ILKOM Jurnal Ilmiah Vol. 15, No. 3, December 2023, pp. 398-404
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

N-gram and Kernel Performance Using Support Vector Machine Algorithm for Fake News Detection System

Deny Jollyta [a,1]; Gusrianty [a,2]; Prihandoko [b,3,*]; Darmanta Sukrianto [c,4]
[a] *Institut Bisnis dan Teknologi Pelita Indonesia, Jl. Jend. Ahmad Yani No.78-88, Pekanbaru, Indonesia*
[b] *Universitas Gunadarma, Jl. Margonda Raya 100, Depok, Indonesia*
[c] *AMIK Mahaputera, Jl. HR. Subrantas No.77, Pekanbaru, Indonesia*
[1] *deny.jollyta@lecturer.pelitaindonesia.ac.id; gusrianty@lecturer.pelitaindonesia.ac.id; prihandoko@gmail.com; darman1407@gmail.com*
** Corresponding author*

Article history: Received July 26, 2023; Revised July 27, 2023; Accepted August 14, 2023; Available online December 16, 2023

Abstract

The modern technological advancements have made it simpler for fake news to circulate online. The researchers have developed several strategies to overcome this obstacle, including text classification, distribution network analysis, and human-machine hybrid methods. The most common method is text categorization, and many researchers offer deep learning and machine learning models as remedies. An Indonesian language fake news detection system based on news headlines was developed in this work using the Support Vector Machine (SVM) kernel and n-gram. The objective of this research is to identify the model that produces the best performance outcomes. The system deployment on the web will employ the model that produces the greatest outcomes. According to the research findings, the linear kernel SVM algorithm produces the best results, with an accuracy value of 0.974. Furthermore, the bigram feature used in the development of a classification model does not increase the precision of fake news identification in Indonesian. Utilizing the unigram function yields the most accurate results.

Keywords: Fake News; Kernel; N-gram; Support Vector Machine; Text Classification.

Introduction

Indonesia is one of the countries experiencing an upward trend in the number of internet users. According to data from a Kepios research done in 2022–2023, there were 10 million more internet users in Indonesia over these two years (+5.2%). According to Data Reportal 2023, there were 212.9 million internet users in Indonesia in January of that year. Users from diverse backgrounds can easily find information on the most recent news on the internet. However, it happens regularly that some of the most recent news being circulated online is fake.

The term “fake news” has numerous definitions, including false information, statements that have no relationship to reality, information that is intentionally created to deceive or influence the public, and news that is created to gain clicks or ideological advantages [1]. Propaganda and conspiracy theories are two other common online sources of misinformation. The distinction between fake information and actual news is blurred when people are exposed to it [2]. The way fake news is presented and made to appear to be true news serves to support this. Additionally, the development of technology in this century has facilitated the widespread dissemination of fake news on the internet. Finding fake news in the digital age is therefore challenge for researchers.

Based on the results of the numerous studies, the researchers developed several of methods, including text classification, distribution network analysis, and human-machine hybrid methods, to assess the veracity of news stories [3]. The study indicates that the predominant method for proposing solutions using deep learning and machine learning models is classification of texts. The objective of this research is to identify the model that produces the best performance outcomes. The system deployment on the web will employ the model that yields the best outcomes.

Text classification is the process of categorizing text based on a specific pattern seen in the data gathered through processing text or known as text mining. Natural language processing, machine learning, and data mining are used to automatically detect these patterns in electronic text [4]. In order to find unique information using machine learning, text mining involves turning unstructured text into a structured manner [5]. A classification algorithm from data mining can be used for classification [6], such as Support Vector Machine (SVM) algorithm. In

this study, a fake news detection system for the Indonesian language based on news titles was developed using the SVM kernel and n-gram. Kernels from the SVM algorithm, including linear, polynomial, radial basis function, and sigmoid kernels applied.

The linear kernel SVM algorithm is employed in research [7] to produce the best predicted results in detecting fake news based on news titles with the Bootstrap Aggregation technique. Additionally, employs different iterations of the Naïve Bayes algorithm and SVM to identify fake Indonesian news that has been gathered by crawlers [5]. The results indicated that the sigmoid kernel SVM achieved the best prediction results, with 95.6% precision, 100% recall, 97.7% f1-score, and 96.5% accuracy. In other studies, merging n-grams can help deep learning models perform better [8]. The outcomes demonstrated that their deep learning model had a 99.88% accuracy rate. According to the research findings from this study, n-grams contribute to improving the precision of fake news identification when utilizing a deep learning algorithm technique.

Method

This research was carried out in stages, commencing with data collecting, preparation, and deployment model for the Indonesian language fake news detection system on the web. **Figure 1** displays the stages of this research.

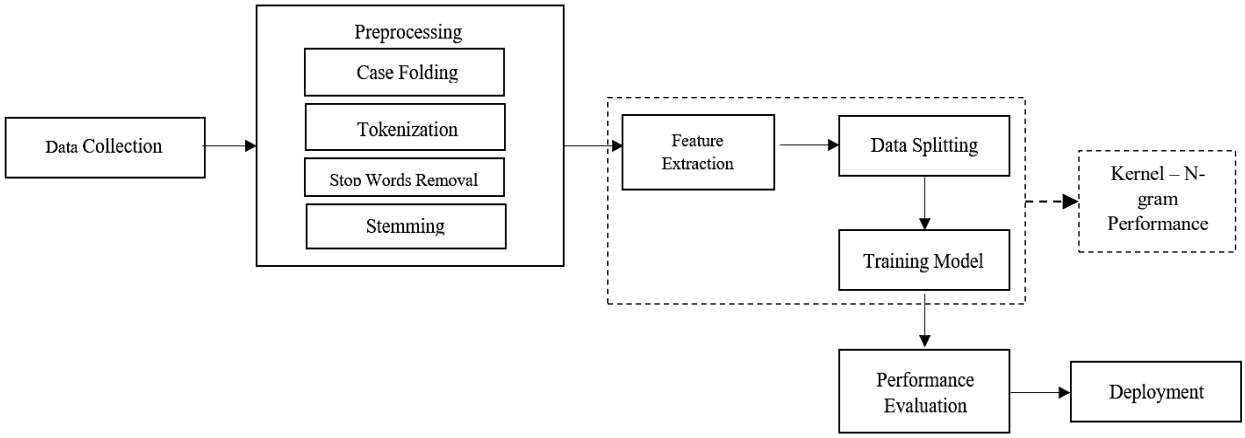


Figure 1. Research stages using SVM

A. Data Collection

The dataset for this study comprises news information obtained through web scraping with the Beautiful Soup library from the Kompas.com website. Reliable news data can be acquired at https://news.kompas.com/search, whilst information for fake news can be found at https://www.kompas.com/tag/hoaks. On the URL for downloading the reliable news, Kompas.com, there are a number of fake news pieces that spread fake news. When reliable news data is pulled, the news will be removed from the results.

There were 1,013 pieces of fake news that appeared during the 2022–2023 timeframe and 1,027 news reports that were withdrawn due to factual errors. As illustrated in **Figure 2**, the news data that will be retrieved includes the title, date, and news tag (HOAX OR FACT). Following the completion of data collection, the news tag will be modified, changing the news tag in the factual news data on the compass to Fact and Fake for Kompas.com fake news data.

Figure 2. Removal of news data

B. Preprocessing

Preprocessing has the ability to enhance the quality of some datasets in general and is especially beneficial for text categorization [9]. The news data is cleaned up and transformed during the preprocessing stage into a format that is simpler for a computer to understand and evaluate [10], [11]. Four steps make up this procedure including case folding, tokenization, stop words elimination, and stemming.

According to [12], the case folding is to recognize an employment of capital letters inconsistently throughout a document, which might make it difficult to classify it, which makes the presence of capitalization problematic. Additionally, noise-producing non-letter characters, including punctuation, numerals, and special characters are eliminated. The study [13] explained that tokenization is a process of cutting strings from each sentence into single words that stand alone. These single words are also known as tokens. The third step of preprocessing is called stop word removal. Stop word usually have little or no meaning, for example, the words "which," "in," "to," "from". The main function of stop word removal is so that stop words do not affect the results of the next process [14]. The final step of preprocessing is stemming. Stemming is the process of decomposing a word into its basic word by removing affixes. In Indonesian language texts, suffixes, prefixes and confines are omitted [15].

C. Feature Extraction

In order to extract the features of news headlines, this study uses Term Frequency-Inverse Document Frequency (TF-IDF) and n-grams. The two elements that make up TF-IDF are Term Frequency (TF) and Inverse Document Frequency (IDF). According to Ramos (2003), a word with a high TF-IDF score indicates that it has a close connection to the text in which it appears. Important terms in certain documents, such news articles, can be found using TF-IDF.

According to research [8], an n-gram is a collection of n items that appear concurrently or continuously in a text or lengthy string of sentences. The term on the news headline is the object of this study. In this study, TF-IDF was used to extract one word (unigram) and two words (bigram) from news data that had passed preprocessing.

D. Data Splitting and Training Model

In this study, the dataset is split into two categories: training and testing data, with a ratio of 60% training data and 40% testing data. To distinguish between fake and actual news, the four kernels of the SVM are used. The kernels are Sigmoid, Linear, Polynomial, and Radial Basis Function (RBF). The equation for the four kernels is shown below [5].

a. Kernel Linear

$$K(x, x_i) = x.x^T \tag{1}$$

b. Kernel Polynomial

$$K(x, x_i) = (1 + x.x_i^T)^d \tag{2}$$

c. Kernel Radial Basis Function (RBF)

$$K(x, x_i) = exp - \gamma \|x - x_i\|^2 \tag{3}$$

d. Kernel Sigmoid

$$K(x, x_i) = \tanh(\gamma x_i^T x_j + r) \tag{4}$$

Where,

K : Kernel Function; γ : Gamma Parameter; d : Degree Parameter; r : Coefficient Parameter; x : Kernel Parameter

The optimal model performance is determined by comparing the outputs of the kernel and n-gram SVM models for text classification. Each kernel works using n-grams. The results of each model are compared to determine the best model performance.

E. Performance Evaluation

This study utilizes the confusion matrix, accuracy, precision, recall, and f1-score to display the performance quality of the four SVM kernel models. The equations for the four performance evaluation tools are displayed in Equation 5 to 7 [7].

a. Accuracy

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

b. Precision

$$precision = \frac{TP}{TP + FP} \tag{6}$$

c. Recall

$$recall = \frac{TP}{TP + FN} \tag{7}$$

d. f1-score

$$f1 - score = \frac{TP + TP}{TP + TP + FP + FN} \quad (8)$$

Where,

TP : True Positive; TN : True Negative; FP : False Positive; FN : False Negative

F. Deployment

Deployment is done online following the conclusion of the performance review. Pickle is a Python library that is utilized for deployment [16]. Pickle is used to store algorithm models and tools for feature extraction, the flask library as a framework for web development, and machine learning libraries like Sklearn, NLTK, and Sastrawi [17], which are used for preprocessing, feature extraction, and news identification.

Results and Discussion

A. Post Preprocessing Word Cloud Dataset Analysis

After preprocessing, word cloud analysis was used to compare the terms found in the fact and fake news datasets. A word is frequently used in a text document if it appears in the word cloud in a large size. The keywords identified in the two datasets in question are displayed in **Figures 3 (a)** and **(b)**.

(a) (b)

Figure 3. Word cloud of fact news (a) and fake news (b)

The keywords discovered in the two datasets under consideration are shown in **Figure 3**. "*buka puasa"* meaning iftar, Indonesia, and "*Eid*" are the terms that stand out in **Figure 3 (a)** in huge type. The word cloud information is the fact news dataset that kompas.com is now using extensively to talk about the fasting month. Referring to **Figure 3 (b)**, the terms "ferdy", "sambo", "hoax", and "police" are displayed in huge type. The dataset of fake news from Kompas.com, which primarily discusses hoaxes with Ferdy Sambo material, is the information that was gleaned from the word cloud.

B. N-gram Analysis on The Dataset

Refers to Equation 1 to 4, the n-gram analysis is carried out. A significant set of n-grams that appear in the fact and fake news databases were searched for using this methodology. By using Equation 5 to 8, the most significant TF-IDF values and unigrams from the feature extraction process are displayed in **Figure 4**. "*Jadwal*" meaning schedule is the unigram with the highest TF-IDF value for factual news, while "*hoaks*" or hoax is the unigram with the highest TF-IDF value for fake news, as seen by the data from **Figure 4**.

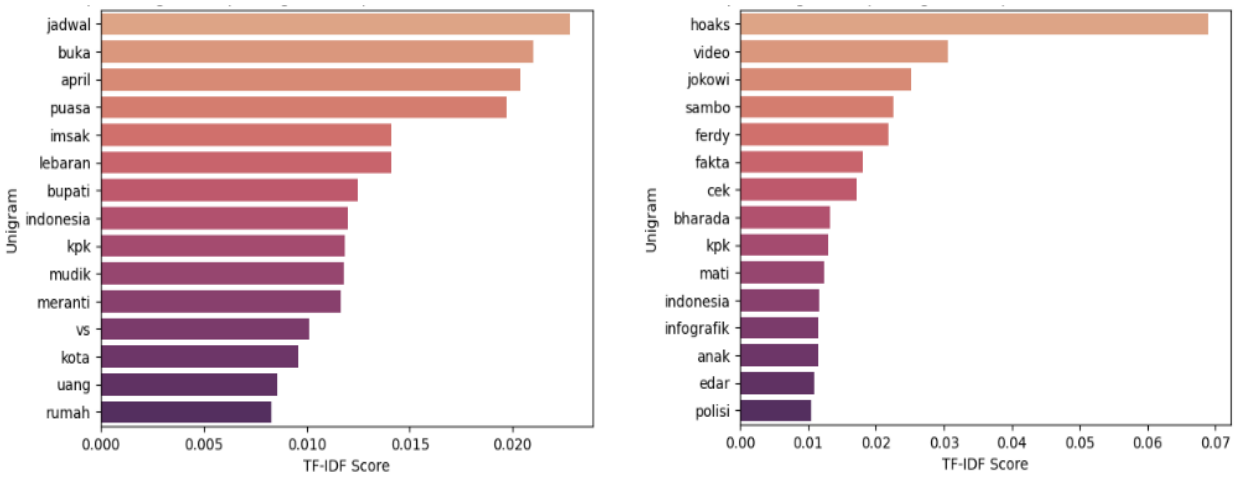


Figure 4. The most important unigrams with TF-IDF values for facts and fakes news (left to right)

After performing the feature extraction method, **Figure 5** displays the most crucial bigram together with the TF-IDF value that refers to Equations 1 to 8. "*Buka puasa*" meaning iftar is the bigram with the greatest TF-IDF value in factual news, and "ferdy sambo" is the bigram with the highest TF-IDF in fake news.

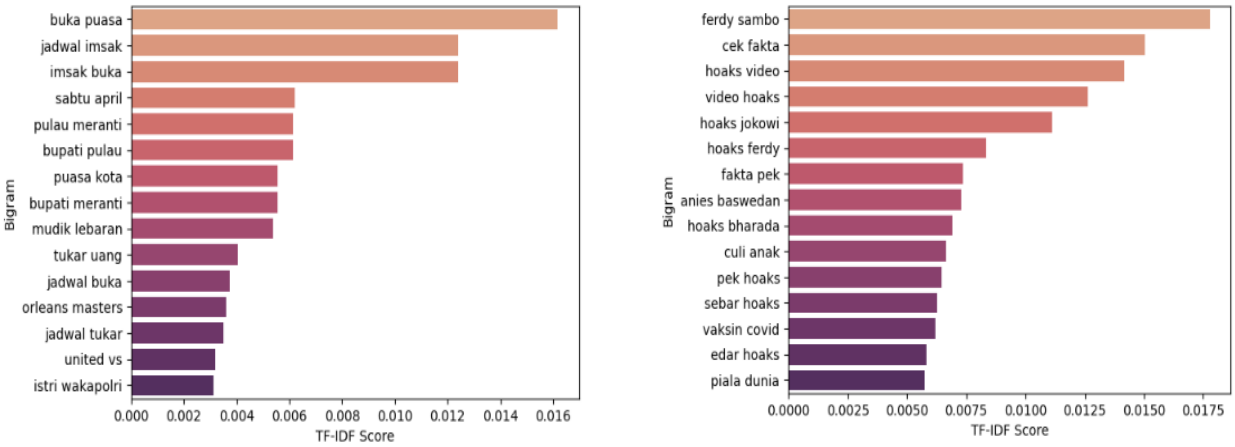


Figure 5. The most important bigrams with TF-IDF values for fact and fake news (left to right)

A further detail of information about news content is discovered based on **Figure 5**. The "Orleans Masters" and "Meranti Island", for instance, are topics covered in news articles. The "World Cup", "Anies Baswedan", "Covid vaccinations", and "Jokowi" are all the subjects of fake news hoaxes.

C. Model Performance Analysis

Following the classification, the four kernel models for the SVM obtain a performance evaluation. Accuracy, precision, recall, and the f1-score are used to evaluate performance. **Table 1** displays the confusion matrix findings for the four SVM kernel models, while **Table 2** displays the accuracy, precision, recall, and f1-score values.

Table 1. Confusion Matrix Data Testing Results for the Four Kernel Models

No	Model	N-gram	TP	FP	FN	TN
1	Linier	Unigram	409	19	2	386
		Bigram	405	124	6	281
2	Polynomial	Unigram	400	88	11	317
		Bigram	409	173	2	232
3	RBF	Unigram	410	23	1	382
		Bigram	408	166	3	239
4	Sigmoid	Unigram	409	21	2	384
		Bigram	406	142	5	263

Table 2. Performance Evaluation for the Four Kernel Models

No	Model	N-gram	News	Accuracy	Precision	Recall	f1-score
1	Linier	Unigram	Fact	0.974	0.99	0.95	0.97
			Fake		0,96	1,00	0,97
		Bigram	Fact	0.841	0,98	0,69	0,81
			Fake		0,77	0,99	0,86
2	Polynomial	Unigram	Fact	0.879	0,97	0,78	0,86
			Fake		0,82	0,97	0,89
		Bigram	Fact	0.786	0,99	0,57	0,73
			Fake		0,70	1,00	0,82
3	RBF	Unigram	Fact	0.971	1,00	0,94	0,97
			Fake		0,95	1,00	0,97

No	Model	N-gram	News	Accuracy	Precision	Recall	f1-score
		Bigram	Fact	0.793	0,99	0,59	0,74
			Fake		0,71	0,99	0,83
4	Sigmoid	Unigram	Fake	0.972	0,99	0,95	0,97
			Fact		0,95	1,00	0,97
		Bigram	Fact	0.820	0,98	0,65	0,78
			Fake		0,74	0,99	0,85

D. Deployment

The linear kernel SVM model with the unigram feature yields the best results, according to the performance evaluation from **Tables 1** and **2**. This implies that models and feature extraction tools can be kept for use online. These two items are kept in street vendor format and then added to the web's coding structure. Installing Python 3.8 as a Path, installing libraries, and setting up a Virtual Environment are all necessary before the system can be launched on the web.

This system comprises two pages including the home page, which loads first when the system is activated, and the news detection page, where news titles are entered, and news detection is performed. The study of fake news detection online technology recognizes news titles supplied as searches. A classification model in the form of a SVM model examines entered queries by utilizing the unigram feature. The web system displays the line "News Title Detection Results According to the System are → Facts" on the news detection page if the news title entered is recognized by the system as a fact news title. In contrast, if the web system recognizes the news headline as a false news title, the phrase will change to False. According to the Indonesian language fake news detection system, **Figures 7** and **8** display page for home page, news detection and news detection results, respectively.

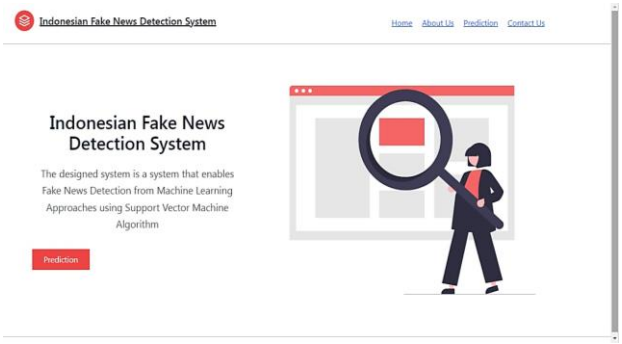
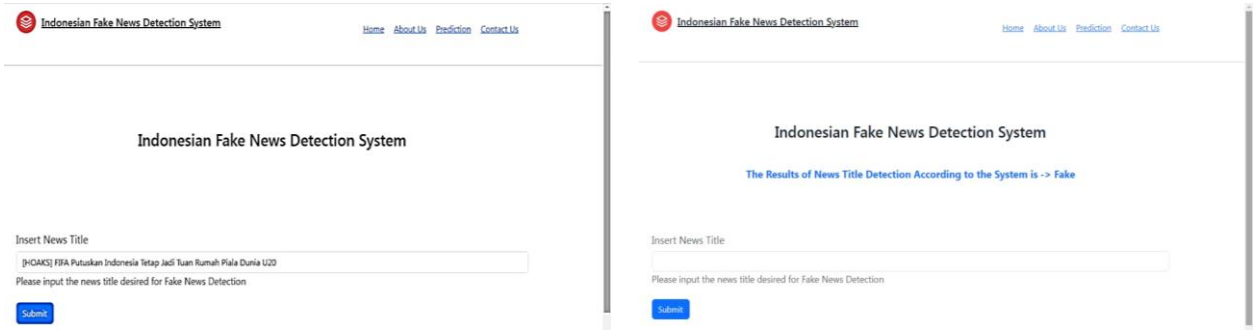


Figure 7. View of Home Page



(a) (b)

Figure 8. View of the News Detection Page (a) and News Detection Results (b)

Conclusion

According to research findings, the linear kernel SVM method performs superior than bigram. The linear kernel, which is equal to 0.974, yields the best accuracy. The accuracy of identifying fake news in Indonesian is not increased by using the bigram feature in the classification model. One could argue that the bigram has little bearing on the identification of fake news. The unigram feature yields the most precise results, allowing for the right web

deployment builds according to this accuracy. Deployment on web is highly interactive and extensible for other classification algorithms.

References

[1] J. P. Baptista, A. Gradim, and E. Correia, "The relationship between the belief in fake news and the strategies to seek information from young Portuguese people," *Observatorio*, vol. 16, no. 3, pp. 203–232, 2022, doi: 10.15847/obsOBS16320222082.

[2] M. Karami, T. H. Nazer, and H. Liu, "Profiling Fake News Spreaders on Social Media through Psychological and Motivational Factors," 2021. doi: 10.1145/3465336.3475097.

[3] P. Nordberg, J. Kavrestad, and M. Nohlberg, "Automatic detection of fake news," in *6th International Workshop on Socio-Technical Perspective in IS Development (STPIS'20)*, 2020, pp. 168–179.

[4] M. Albahar and J. Almalki, "Deepfakes: Threats and countermeasures systematic review," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 22, pp. 3242–3250, 2019.

[5] R. R. Sani, Y. A. Pratiwi, S. Winarno, E. D. Udayanti, and F. Alzami, "Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia," *J. Masy. Inform.*, vol. 13, no. 2, pp. 85–98, 2022, doi: 10.14710/jmasif.13.2.47983.

[6] D. Jollyta, G. Gusrianty, and D. Sukrianto, "Analysis of Slow Moving Goods Classification Technique: Random Forest and Naïve Bayes," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 5, no. 2, pp. 134–139, 2019, doi: 10.23917/khif.v5i2.8263.

[7] B. Probierz, P. Stefanski, and J. Kozak, "Rapid detection of fake news based on machine learning methods," *Procedia Comput. Sci.*, vol. 192, no. January, pp. 2893–2902, 2021, doi: 10.1016/j.procs.2021.09.060.

[8] T. Chauhan and H. Palivela, "Optimization and improvement of fake news detection using deep learning approaches for societal benefit," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 11, 2021, doi: 10.1016/j.jjimei.2021.100051.

[9] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, pp. 1–22, 2020, doi: 10.1371/journal.pone.0232525.

[10] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," *Mach. Learn. with Appl.*, vol. 4, no. March, p. 12, 2021, doi: 10.1016/j.mlwa.2021.100032.

[11] V. Kumar, A. Kumar, A. K. Singh, and A. Pachauri, "Fake News Detection using Machine Learning and Natural Language Processing," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 844–847, 2019, doi: 10.1109/ICTAI53825.2021.9673378.

[12] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *Resti*, vol. 1, no. 1, pp. 19–25, 2017.

[13] R. Rahmaddeni, M. K. Anam, Y. Irawan, S. Susanti, and M. Jamaris, "Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination," *Ilk. J. Ilm.*, vol. 14, no. 1, pp. 32–38, 2022, doi: 10.33096/ilkom.v14i1.1090.32-38.

[14] A. E. Budiman and A. Widjaja, "Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 3, pp. 475–488, 2020, doi: 10.28932/jutisi.v6i3.2892.

[15] L. Agusta, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia," in *Konferensi Nasional Sistem dan Informatika 2009*, 2009, no. KNS&I09-036, pp. 196–201.

[16] P. Kanani and M. Padole, "Deep learning to detect skin cancer using google colab," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 2176–2183, 2019, doi: 10.35940/ijeat.F8587.088619.

[17] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 874, no. 1, pp. 1–7, doi: 10.1088/1757-899X/874/1/012017.