Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (185-193)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.34084

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Hoax News Detection Using Passive Aggressive Classifier And TfidfVectorizer

**Maulana Fajar Lazuardi[1], Renaldy Hiunarto[2],
Kareena Futri Ramadhani[3], Noviandi[*4], Riya Widayanti[5], Muhamad Hadi Arfian[6]**

[1,2,3,4,6]Informatics Engineering, Faculty of Computer Science, Esa Unggul University
[5]Information System, Faculty of Computer Science, Esa Unggul University

[1,2,3,4,5,6]Arjuna Utara, Duri Kepa, Kebon Jeruk, Jakarta Barat, 11510.

Article:
Accepted: July 23, 2023
Revised: June 02, 2023
Issued: October 28, 2023

***Correspondence Address:**
noviandi@esaunggul.ac.id

ABSTRACT

Indonesia has one of the highest numbers of social media users. Million social media users in Indonesia is expected to reach 167 million by January 2023. These users are spread across various social media platforms including Twitter, with 24 million users. The high number of social media users on Twitter has made the information validation process even more neglected. Moreover, the trend in news interest read by social media users was adjusted only to their individual tastes. This phenomenon is evidenced by the large amount of fake news (hoaxes) circulating in society, which are spread through social media. Therefore, an accurate machine learning model is required to classify "real" and "hoax" news. This study uses the TfidfVectorizer algorithm and Passive Aggressive Classifier for datasets shared through the Kaggle site. The contents of the dataset were sourced via social media Twitter over a span of five years, namely 2015-2020. At the preprocessing stage, to create the Confusion Matrix, the machine learning model shows that it can work well as expected, namely obtaining Accuracy, Precision, and Recall scores of 82.44%, 80.66%, and 82.44%, respectively. In addition, the results of the confusion matrix show that in the dataset used, there is more "real" news than "hoaxes," that is, the model is able to predict 1059 real news and 211 hoax news, with actual conditions 1106 real news and 164 hoax news.

Keywords: *machine learning, hoax news, social media, classification, tfidfvectorizer, passive aggressive classifier*

1. INTRODUCTION

In the contemporary era of information dissemination, the emergence of digital platforms and social networks has facilitated the rapid spread of news, giving rise to deep concerns – the dissemination of false information and fake news. The ability to discern between authentic news and fabricated content has become crucial to maintain the integrity of information sources and ensure an informed society, especially amidst the popularity of social media as a primary source of information for many individuals.

Social media is a popular communication medium today. This is inseparable from the convenience provided by social media to disseminate information. The ease of interacting with other people is an attraction for users [1]. Users need only create an account to start using these facilities. It is no wonder that many Indonesians actively use social media.

Indonesia has one of the largest populations of social media users worldwide, with 167 million users staggering as of January 2023. These users flock to a variety of social media platforms, including Twitter, which boasts of 24 million users in the country [2].

The high number of social media users on Twitter makes the information validation process even more neglected. Moreover, the trend of interest in news read by social media users is only adjusted to the tastes of each individual, making the truth of the information consumed by the public even more misleading. This phenomenon is demonstrated by the circulation of fake news (hoaxes) in society, which spreads through social media [3]. Between January and March, 2023, the Ministry of Communication and Informatics identified 425 instances of fake news. It shows an increase from the previous year during the same period, which saw 393 instances of fake news [4]. This rise in fake news could potentially result in misinformation spreading throughout society.

Fake news contains misleading news and is generally spread through social and other online media. Done to advance or impose certain ideas with their aims and objectives. In the Big Indonesian Dictionary (KBBI), the word hoax is defined as false or lying [5]. This type of news may contain inaccurate or exaggerated claims, and has the potential to go viral through algorithms, potentially trapping unsuspecting users. For example, medical research in Taiwan showed that the spread of hoaxes about the COVID-19 vaccine resulted in fewer vaccine doses absorbed by society because those who are exposed to hoaxes become hesitant and even anti-vaccine [6].

Numerous algorithms have been used to detect fake news through classification, including the Support Vector Machine (SVM) [7], K-Nearest Neighbor (KNN) [8], Naive Bayes [9], and Random Forest [10].

The classification methods commonly used in previous studies for hoax news include Naive Bayes and SVM. Naive Bayes is a classification method based on probability and Bayesian Theorem [11]. Each method has its own strengths and weaknesses. Naive Bayes has the advantage of being implementable on large datasets and capable of handling missing values [12]. However, Naive Bayes assumes that features are independent or not related to each other [13]. In its use, these features may be interconnected, making them unable to be encoded in Naive Bayes. On the other hand, SVM has the drawback of being challenging to use for large-sized data and is specifically designed for binary classification [14].

Several studies have been conducted on fake news classification using different algorithms. Yonathan et al. used the SVM algorithm and achieved an accuracy rate of 78.02% [15]. Dhaneswara et al. conducted a study using the KNN algorithm and achieved an accuracy rate of 72.5% [16]. Soleman used the Naïve Bayes algorithm and achieved an accuracy of 72% [17]. Khalil et al. conducted research using the Random Forest algorithm with Particle Swarm Optimization (PSO) and found that it has an accuracy rate of 73% [18].

This study uses a combination of a Passive Aggressive Classifier and TfidfVectorizer. The TfidfVectorizer algorithm calculates the product of Term Frequency and Inverse Document Frequency. Term frequency measures the number of times a term appears in a document, whereas Inverse Document Frequency reduces the weight of terms that frequently appear across various documents, considering the frequency of documents containing the word [19]. In contrast, the Passive Aggressive Algorithm is an online learning algorithm that remains passive when classification results are accurate. However, when an error occurred, the model weights were updated more aggressively to rectify the fault. This enables the model to adapt quickly to

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (185-193)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.34084

changes in data that are presented sequentially [20].

The aim of this research is to distinguish between "Real" and "Fake" news that has been posted on Twitter over a span of 5 years (2015-2020). The researcher took the dataset from Kaggle.com and used the TfidfVectorizer and Passive Aggressive Classifiers. Researchers have also used the Python programming language and Jupyter Notebook as IDEs for data analysis and building machine-learning models.

The significance of this research not only in its potential to contribute to the development of advanced tools for fake news detection but also in addressing the social impact of misinformation or hoax exacerbated by the existence of social media. As the digital landscape continues to evolve and the role of social media becomes increasingly dominant, staying ahead in identifying and combating fake news on these platforms is crucial to maintain the integrity of information sources and foster a well-informed society in this digital era. This study aims also to make a meaningful contribution to ongoing efforts aimed at strengthening the foundations of reliable and trustworthy information dissemination in a digital era dominated by social media.

2. METHODS

The method used in this study is quantitative because researchers are trying to measure, analyze, and interpret data that are measurable and can be calculated. This method relies on statistical models and machine-learning algorithms to analyze data, identify patterns, and predict behavior based on the numerical features of the data. The steps of the quantitative method are illustrated in Figure. 1.

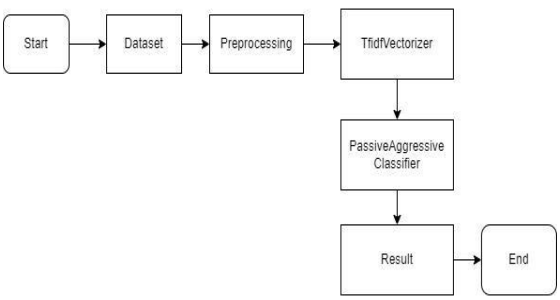

Figure 1.Research framework

2.1. Dataset

The dataset named "Indonesia False News (Hoax)" was acquired from Kaggle and consists of various columns, including ID, label, date, news title, news narrative, and image file name. Figure 2 displays the dataset used, which comprises 4231 rows with different headlines.

	ID	label	tanggal	judul
0	71	1	17-Aug-20	Pemakaian Masker Menyebabkan Penyakit Legionna...
1	461	1	17-Jul-20	Instruksi Gubernur Jateng tentang penilangan ...
2	495	1	13-Jul-20	Foto Jim Rohn: Jokowi adalah presiden terbaik ...
3	550	1	8-Jul-20	ini bukan politik, tapi kenyataan Pak Jokowi b...
4	681	1	24-Jun-20	Foto Kadrun kalo lihat foto ini panas dingin

Figure 2.Dataset

The next step was to split the dataset into two parts, 70% for training and 30% for testing. This division aids in creating a model that can be applied to new data more effectively and with a greater accuracy.

2.2. Preprocessing

During the preprocessing stage, various methods are used, including data cleaning, normalization, and encoding.

a. Data Cleaning

At this stage, the dataset is cleaned using null data and unused columns. This is important because data with null or missing values can interfere with the consistency of the analysis and produce inconsistent results. When there are null values in a dataset, analytical methods such as statistical calculations, visualization, or machine learning models can produce invalid or ambiguous results.

The dataset being used has one column that is not utilized during the preprocessing stage, which is the "image file name" column. Unused fields can add unnecessary noise to a model. The quality of the developed model can be improved by focusing on the important variables and clean data. The removal of unused columns can enhance the quality of the model and generate more precise and dependable predictions. Figure 3 displays the outcome of the null value check and the removal of columns.

```
ID          0
label       0
tanggal     0
judul       0
narasi      0
```

Figure 3.Null values

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (185-193)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.34084

b. Normalization

To improve the analysis of the machine learning model, it is important to change the data type of the description in the label column from an integer consisting of 0 and 1 to a string data type with the values "Real" and "Hoax." This avoids ambiguity and makes it easier to interpret the results later.

	ID	label	tanggal	judul
0	71	Hoax	17-Aug-20	Pemakaian Masker Menyebabkan Penyakit Legionna...
1	461	Hoax	17-Jul-20	Instruksi Gubernur Jateng tentang penilangan ...
2	495	Hoax	13-Jul-20	Foto Jim Rohn: Jokowi adalah presiden terbaik ...
3	550	Hoax	8-Jul-20	ini bukan politik, tapi kenyataan Pak Jokowi b...
4	681	Hoax	24-Jun-20	Foto Kadrun kalo lihat foto ini panas dingin

Figure 4.Change the column label category

c. Encoding (TfidfVectorizer)

Once pre-processing is complete, the next step involves encoding or transforming the text into a numeric format. One method for achieving this is through the use of a TfidfVectorizer, which is formulated using the following equation:

$$TF - IDF = TF \log\left(\frac{n}{DF}\right)$$

This equation calculates the frequency of a term (word) in a document using TF, the number of documents containing the term (word) using DF, and N the total number of documents [17].

For further clarity, here is an example of a sentence taken from the dataset then converted from text to numeric using TfidfVectorizer and displayed with a Boxplot.

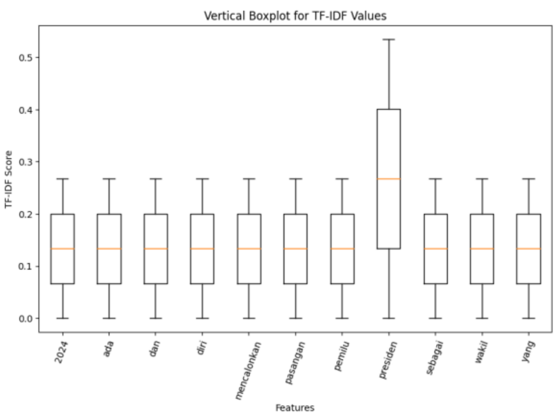

Figure 5.Boxplot for tf-idf values

This transformation depicts the process of converting the original text into a vector representation using the TF-IDF (Term Frequency-Inverse Document Frequency) method. TF-IDF is a technique used to assign weights to words in a document based on how often the word appears in the document and how common the word is across the entire document collection.

In this example, the original text "Pemilu 2024 ada 3 pasangan yang mencalonkan diri sebagai presiden dan wakil presiden" is transformed into a vector with TF-IDF values. Each word in the text has a TF-IDF value reflecting the importance of that word in the document and how common or unique it is across the document collection. Here's an explanation for some elements of the transformation vector:

- "presiden": A higher TF-IDF value (0.53) indicates that the word "presiden" carries more weight in the document, signifying its importance in the original text.
- Other words such as "2024," "ada," "dan," "diri," "mencalonkan," "pasangan," "pemilu," "sebagai," "wakil," and "yang" have lower TF-IDF values (0.27), indicating their presence in the document but with lower weights.
- TF-IDF values reflect the relationships between words in the document and help extract important or unique information from the original text.

By using the TF-IDF vector representation, text can be transformed into a numeric format suitable for various natural language processing and machine learning tasks, such as text classification, sentiment analysis, or document clustering.

2.3. Passive Aggressive Classifier

Passive-Aggressive algorithms are commonly used in large-scale learning and for classification tasks. It is one of the few effective and efficient online learning algorithms for certain applications. Passive-Aggressive algorithms function by acting passively to produce accurate classifications and aggressively for any errors in other machine learning algorithms.

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (185-193)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.34084

The algorithm is trained incrementally using data sets either singly, sequentially, or in small groups, known as mini-sets. This process of improvement is done incrementally as smaller data sets are collected.

Using this algorithm, models trained and deployed in online learning productions can continue to learn as new data sets are collected. As a result, we can conclude that algorithms like Passive-Aggressive Classifier are the most effective for systems that consume data in a continuous stream.

This is especially helpful in situations where there is a lot of data because due to the sheer volume of data, training the complete data set is computationally impossible. Online learning algorithms will simply collect the training dataset, update the classifier, and then discard it. Incremental learning performs one-by-one checks rather than checking all the training samples at once. The obvious benefit is having a smaller memory footprint, which is a huge advantage[21].

After transforming the text, it was classified in the next step. The Passive Aggressive Classifier algorithm was used for classification at this stage. This algorithm categorizes data samples into appropriate classes based on the features provided. The Passive Aggressive Classifier algorithm functions by making iterative adjustments to the weight vector used for the data classification. If the algorithm encounters a misclassified training-data sample, it aggressively updates its weight vector to correct the error. However, if a data sample can be classified correctly, the algorithm maintains its weight vector.

The basic formula for the Passive Aggressive Classifier algorithm is as follows:

$$w\_new = w\_old + (loss / ||x||^2) * y * x$$

This equation displays the process of updating the weight vector (w_new) after analyzing the training data. The weight vector prior to updating is denoted as w_old. The loss or error that is a result of incorrect classification is represented by "loss." Additionally, "||x||^2" symbolizes the square of the Euclidean norm of the feature vector "x." "y" refers to the appropriate class label for the sample data, while "x" denotes the feature vector of the data sample.

This formula shows that the adjustment of the weight vector (w_new - w_old) depends on the loss, feature norm (||x||^2), correct class label (y), and feature vector (x). The constant value (loss / ||x||^2) regulates the extent to which the weights are updated.

The relationship between the preprocessing steps and classifier is depicted in the diagram below.

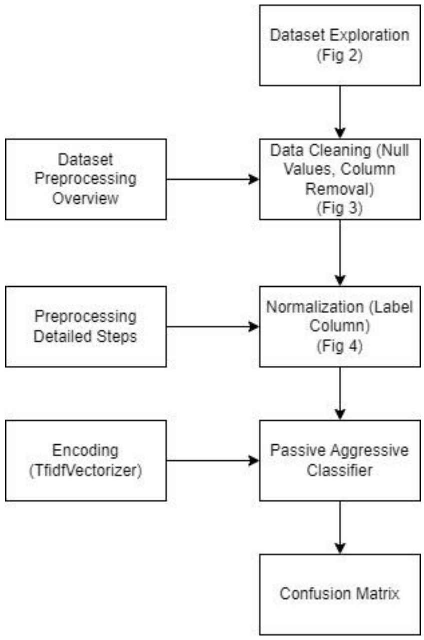

Figure 6.Preprocessing and classification

This iterative process ensures adaptation to sequential data changes and is a crucial part of the overall methodology. The structured methodology, complemented by a visual overview, aimed to provide a clearer understanding of the relationships between the different steps in the study.

3. RESULT AND DISCUSSION

After conducting thorough research using evaluation methods and a matrix, it was found that there were no null data during the preprocessing stage. The news content was successfully transformed into a numeric form using the TfidfVectorizer method. Additionally, there was a normalization stage where the data labels were categorized as "Real" or "Hoax."

When evaluating the performance of the classification model, we used three important metrics: Accuracy, Precision, and Recall. Accuracy measures the percentage of correct predictions made by the model compared with the actual data in the test dataset. Precision is

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (185-193)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.34084

the ratio of the correctly predicted positive results to the total positive predictions made by the system. This helps to determine the percentage of truly fake news among all news stories predicted to be fake. Recall, however, is the ratio of correctly predicted positive results to all actual positive data. It represents the percentage of news stories predicted to be fake out of all the news stories that are actually fake. These three values were calculated using the Confusion Matrix, which shows the predicted results of the system and the actual conditions (actual amount of data) from the data tested by the system as follows:

$$Accuracy: \frac{TP + TN}{TP + TN + FP + FN} x\ 100\%$$

$$Precision: \frac{TP}{(TP + FP)}$$

$$Recall: \frac{TP}{(TP + FN)}$$

3.1. Score Evaluation Results

Score evaluation results refer to the assessment results or scores that result from the evaluation of a model or system. In the context of statistics or machine learning, the metrics or scores used to measure model performance usually vary depending on the specific task, such as classification, regression, or other tasks.

In other words, evaluation scores indicate the values that result from evaluating a model or system, the metrics are chosen based on the objectives of a particular task. The evaluation score helps determine how well the model can meet the set needs or goals[22].

Table 1.Summary of evaluation scores

Accuracy	Precision	Recall
82.44%	80.66%	82.44%

After analyzing the scores in Table 1, it can be inferred that the model has fairly good accuracy and is consistent with relatively high precision and recall. The Accuracy, Precision, and Recall scores were 82.44 %, 80.66 %, and 82.44 %, respectively. Nevertheless, it is essential to consider the specifics of the problem in question and take a closer look at the evaluation matrix, such as the confusion matrix, for a more comprehensive understanding of how the model performs in each category or class.

3.2. Confusion Matrix

The confusion matrix table is used to measure the performance of a classification model because it consists of many rows of test data that are predicted correctly and incorrectly by the classification model. Confusion matrix are a useful tool to analyze how well a clustering and classifier can recognize tuples from different classes [23].

In the development of machine learning models, modeling requires effective performance measurements to evaluate the suitability and level of accuracy of the model in detecting objects or data in real-time. One method for testing the accuracy and precision of modeling in machine learning is the confusion matrix. The confusion matrix is a table that describes four combinations of predicted values and actual values:

- True Positive (TP) has a positive predictive value for positive actual conditions.
- False Positive (FP) has a positive predictive value for negative actual conditions.
- True Negative (TN) has a negative predictive value with actual negative conditions.
- False negatives (FN) have a negative predictive value with positive actual conditions.

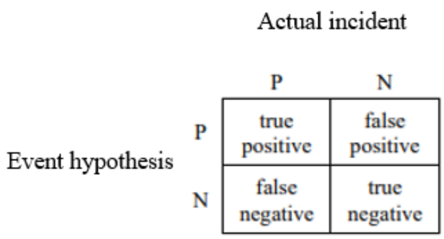

Figure 7.Confusion matrix

Recall is used to measure the fraction of positive patterns that will be correctly classified. Meanwhile, the precision function is the recall of all predicted patterns in the positive class, F-Measure is a measure of the balanced average between recall and precision, and accuracy is the ratio of correct predictions to the overall evaluated sample.

All possible positive true events (P) and all possible negative true events (N) are found in the matrix, as shown in the table above. These values include "True Positive" (TP), "True Negative" (TN), "False Positive" (FP), and "False Negative" (FN) [24]. Furthermore, the equation can be used to calculate the accuracy based on these values (1)

Jurnal Teknik Informatika Vol. 16 No. 2, October 2023 (185-193)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.34084

$$accuracy = \frac{(TP + TN)}{P + N} \qquad (1)$$

The accuracy is then used as parameter as to whether or not a model in performing classification.

Meanwhile, to calculate the level of precision prediction of events can be used equation (2)

$$prediction \text{ precision} = \frac{TP}{TP + FP} \qquad (2)$$

The precision results mentioned above are used to show how precisely a model predicts positive events in various prediction activities.

In addition to precision and accuracy, recall or sensitivity of the system to the class can also be viewed to see more details of the system's performance. Recall can be calculated using equation (3)

$$prediction \text{ sensitivity} = \frac{TP}{TP + FP} \qquad (3)$$

Figure 6 shows the result of the confusion matrix from the simulation that was performed previously.

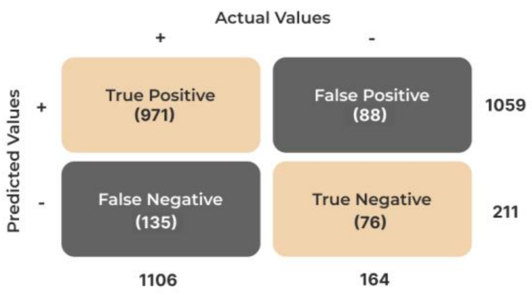

Figure 8.Confusion matrix with values

According to the confusion matrix, the model successfully predicted 1059 instances of real news and 211 instances of hoax news out of a total of 1106 real news and 164 hoax news. The lower the false positive (FP) and false negative (FN) values, the more accurate and reliable is the model [25].

3.3. Discussion of Results

The primary objective of this research was to distinguish between "Real" and "Hoax" news on Twitter over a span of five years (2015-2020). The model, employing the combined method of Passive Aggressive Classifier and TfidfVectorizer, showed promising results with an accuracy of 82.44%.

These results align with the overarching goal of mitigating the spread of fake news on social media platforms, particularly Twitter, where misinformation can rapidly proliferate. While the model demonstrates high accuracy, the research discussion must also consider the potential societal impact and specific nuances of the classification results.

In conclusion, the model provides a solid foundation for further exploration of combatting the challenges posed by fake news in the age of social media. The nuanced discussion of the results underscores the relevance and implications of the findings in the context of the research objectives outlined in the introduction.

CONCLUSION

This research successfully developed and trained a machine learning model to detect hoax news using the TfidfVectorizer and the Passive Aggressive Classifier algorithm. The model was evaluated on the " Indonesian False News (Hoax)" dataset obtained from Kaggle, which contains 4231 rows of news from Twitter over five years.

During the preprocessing stage, data cleaning, normalization, and encoding techniques were employed to prepare the dataset. The resulting model achieved favorable performance metrics with accuracy, precision, and recall scores of 82.44%, 80.66%, and 82.44%, respectively.

The confusion matrix analysis indicated the model's ability to effectively predict real and hoax news with minimal False Positive and False Negative values. These findings demonstrate the improved feasibility and accuracy of the proposed model.

In conclusion, this study shows a promising approach using machine learning to combat hoax news, contributing to the fight against misinformation in the realm of social media. The success of the model highlights its potential to ensure the credibility of news sources and enhance the dissemination of accurate information. Further advancements and refinements in this area can lead to more robust and reliable models for detecting fake news in the future.

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v16i2.34084

REFERENCES

[1] Y. Rohmiyati, "Analisis Penyebaran Informasi Pada Sosial Media," *ANUVA*, vol. 2, no. 1, pp. 29–42, 2018.

[2] "Digital 2023: Indonesia — DataReportal – Global Digital Insights." Accessed: May 19, 2023. [Online]. Available: https://datareportal.com/reports/digital-2023-indonesia

[3] M. Masril and F. W. Lubis, "Analisis Penggunaan Media Sosial dan Penyebaran Hoax Di Kota Medan," *JURNAL SIMBOLIKA: Research and Learning in Communication Study*, vol. 6, no. 1, pp. 11–22, Apr. 2020, doi: 10.31289/simbollika.v6i1.2937.

[4] Kementrian Komunikasi dan Informatika, "Triwulan Pertama 2023, Kominfo Identifikasi 425 Isu Hoaks," SIARAN PERS NO. 50/HM/KOMINFO/04/2023. Accessed: May 20, 2023. [Online]. Available: https://www.kominfo.go.id/content/detail/48363/siaran-pers-no-50hmkominfo042023-tentang-triwulan-pertama-2023-kominfo-identifikasi-425-isu-hoaks/0/siaran_pers

[5] KBBI Daring, "Hasil Pencarian - KBBI Daring." Accessed: Jul. 09, 2023. [Online]. Available: https://kbbi.kemdikbud.go.id/entri/hoaks

[6] Y.-P. Chen *et al.*, "The Prevalence and Impact of Fake News on COVID-19 Vaccination in Taiwan: Retrospective Study of Digital Media," *J Med Internet Res*, vol. 24, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:247954679

[7] I. Afrianty, D. Nasien, and H. Haron, "Performance Analysis of Support Vector Machine in Sex Classification of The Sacrum Bone in Forensic Anthropology," *JURNAL TEKNIK INFORMATIKA*, vol. 15, no. 1, pp. 63–72, Jun. 2022, doi: 10.15408/jti.v15i1.25254.

[8] V. Alvian, D. Hidayatullah, A. Nilogiri, H. Azizah, and A. Faruq, "Klasifikasi Siswa Berprestasi Menggunakan Metode K-Nearest Neighbor (KNN) Pada SMA Negeri 2 Situbondo," 2022. [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST

[9] H. Susana and N. Suarna, "PENERAPAN MODEL KLASIFIKASI METODE NAIVE BAYES TERHADAP PENGGUNAAN AKSES INTERNET," *Jurnal Sistem Informasi dan Teknologi Informasi)*, vol. 4, no. 1, pp. 1–8, 2022.

[10] R. Supriyadi, W. Gata, N. Maulidah, A. Fauzi, I. Komputer, and S. Nusa Mandiri Jalan Margonda Raya No, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," vol. 13, no. 2, pp. 67–75, 2020, [Online]. Available: http://journal.stekom.ac.id/index.php/E-Bisnis■page67

[11] M. G. Hussain, M. R. Hasan, M. Rahman, J. Protim, and S. Al Hasan, "Detection of Bangla Fake News using MNB and SVM Classifier," May 2020, [Online]. Available: http://arxiv.org/abs/2005.14627

[12] T. Arifin and D. Ariesta, "PREDIKSI PENYAKIT GINJAL KRONIS MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER BERBASIS PARTICLE SWARM OPTIMIZATION," *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 26–30, Apr. 2019, doi: 10.36787/jti.v13i1.97.

[13] S. Suprianto, "Implementasi Algoritma Naive Bayes Untuk Menentukan Lokasi Strategis Dalam Membuka Usaha Menengah Ke

Bawah di Kota Medan (Studi Kasus: Disperindag Kota Medan)," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 1, no. 2, p. 125, Jan. 2020, doi: 10.30865/json.v1i2.1939.

[14] F. Hilmiyah, "PREDIKSI KINERJA MAHASISWA MENGGUNAKAN SUPPORT VECTOR MACHINE UNTUK PENGELOLA PROGRAM STUDI DI PERGURUAN TINGGI (Studi Kasus: Program Studi Magister Statistika ITS)," 2017.

[15] A. Yonathan, H. Sujaini, E. Esyudha Pratama, and Nawawi, "Perbandingan Algoritma Klasifikasi dalam Pendeteksian Hoax pada Media Sosial," 2022, doi: 10.26418/juara.v1i1.53126.

[16] C. Dhaneswara, Y. Azhar, and N. Hayatin, "DETEKSI BERITA HOAX PADA DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN METODE MODIFIED K- NEAREST NEIGHBOR," 2021.

[17] S. Soleman, "Pemanfaatan Metode Klasifikasi Naïve Bayes Untuk Pendeteksi Berita Hoax Pada Artikel Berbahasa Indonesia," *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*, vol. 7, no. 2, p. 83, Dec. 2021, doi: 10.24014/coreit.v7i2.14290.

[18] T. M. Khalil, M. Fmipa, U. Sahid, P. Matematika, and F. Uny, "Klasifikasi informasi hoaks pada media sosial twitter menggunakan algoritma random forest berbasis particle swarm optimization," 2022. [Online]. Available: http://journal.student.uny.ac.id/ojs/index.php/jktm:

[19] M. Nurjannah, I. Fitri Astuti, and D. Program Studi, "PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING Mahasiswa S1 Program Studi Ilmu Komputer FMIPA Universitas Mulawarman 2,3)," 2013.

[20] K. Crammer, O. Dekel, and J. Keshet, "Online Passive-Aggressive Algorithms Shai Shalev-Shwartz Yoram Singer †," 2006.

[21] Kumar, P. Niranjan. "Detection of Textual Propaganda Using Passive Aggressive Classifiers." International Journal of Advanced Trends in Computer Science and Engineering, vol. 12, no. 2, March-April 2023, http://www.warse.org/IJATCSE/static/pdf/file/ijatcse071222023.pdf, https://doi.org/10.30534/ijatcse/2023/071222023.

[22] Karyawati, AAIN Eka et. all. "Comparison of SVM and LIWC for Sentiment Analysis of SARA." International Journal on Cybernetics & Informatics (IJCCS), 2022, https://doi.org/10.22146/ijccs.69617

[23] Ayu, Tsabitah et. all. "Pendiagnosa Daun Mangga Dengan Model Convolutional Neural Network." Jurnal Computational and Applied Mathematics (CESS), vol. 6, no. 2, 2021, https://doi.org/10.24114/cess.v6i2.22857.

[24] S. Shalev-Shwartz and S. Ben-David, "Understanding Machine Learning: From Theory to Algorithms," 2014. [Online]. Available: http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

[25] Y. Wang, Y. Jia, Y. Tian, and J. Xiao, "Deep reinforcement learning with the confusion-matrix-based dynamic reward function for customer credit scoring," *Expert Syst Appl*, vol. 200, p. 117013, Aug. 2022, doi: 10.1016/J.ESWA.2022.117013.