III. RESULTS

The experiments were conducted to determine which model among the proposed models was the most suitable for achieving high classification performance.

A. Performance Measures

We investigated model candidates using 10-fold CV in the first experimental group. The process involved repeating the iteration 10 times for training and testing. In each iteration, one subset was excluded from testing, whereas the remaining subsets were used for training. The purpose of using a 10-fold CV for testing each model was to perform 10 iterations to avoid overfitting. Additionally, CV was used to estimate the performance of the model in ML using data that had not been previously reported. Table 2 presents the results of the study.

Table 2. Result of 10-fold cross-validation test

Model	Measurement (%)				Time (ms)	
	Acc	Pre	Rec	F1	Train	Test
SVMs	90.7	92.7	88.1	90.3	1.196	0.598
NB	83.6	76.8	96.3	85.4	1.562	1.562
LR	91.7	91.8	91.5	91.5	24.029	4.032
MLP	91.9	92.7	91.7	92	788.035	6.001
RF	91.7	92.9	90.5	91.6	37.503	1.795
AB	92.5	93.9	90.5	92.1	13.851	1.561
BP	90.7	91.4	90.1	90.4	50.962	5.96
GB	92.4	93.2	91.5	92.3	89.007	5.006

The results in Table 2 indicate that AB and GB have the best accuracies of 92.5% and 92.4%, respectively. For precision, AB achieved the best results (94%), whereas that of GB was slightly lower (93.2%). Furthermore, NB achieved the best recall but the worst accuracy and precision. This means that NB can detect the first class (fake) better than the other models. However, because its precision was low (76.8%), many mistakes were made when detecting the second class (genuine). This causes inconvenience to Instagram users because NB detects many genuine accounts as fake.

B. Underfitting or Overfitting Test

To analyze the performance of ML models, we must evaluate whether the models are overfitting or underfitting. Overfitting is a condition in which the trained model performs extremely well on the training data and does not fit well with the testing data. Therefore, when the error rates are low for the training dataset and high for the testing dataset, overfitting occurs [33]. Underfitting occurs when the trained model performs poorly on the training and testing data. Technically, underfitting occurs when the error rates are high for both the training and testing data [33]. An over- or underfitting model cannot be considered a good fit. Each model's mean squared error (MSE) was examined when tested with both training and testing data to determine whether the model candidates were over- or underfitting. Table 3 presents the results of the study.

Table 3. Over- and under-fitting tests on model candidates

Model	MSE Score		
	Train Data	Test Data	Difference
SVMs	0.089	0.101	0.012
NB	0.163	0.145	-0.018
LR	0.07	0.101	0.031
MLP	0.065	0.072	0.007
RF	0.046	0.087	0.041
AB	0.069	0.087	0.018
BP	0.054	0.072	0.018
GB	0.059	0.029	-0.03

Table 3 shows that all model candidates' MSE scores of testing with both training and testing data are low. Nearly all of them are below 0.01, except for the MSE of NB and the MSE of SVMs testing data, which are slightly higher. This implies that all model candidates did not suffer from underfitting because they all fit well to the problem. Furthermore, the difference in MSE scores between training and testing data is minimal (below 0.05). Therefore, we can conclude that none of the model candidates experience overfitting and that they can learn the data to determine the patterns accurately.

C. Criteria for Fake Account Detection

By determining the criteria for identifying fake accounts on Instagram, we tested and analyzed the importance of features using the AB model. The AB was selected because it demonstrated superior performance among the tested model candidates. Feature Importance (FI) is a technique that calculates the scores for all model input features. A higher score indicates that a feature significantly affects the model in predicting a specific variable. The feature importance in AB is illustrated in Fig. 3.

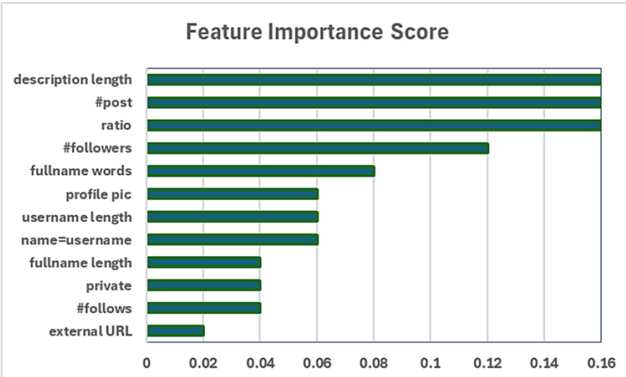

Fig. 3. Feature importance scores of AB input features.

In Fig. 3, we can observe that the three features with the highest scores are description length, number of posts (#post), and ratio, with a score of 0.16, followed by the number of followers (#followers) with a feature importance score of 0.12. Using these feature importance scores, we conducted a set of experiments to test the impact of FI scores on AB performance. Several tests were conducted using input features with feature importance scores >0.05, >0.06, >0.08, >0.12, and all features. The performance comparison results are shown in Fig. 4.

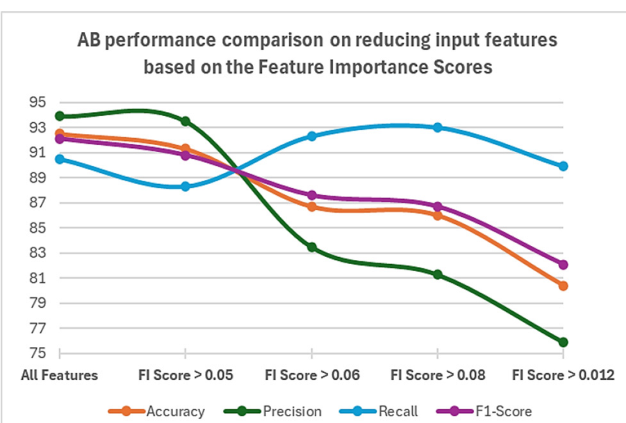

Fig. 4. Comparison of AB performance on reducing input features based on the feature importance scores.

As shown in Fig. 4, the more features included, the better the performance of the AB model. However, the best recall results were obtained when we used input features with FI scores greater than 0.08. This indicates that using all the features provides the best performance for this problem, except for recall. Therefore, if we focus on creating a model that can correctly detect fake and genuine accounts, we can train the model using all the features and attributes. However, we assume that the focus is on maximizing the model's ability to detect fake accounts, undermining a few misdetected genuine accounts. In this case, we can train with features having feature importance scores greater than 0.08. The AB model trained using input features with an FI score >0.08 achieves the highest recall. The highest recall is the highest ratio of correctly predicted positive observations to all positive observations. In this case, the model with this configuration will detect fake accounts more successfully than other models.

D. Testing on Other Datasets

In the final experiment, we tested our best models (AB and GB) on two public datasets created by Jafari[2)] and Purba[3)]. Jafari's dataset comprises 785 records, with 692 records labeled as fake and 93 records labeled as genuine. Purba's dataset comprises two parts, 2-class and 4-class. This dataset was used in Purba's research [5]. The 2-class part comprises 65326 records, 32866 of which are fake account records, and the rest are genuine accounts. The 4-class part of Purba's dataset comprises 43307 records, with real, active-fake, inactive-fake, and spammer-fake accounts for 10441, 12054, 10549, and 10263, respectively. The results of these experiments are presented in Table 4.

Table 4. Results of 10-fold CV test on other datasets

Dataset	Model	Measurement (%)			
		Acc	Pre	Rec	F1
Jafari's	GB	94	96	97	97
	AB	94	96	97	96
Purba's (2-class)	GB	88	92	82	87
	AB	85	85	85	85
	Purba et al.'s RF [5]	90	90	90	90
Purba's (4-class)	GB	90	90	90	90
	AB	64	61	64	60
	Purba et al.'s RF [5]	92	92	92	92

As listed in Table 4, the performances of our best models were fairly good, and we used parameters that were optimized for Bakhshandeh's dataset. However, the AB did not perform well when applied to Purba's 4-class dataset. For the heavily imbalanced Jafari dataset, AB and GB performed better than when applied to the Bakhshandeh dataset (Table 2). We assume that this is because the total number of fake accounts is considerably higher than that of real accounts in Jafari's dataset.

2) https://www.kaggle.com/datasets/rezaunderfit/instagram-fake-and-real-accounts-dataset
3) https://www.kaggle.com/datasets/krpurba/fakeauthentic-user-instagram

Therefore, these models can be easily generalized and used to detect fake accounts. For Purba's dataset, the performances of AB and GB were worse than the best results of Purba et al. [5], although they are generally satisfactory. All performance measurements were above 80% except for AB, which performed poorly on Purba's 4-class dataset. However, the precision of the GB on Purba's 2-class dataset was better than that of Purba's RF. This means that the GB model demonstrates a higher confidence level in predicting the positive class (fake accounts), but it may potentially disregard some actual positive cases.

IV. DISCUSSION AND CONCLUSIONS

Individuals create fake accounts to express themselves on social media, without revealing their identities. However, creating fake accounts can harm the reputation of individuals and businesses, thereby decreasing genuine likes and followers. Based on the test results of the model candidates for detecting fake accounts, AB and GB exhibited the superior performances, with an accuracy greater than 92%, a precision greater than 93%, a recall greater than 90%, and an F1-score greater than 92%. These facts indicate that among the model candidates tested, boosting ensemble models, such as AB and GB, outperform other candidates; therefore, boosting techniques are more suitable for fake account detection. However, to detect fake accounts accurately and disregard genuine accounts, naïve Bayes is the best because it has a recall of 96%. When the input features were tested for AB, all input features provided the best accuracy and precision, but using input features with an importance score >0.08 provided the best recall. The GB model performed well on the two other Instagram datasets. This indicates that fake Instagram accounts can be detected effectively.

REFERENCES

[1] S. Sheikhi, "An efficient method for detection of fake accounts on the instagram platform," *Revue d'Intelligence Artificielle,* vol. 34, no. 4, pp. 429-436, Aug. 2020. DOI: 10.18280/ria.340407.

[2] J. Ezarfelix, N. Jeffrey, and N. Sari, "Systematic literature review: instagram fake account detection based on machine learning," *Engineering, Mathematics and Computer Science (EMACS) Journal,* vol. 4, no. 1, pp. 25-31, Feb. 2022. DOI: 10.21512/emacsjournal.v4i1.8076.

[3] T. S. Nivas, P. Sriramkrishna, S. S. K. Reddy, P. V. R. G. Rao, and R. S. P. Komali, "Fake account detection on instagram using machine learning," *International Journal of Research in Engineering, Science and Management,* vol. 7, no. 5, pp. 24-26, 2024.

[4] R. Herlinda, M. Diqi, M. E. Hiswati, and P. Wanda, "Re-Fake: Klasifikasi Akun Palsu di Sosial Media Online menggunakan Algoritma RNN," *Prosiding Seminar Nasional Sains Teknologi dan Inovasi Indonesia (SENASTINDO),* vol. 3, pp. 191-200, Nov. 2021. DOI: 10.54706/senastindo.v3.2021.139.

[5] K. R. Purba, D. Asirvatham, and R. K. Murugesan, "Classification of Instagram fake users using supervised machine learning algorithms," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 10, no. 3, pp. 2763-2772, 2020. DOI: 10.11591/ijece.v10i3.pp2763-2772.

[6] S. C. Boerman, "The effects of the standardized instagram disclosure for micro- and meso-influencers," *Computers in Human Behavior,* vol. 103, pp. 199-207, 2020. DOI: 10.1016/j.chb.2019.09.015.

[7] I. Restuningrum Pamungkas and N. Lailiyah, "Presentasi diri pemilik dua akun instagram di akun utama dan akun alter," *Interaksi Online,* vol. 7, no. 4, pp. 371-376, Oct. 2019.

[8] M. B. Albayati and A. M. Altamimi, "Identifying fake facebook profiles using data mining techniques," *Journal of ICT Research and Applications,* vol. 13, no. 2, pp. 107-117, 2019. DOI: 10.5614/itbj.ict.res.appl.2019.13.2.2.

[9] L. E. Peterson, "K-nearest neighbor," *Scholarpedia,* vol. 4, no. 2, p. 1883, 2009. DOI: 10.4249/scholarpedia.1883.

[10] C. Campbell and Y. Ying, *Learning with Support Vector Machines,* 1th ed. Berlin :Springer, 2011.

[11] J. R. Quinlan, "Induction of decision trees," *Machine Learning,* journal article vol. 1, no. 1, pp. 81-106, Mar. 1986. DOI: 10.1007/BF00116251.

[12] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, California, USA, vol. 5, pp. 281-298, 1967.

[13] L. Kaufman and P. J. Rousseeuw, "Partitioning Around Medoids (Program PAM)," in *Finding Groups in Data: An Introduction to Cluster Analysis*(Wiley Series in Probability and Statistics, 2th ed. Hoboken, NJ:John Wiley & Sons, 1990. DOI: 10.1002/9780470316801.

[14] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," presented at the Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, USA, 2001. DOI: 10.1145/502512.502529.

[15] L. Breiman, "Random forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, Oct. 2001. DOI: 10.1023/A:1010933404324.

[16] C. D. Manning, P. Raghavan, and H. Schuetze, "Naïve bayes text classification," in *Introduction to Information Retrieval*: Cambridge University Press, pp. 234-265, 2008.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, pp. 318-362, 1986.

[18] L. Breiman, "Bagging predictors," *Machine Learning,* vol. 24, no. 2, pp. 123-140, Aug. 1996. DOI: 10.1007/BF00058655.

[19] S. Menard, *Logistic Regression: From Introductory to Advanced Concepts and Applications*. 1th ed. Los Angeles: SAGE, 2010.

[20] G. S. Budhi and R. Chiong, "A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction," *ACM Transactions on Internet Technology,* vol. 23, no. 1, pp. 1-24, no. 16, Apr. 2023. DOI: 10.1145/3568676.

[21] G. S. Budhi, R. Chiong, I. Pranata, and Z. Hu, "Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis," *Archives of Computational Methods in Engineering,* vol. 28, pp. 2543-2566, Jan. 2021. DOI: 10.1007/s11831-020-09464-8.

[22] A. Alwosheel, S. van Cranenburgh, and C. G. Chorus, "Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis," *Journal of Choice Modelling,* vol. 28, pp. 167-182, Sep. 2018. DOI: 10.1016/j.jocm.2018.07.002.

[23] J. Han, M. Kamber, and J. Pei, "3-Data Preprocessing," in *Data Mining (Third Edition)*, J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 83-124.
[24] F. K. Oksuzoglu, What is a good Instagram follower ratio?, 2024, [Online] Available: https://circleboom.com/blog/what-is-a-good-instagram-follower-ratio/.
[25] C. King, What Is A Good Follower to Following Ratio?, 2023, [Online] Available: https://hypeauditor.com/blog/what-is-a-good-follower-to-following-ratio/#what-is-a-good-follower-to-following-ratio-on-instagram.
[26] H. Stepanek, *Thinking in Pandas: How to Use the Python Data Analysis the Right Way*, 1st ed. Portland, OR: Apress, 2020.
[27] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, pp. 1-27, May 2011. DOI: 10.1145/1961189.1961199.
[28] T. F. Chan, G. H. Golub, and R. J. LeVeque, "Updating formulae and a pairwise algorithm for computing sample variances," *COMPSTAT 1982 5th Symposium held at Toulouse 1982,* pp. 30-41, 1982. DOI: 10.1007/978-3-642-51461-6_3.
[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of Thirteenth International Conference on Artificial Intelligence and Statistics*, Sardinia, IT, 2010, vol. 9, pp. 249-256, 2010.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations*, San Diego, USA, pp. 1-15, 2015. DOI: 10.48550/arXiv.1412.6980.
[31] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost," *Statistics and Its Interface,* vol. 2, pp. 349-360, 2009.
[32] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics,* vol. 29, no. 5, pp. 1189-1232, 2001.
[33] C. Aliferis and G. Simon, "Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI," in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, 2024th ed, Berlin: Springer, pp. 477-524, 2024.

Yulia (1st Author)
Yulia was born in Pasuruan, Indonesia, on 31 July 1976. She obtained her bachelor's degree in 1998 from the Informatics department, Surabaya University, Indonesia. Her master's degree in 2005 from the Information Technology department, University of Indonesia, Jakarta. Her research interests include Data Analysis and Information Systems.

Hendy Gunawan (2nd Author)
Hendy was born in Banjarmasin, Indonesia, on 19 January 2001. He obtained his bachelor's degree in 2022 from the Informatics department, Petra Christian University, Surabaya, Indonesia. His research interests include Data Mining and Data Analysis.

Gregorius Satia Budhi (3rd & Corresponding Author)
Greg was born in Surabaya, Indonesia, on 7 March 1971. He obtained his bachelor's degree in 1993 from the Electrical Engineering department - Computer Engineering program, Adhi Tama Institute of Technology Surabaya, Indonesia. His master's degree in 2001 from the Informatics department, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia. He obtained his PhD degree in 2022 from the School of Information and Physical Sciences, College of Engineering, Science and Environment - Doctor of Philosophy (Information Technology) program, The University of Newcastle, Australia. His research interests include Data and Text Mining, Artificial Intelligence and Deep Learning.

Kartika Gunadi Kartawidjaja (4th Author)
Gunadi was born in Blitar, Indonesia, on 6 June 1962. He obtained his bachelor's degree in 1987 from the Civil Engineering department, Petra Christian University, Surabaya, Indonesia. His master's degree in 1997 from the Informatics department, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia. His research interests include Artificial Intelligence and Data Science.