Detecting Hoax News Regarding the Covid-19 Vaccine using Levenshtein Distance

Gilang Brilians Firmanesha[1*], Sri Suryani Prasetyowati [2], Yuliant Sibaroni[3]
[1,2,3]Telkom University, Indonesia
gilangbrilians@student.telkomuniversity.ac.id[1*], srisuryani@telkomuniversity.ac.id[2],
yuliant@telkomuniversity.ac.id[3]

Submitted: 16 June 2022, Revised: 23 June 2022, Accepted: 14 July 2022

Abstract *– The internet is a communication tool that we often use. The internet itself has brought many benefits. However, some people misuse it, for example, individuals or a group of people who spread hoaxes or fake news to incite and lead the publics' opinions to their desired side. When COVID-19 spread in Indonesia and the government implemented mandatory vaccine obligations, the total of hoaxes on vaccination increased rapidly. Due to a large number of hoaxes on the Internet on COVID-19 vaccinations, As for several studies on the creation of a hoax detection system with various methods to try to overcome this problem, one of the studies with a system that detects hoax news and uses several methods, one of these methods is Levenshtein, getting a fairly low-performance result of 40% compared to other methods used. Therefore. Researchers are motivated to develop a hoax detection system with a similar method by adding Feature Extraction which aims to improve system performance from the previous research. In this study, 2 main experiments were conducted using Levenshtein distance as the main classification method, the results showed the best results in experiment-2 with an f1-score of 70.2% which was an increase compared to previous studies due to adding feature extraction using tf-idf.*

Keywords*: Hoax Detection, Covid-19, Vaccination, Levenshtein Distance, TF-IDF*

1. Introduction

Corona Virus (Covid-19) is a virus that first appeared in the city of Wuhan, China, in December 2019 and now has spread around the world [1]. According to the data conveyed by the government of Indonesia, up until today, 4.244.761 people have tested positive for Covid-19 [2]. The Indonesian government has made various efforts to push the number of Covid-19 spread; one is to vaccinate Indonesian citizens.

The action made by the government reaped Pros and Cons from various parties; some groups or societies disagree with the policy. Fake news emerges from those who disagree with the policy, which we often call Hoaxes. Hoaxes arise everywhere, especially on the internet; the parties who spread the news aim to incite or lead the public's opinions to their desired side [3]. During the pandemic period, there was a significant increase in the number of hoaxes about Covid-19 and vaccines, with the number reaching 1733 news [4]. Similar to Covid-19, hoaxes themselves can also be said to be an epidemic because of their swift spread on the internet.

In order to know whether the news that spreads on the internet is accurate or not, a hoax detection system is needed. One of the researches about hoaxes detection was conducted by Madani Y. [5], and in the research, the data used are Twitter tweets with hoax and non-hoax elements. The approach used is Artificial Intelligence by comparing the Decision Tree, Random Forest, and Logistic Regression methods, where the best results are obtained with the Random Forest method. Similar research was conducted by Aldwairi M. [6], who used the same methods with Accuracy, Recall, Precision, and F-Measure results above 90%. Still, this study has a weakness in that the data used is not shown what it looks like, and the amount of data used is not explained. The researcher has also conducted a comparison between two different classification methods, there is Support Vector Machine (SVM) and Stochastic Gradient Descent (SGD), by weighting the TF-IDF feature to get the best results with the SGD method, but in this research,

the data used is still very small [7]. Another research by Adzlan Ishak [8] used an approach of Distance-Based with a method of Levenshtein Distance, which is unique because of its way of detecting hoaxes by looking at the similarity value from the structure of the words to do classification. This research obtained a quite good result, namely 71% on sensitivity/recall, in which this research did not use any data on either social media or online news, it only used e-mail.

In most of the research, the problem that is usually researcher faced when building this hoax detection system is the collecting data about the hoax and also determining hoax from their sentence structure and their similarity. In the research above, many of them had discussed standard classification methods such as Decision Tree, SVM, SGD, Random Forest, and then there are Levenshtein Distance methods that can help detect hoaxes based on their similarity. Therefore, this research purposed a Hoaxes Detection system using the Levenshtein Distance method because it has been tested in previous studies and has potential if it is conducted in the following research on the hoaxes detection system using the method. The contribution of this research is using a larger number of data and adding tf-idf, data used in this research were Indonesian news collected from a reliable Indonesian government website that is labeled manually, by adding Feature Extraction TF-IDF and using the method of K-Fold Cross-Validation to Measure performance system. The purpose to use this method is to determine the hoax news is more accurate based on their similarities and the structure of sentences on the corpus/dataset.

1.1. Related Works

Research on the system of hoax detection from a problem spread on the internet has been done a lot before; one of them is the research conducted by S. Y. Yuliani [9]. Regarding the hoax detection case, the researcher compared the methods, namely Similarity Method, Smith-Waterman, and Damerau Levenshtein. Determination of whether news contains a hoax or not can be determined based on the structure of writing words or sentences. From this research, it can be obtained that the method of Smith-Waterman has the highest result among the others, with an accuracy of 99.29%. However, this algorithm has a problem in that the more words used, the more time needed to detect hoaxes. Unfortunately, in this research, the performance of Levenshtein Distance is not quite good because the result obtained is 40%.

Another research about hoax detection that added the method TF-IDF as a feature extraction was conducted by B. Lalitha Devi [10]. In this research, the method Naïve Bayes is used as a classifier by adding the feature of TF-IDF and by using a dataset in the form of both news and articles from the internet that has been labeled hoax or not. The obtained result of this research is quite good.

There is also research that uses Twitter posts as the data used, and the research was conducted by Titi Widarenta [11]. This research has aimed to detect hoaxes on Twitter according to the context of the posts. The method used in this research is a Support Vector Machine by adding Doc2Vec as feature extraction. Before experimenting by using Doc2Vec, this research tested the method that was developed by Afriza Research [12]. By using an approach of Frequency-Based with the same dataset, a not very good result was obtained, namely 65%. The result of the research using the feature extraction of Doc2Vec obtained a better accuracy than the trial using the Frequency-based, which is 93.4%.

Another research was conducted by Adzlan Ishak [8]. The case examined was a hoax detection system on the message received through e-mail. In this research, the comparison between the structure of words or sentences is also conducted to determine if the message on the e-mail is a hoax. The result of this research shows that the method used is, Levenshtein Distance produced a performance of 74% on sensitivity (recall).

Based on the research above, this research will also use Levenshtein Distance because this method can still be developed to obtain a more optimal result. By developing the previous researches about the hoax detector using Levenshtein Distance [8] that only tried to detect email

messages by adding the expansion feature of TF-IDF and also did K-Fold Cross Validation because previewing the other research that feature extraction conducted by the method of TF-IDF can increase the result of system performance [10].

2. Research Method

This section explains the research method used and also the proposed system model of this research that can be seen in Figure 1.

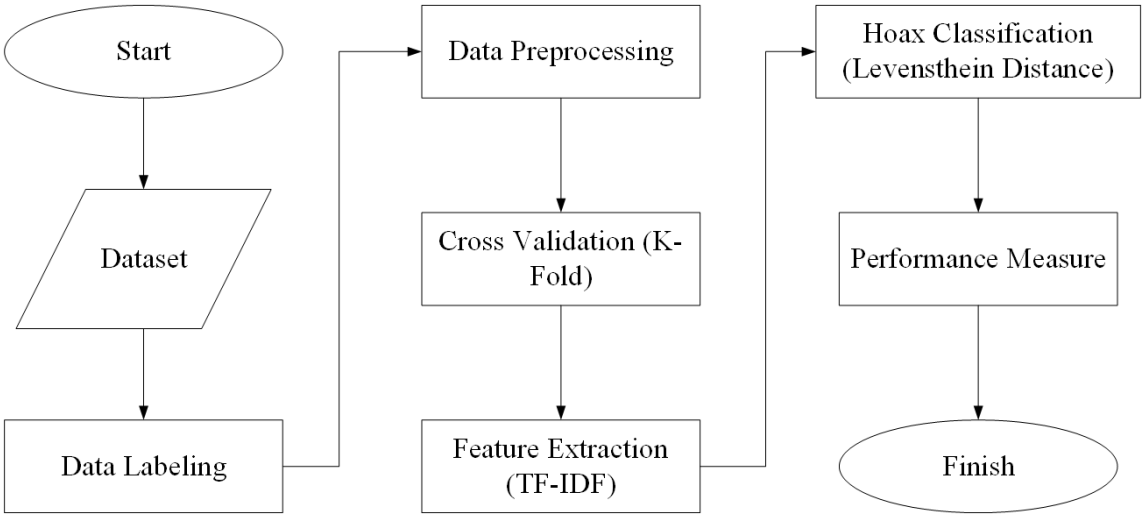

Figure 1. System Diagram of hoax detection.

2.1. Dataset

The dataset used in the making of the system is news that is obtained from a reliable government site, namely covid19.go.id. This news has a specific page that collects hoax and non-hoax news about Covid-19 and also vaccination that spread on the internet. Thus, the data obtained on this site has been labeled whether hoax or non-hoax according to the search results and research from the experts. Table 1 shows the data that will be used.

Table 1. Example Data Vaccine News Dataset

No	News	Label
1	Vaksin COVID-19 Menyebabkan Kanker Kambuh	1 (Hoax)
2	Pasien Kanker Dapat Menerima Vaksin COVID-19 Asal Tetap Dalam Pengawasan Medis	0 (Non-Hoax)
3	Sambut Kedatangan Vaksin Moderna dan AstraZenecca, Menlu : Mekanisme dose-sharing Vaksin Penting agar Dunia dapat Keluar dari Pandemi	0 (Non-Hoax)
4	Vaksin Covid-19 Mengandung Microchip Magnetik	1 (Hoax)

2.2. Data Preprocessing

Data Preprocessing is a series of steps of making Raw Data into Clean, efficient, and ready-to-use data. Some of the steps from Data Preprocessing are handling empty data, data duplication, removing punctuations in a sentence, and removing unknown terms [13]. The process of Data Preprocessing is at the beginning of the system after the data is collected and labeled. The steps of Preprocessing Data that will be used in the making of the system can be seen in Figure 2.

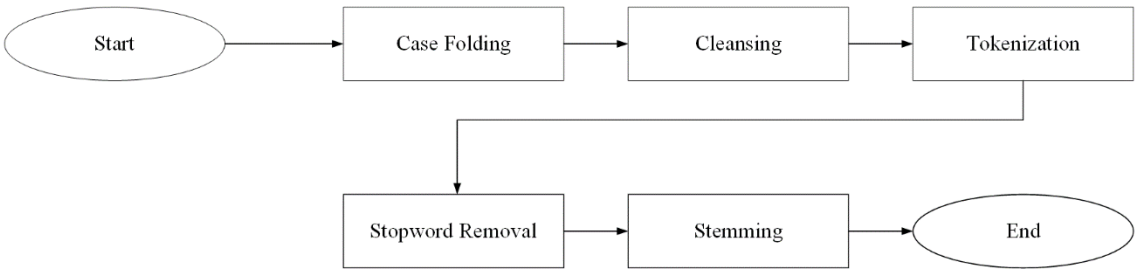

Figure 2. Preprocessing Process Diagram.

A. Case Folding

Case folding is the process of changing a word or sentence with a string type into lowercase letters, generally, a word in a dataset is not structured and consistent in the use of capital letters, the case-folding purpose itself is to equate the use of capital letters, when data is equated is easier for the system to classify the hoax data when detecting. The case folding process can be seen in Figure 3 below.

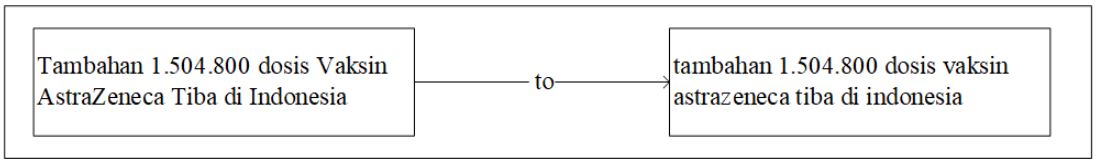

Figure 3. Case Folding Process.

B. Cleansing

Process of removing numbers and punctuations in a sentence. Same like case folding, the cleansing purpose is to remove numbers or punctuation in words to equate data form, so the data in the dataset will have the same structure, in other words, all data will have no numbers and punctuation. The cleansing process can be seen in Figure 4.

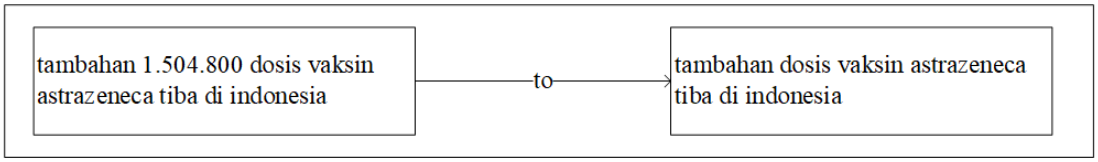

Figure 4. Cleansing Process.

C. Tokenization

Tokenization is the process of changing a sentence into tokens per word, the purpose is to help analyze word in other preprocessing processes such as stopword removal, is easier for the system to filter word when is already in tokens. The example of this process can be seen in Figure 5.

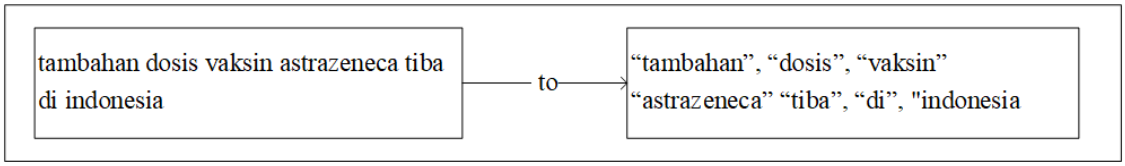

Figure 5. Tokenization Process.

D. Stopword Removal

Stopword Removal is the process of data removing words that occurs commonly in a dataset. this process is to filter important words and words that have no meaning or occur commonly dataset, words that have no meaning themselves can slow the system during classification. In this process will be used sastrawi library, Unfortunately, sastrawi only contains general Indonesian words, so modifications will be made to add words for the stopword corpus, One example that has been modified is the word "dosis", this word modification where made because “Dosis” word itself appears quite often in a hoax and non-hoax class data, which makes the word irrelevant to be used for the classification process. An example of a stopword process can be seen in Figure 6.

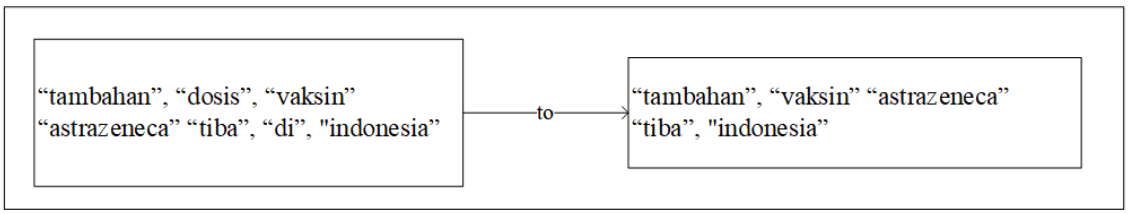

Figure 6. Stopword Removal Process.

E. Stemming

Stemming is the process of changing a word into its basic. changing a word into its basic form is to group words that have the same meaning but different forms, the different forms in a word are occurs because of affixes, This process will use the same library as the stopword process, namely Sastrawi, Sastrawi itself contains Indonesia basic form words for stemming. An example of this process can be seen in Figure 7.

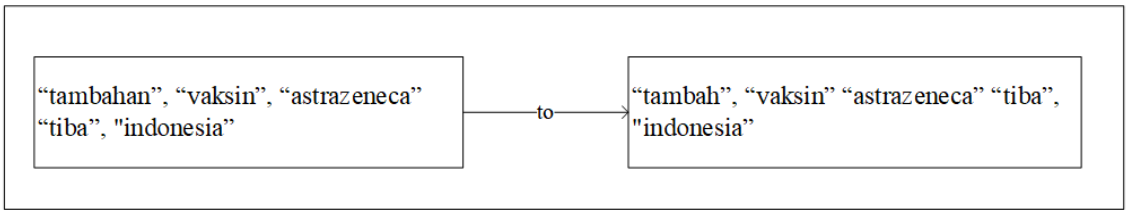

Figure 7. Stemming Process.

2.3. Feature Extraction

After the data were cleaned by some process of Preprocessing, then the Feature Extraction is conducted by using TF-IDF. TF-IDF or Term Frequency Inverse Document Frequency is a method to see how often the frequency of a word arises on the document [14]. The way of work of the TF-IDF algorithm is by changing a document into a matrix per word to calculate the weight.

The more often the word appears, the greater the weight of the word, which means the word is important. TF (Term Frequency) calculate the frequency of word that appear in the dataset or corpus, with the following equation.

$$tf(i) = \frac{frequency(i)}{\sum frequency(t)} \quad (2)$$

Where $frequency(i)$ is the word frequency that occurred on the document, and $\sum frequency(t)$ is the total of a word in the document. IDF (Inverse Document Frequency) is a measure of how frequently or infrequently a word appears in the full dataset or document, the less frequently a term or word appears in the document, the higher the IDF number. On the other hand, if the word appears often enough, the resulting IDFvalue will be small, with the following equation.

$$idf(i) = \log \frac{t}{dfi} \quad (2)$$

Where t is the number of documents, and dfi is the number of documents containing words. After calculating TF and IDF, the TF-IDF value can be determined by the following equation.

$$tfidf(i) = \text{tf(i)} * idf(i) \quad (3)$$

2.4. K-Fold Cross Validation

K-Fold Cross Validation is a technique that aids in the assessment of the system's prediction output. Subsets of the k iterations are created by dividing dataset into subset number of k iteration. Each subset has both train data and test data, with test data making up a larger proportion of the train data. There will be an equal quantity of data between each subgroup. K-Fold Cross Validation is frequently utilized to avoid or reduce system overfitting [15]. The K-Fold Cross Validation procedure with a number of iterations with k = 5 is demonstrated in the example below.

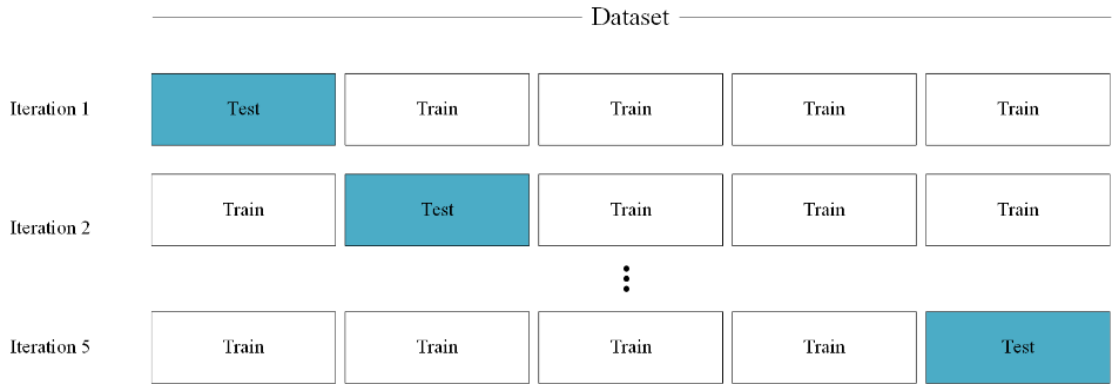

Figure 7. K-fold cross validation

It can be seen in the figure above that the total of the data test is approximately 20%, while the data train is higher, which is about 80% of the dataset used.

2.5. Levensthein Distance

Furthermore, calculations are carried out using the Levenshtein Distance method, Levenshtein distance is a method that compares 2 string or in this case news data similarity and structure to classify data, this method use matrix to calculate the distance between those 2 string.

The distance from these two strings is determined by several operations, namely insertion, deletion, and substitution [16].

$$lev_{a,b}(i,j) = \begin{cases} max\ (i,j) \\ 1 + min \begin{cases} lev_{a,b}(i-1,j) \\ lev_{a,b}(i,j-1) \\ lev_{a,b}(i-1,j-1) + (a_i \neq b_j) \end{cases} \end{cases} \quad (4)$$

In which the first minimum $lev_{a,b}(i-1,j)$ is the calculation of the deletion operation, then the second is the calculation $lev_{a,b}(i,j-1)$ for the insertion operation, and the last one $lev_{a,b}(i-1,j-1) + (a_i \neq b_j)$ is the calculation of the substitution. After performing the calculation operation, after calculate those operation Levenshtein Distance will find the similarity value between each word using the following equation [17].

$$similarity\ (str1, str2) = 1 - \frac{lev_{a,b}(i,j)}{maxLenght(str1,str2)} \quad (5)$$

Where $str1, str2$ are the first and second string or hoax news data, those strings will be compared with each other's to see their similarity based on their word structure and character arrangment, the process will be carried out by $lev_{a,b}(i,j)$, $maxLenght(str1, str2)$ is used to find the largest string length between $str1, str2$. Assuming that the similarity value spans from 0 to 1, with 1 being the greatest, the value of 1 implies that the two words are highly similar or, in other words, identical. this method can measure the similarity value based on the structure of the word character arrangement.

2.6. Performance Measure

In this research, F1-Score is the main metric used to measure the performance of the system. F1-score is quite often used to evaluate the system performance made after the models are formed. The confusion matrix is a method to measure and evaluate the performance result from artificial intelligence systems that are used explicitly in classification models [9]. In the Confusion Matrix, there are four outputs there are Accuracy, Precision, F1 Score, and Recall, then in this research will use the four outputs with the main metric will f1 score, with the following equation.

$$F1\ Score = \frac{2 * precison * Recall}{(Precision + Recall)} \quad (6)$$

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (7)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (8)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (9)$$

TP (True Positive) is where the system predicts correctly on true or positive class, TN (True Negative) is the same like but the system predict correctly on false or negative class, FP (False Positive) is the opposite of TP where the system incorrectly predicts the positive class, and FN (False Negative) is where the system incorrectly predicts false or negative class.

3. Result and Discussion

The experiment in this research will be conducted by the data collected as many as 2.429 about the news of Covid-19 vaccinations that had been labeled when doing data crawling between hoax and non-hoax with the distribution of hoax and non-hoax, on this system will be tested using 2 experiments where the experiment-1 where the system will not use TF-IDF and the other experiment will use TF-IDF. These two Experiment are designed to examine the impact of TF-IDF on system performance. Table 2 displays the outcomes of both tests.

Table 2. Experiment Result

k	**Experiment-1 without using TF-IDF**		**Experiment-2 using TF-IDF**	
	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)
1	57.5	60.5	69.2	57.7
2	56.2	59.4	70	58
3	57.2	60.7	69.7	57.5
4	56.3	59.6	70.2	58.1
5	57.9	62	68	57.1

The results in this experiment-1 without using TF-IDF show that the best results are at k = 5 with a value of f1-score 57.9%, with results that are not much different from other k iterations. The f1-score value of each k iteration is relatively low, ranging from 56% to 57 %. On The other hand, the results from Experiment-2 show quite different results from experiment-1, where the results this experiment show a significant increase in f1-score, best results for experiment-2 are at k = 4 with an f1-score value of 70.2%, the results from each iteration are not much different from each other as happened in the first experiment.

With the results obtained, it can be stated that there is an improvement from one of the previous studies, previous studies obtained results of 40% without using feature extraction [9], it can be seen in this study that has tried to solve the shortcomings of the previous research and get good enough results, experiment-1 there is an increase in accuracy, namely 57% improved outcomes from experiment-1, and a more substantial improvement is seen in experiment-2, which is 70.2% with the inclusion of feature extraction TF-IDF, we can conclude that TF-IDF feature extraction improves system performance using the Levenshtein distance approach.

4. Conclusion

In this research, a hoax detection system has been built by conducting two experiments, namely experiment-1 without using TF-IDF and using TF-IDF. The purpose of these two experiments is to observe the performance of the results of experiments that use TF-IDF and those that do not, experiment-2 resulted in a better model performance than experiment-1, where experiment-1 yielded much better precision than experiment-2, which was 69.2, on the other hand, experiment-2 had much better F1-score results than experiment 1, which was 70.2% for the f1-score. The difference in the results was surely influenced by the feature extraction using TF-IDF. The word weighting performed by TF-IDF greatly affects the results of the system f1-score and recall.

For future work, the system can be improved by increasing the amount of data used and adding other features such as news sources. With the addition of the amount of data and other features, there will surely be an increase in system performance results. The dataset used in this study has an uneven distribution of hoax and non-hoax label data, so datasets that have an even distribution of data can improve results.

Reference

[1] O. D. Apuke and B. Omar, "Fake news and COVID-19: modelling the predictors of fake news sharing among social media users," *Telematics and Informatics*, vol. 56, Jan. 2021, doi: 10.1016/j.tele.2020.101475.

[2] Indonesia Goverment, "Positive Covid-19 indonesia." https://covid19.go.id/peta-sebaran (accessed Nov. 01, 2021).

[3] Kuntarto and R. Widyaningsih, "Motivasi Penyebaran Berita Hoax," *Seminar Nasional Pengembangan Sumber Daya Perdesaan dan Kearifan Lokal Berkelanjutan LPPM UNSOED*, pp. 209–215, Oct. 2020.

[4] Indonesia Goverment (Kominfo), "Hoax in indonesia." https://aptika.kominfo.go.id/2021/05/kominfo-catat-1-733-hoaks-covid-19-dan-vaksin/ (accessed Nov. 01, 2021).

[5] Y. Madani, M. Erritali, and B. Bouikhalene, "Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets," *Results in Physics*, vol. 25, Jun. 2021, doi: 10.1016/j.rinp.2021.104266.

[6] M. Aldwairi and A. Alwahedi, "Detecting fake news in social media networks," in *Procedia Computer Science*, 2018, vol. 141, pp. 215–222. doi: 10.1016/j.procs.2018.10.171.

[7] P. Agung B, I. R. Rizal, E. Dania, S. Yosua Alvin Adi, A. M, and S. Aghus, *Hoax Detection System on Indonesian News Sites Based on Text Classification using SVM and SGD*. Semarang: Proc. of 2017 4th Int. Conf. on Information Tech., Computer, and Electrical Engineering (ICITACEE), 2017.

[8] Adzlan Ishak, Y.Y. Chen, and Suet-Peng Yong, "Distance-based Hoax Detection System," *2012 International Conference on Computer & Information Science (ICCIS)*, p. 1132, 2012.

[9] S. Y. Yuliani, S. Y. Yuliani, S. Sahib, M. F. Abdollah, Y. S. Wijaya, and N. H. M. Yusoff, "Hoax news validation using similarity algorithms," in *Journal of Physics: Conference Series*, Jun. 2020, vol. 1524, no. 1. doi: 10.1088/1742-6596/1524/1/012035.

[10] B. L. Devi, A. Soni, S. S. Kapkoti, and S. Shankar, "Fake News Detection Based on Machine Learning by using TFIDF," *International Journal of Engineering Science and Computing IJESC*, 2019.

[11] T. Widaretna and J. Tirtawangsa, "Indonesian Hoax Identification on Tweets Using Doc2Vec," *Telkom University*, 2021.

[12] A. Afriza and J. Adisantoso, "Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis", [Online]. Available: http://journal.ipb.ac.id/index.php/jika

[13] S. García, J. Luengo, and F. Herrera, "Intelligent Systems Reference Library 72 Data Preprocessing in Data Mining," 2015. [Online]. Available: http://www.springer.com/series/8578

[14] Y. T. Zhang, L. Gong, and Y. C. Wang, "Improved TF-IDF approach for text classification," *Journal of Zhejiang University: Science*, vol. 6 A, no. 1, pp. 49–55, Jan. 2005, doi: 10.1631/jzus.2005.A0049.

[15] D. Berrar, "Cross-validation," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.

[16] B. P. Pratama and S. A. Pamungkas, "Analisis Kinerja Algoritma Levenshtein Distance Dalam Mendeteksi Kemiripan Dokumen Teks," *Jurnal Matematika "Log!k@,"* vol. 6, no. 2, pp. 131–143, 2016.

[17] D. Winarsono, D. O. Siahaan, and U. Yuhana, "Sistem Penilaian Otomatis Kemiripan Kalimat Menggunakan Syntactic-Semantic Similarity Pada Sistem E-Learning," *Jurbal Ilmiah Kursor*, vol. 5, no. 2, 2009.