Deep neural networks approach with transfer learning to detect fake accounts social media on Twitter

Arif Ridho Lubis[1], Santi Prayudani[1], Muhammad Luthfi Hamzah[2], Yuyun Yusnida Lase[1], Muharman Lubis[3], Al-Khowarizmi[4], Gabriel Ardi Hutagalung[1]
[1]Department of Computer Engineering and Informatics, Politeknik Negeri Medan, Medan, Indonesia
[2]Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
[3]School of Industrial Engineering, Telkom University, Bandung, Indonesia
[4]Department of Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received Jul 16, 2023
Revised Oct 19, 2023
Accepted Nov 6, 2023

Keywords:

Convolutional neural network
Deep neural network
Fake account
Transfer learning
Twitter

ABSTRACT

The massive use of social media makes people take actions that have a negative impact on cyberspace, such as creating fake accounts that aim to commit crimes such as spam and fraud to spread false information. Fake accounts are difficult to detect in the traditional way because fake accounts always use photos, names, and unreal information, there are several criteria that can identify a fake account such as no information, few followers, and minimal activity. In the traditional model, it is difficult to detect fake accounts on many Twitters social media accounts, so the application of the deep learning model with the convolutional neural network (CNN) algorithm and the application of deep learning can help detect fake accounts. This study will use data on Twitter social media so that this research produces good accuracy for the scenarios described at the methodology stage. This research produces an accuracy of 86% for the deep learning model with the CNN algorithm, and with the traditional model, it produces an accuracy of 51% while the use of transfer learning produces an accuracy of 93.9%.

Corresponding Author:

Arif Ridho Lubis
Department of Computer Enginnering and Informatics, Politeknik Negeri Medan
Medan, Indonesia
Email: arifridho@polmed.ac.id

1. INTRODUCTION

Social media is one of the most popular communication services, one of which is Twitter which has 321 million users in 2021 [1], [2], this is due to the development of the internet and technology [3], [4] the massive use of social media makes people take action actions that have a negative impact on cyberspace, such as creating fake accounts that aim to commit crimes such as spam and fraud to spread false information [5], [6]. Twitter social media always takes action in overcoming crime problems, one of which is letting users report activities that are considered to be detrimental to others [7]. A fake account is an account that is deliberately created with the aim that other people do not know the real identity so that activities to commit fraud and influence other people cannot be detected or known [8]–[10]. Fake accounts are difficult to detect in traditional ways because fake accounts always use photos, names and information that are not real [11], [12], but there are several criteria that can help in identifying fake accounts such as no information, few followers and minimal activity. However, the use of traditional techniques used to detect fake accounts is usually done by using manual features which have limitations with fake accounts that are growing [13], [14]. Research that has been conducted by Sahoo and Gupta [15] detects fake accounts in real time which has drawbacks with a very large amount of data and the training process takes a long time, another research conducted by Kondenti *et al.* [16]

detects fake accounts using machine learning which where this research has weaknesses in data training with a long time so this research tries to build a model in detecting fake accounts with a deep learning approach that can be done in detecting fake accounts, deep learning is a field of science that conducts learning using architecture in extracting features through the matrix representation of the data [17]–[19].

The problem in this study is the difficulty of detecting fake accounts on Twitter social media so it requires a computational approach, namely deep learning and transfer learning to be able to detect fake accounts, the traditional method approach has been carried out but has experienced accuracy problems resulting in account detection. Fakes are too small as done by Akyon and Kalfaoglu [20] and research conducted by Agarwal dan Dixit [21] which has an average accuracy of 78% so it requires transformers models that can improve accuracy in detecting fake accounts. The use of deep learning in this study aims to compare with the use of transfer learning, the difference between deep learning takes a long time in the model training process so that it requires large computational resources [22], [23] but with transfer learning it does not have to be computational resources large, because transfer learning will share information and knowledge that has been previously trained so that it can be used in relatively small computations [24].

With transfer learning it can select features based on information and knowledge from previous training. So that the application of deep learning models with transfer learning can optimally detect fake accounts on Twitter social media [25], [26]. The contribution to this research is collecting datasets that contain real and fake accounts on Twitter social media, providing identification of features that can detect fake accounts, and providing model parameter adjustments to the data to detect fake accounts so as to achieve maximum accuracy.

2. METHOD

Transfer learning will utilize the representation of features with a model that has been trained first. In the training process, large amounts of data will be used according to the required variables or parameters. Then, after the training process, a model will be formed that can perform the tasks that have been synchronized with the new model. The following is the process of transfer learning in Figure 1.

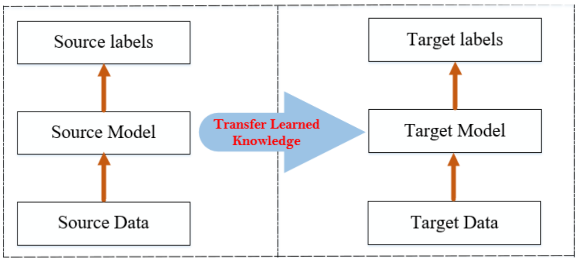


Figure 1. Method transfer learning

2.1. Datasets

This study will use the dataset contained in the study [27]. The dataset contains a collection of data that contains important features in detecting fake accounts such as profile photos, username character length, full name character length, number of followers, number of followers, and number of posts. The entire feature will be processed to find out the feature value of each interest. This research will apply models from deep learning and transfer learning which will look at approaches to identify the most important features of the entire data, then using a combination of parameters in the model is also needed to identify fake accounts.

2.2. Framework

In the framework there is a dataset that will be used, where the dataset has features that can be of important value in implementing deep learning algorithms with transfer learning models by grouping fake accounts and non-fake accounts in training data. This research has explained the features that will be used so that the deep learning and transfer learning stages consist of the data acquisition process, data cleaning to modeling to the evaluation stage. In terms of data acquisition, this research uses data obtained from research [27] then this research performs the stages of cleaning the data by making changes to the text data on the case size, then the data will go through a transformation process and then it will be processed to find out the most important feature extraction by using Word2Vec and GloVe features. The following is the framework contained in Figure 2.

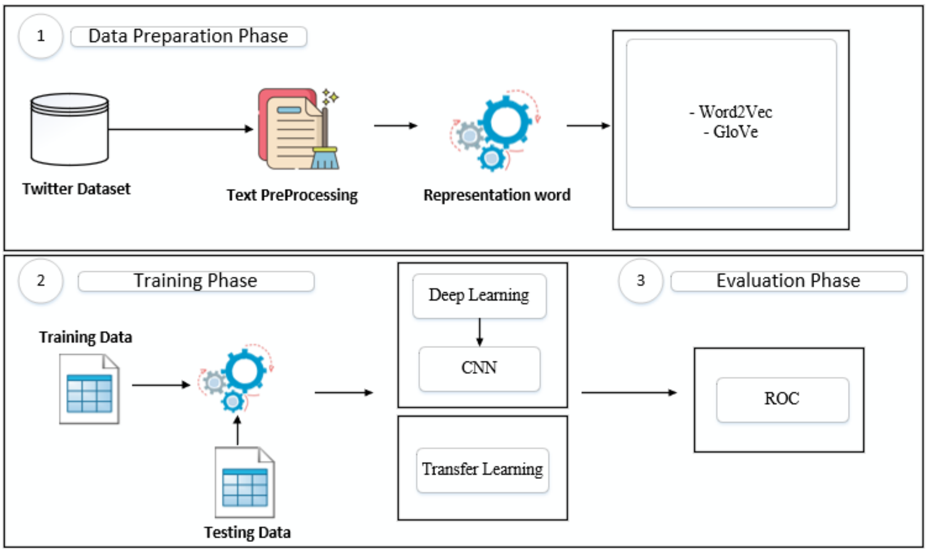


Figure 2. Research framework

The information in Figure 2 will explain the framework of this research as:

- In this study, data from Twitter social media will be carried out which will then be carried out in the text preprocessing process to make semantic changes from unstructured words to structured words by going through the case folding, stopword, and tokenization stages.
- Then after the text preprocessing stage is carried out, word representation will be applied using the Word2Vec and GloVe methods which can convert words into vectors which will then use the polarity technique to determine whether or not the label is from a fake account.
- The next stage is the data training process which will apply the deep learning approach with transfer learning.
- The next stage is to carry out the data testing process with the model that has been formed which will be evaluated using the receiver operating characteristic.

2.3. Model testing scenarios

In this scenario there are 2 types of scenarios that will exist in this study: the first scenario will test the deep learning model approach with the convolutional neural network (CNN) algorithm to see performance in detecting fake accounts. Second scenario is that the data will be transformed and then will apply the model of the transfer learning that already has knowledge and information through training data with large amounts of data that will detect fake accounts on Twitter social media. After implementing the transfer learning model, we will evaluate whether deep learning performance with CNN can be optimal or whether transfer learning performance is more optimal.

3. RESULT AND DISUSSION

At this stage, a discussion of the results of the research that has been carried out will be carried out according to the stages of the research methodology that has been described. In this study, which will show the results of implementing deep learning with transfer learning in detecting fake accounts on Twitter social media, in detecting fake accounts this research uses features that are considered important in completing the need for detecting fake accounts on social media. The dataset contains a collection of data that contains important features in detecting fake accounts such as profile photos, username character length, full name character length, number of followers, number of followers and number of posts.

3.1. Descriptive data analysis

Descriptive statistics will refer to the process of examining the descriptive statistics of a dataset. Descriptive statistics is a method that is often used to see the characteristics of the data so that it can easily understand patterns and trends that can become data for detecting fake accounts on social media Twitter. In the dataset used descriptive statistics will display data on the number of counts, the number of means, and the number of STDs. The following are the results of the descriptive statistics contained in Table 1.

Table 1. Descriptive data statistics

	Username	Full name	Description	Post	Followers	following
Count	576	576	576	576	576	576.00
Mean	0.16	1.46	22	107.49	85307	508.38
STD	0.21	1.05	37	402.03	910148	917.98
Min	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	1.00	0.00	0.00	39.00	57.50
50%	0.00	1.00	0.00	9.00	150.50	229.00
75%	0.31	2.00	34.00	81.50	716.00	589.00
Max	0.92	12.00	150.00	7389	153338538	7500

3.2. Data visualization

Data visualization is a display that will display the distribution of fake and non-fake account data. The data distribution will be used for analysis to be applied to the model according to the planned scenario. The purpose of data visualization is to see the character of the data that will be used to apply the model, when applying the model to detect fake accounts, it will recognize data patterns based on the distribution data in Figure 3. In Figure 3 there is a visualization of fake accounts and non-fake data as follows.

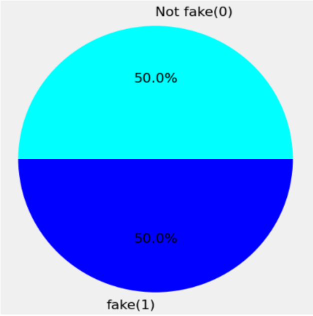
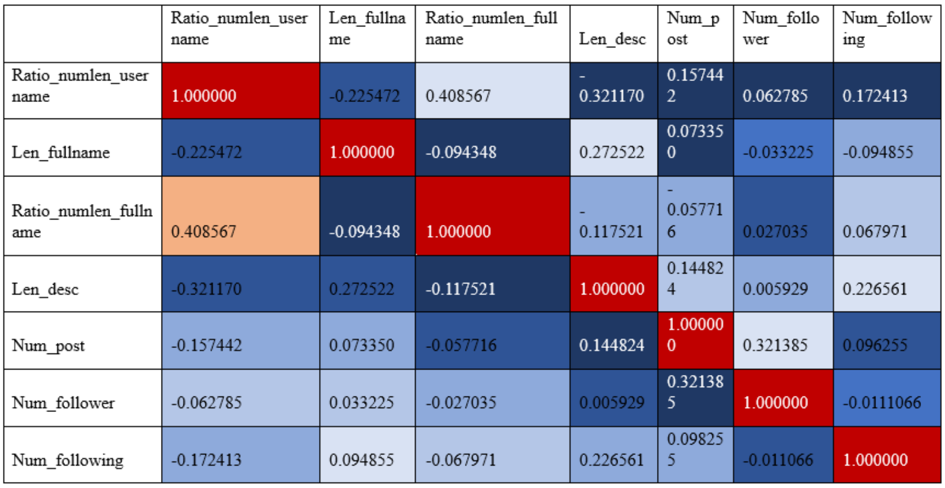


Figure 3. Data visualization

3.3. Heat map analysis

Heat map analysis is a matrix that correlates with features in datasets or documents. The heatmap usually consists of using a correlation matrix to show that each feature has a relationship in the dataset. On the heat map, each cell in the matrix is colored based on the degree of correlation. In Figure 4 there is a correlation between columns and rows, such as the character length ratio column in the username which will illustrate the ratio of the number of number characters to the total number of characters in the words in the username. The following is a visualization of the heatmap in Figure 4.



	Ratio_numlen_username	Len_fullname	Ratio_numlen_fullname	Len_desc	Num_post	Num_follower	Num_following
Ratio_numlen_username	1.000000	-0.225472	0.408567	-0.321170	0.157442	0.062785	0.172413
Len_fullname	-0.225472	1.000000	-0.094348	0.272522	0.073350	-0.033225	-0.094855
Ratio_numlen_fullname	0.408567	-0.094348	1.000000	-0.117521	-0.057716	0.027035	0.067971
Len_desc	-0.321170	0.272522	-0.117521	1.000000	0.144824	0.005929	0.226561
Num_post	-0.157442	0.073350	-0.057716	0.144824	1.000000	0.321385	0.096255
Num_follower	-0.062785	0.033225	-0.027035	0.005929	0.321385	1.000000	-0.0111066
Num_following	-0.172413	0.094855	-0.067971	0.226561	0.098255	-0.011066	1.000000

Figure 4. Heat map visualization

3.4. Pair plot analysis

Pair plot analysis is a visualization technique that will describe the relationship between numerical variables and features in the dataset. This technique will spread the overall visualization simultaneously using a scatter plot between all pairs of variables in the dataset. In this study, pair plots will look at the relationship between one feature and another. Figure 5 will explain the pair plot above, fake accounts are more spread out around the average feature pad. Real account values, in this case, are very clustered around the average. Most fake accounts have fewer description words in their bio because the average description words for fake accounts are fewer than for real accounts, ratio_numlen_fullname and ratio_numlen_username correlate with each other. The correlation matrix shows a value of +0.4085 for this correlation. The average number of posts and followers on fake accounts is close to zero. The following is a pair plot analysis in Figure 5.

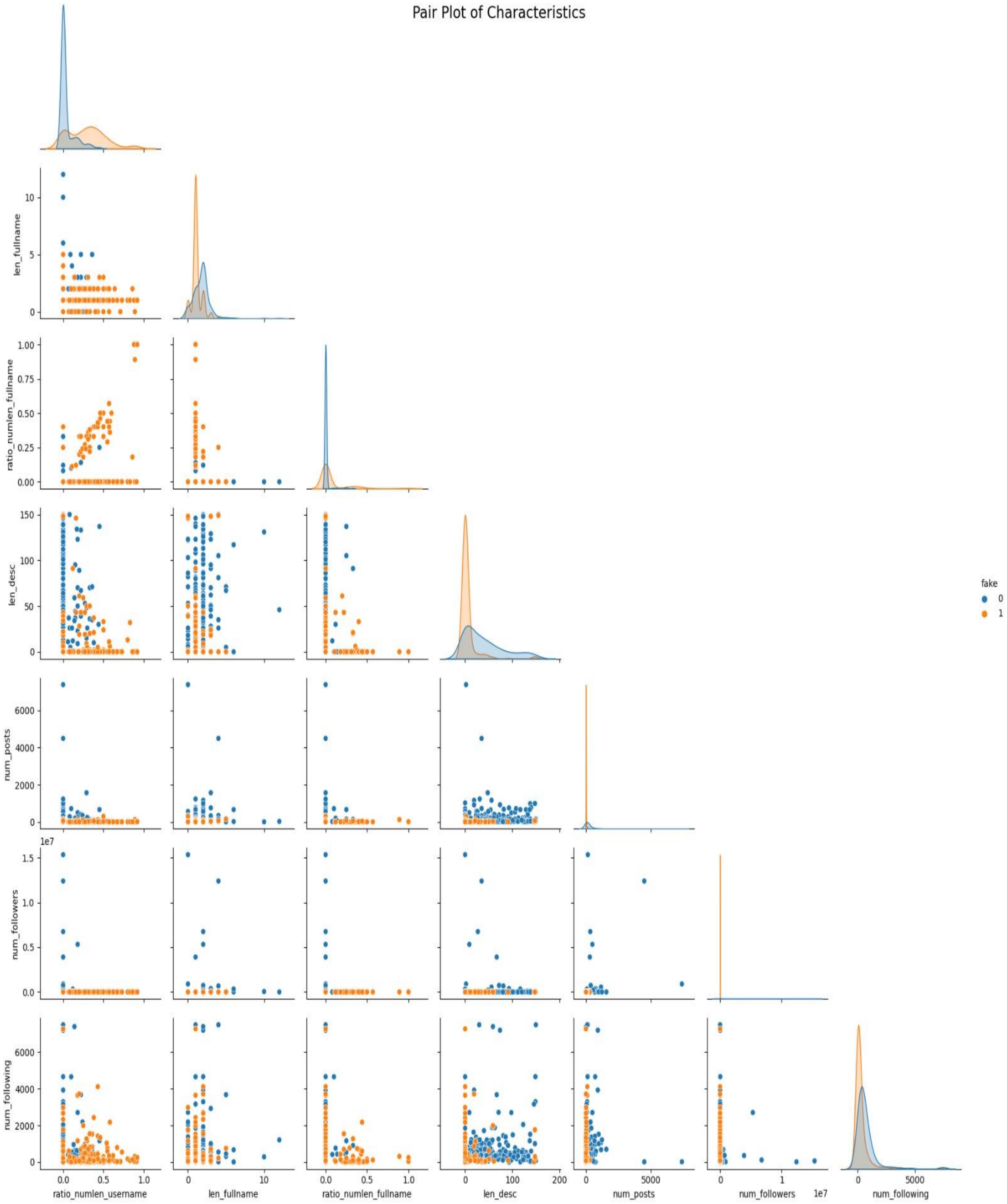


Figure 5. Visualization of pair plots

3.5. Deep learning evalution

After the training process, an evaluation of the model with test data will be carried out for deep learning models with transfer learning in detecting fake accounts. The receiver operating characteristic (ROC) curve displays the relationship between the true positive rate (TPR) and the false positive rate (FPR) at various threshold values. The purpose of this evaluation is to see the model accuracy of the deep learning approach. The following are the results of the evaluation with the ROC curve in Figures 6 and 7.

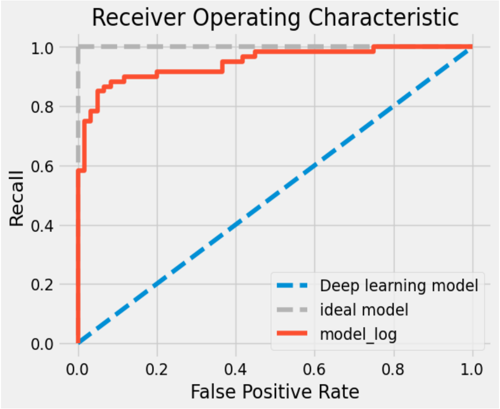


Figure 6. ROC curve evaluation in deep learning

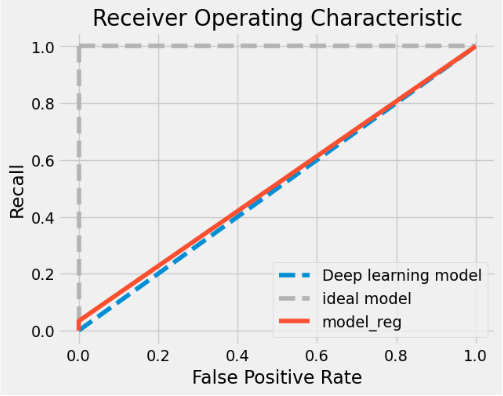


Figure 7. Evaluation of the ROC curve without deep learning

Based on Figure 6 there is a red curve which is the result of a deep learning model with an accuracy of 86%. From recalls about 42% and from false positives rate approximate 19%, barely deviating from the ideal. So, the general model looks really good because it follows the gray ideal rather than the blue price line. But it is not perfect, while the ROC curve using the traditional model will be shown in Figure 7 which will be used to visualize the behavior of the positive rate and the false positive rate. The ROC curve of the model without deep learning deviates slightly from the ideal gray dotted line of the model without regularization (model_log). This shows that the model without deep learning has a slight accuracy, which is 51% worse than the model using deep learning.

3.6. Evaluation transfer learning

At this stage, testing of the deep learning model with transfer learning will be carried out to determine the resulting accuracy, in Figure 6 previously it has shown a model with deep learning which produces an accuracy of 86%, so this research will use transfer learning to increase the resulting accuracy in detecting accounts. Fake, the following is an evaluation of the ROC with a deep learning model using transfer learning which is shown in Figure 8. Based on Figure 8, the accuracy value with ROC evaluation with a value of 93.9% is very good. This result proves that deep learning models with transfer learning can be more optimal in detecting fake accounts.

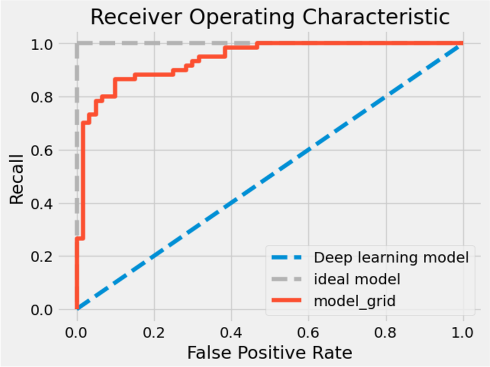


Figure 8. Evaluation of the ROC curve on transfer learning

4. CONCLUSION

In this study it produces optimal performance in detecting fake accounts so that it can distinguish which accounts are fake and not fake based on the features used, this research has also succeeded in comparing the performance of deep learning with the CNN and transfer learning algorithms. This research produces good accuracy against scenarios that have been described at the methodological stage. This research produces an accuracy for the deep learning model with the CNN algorithm of 86%, and with the traditional model it produces an accuracy of 51% while the use of transfer learning produces an accuracy of 93.9%.

ACKNOWLEDGEMENTS

We would like to thank Politeknik Negeri Medan, for supporting this research through grant no B/319/PL5/PT.01/04/2023.

REFERENCES

[1] A. R. Lubis, S. Prayudani, M. Lubis, and O. Nugroho, "Sentiment analysis on online learning during the COVID-19 pandemic based on opinions on Twitter using KNN method," in *2022 1st International Conference on Information System and Information Technology, ICISIT 2022*, 2022, pp. 106–111, doi: 10.1109/ICISIT54091.2022.9872926.

[2] S. Kurniawan and I. Budi, "Indonesian tweets hate speech target classification using machine learning," in *2020 5th International Conference on Informatics and Computing, ICIC 2020*, Nov. 2020, pp. 1–5, doi: 10.1109/ICIC50835.2020.9288515.

[3] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "The effect of the TF-IDF algorithm in times series in forecasting word on social media," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 22, no. 2, pp. 976–984, May 2021, doi: 10.11591/ijeecs.v22.i2.pp976-984.

[4] P. Sharm, "Internet of things and blockchain," in *Blockchain for Business: How it Works and Creates Value*, vol. 6, no. 6, Wiley, 2021, pp. 295–335.

[5] A. Gupta and R. Kaushal, "Towards detecting fake user accounts in Facebook," in *ISEA Asia Security and Privacy Conference 2017, ISEASP 2017*, 2017, pp. 1–6, doi: 10.1109/ISEASP.2017.7976996.

[6] V. Singh, R. Shanmugam, and S. Awasthi, "Preventing fake accounts on social media using face recognition based on convolutional neural network," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 57, Springer, 2021, pp. 41–53.

[7] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in *SMAP 2020 - 15th International Workshop on Semantic and Social Media Adaptation and Personalization*, 2020, pp. 1–6, doi: 10.1109/SMAP49528.2020.9248443.

[8] M. M. Swe and N. N. Myo, "Fake accounts detection on Twitter using blacklist," in *Proceedings - 17th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2018*, 2018, pp. 562–566, doi: 10.1109/ICIS.2018.8466499.

[9] L. Caruccio, D. Desiato, and G. Polese, "Fake account identification in social networks," in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2019, pp. 5078–5085, doi: 10.1109/BigData.2018.8622011.

[10] A. Kumar and N. Sachdeva, "Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data," *Multimedia Systems*, vol. 28, no. 6, pp. 2027–2041, 2022, doi: 10.1007/s00530-020-00672-7.

[11] S. R. Krishna, K. U. Reddy, T. A. Reddy, A. Saiteja, and R. Sumanjali, "Detection of fake and clone accounts in Twitter using classification and distance measure algorithms," in *Smart Innovation, Systems and Technologies*, 2022, vol. 265, pp. 391–399, doi: 10.1007/978-981-16-6482-3_39.

[12] M. Heidari *et al.*, "BERT model for fake news detection based on social bot activities in the COVID-19 pandemic," in *2021 IEEE 12th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2021*, 2021, pp. 103–109, doi: 10.1109/UEMCON53757.2021.9666618.

[13] J. P. Baptista and A. Gradim, "Understanding fake news consumption: A review," *Social Sciences*, vol. 9, no. 10. MDPI, pp. 1–22, 2020, doi: 10.3390/socsci9100185.

[14] M. S. Al-Zaman, "Social media fake news in India," *Asian Journal for Public Opinion Research*, vol. 9, no. 1, pp. 25–47, 2021, doi: 10.15206/ajpor.2021.9.1.25.

[15] S. R. Sahoo and B. B. Gupta, "Real-time detection of fake account in twitter using machine-learning approach," in *Advances in Intelligent Systems and Computing*, 2021, vol. 1086, pp. 149–159, doi: 10.1007/978-981-15-1275-9_13.

[16] P. Kondeti, L. P. Yerramreddy, A. Pradhan, and G. Swain, “Fake account detection using machine learning,” in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 53, Springer, 2021, pp. 791–802.
[17] P. Majerczak and A. Strzelecki, “Trust, media credibility, social ties, and the intention to share information verification in an age of fake news,” *Behavioral Sciences*, vol. 12, no. 2, p. 51, 2022, doi: 10.3390/bs12020051.
[18] C. M. Pulido, L. Ruiz-Eugenio, G. Redondo-Sama, and B. Villarejo-Carballido, “A new application of social impact in social media for overcoming fake news in health,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, p. 2430, 2020, doi: 10.3390/ijerph17072430.
[19] L. Tian, X. Zhang, Y. Wang, and H. Liu, “Early detection of rumours on Twitter via stance transfer learning,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12035 LNCS, pp. 575–588, doi: 10.1007/978-3-030-45439-5_38.
[20] F. C. Akyon and M. E. Kalfaoglu, “Instagram Fake and Automated Account Detection,” in *Proceedings - 2019 Innovations in Intelligent Systems and Applications Conference, ASYU 2019*, 2019, pp. 1–7, doi: 10.1109/ASYU48272.2019.8946437.
[21] A. Agarwal and A. Dixit, “Fake news detection: An ensemble learning approach,” in *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, 2020, pp. 1178–1183, doi: 10.1109/ICICCS48265.2020.9121030.
[22] Y. Tashtoush, B. Alrababah, O. Darwish, M. Maabreh, and N. Alsaedi, “A Deep Learning Framework for Detection of COVID-19 Fake News on Social Media Platforms,” *Data*, vol. 7, no. 5, p. 65, 2022, doi: 10.3390/data7050065.
[23] L. Tang and Q. H. Mahmoud, “A survey of machine learning-based solutions for phishing website detection,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 3. MDPI, pp. 672–694, 2021, doi: 10.3390/make3030034.
[24] A. Zhang *et al.*, “Transfer learning with deep recurrent neural networks for remaining useful life estimation,” *Applied Sciences (Switzerland)*, vol. 8, no. 12, p. 2416, 2018, doi: 10.3390/app8122416.
[25] Z. Huang, Z. Pan, and B. Lei, “Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data,” *Remote Sensing*, vol. 9, no. 9, p. 907, 2017, doi: 10.3390/rs9090907.
[26] T. Rahman *et al.*, “Transfer learning with deep Convolutional Neural Network (CNN) for pneumonia detection using chest X-ray,” *Applied Sciences (Switzerland)*, vol. 10, no. 9, p. 3233, 2020, doi: 10.3390/app10093233.
[27] S. Cresci, A. Spognardi, M. Petrocchi, M. Tesconi, and R. Di Pietro, “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race,” in *26th International World Wide Web Conference 2017, WWW 2017 Companion*, 2017, pp. 963–972, doi: 10.1145/3041021.3055135.

BIOGRAPHIES OF AUTHORS

Arif Ridho Lubis He received a doctoral degree from Universitas Sumatra Utara in 2023, a master’s degree from Universiti Utara Malaysia in 2012, and graduated from Universiti Utara Malaysia in 2011, both in Information Technology. He is a Lecturer at the Department of Computer Engineering and Informatics, Medan State Polytechnic in 2015. His research interests include networking, project management science and computer science. He can be contacted at email: arifridho@polmed.ac.id.

Santi Prayudani was born in the Municipality of Binjai, North Sumatra on March 28, 1986. The author studied Bachelor's degree in the Computer Science Study Program, at Universitas Sumatera Utara in 2004. Then continued his master's degree in the Informatics Engineering Study Program, at Universitas Sumatera Utara in 2011. Started his career as a teacher at SDS Al Azhar Medan in 2010. Then he also taught as a lecturer at AMIK Harapan and Panca Budi Development University from 2011 to 2014. Currently, the author has been given a mandate by the state to serve as a lecturer at the Medan State Polytechnic since 2015. She can be contacted at email: santiprayudani@polmed.ac.id.

Muhammad Luthfi Hamzah He is a Lecturer at the Information Systems Department, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia. He received the graduated in Software Engineering from Universiti Utara Malaysia in 2011, master’s degree in Computer Science from Universitas Putra Indonesia YPTK Padang in 2016, and doctoral degree. in Technical and Vocational Education at the Engineering Faculty of Universitas Negeri Padang in 2021. His research interests include software engineering, blended learning, management information systems and artificial intelligence. He can be contacted at email: muhammad.luthfi@uin-suska.ac.id.

Yuyun Yusnida Lase holds a master from Universitas Sumatera Utara in 2012. Currently, she is lecture at Diploma-4 Software Engineering Technology, Department of Computer and Informatics Engineering, Politeknik Negeri Medan, Indonesia. Her research includes data mining, machine learning, information systems, and artificial intelligence. She can be contacted at email: yuyunlase@polmed.ac.id.

Muharman Lubis has finished his Doctoral degree recently in Information Technology at 2017 in International Islamic University Malaysia, he also received his Master degree from same university at 2011 and Bachelor degree from University Utara Malaysia at 2008, both in Information Technology. He joined as a Lecturer in the School of Industrial Engineering, Telkom University, in 2017. His research interests include privacy protection, information security awareness, knowledge management and project management. He can be contacted at email: muharmanlubis@telkomuniversity.ac.id.

Al-Khowarizmi was born in Medan, Indonesia, in 1992. He is a Dean in Faculty of Computer Science and Information Technology at Universitas Muhammadiyah Sumatera Utara (UMSU). He got Doctoral Degree from Universitas Sumatera Utara in 2023 field Computer Science. His main research interest is data science, big data, machine learning, neural network, artificial intelligence and business intelligence. He can be contacted at email: alkhowarizmi@umsu.ac.id.

Gabriel Ardi Hutagalung was born in Medan in 1992. He has served as a Lecturer at the Politeknik Negeri Medan from 2019 until now. He received a master's degree in Informatics Engineering from the Universitas Sumatera Utara in 2017. Currently pursuing a Doctoral degree in computer science at the Universitas Sumatera Utara. He can be contacted at email: gabrielhutagalung@polmed.ac.id.