Figure 6 displays the credibility confusion matrix for scenario II. Which is known that users are not credible on actual data, it is predicted that as many as 306 users are not credible on the system and it is predicted that as many as 237 are credible on the system. Furthermore, based on actual data, it is predicted that as many as 272 users are not credible on the system and 359 users are credible on the system.

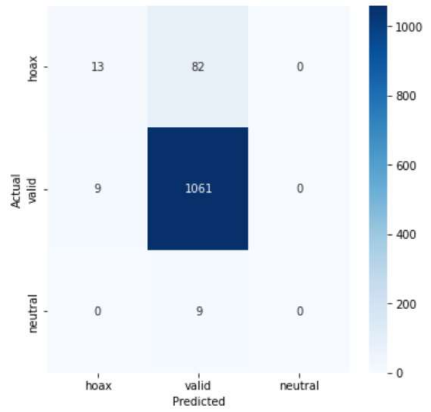

Figure 7. Confusion Matrix Hoax (Scenario II)

Figure 7 displays the hoax confusion matrix for scenario II. Which is known that hoaxes in actual data, predicted as many as 13 data as hoaxes in the system, predicted 82 data as valid in the system, and predicted as many as 0 data as neutral in the system. Furthermore, valid on actual data, predicted as many as 9 data as hoaxes on the system, predicted 1061 data as valid on the system, and predicted as many as 0 data as neutral on the system. After that, neutral on actual data, predicted as many as 0 data as hoaxes on the system, predicted 9 data as valid on the system, and predicted as many as 0 data as neutral on the system.

Table 11. Value of Confusion Matrix Credibility (Scenario II)

	Precision	Recall	F1-score
0	0.53	0.56	0.55
1	0.60	0.57	0.59
accuracy			0.57

Table 12. Value of Confusion Matrix Hoax (Scenario II)

	Precision	Recall	F1-score
1	0.59	0.14	0.22
2	0.92	0.99	0.95
3	0.00	0.00	0.00
accuracy			0.91

Tables 11 and 12 show the credibility and hoax confusion matrix values in scenario II. Where the precision value is obtained from a comparison of the amount of relevant information obtained by the system with the total amount of information retrieved by the system. Furthermore, the recall value is obtained from a comparison of the amount of relevant information obtained by the system with the total amount of relevant information contained in the information, whether retrieved or not retrieved by the system. Then, the f1-score value is obtained from the average harmonic result between the precision and recall values. Meanwhile, the accuracy value indicates the effectiveness of the test based on the effectiveness between the predicted value and the actual value.

4. Conclusion

Based on the results and discussion, it can be concluded that the detection of fake news with tweets containing the #covid19 hashtag based on author credibility using the Information Gain and KNN (K-Nearest Neighbor) methods was successfully carried out with an accuracy value of 91%, a correlation value between credibility and hoaxes of 0.115, and a p-value <0.05. This proves that the system is 91% accurate and there is a significant correlation between credibility and hoaxes. Scenario division into 2 scenarios also has a significant impact on precision, recall, and f1-score. Where in the scenario I, precision is 0.60, recall is 0.59, and f1-score is 0.59 for credibility. And precision is worth 0.50, recall is worth 0.37, and f1-score is worth 0.38 for hoaxes. Meanwhile, in scenario II, precision is 0.57, recall is 0.57, and f1-score is 0.57 for credibility. And precision is worth 0.50, recall is worth 0.38, and f1-score is worth 0.39 for hoaxes.

For research development, it is necessary to do other news classifications, not only COVID-19 in Indonesia. In addition, it is necessary to research whether the calculation of the confusion matrix between credibility and hoaxes can be combined. In addition, the addition of using feature selection (not only Information Gain) and using other classification methods to improve system performance.

References

[1] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in Social Media Research: Past, Present, and Future," *Inf. Syst. Front.*, vol. 20, no. 3, pp. 531–558, Jun. 2018, doi: 10.1007/s10796-017-9810-y.

[2] N. S. Mudawamah, "Internet User Behavior: Case Study of Library and Information Science Department Students of UIN Maulana Malik Ibrahim".

[3] S. Alhabash and M. Ma, "A Tale of Four Platforms: Motivations and Uses of Facebook, Twitter, Instagram, and Snapchat Among College Students?," *Soc. Media Soc.*, vol. 3, no. 1, p. 205630511769154, Jan. 2017, doi: 10.1177/2056305117691544.

[4] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, "An Exploratory Study of Covid-19 Misinformation on Twitter," *Online Soc. Netw. Media*, vol. 22, p. 100104, Mar. 2021, doi: 10.1016/j.osnem.2020.100104.

[5] L. Garrett, "Covid-19: the Medium is the Message," *The Lancet*, vol. 395, no. 10228, pp. 942–943, Mar. 2020, doi: 10.1016/S0140-6736(20)30600-0.

[6] J. Zarocostas, "How to Fight an Infodemic," *The Lancet*, vol. 395, no. 10225, p. 676, Feb. 2020, doi: 10.1016/S0140-6736(20)30461-X.

[7] J. Zhang, B. Dong, and P. S. Yu, "Fake Detector: Effective Fake News Detection with Deep Diffusive Neural Network."

DOI: https://doi.org/10.29207/resti.v7i1.4871

arXiv, Aug. 10, 2019. Accessed: Jan. 14, 2023. [Online]. Available: http://arxiv.org/abs/1805.08751

[8] T. R. A. Pangaribuan, "Social Media Credibility in Reporting the Jakarta Governor Election," vol. 18, no. 2.

[9] R. I. Pristiyanti, M. A. Fauzi, and L. Muflikhah, "Sentiment Analysis Summarizing Film Reviews Using Information Gain and K-Nearest Neighbor Methods".

[10] I. Maulida, A. Suyatno, and H. R. Hatta, "Feature Selection in Abstract Indonesian Text Documents Using the Information Gain Method," *J. SIFO Mikroskil*, vol. 17, no. 2, pp. 249–258, Oct. 2016, doi: 10.55601/jsm.v17i2.379.

[11] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, "Attribute Reduction Using Information Gain for Cluster Optimization of the K-Means Algorithm," *J. Edukasi Dan Penelit. Inform. JEPIN*, vol. 6, no. 1, p. 48, Apr. 2020, doi: 10.26418/jp.v6i1.37606.

[12] R. I. Perwira, B. Yuwono, R. I. P. Siswoyo, F. Liantoni, and H. Himawan, "Effect of Information Gain on Document Classification Using K-Nearest Neighbor," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 8, no. 1, p. 50, Jan. 2022, doi: 10.26594/register.v8i1.2397.

[13] S. Tang, S. Yuan, and Y. Zhu, "Data Preprocessing Techniques in Convolutional Neural Network Based on Fault Diagnosis Towards Rotating Machinery," *IEEE Access*, vol. 8, pp. 149487–149496, 2020, doi: 10.1109/ACCESS.2020.3012182.

[14] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The Impact of Features Extraction on the Sentiment Analysis," *Procedia Comput. Sci.*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[15] O. Caelen, "A Bayesian Interpretation of the Confusion Matrix," *Ann. Math. Artif. Intell.*, vol. 81, no. 3–4, pp. 429–450, Dec. 2017, doi: 10.1007/s10472-017-9564-8.

[16] Isman, Andani Ahmad, and Abdul Latief, "Comparison of KNN and LBPH Methods in Classification of Herbal Leaves," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 5, no. 3, pp. 557–564, Jun. 2021, doi: 10.29207/resti.v5i3.3006.

[17] G. B. Firmanesha, S. S. Prasetyowati, and Y. Sibaroni, "Detecting Hoax News Regarding the Covid-19 Vaccine Using Levenshtein Distance".

[18] A. Essra, "Analysis of Information Gain Attribute Evaluation for Classification of Intrusion Attacks," 2016.

[19] M. I. A. Ismandiya and Y. Sibaroni, "Indonesian News Classification Using Weighted K-Nearest Neighbour".

[20] I. J. A. Cici Apriza Yanti, "Pearson, Spearman, and Kendall Tau Correlation Test Differences in Analyzing the Incidence of Diarrhea," *J. Endur.*, vol. 6, no. 1, pp. 51–58, Jun. 2022, doi: 10.22216/jen.v6i1.137.

.

DOI: https://doi.org/10.29207/resti.v7i1.4871