ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

COMPARISON OF K-NN, SVM, AND RANDOM FOREST ALGORITHM FOR DETECTING HOAX ON INDONESIAN ELECTION 2024

Indra[1], Agus Umar Hamdani[2], Suci Setiawati[3], Zena Dwi Mentari[4]
, Mauridhy Hery Purnomo[5]
[1,2,3,4]Teknik Informatika, Universitas Budi Luhur
[5] Department of Computer Engineering, Institut Teknologi Sepuluh Nopember

email: indra@budlihuhur.ac.id[1], agus.umarhamdani@budiluhur.ac.id[2], 1811520079@student.budiluhur.ac.id[3], 1911500344@student.budiluhur.ac.id[4], hery@ee.its.ac.id[5]

Abstract

During the year 2022, The Indonesian National Police (POLRI) received 113 reports related to the spread of hoax news related to 2024 Indonesian Election (PEMILU). There are still relatively few hoax detection tools that already exist in Indonesia. This research creates a system that can detect hoax news in Indonesian tweets about the Indonesian Election (PEMILU) 2024 by comparing three methods, namely K-NN, SVM, and Random Forest. The process of labeling (create model) using validation on ground truth data, namely cekfakta.tempo, cekfakta.kompas, and turnbackhoax.id. In this research, we also check the differences between different types of distance measurements in applying the K-NN algorithm. The method used for feature extraction in this research is TF-IDF. The results of experiments show that the highest accuracy results are obtained using the SVM and K-NN algorithms with distance measurements using Euclidean Distance, which is 86.36%. The best precision value is obtained using the K-NN algorithm with distance measurements using Manhattan Distance, which is 86.95%.

Keywords : Indonesian Election 2024, TF-IDF, K-NN, Tweet, Hoax Detection

Received: 01-03-2024 | **Revised:** 16-03-2024 | **Accepted:** 19-03-2024
DOI: https://doi.org/10.23887/janapati.v13i1.76079

INTRODUCTION

Misleading news that deliberately misleads people and has a political agenda is considered hoax news [1]. Hoax news is spread in any form, such as text, images, and videos. The spread of hoax news continues to increase, disturbing the community because it harms many parties. The incident occurred because some people needed to make sure when receiving information obtained through social media. Hoax news about the Indonesian election (PEMILU) 2024 has been circulating in the community, which can be unsettling and lead to wrong views facing the Indonesian election (PEMILU) 2024. The Indonesian National Police (POLRI) has received 113 reports about the spread of hoax news about the Indonesian election (PEMILU) 2024 in the 2022 period. The Ministry of Communication and Informatics recorded 9,417 findings of hoax issues from August 2018 - February 16, 2023. Twitter became one of the platforms for the spread of hoax information. Considering these problems, this research aims to solve the problem by classifying tweets into hoaxes and facts.

Previous research that compared Naïve Bayes and SVM to detect hoax news shows that using Naïve Bayes method produces an accuracy of 78% and using the SVM method produces an accuracy of 80%. The data used was obtained by crawling data using the tweepy library with keywords related to covid-19 [2]. In previous research by [3], using the Naive Bayes algorithm for detecting hoax news on Indonesian tweets obtained an accuracy of 72.06%. The tweet data used is obtained by crawling data based on keywords by looking at hashtags that are thought to contain hoax news in the period October 2019 - March 2020. Another research by [4], it was the first work to prove the performance of brief fake news detection and topic classification simultaneously obtained a better accuracy value than the latest methods by using a new model of multi-task learning (FDML) fake news detection. In another study by [5], detecting hoax news from three public datasets namely Weibo, Twitter, and PHEME by applying Human Cognition-based Consistency Inference Network (HCCIN) to comprehensively explore consistent and inconsistent semantics to detect multi-modal

fake news reveals the advantages of HCCIN. In another study by [6], by proposing a new model to improve the accuracy of fake news detection, namely extracting and combining global, spatial, and temporal features from text using TF-IDF, CNN, and BiLSTM methods simultaneously. Then in a fast classifier using a fast learning network (FLN) to classify these features efficiently. Previous research by [7] on Novel Blockchain-Based Deepfake Detection Method Using Federated Model and Deep Learning shows that the research is superior in terms of accuracy and AUC compared to recent works, the datasets in the research are obtained from FaceForensik (FF ++), DeepFakeTIMIT, Pratinjau Tantangan Deteksi DeepFakes (DFDCpre), and CelebDF. The previous research by [8] on the Classification of Covid-19 Hoax News Using a Combination of the K-NN and Information Gain Methods gave an accuracy result of 95%, the dataset in the research was obtained in the form of news related to Covid-19. Another research by [9], Detecting Hoaxes in Indonesian News Using TF/TDM and K Nearest Neighbor gave an accuracy up to 83.6%. The hoaxes data in this research was retrieved from an Indonesian hoax-debunking community website published between July 31, 2015 - November 22, 2017. The hoaxes data were then being compared against real news from various reputable news websites in Indonesia within a similar range of publication dates.

The objective of the research is to comparative analysis of algorithmic performance utilizing Twitter data, with a focus on the Indonesian context during the 2024 election (PEMILU). The study compares three methods - K-NN, SVM, and Random Forest - for detecting hoax news on Indonesian tweets. Additionally, it examines the impact of different distance measurement methods within the K-NN algorithm. Data collection involves crawling tweet data using the tweepy library and online news media, validated against credible sources such as cekfakta.tempo.co and turnbackhoax.id. By modeling hoaxes in the Indonesian language using ground truth from local news sources, the research aims to enhance accuracy, precision, and recall values in hoax detection.

The evaluation was conducted using the Confusion Matrix method by calculating the accuracy, precision, and recall values to assess the performance of the three methods. The contribution of this research includes the development of a fake news detection application that can be used practically by desktop users.

The innovative approach in this research lies in its application of machine learning algorithms, specifically K-NN, SVM, and Random Forest, to the task of detecting fake news on the Indonesian-language Twitter platform during the 2024 election (PEMILU). While similar studies may have been conducted in other contexts or languages, the specific focus on Indonesian tweets related to a significant political event like PEMILU is unique. This approach allows for a targeted investigation into the effectiveness of different algorithms in a specific cultural and linguistic context, providing insights that may not be readily available from studies conducted in other settings. The data collection process involved crawling tweet data using the tweepy library and online news media. The collected data were validated against credible sources such as cekfakta.tempo.co and turnbackhoax.id. Hoaxes were modeled in the Indonesian language using ground truth from local news sources, and manual labeling was conducted to ensure accuracy and reliability. As for the choice of algorithms, K-NN, SVM, and Random Forest are commonly used in classification tasks and have demonstrated effectiveness in various domains, including text classification. By comparing these algorithms, the research aims to identify which method performs best in the context of fake news detection on Indonesian-language tweets, thereby contributing to the advancement of knowledge in this field.

The selection of the SVM, Random Forest and KNN algorithms for this study is motivated by previous research findings indicating their hight accuration [2][10]. In previous studies, SVM, Random Forest and KNN algorithms have consistently demonstrated high accuracy and reliability in various classification tasks for hoax detection. Leveraging the success of these algorithms in previous research endeavors, we aim to further investigate their effectiveness in the specific context of our study. By adopting SVM, Random Forest and KNN as our primary algorithms, we seek to build upon existing knowledge and potentially uncover new insights into their applicability and performance within our dataset. This strategic choice enables us to capitalize on established methodologies while exploring novel avenues for classification and prediction.

The previous study using of SVM, Random Forest, and KNN algorithms in the study by [10], which achieved accuracies above 90% using COVID-related English datasets, on the other hand with Suci's research [2], where SVM outperformed NBC with an accuracy exceeding 80% on COVID-related Indonesian

datasets. Given these findings, our study leverages this precedent to inform hoax detection for the 2024 indonesian election dataset. By drawing from the strengths observed in previous research, particularly the robust performance of SVM in an Indonesian context, we aim to apply these insights effectively to our dataset, thereby enhancing the reliability and effectiveness of our hoax detection framework.

METHOD

Figure 1 is the method stage in building a hoax detection system. In the first stage, it is the collection of tweet data (crawling) using the tweepy library which will be used as a dataset, then the proportion of the dataset is 80% for train data and 20% for test data stored in the form of Excel files. The use of this proportion is based on several factors. Firstly, the allocation of 80% of the data for training ensures that the model has a sufficient amount of diverse data to learn the underlying patterns. The more data available for training, the better the model will perform. Secondly, by setting aside 20% of the data as test, we can evaluate the performance of the model independently. This approach has support from the research done by [11] and One such study that provides insights into the effectiveness of the 80:20 split ratio by [10]. In this paper, the authors likely discuss the rationale behind selecting this particular proportion and its impact on the accuracy of the model. In the second stage, after the dataset is collected, a labeling process is carried out on the training data manually with 2 types of labels, that is hoax and non-hoax. In the third stage, the dataset then goes into the preprocessing process to filter or clean the data to produce clean data. In the fourth stage, namely, feature extraction or word weighting using the TF-IDF method on each clean data. In the fifth stage, the feature extraction results are used at the classification stage using the K-Nearest Neighbor (K-NN), SVM, and Random Forest algorithm, then the classification results are obtained in the form of prediction labels from each testing data.

Dataset Collection

The data used in our research are tweet data get from Twitter by using API keys obtained by way of registering developer accounts through developed Twitter. The data collection of tweets is done using the tweepy library. The tweet data is then exported into an Excel file and combined with news data obtained from the online news media portal. The keywords used when crawling data are related to the Indonesian election (PEMILU) 2024. The data was collected from June 1, 2023, to February 8, 2024, which is 220.

The selected date range of June 1, 2023, to February 8, 2024, was chosen to align with the period leading up to the 2024 presidential election in Indonesia, including the nomination of presidential and vice-presidential candidates and the campaign period from November 28, 2023, to February 10, 2024. This timeframe ensures that the data collected is contextually relevant and captures significant events and trends related to the research topic.

Manual Labeling

Labeling tweet data is done manually by giving a value of 1 (one) to tweet data containing non-hoax information and 0 (zero) to tweet data containing hoax information. In the manually labeled stage of the training dataset, it is done by voting by considering the opinions of 3 (three) volunteers [2][12]. The volunteer must give an idea of whether each piece of data includes hoax or non-hoax information. The labeling decision is obtained if 3 (three) people think that the tweet data include hoax information or has a label of 0 (zero), then the data will be labeled 0 (zero), and otherwise. In determining the label, volunteers already know what hoax news is and understand the flow of checking information based on Figure 2, and have received references from trusted news websites such as TurnBackHoax.ID, cekfakta.kompas.com, and cekfakta.tempo.co.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

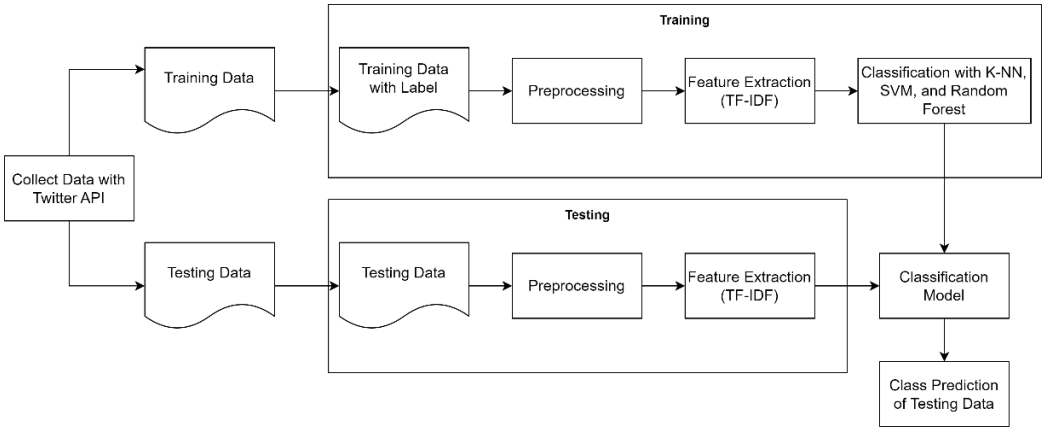


Figure 1. Stages of The Method

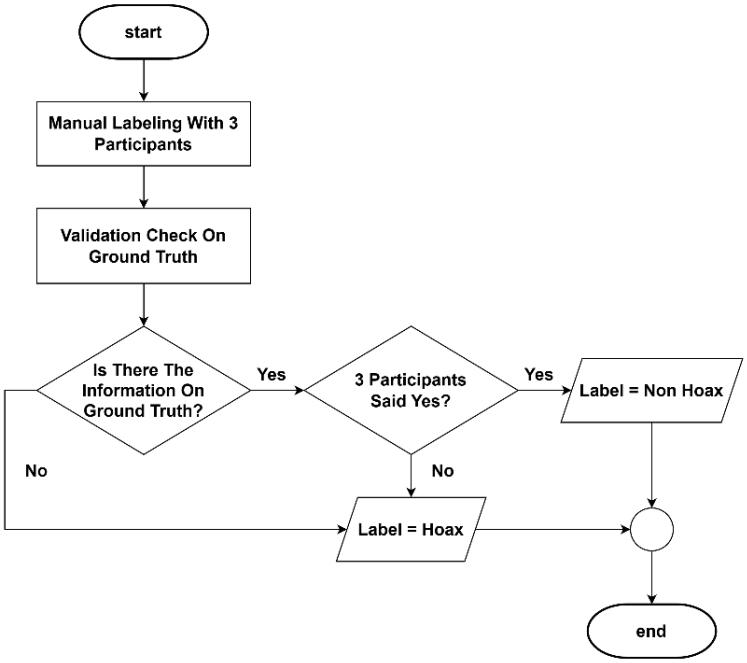


Figure 2. Stage of Manual Labeling

Preprocessing

In the preprocessing phase, the tweet data cleaning process is carried out to remove meaningless words. The results of this phase will provide more structured and clear tweet data, or what is commonly called clean text. The process carried out in the preprocessing phase has the following order:

a. Case Folding: the stage to convert all the big letters or capital in the data into small letters or vice versa.
b. Cleansing: The phase cleans data by removing characters other than a to z or deleting components that have no relationship with the information on the data, such as usernames, emoticons, symbols, numbers, reading marks, mentions, hashtags, and URLs.
c. Slangword: a phase of changing non-originate words on each data. Words that do not contain abbreviations or Gaul language. The process of changing this word is based on the slang word dictionary contained in the CSV document. The dictionary is derived from previous research that can be downloaded on the GitHub site (https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection).
d. Stopword Removal: The phase removes words that have no meaning that usually appear in large numbers. The word

removal process is done using the stopword removal library in Python.

e. Stemming: the process of converting a word into a basic word form. The goal of this stage is to clear a word with an inaccurate inscription by removing all inscriptions on each word. This process is done using the Stemmerfactory library available in Python. For the word "menggunakan," the stemming process would yield the base word "guna" by removing the prefix "meng-" and the suffix "-kan."

Term Frequency-Inverse Document Frequency (TF-IDF)

In the TF-IDF method, the weight of a word indicates its relevance in a document; the higher the weight value, the more significant the word's contribution to document formation. The TF-IDF method calculates the weight of each word in a document or even a set of documents [13]. The stages of TF-IDF are as follows [8]:

a. Compute the number of occurrences of the term i in the j $(tf_{i,j})$paper.
b. Compute the number of documents containing the term i (df_i).
c. Compute the weight value of the inverse document frequency (IDF) using the equation:

$$idf_i = \log\left(\frac{N}{df_i}\right) \quad (1)$$

Description:
N = total number of documents.

d. Compute the weight value of TF-IDF using the compound:

$$w_{i,j} = tf_{i,j}\, x\, idf_i \quad (2)$$

Description:
$w_{i,j}$ = the weight of the term i against the document j
$tf_{i,j}$ = the frequency of the term i in the document j
idf_i = the weight value of IDF term i

K-Nearest Neighbor (K-NN)

KNN is a technique that performs categorization based on training data or learning data observed from the closest distance to the object depending on the k value. This technique attempts to categorize new objects based on characteristics and training data. Determine the training data and test data before taking measurements using the K-NN method. Then the measurement procedure is performed by determining the distance. The neighbor's K value must be set before calculating the distance of the data from its neighbors. The K-NN method can use several distance measurements including the following [14]:

a. Euclidean Distance

$$d(x_i x_j) = \sqrt{\sum_{r=1}^{n}\left(a_r(x_i) - a_r(x_j)\right)^2} \quad (3)$$

Description:
$d(x_i x_j)$: euclidean distance
$(x_i), (x_j)$: record i, record j
(a_r) : data r
i, j : 1,2,3,…n
n : object dimensions

b. Jaccard Distance

Jaccard distance is the magnitude of the intersection that divides the magnitude of the union into two sets: how it is defined. The Jaccard similarity formula is as follows [15]:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \times 100 \quad (4)$$

Description:
X = Data 1
Y = Data 2
|X ∩ Y| = The set of all members of X and also includes members of Y
|X ∪ Y| = The set of all members of X or Y or both

c. Manhattan Distance

Manhattan distance is the variance between the two absolute coordinates determined by the Manhattan distance [15].

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (5)$$

Description [16]:
d(x,y) = Distance
x = Location coordinates 1
y = Location coordinates 2

Support Vector Machine (SVM)

SVM is a method for classifying data by finding the best boundary between categories, be it a straight line or complex, based on the distance between important points in the data. SVM can also improve classification by moving the data to a higher dimension.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

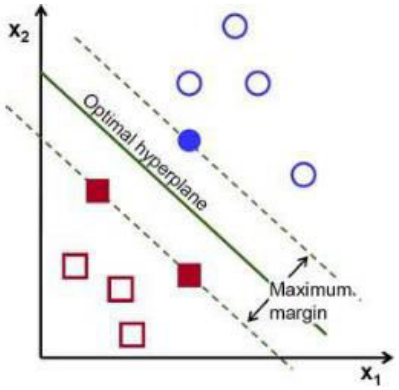


Figure 3. Illustration of SVM Method

In Figure 3, you can see a hyperplane used by SVM to distinguish and categorize two classes of data effectively. SVM can handle datasets with unlimited dimensions through the utilization of kernel techniques. SVM uses only a limited number of contributed data points (support vector) to form a model that will be used in the classification process. Here is the SVM equation [2]:

$$f(x) = \sum_{i=1}^{m} a_i y_i K(x, x_i) + b \qquad (6)$$

Description:
w = hyperplane parameter being searched for (the perpendicular line between hyperplane line and support vector point)
x = SVM input data point
a_i = weight value of each data point
$K(x, x_i)$ = kernel function
b = hyperplane parameter sought (bias value)

Random Forest

The random forest classifier is made up of numerous decision trees created from randomly chosen portions of the training dataset. It aggregates the decisions from these diverse trees to ascertain the ultimate classification for a test item [17]. This method requires attributes and data randomly according to the conditions imposed to build a decision tree consisting of root nodes, internal nodes, nodes, and leaf nodes. The root node is the highest node, often called the tree's root, in a decision tree. An internal node is a splitting node, having at least two branches as outputs and just one as input. A leaf node marks the end, possessing only input and no outgoing branches. The decision tree starts by calculating the entropy value as a determinant of the impurity level of the attribute and the information gain value. Equation 7 is the formula for calculating the entropy value, while equation 8 is the formula for calculating the information gain value [18].

$$Entropy\ (Y) = -\sum_i p(c|Y)\ log_2\ p\ (c|Y) \qquad (7)$$

Description:
Y = the set of cases
p(c|Y) = the proportion of Y values to class c.

$$Information\ Gain\ (Y, a) = Entropy(Y) - \sum_{veValues} \frac{|Yv|}{|Ya|} Entropy(Yv) \qquad (8)$$

Description:
Values (a) = all possible values in the set of cases a
Y_v = a subclass of Y with class v corresponding to class a
Y_a = all values corresponding to a.

Confusion Matrix

After obtaining the classification results in the form of prediction labels from each test data, then visualized into a confusion matrix table. The Confusion Matrix is a common technique employed in data mining to compute accuracy. The confusion matrix is illustrated with a table stating the amount of properly classified test data and the amount of incorrectly typed test data [19]. The Confusion matrix table can be viewed in Table 1.

Table 1. Confusion Matrix

	Predicted Class Yes (Hoax)	Predicted Class No (Non-Hoax)
Actual Class Yes (Hoax)	True Positive (TP)	False Negative (FN)
Actual Class No (Non-Hoax)	False Positive (FP)	True Negative (TN)

True Positive (TP) is predicted class and actual class are both hoaxes. False Positive (FP) is indicated class is a hoax and the actual class is non-hoax. True Negative (TN) is the predicted class and the actual class is both non-hoax. False Negative (FN) is a predicted class is non-hoax and the actual class is a hoax.

At this stage, the performance value of the model that has been made will be measured through the process of calculating accuracy, precision, and recall [3]. The formula for calculating accuracy, precision, and recall are defined as follows:

1. Accuracy

   The accuracy value is calculated from the number of correct predictions divided by the total number of predicted documents [20].

$$Accuracy = \frac{tp + tn}{tp + fp + +fn + tn} \qquad (9)$$

2. Precision

The precision value is obtained from the number of correct predictions (tp) divided by the number obtained (tp + fp) [20].

$$Precision = \frac{tp}{tp + fp} \quad (10)$$

3. Recall

The recall value is obtained from the number of correct predictions divided by the sum of all existing data [20].

$$Recall = \frac{tp}{tp + fn} \quad (11)$$

RESULT AND DISCUSSION

Dataset Collection

The data collected was sourced from one of the social media, namely Twitter. The keywords used when crawling data are related to the Indonesian election (PEMILU) 2024 such as pemilu 2024, pilpres 2024, the names of presidential candidates, vice presidents, and political parties. The process of collecting tweet data uses the tweepy library. Crawling data is done by entering the API Key, API Key Secret, Access Token, and Access Token Secret obtained by creating an account on the Twitter Developer site. The crawled tweet data is then saved into an Excel form using the panda's library. The crawled tweet data is combined with news data obtained from online news portal media such as turnback hoaxes.id, cekfakta.kompas.com, and cekfakta.tempo.co. The data was collected from June 1, 2023, to February 8, 2024, which is 220 data. Sample data can be viewed in Table 2:

Table 2. Sample Data of Dataset Collection Result

No.	Source	Text
1.	Twitter	Selain Menangkan Ganjar di Pilpres 2024, PDIP dan PPP Juga Jalin Kerja Sama di Pileg 2024 https://t.co/GB9Uv5nUFr (In addition to winning Ganjar in the presidential election 2024, PDIP and PPP also cooperate in the legislative election 2024 https://t.co/GB9Uv5nUFr)
2.	Tempo.co	Dari 1.902 Bacaleg, KPU DKI Jakarta Sebut Hanya 226 yang Penuhi Syarat (Out of 1,902 Bacaleg, DKI Jakarta KPU Mentions Only 226 Who Meet the Requirements)
3.	Twitter	Gerindra akan Minta Saran Presiden Jokowi soal Cawapres Pendamping Prabowo Subianto - http://Tribunnews.com #Prabowo #BangkitBersama (Gerindra will ask President Jokowi for advice on Prabowo Subianto's running mate - http://Tribunnews.com #Prabowo #RiseTogether)
4.	Turnbackhoax.id	Ganjar Terbukti Terlibat Kasus Korupsi E-KTP Hingga Memicu Kemarahan Megawati (Ganjar Proven Involved in E-KTP Corruption Case, Triggering Megawati's Anger)
5.	Kompas.com	[HOAKS] Koalisi Perubahan Telah Deklarasikan Khofifah sebagai Cawapres ([HOAKS] Coalition of Change Has Declared Khofifah as Vice Presidential Candidate)

Manual Labeling

After the data is collected, it will then be labeled manually. This labeling stage is carried out by validating data on online news such as turnbackhoax.id and cekfakta.kompas.com. The labels "hoax" and "non hoax" were chosen because the main focus of labeling is to identify whether information is a hoax or not. By using these two labels, we can clearly distinguish between content that needs to be cautioned as untrue and content that can be trusted. In addition, the use of these simple labels also facilitates understanding for end-users who will use the labeling results to make decisions or consume information. The "neutral" label was not chosen because it tends not to provide clear information about the truth or untruth of the information. The determination of labels refers to research conducted [21]. The following is a sample of manual labeling data can be viewed in Table 3:

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

Table 3. Sample Data of Labeling Manual Result

No.	Data	Label
1.	gerindra gelar kampanye akbar usung anies baswedan (gerindra holds grand campaign endorse anies baswedan)	0 (Hoax)
2.	mahfud resmi damping anies pilih presiden restu jokowi (mahfud official accompany anies choose president bless jokowi)	0 (Hoax)
3.	golkar resmi gabung koalisi anies cek fakta via (golkar official join anies coalition fact check via)	0 (Hoax)
4.	elite partai politik tolak pilih tunda agus harimurti yudhoyono prabowo (political party elite refuse vote delay agus harimurti yudhoyono prabowo)	1 (Non Hoax)
5.	daftar partai politik lolos tahap verifikasi calon serta pilih (list political party pass elect participant candidate verification stage)	1 (Non Hoax)
6.	gembira sistem pilih buka partai adil sejahtera raya calon legislatif indonesia (happy open elect system adil sejahtera party legislative candidates feast Indonesia)	1 (Non Hoax)

Preprocessing

After the data is collected and labeled, then the data will undergo preprocessing procedures. The purpose of this stage is to produce clean data free from indicators that can interfere during the implementation process of word weighting with TF-IDF. An example of the results of each stage in preprocessing can be viewed in Table 4 below:

Table 4. Example of Preprocessing Result

Steps	Result
Original Data	Cawe cawe Presiden pd pemilu 2024 tergolong inkonstitusional (bukan dlm tugas dan kewenangan presiden) & perbuatan yg tercela. #MakzulkanPresidenCawe2 (President interference in election 2024 is classified as unconstitutional (not within the duties and authority of the president) & a despicable act. #MakzulkanPresidentInterference)
Case folding	cawe cawe presiden pd pemilu 2024 tergolong inkonstitusional (bukan dlm tugas dan kewenangan presiden) & perbuatan yg tercela. #makzulkanpresidencawe2
Cleansing	cawe cawe presiden pd pemilu tergolong inkonstitusional bukan dlm tugas dan kewenangan presiden perbuatan yg tercela
Slangword	cawe cawe presiden pada pemilihan umum tergolong inkonstitusional bukan dalam tugas dan kewenangan presiden perbuatan yang tercela
Stopword Removal	cawe cawe presiden pemilihan umum tergolong inkonstitusional bukan dalam tugas kewenangan presiden perbuatan tercela
Stemming	cawe cawe presiden pilih umum golong inkonstitusional bukan dalam tugas wenang presiden buat cela

Term Frequency-Inverse Document Frequency (TF-IDF)

After passing the labeling and preprocessing stages, the next stage in this research is word weighting with the TF-IDF method. According to [22], the purpose of this process is to find a representation of the value of each document from training data where a vector will be formed between documents with terms which then for similarities between documents with clusters will be determined by a vector prototype also called cluster centroid. The results of the TF-IDF vector in the sample data can be seen in Table 5.

K-Nearest Neighbor Classification

After the word weighting stage is completed, the next stage is classification using the K-NN algorithm by determining the Euclidean distance. This stage aims to obtain label prediction results for each test data by matching them to the training data. An example of the results of calculating Euclidean Distance to calculate the distance between test data and training data can be seen in Table 6.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

Table 5. TF-IDF Vector Result of Sample Data

Data Type	Data	Label	Result of vector TF-IDF
Test Data	partai demokrasi indonesia juang bangun jis agenda kampanye murah anies baswedan (democratic struggle indonesia party builds jis cheap campaign agenda anies baswedan)	No labeling yet	[0; 0; 0,3; 0; 0; 0,3; 0,3; 0; 0,3; 0; 0; 0; 0; 0; 0; 0]
Train Data 1	gerindra gelar kampanye akbar usung anies baswedan (gerindra holds grand campaign endorse anies baswedan)	Hoax (0)	[0; 0,3; 0,3; 0,3; 0,3; 0,3; 0,3; 0; 0; 0; 0; 0; 0; 0; 0; 0]
Train Data 2	bacaleg partai gerindra sragen bekuk polisi libat edar narkoba pilih (gerindra party candidate sragen arrested police drug distributin elect)	Non Hoax (1)	[0; 0; 0; 0; 0; 0; 0; 0,3; 0,3; 0,3; 0,3; 0,3; 0,3; 0,3; 0,3; 0,3]

Table 6. Measurement Result of Euclidean Distance

Euclidean Distance (data (test i, train i))
$d_{(1,1)}$ = 3,63
$d_{(1,2)}$ = 2,29
$d_{(1,3)}$ = 3,67
$d_{(1,4)}$ = 3,72
$d_{(1,5)}$ = 3,58
$d_{(1,6)}$ = 3,63

If the Euclidean distance has been obtained and sorted ascending, then the closest distance value based on the predetermined number of K can be taken. As an example of K value = 3, then the results obtained based on the specified K value can be observed in the following Table 7:

Table 7. Result of K Nearest Neighbor Classifier

Euclidean Distance (data (test i, train i))	Label
$d_{(1,2)}$ = 2,29	Hoax (0)
$d_{(1,5)}$ = 3,58	Non Hoax (1)
$d_{(1,1)}$ = 3,63	Hoax (0)

Based on Table 7, it can be seen that the nearest neighbor data are Training 2, Training 5, and Training 1. The number of label comparisons 0 (hoax) and 1 (non hoax) is 2: 1, so the results of testing on test data 1 have a label of 0 (hoax).

Evaluation Method

In this research, evaluation was carried out by calculating the accuracy, precision, and recall of the classification results using the K-NN, SVM, Random Forest, and CNN algorithm in determining the label for the test data.

1. Testing with KNN Algorithm

   Testing in this research also compares classification results with various K values based on those provided in the system, respectively K = 3; K = 5; K = 7; K = 9; and K = 11. The K value represents the number of K nearest neighbors between the training and testing data to determine the final result of labeling the test data.

   a. The test results in Table 8, show that the application of Euclidean Distance in the KNN method using several variations of the K value gets the highest value in the recall test results using the K = 3, 5, 7, 9, and 11, with a value of 100%. Based on the recall value that gets the highest results in classifying the 2024 election hoax news, Euclidean Distance tends to have classification results that get true positives and do not tend to have false positive classification results.

Table 8. Results with Euclidean Distance

K Value	Euclidean Distance		
	Accuracy	Precision	Recall
3	79,54%	72,72%	100%
5	72,72%	66,66%	100%
7	63,63%	60%	100%
9	59,09%	57,14%	100%
11	54,54%	54,54%	100%
Average	65,90%	62,21%	100%

The results of testing Table 8 show that the application of Euclidean Distance gets

the average of all variants of the K value with the correct prediction ratio of the entire data with an accuracy value of 65,90%, while for the ratio of positive correct predictions of the overall positive prediction results with a precision value of 62,21%, and the ratio of positive correct predictions of the overall positive correct data has a recall value of 100%. Furthermore, the following K-NN test results are displayed with the Jaccard model in Table 9.

b. The highest Jaccard Distance test in Table 9, obtained in the precision test with the highest value at k = 11 with a value of 68.18%. Based on the highest precision value in classifying the 2024 Election hoax news, the Jaccard Distance tends to get true positives and does not tend to produce false positive classification results.

Table 9. Results with Jaccard Distance

K Value	Jaccard Distance		
	Accuracy	Precision	Recall
3	54,54%	59,09%	54,16%
5	56,81%	64,70%	45,83%
7	54,54%	60%	50%
9	52,27%	57,14%	50%
11	63,63%	68,18%	62,5%
Average	56,36%	61,82%	52,50%

The results of testing Table 9 show that the application of Jaccard Distance gets the average of all variants of the K value with the correct prediction ratio of the entire data with an accuracy value of 56.36%, while for the ratio of positive correct predictions of the overall positive prediction results with a precision value of 61,82%, and the ratio of positive correct predictions of the overall positive correct data has a recall value of 52,50%. Furthermore, the following K-NN test results are displayed with the Manhattan model in Table 10.

c. The highest Manhattan Distance test Table 10, obtained in the recall test with a value of k = 3 with a value of 100%. Based on the highest recall value in classifying the 2024 Election hoax news, Manhattan Distance tends to get true positives and does not tend to produce false positive classification results.

Table 10. Results with Manhattan Distance

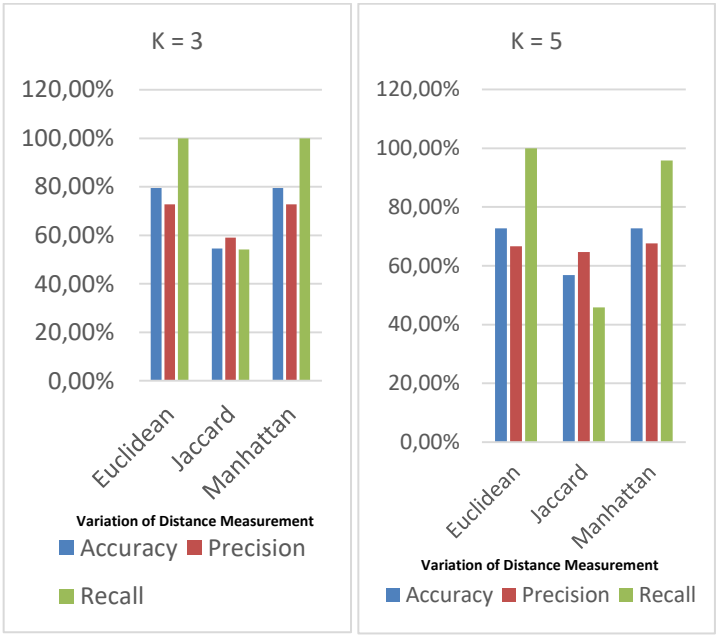
K Value	Manhattan Distance		
	Accuracy	Precision	Recall
3	79,54%	72,72%	100%
5	72,72%	67,64%	95,83%
7	68,18%	63,88%	95,83%
9	61,36%	58,97%	95,83%
11	59,09%	57,49%	95,83%
Average	68,18%	64,14%	96,66%

The results of testing Table 10 show that the application of Manhattan Distance gets an average of the entire variant of the K value with the correct prediction ratio of the entire data with an accuracy value of 68,18%, while for the ratio of correct positive predictions of the entire positive prediction results with a precision value of 64,14%, and the ratio of correct positive predictions of the entire correct positive data has a recall value of 96,66%.

The following Figure 4 illustrates the performance of the KNN method by comparing the accuracy, precision, and recall values using variation of K values and variation of distance measurement methods.

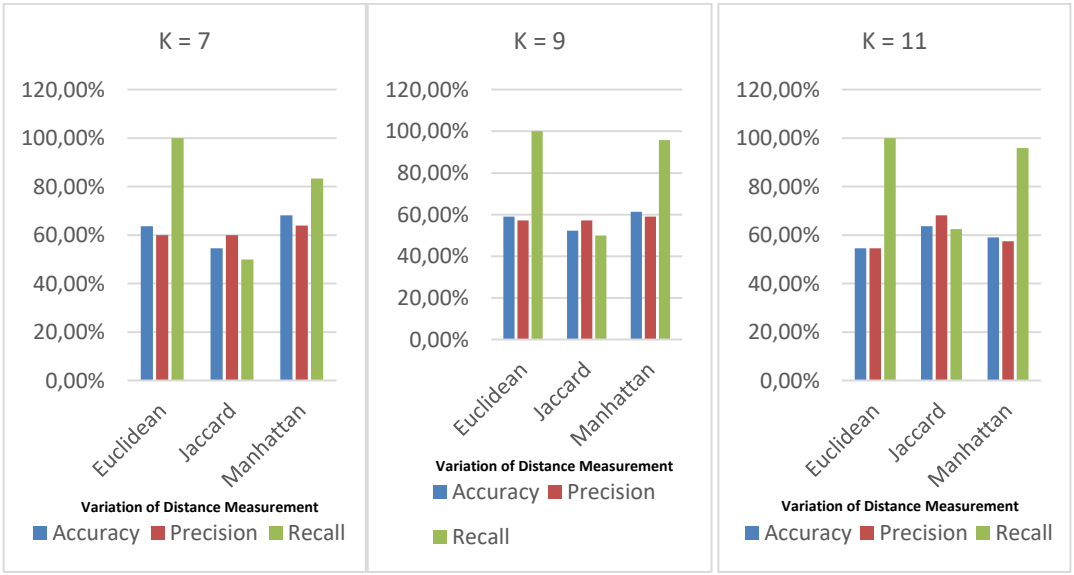
Figure 4a, 4b, and 4c shows that Euclidean distance produces higher accuracy using values of K = 3, and 5, while Figure 5a and 5b depicts that Manhattan distance produces higher accuracy using values of K = 7 and 9. In Figure 4a, 4b, 4c, 5a, and 5b shows that Manhattan distance produces higher precision using K values = 3, 5, 7, and 9, while Figure 5c shows that Jaccard distance produces higher accuracy using K value = 11. In Figure 4a, 4b, 4c, 5a, 5b, and 5c depicts that Euclidean distance produces higher recall using K values = 3, 5, 7, 9, and 11.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024



(a) K=3 (b) K=5

Figure 4. Graph of KNN Method with K= 3, and 5



(c) K=7 (d) K=9 (e) K=11

Figure 5. Graph of KNN Method with K=7, 9, and 11

In our research, the highest accuracy value was obtained using a value of K = 1 with a variation of Euclidean distance. This is different from the research conducted by [23] which produced the highest accuracy value using the value of K = 4 and Jaccard distance. On the other hand, the highest precision value in this research was obtained using a value of K = 1 with Manhattan distance, in contrast to research conducted by [23] which produced the highest precision value using a value of K = 4 with Jaccard distance. The highest recall value in our research was obtained using the values of K = 3,5,7,9, and 11 with Euclidean distance, in contrast to research conducted by [23] which produced the highest precision value using the value of K = 4 with Jaccard distance.

2. Testing with KNN, SVM, Random Forest Algorithm

Table 11. Result of Method Comparison

Metode	Accuracy	Presicion	Recall
SVM	86,36%	85%	88%
Random Forrest	81,82%	80%	83%
KNN	79,54%	72,72%	100%

Based on Table 11, it can be concluded that the test results get the highest accuracy of 86.36% by using the SVM and K-NN algorithms (using K = 3 and Euclidean distance). In testing experiments using various K values, the best accuracy value is obtained at each value of K = 3 and 5 with an Euclidean distance of 72.72% - 79,54%. While the best precision value is obtained at each value of K = 1,3, 5, 7, and 9 with Manhattan distance, which is 58.97% - 86.95%. And the best recall value is obtained at each value of K = 1 - 11 with Euclidean distance, which is 91.66% - 100%.

From the provided data, it can be inferred that both the KNN and SVM models exhibit the same accuracy rate of 86.36%, yet KNN demonstrates higher precision and recall compared to SVM. The Random Forest model, despite having a slightly lower accuracy rate at 81.82%, also displays lower precision and recall compared to the other two models. Specifically, KNN stands out with the highest recall value of 91.66%, indicating its ability to correctly identify a large portion of positive instances. Although SVM shares the same accuracy rate as KNN, its overall performance is slightly inferior, underscoring the importance of incorporating precision and recall evaluations in assessing classification model performance.

The variation in performance metrics among the models can be attributed to their inherent algorithms and underlying assumptions. Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) have different decision boundary constructions and distance metrics, leading to variations in their predictive capabilities. While SVM aims to find the hyperplane that maximizes the margin between classes, KNN relies on the proximity of data points to determine classification. Additionally, the choice of hyperparameters, such as the kernel function for SVM or the number of neighbors for KNN, can significantly impact model performance. Random Forest, on the other hand, employs an ensemble approach by constructing multiple decision trees and aggregating their predictions, which introduces another layer of complexity and potential variability in performance. Furthermore, the nature of the dataset itself, including its size, class distribution, and feature characteristics, can influence how well each model generalizes to unseen data. Therefore, the differences in accuracy, precision, and recall reflect the unique strengths and weaknesses of each model in handling the specific task and dataset.

In the research conducted by [Suci], the SVM algorithm yielded high accuracy values, demonstrating the robustness of their data collection process and providing valuable insights into the performance of various classification algorithms on the dataset. Additionally, in the study by [Alhadeed], the overall results validate the integrity of our baseline truth data and offer significant insights into the performance of different classification algorithms on the dataset. Particularly noteworthy are the superior results obtained from classifiers such as NN, DT, and LR. LR demonstrates efficacy in binary classification problems and can be regarded as a single-layer NN. Moreover, it is observed that the results from LR and Perceptron are similar, as LR essentially functions as a Perceptron with a sigmoidal activation function. The final configuration of the detection system will depend on the classification algorithm that produces the best results in building the ensemble detection model.

CONCLUSION

In conclusion, the research demonstrated that SVM and K-NN methods with Euclidean Distance measurement achieved the highest accuracy, at 86.36%. Additionally, the K-NN method with Manhattan Distance measurement attained the best precision, with a value of 86.95%. These findings indicate the effectiveness of these methods in detecting hoaxes in Indonesian tweets related to the 2024 election. Although the Random Forest method yielded slightly lower accuracy, it still provided competitive results. Further analysis is warranted to explore factors influencing the relative performance of each method and to identify strategies for enhancing overall accuracy and precision.

To address the difficulties encountered in recognizing sentences containing abbreviated words or slang, as well as the limited diversity in the dataset, several recommendations can be made. Firstly, expanding the dataset by

collecting more diverse and representative samples of Indonesian tweets related to the election 2024 can help improve the accuracy and generalization of the classification model. Additionally, implementing more robust preprocessing techniques, such as incorporating stemming variations and handling slang words more effectively, can enhance the system's ability to process and classify text data accurately. Moreover, exploring the use of additional feature extraction methods beyond TF-IDF, such as the bag of words approach, can provide alternative perspectives on text representation and potentially improve classification performance. Lastly, incorporating data from other social media platforms like Instagram and YouTube can enrich the dataset and broaden the scope of the analysis, leading to more comprehensive insights into hoax detection across different online channels.

Regarding the discussion on system/application development, it's important to clarify the research objectives. If the primary aim is to develop a practical hoax detection application, then discussing potential enhancements or iterations to the system architecture, user interface, and functionality could be beneficial. This could include improving the real-time processing capabilities, enhancing user experience features, and integrating feedback mechanisms for continuous improvement. Alternatively, if the focus is solely on evaluating the performance of different classification algorithms, then the discussion may primarily revolve around methodological aspects, such as algorithm selection, parameter tuning, and evaluation metrics. Clarifying the specific goals of the research will help guide the discussion and provide a clear direction for future work.

REFERENCES

[1] Nurhayati and A. Pasaribu, "Perancangan Sistem Pendeteksi Berita Hoax Menggunakan Algoritma Levenshtein Distance Berbasis Php," *J. SAINTIKOM (Jurnal Sains Manaj. Inform. dan Komputer)*, vol. 19, no. 2, p. 74, 2020, doi: 10.53513/jis.v19i2.2601.

[2] Indra, S. Setiawati, S. Vaddhana, and A. Septiarini, "Comparison of Naive Bayes and Support Vector Machine for Detecting Hoax in Indonesian Tweet Case Study of Tweet Covid-19," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 2022-Octob, no. October, pp. 61–66, 2022, doi: 10.23919/EECSI56542.2022.9946515.

[3] C. S. Sriyano and E. B. Setiawan, "Pendeteksian Berita Hoax Menggunakan Naive Bayes Multinomial Pada Twitter dengan Fitur Pembobotan TF-IDF," *e-Proceeding Eng. Vol.8, No.2*, vol. 8, no. 2, pp. 3396–3405, 2021.

[4] Q. Liao *et al.*, "An Integrated Multi-Task Model for Fake News Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5154–5165, 2022, doi: 10.1109/TKDE.2021.3054993.

[5] L. Wu, P. Liu, Y. Zhao, P. Wang, and Y. Zhang, "Human Cognition-Based Consistency Inference Networks for Multi-Modal Fake News Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 211–225, 2024, [Online]. Available: https://ieeexplore.ieee.org/document/10138033

[6] A. H. J. Almarashy, M.-R. Feizi-Derakhshi, and P. Salehpour, "Enhancing Fake News Detection by Multi-Feature Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 11, pp. 139601–139613, 2023, doi: 10.1109/ACCESS.2023.3339621.

[7] A. Heidari, N. J. Navimipour, H. Dag, S. Talebi, and M. Unal, "A Novel Blockchain-Based Deepfake Detection Method Using Federated and Deep Learning Models," *Cognit. Comput.*, no. 0123456789, 2024, doi: 10.1007/s12559-024-10255-7.

[8] M. Audina, A. E. Karyawati, I. W. Supriana, I. K. G. Suhartana, I. G. S. Astawa, and I. W. Santiyasa, "Klasifikasi Berita Hoaks Covid-19 Menggunakan Kombinasi Metode K-Nearest Neighbor dan Information Gain," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 10, no. 4, p. 319, 2022, doi: 10.24843/jlk.2022.v10.i04.p02.

[9] E. Zuliarso, M. T. Anwar, K. Hadiono, and I. Chasanah, "Detecting Hoaxes in Indonesian News Using TF/TDM and K Nearest Neighbor," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 835, no. 1, pp. 0–6, 2020, doi: 10.1088/1757-899X/835/1/012036.

[10] M. K. Elhadad, K. F. Li, and F. Gebali, "Detecting misleading information on COVID-19," *IEEE Access*, vol. 8, pp. 165201–165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[11] I. L. Kharisma, D. A. Septiani, A. Fergina, and K. Kamdan, "Penerapan Algoritma Decision Tree untuk Ulasan Aplikasi Vidio di Google Play," *J. Nas. Teknol. dan Sist. Inf.*, vol. 9, no. 2, pp. 218–226, 2023, doi:

10.25077/teknosi.v9i2.2023.218-226.

[12] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, no. October, pp. 233–237, 2017, doi: 10.1109/ICACSIS.2017.8355039.

[13] F. G. Weddiningrum, "Deteksi Konten Hoax Berbahasa Indonesia Pada Media Sosial Menggunakan Metode Levenshtein Distance," *Perpust. Univ. Islam Neger Sunan Ampel*, pp. 1–78, 2018.

[14] P. D. Nugraha, S. al Faraby, and Adiwijaya, "Klasifikasi Dokumen Menggunakan Metode Knn Dengan Information Gain," *eProceedings Eng.*, vol. 5, no. 1, pp. 1541–1550, 2018.

[15] M. Addanki, "Integrating Sentiment Analysis in Book Recommender System by using Rating Prediction and DBSCAN Algorithm with Hybrid Filtering Technique," 2023.

[16] Y. Miftahuddin, S. Umaroh, and F. R. Karim, "Perbandingan Metode Perhitungan Jarak Euclidean, Haversine, Dan Manhattan Dalam Penentuan Posisi Karyawan (Studi Kasus : Institut Teknologi Nasional Bandung)," *J. Tekno Insentif*, vol. 14, no. 2, pp. 69–77, 2020, [Online]. Available: https://jurnal.lldikti4.or.id/index.php/jurnaltekno/article/view/270

[17] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," *Big Data Min. Anal.*, vol. 4, no. 2, pp. 116–123, 2021, doi: 10.26599/BDMA.2020.9020016.

[18] V. W. Siburian and I. E. Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," *Annu. Res. Semin.*, vol. 4, no. 1, pp. 144–147, 2018.

[19] M. F. Rahman, D. Alamsah, and M. I. Darmawidjadja, "Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)," *J. Inform.*, vol. 11, no. 1, p. 36, 2017, doi: 10.26555/jifo.v11i1.a5452.

[20] F. Rahutomo, I. Y. R. Pratiwi, and D. M. Ramadhani, "Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia," *J. Penelit. Komun. Dan Opini Publik*, vol. 23, no. 1, 2019, doi: 10.33299/jpkop.23.1.1805.

[21] F. Prasetya and F. Ferdiansyah, "Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes," *J. Sist. Komput. dan Inform.*, vol. 4, no. 1, p. 132, 2022, doi: 10.30865/json.v4i1.4852.

[22] N. K. Widyasanti, I. K. G. Darma Putra, and N. K. Dwi Rusjayanthi, "Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 6, no. 2, p. 119, 2018, doi: 10.24843/jim.2018.v06.i02.p06.

[23] W. Hidayat, E. Utami, A. F. Iskandar, A. D. Hartanto, and A. B. Prasetio, "Perbandingan Performansi Model pada Algoritma K-NN terhadap Klasifikasi Berita Fakta Hoaks Tentang Covid-19," *Edumatic J. Pendidik. Inform.*, vol. 5, no. 2, pp. 167–176, 2021, doi: 10.29408/edumatic.v5i2.3664.