

Analysis of Hoax News in News Media Using the K-Nearest Neighbor (K -NN) Method

Putri Nabila ^{1 *}, Bustami ², Wahyu Fuadi ³

^{1,2,3} Malikussaleh University

putri.20170098@mhs.unimal.ac.id ^{1 *}, bustami@unimal.ac.id ^{2 *}, wahyu.fuadi@unimal.ac.id ³

Abstract

The rapid development of internet technology has facilitated the dissemination of information within the community through various news media. However, this convenience has also fueled the widespread dissemination of unverified fake news (hoaxes), potentially causing social unrest, undermining the credibility of information, and manipulating public opinion. Therefore, a computing system capable of automatically filtering and categorizing news is needed to help readers distinguish between facts and hoaxes. This study aims to develop an automatic classification model to identify hoax news using the *K-Nearest Neighbor* (KNN) method and measure its accuracy.

This study uses the *Euclidean Distance approach* with a value of $k = 3$. This study uses 2,000 news data that have been collected and verified through official fact-checking sites. The data was divided using the *train-test split method* with a ratio of 80:20 to obtain 1,600 *training data* and 400 *testing data*. Based on the results of the *confusion matrix test* at a value of $K = 3$, the *True Positive* (TP) value was 104, *False Positive* (FP) was 0, *False Negative* (FN) was 95, and *True Negative* (TN) was 201. The model performance evaluation showed a *precision* value of 100%, *recall* of 52.26%, and *accuracy* of 76.25%. These results indicate that the K-NN method is able to classify news well and has a high level of precision, although the model's ability to detect all hoax news (*recall*) still needs to be improved in further research. This research is expected to contribute as a technology-based tool (*decision support system*) for the public and academics in monitoring, filtering, and tackling the spread of fake news in digital media.

Keywords: Hoaks, Berita, K-Nearest Neighbor (KNN), Text Mining, Euclidean Distance

1. Introduction

The internet is considered a form of technological advancement that has now become a human necessity, having penetrated the cultural landscape of society, from lifestyle and education to research and research, to the commercial sphere. One of the benefits of the internet is its rapid growth in availability and ease of use by all groups, anywhere and at any time. News is often presented to third parties or groups of people through print media, online platforms, or word of mouth. The dissemination of this information is unfiltered, and anyone can spread unclear news, even containing hoaxes known as "fake news." Unclear news and hoaxes can cause public unrest, especially if the content circulated is sensitive to human life. [1].

A hoax is news and substance that can deviate from someone's acceptance by spreading false news as valid. It can damage image and credibility and influence many people. Hoaxes are often misleading news because they lack reliable sources and clear indications. [2]. Hoax news has been planned for publication by certain individuals for their own benefit. Currently, hoaxes can spread widely via the internet. The rapid development of telecommunications has given rise to the circulation of news on the internet that cannot be handled, for example, such as information about documents containing hoaxes. Therefore, efforts that can be realized immediately and can be understood by the recipient of the news are to group the news according to the category of hoax news or not. The purpose of this grouping is to calculate the level of a target whose class and characteristics are not found. To facilitate the classification of a news, as a first stage, researchers will test the classification of news titles by using *K -Nearest Neighbor*. [3].

The K-Nearest Neighbor algorithm is a *supervised learning algorithm*. [4]. *K-Nearest Neighbor* (KNN) is one of the classification methods commonly used in data mining, with the advantage of executing data and shorter time, which is very useful for executing the distribution of news whose contents are unclear, even news that contains hoax elements. The application of the KNN algorithm was found in a number of studies that have been conducted, namely by Siti Ernawati and Risa Wati regarding sentiment analysis of travel agent reviews using the *K-Nearest Neighbor algorithm*, which obtained accuracy with results reaching 87.00%. [5]. From the results of another study by Pratama Dwi Nugraha et al regarding sentiment analysis of travel agent reviews using the *K-Nearest Neighbor algorithm* and *Information Gain*

feature selection, it shows that the *K-Nearest Neighbor algorithm without Information Gain* feature selection for all training documents using various parameters has a high level of accuracy of 93.94438%.

Based on the explanation above, it is known that classification using the KNN algorithm provides high accuracy and performance yet is simple. Therefore, the *K-Nearest Neighbor* (KNN) algorithm will be applied in this study to classify hoax news headlines. The goal is to determine the performance level of the KNN algorithm in classifying hoax news headlines in Indonesian. This is expected to help news readers distinguish between hoax news and facts.

2. Research methods

This research uses a quantitative approach with text mining and classification methods using the *K-Nearest Neighbor* (KNN) algorithm. The research steps in analyzing hoax news using the *K-Nearest Neighbor* (KNN) method include:

1. Data collection
The authors collected online comment data from credible news media and fact-checking websites (indicated to be hoaxes). This process was carried out carefully to ensure the data obtained was relevant, valid, and worthy of analysis and classification using the *K-Nearest Neighbor* (KNN) method.
2. Data analysis
The analysis was conducted in a structured manner using historical or *real-time data* from social media to generate evidence-based *insights*. To test the system, the authors compared the results of the system analysis with manual calculations to achieve the appropriate level of accuracy.
3. System Design
The system design was conducted to illustrate the system's workflow in analyzing and classifying hoax news. The system was designed to accept input in the form of news text, then process it through *preprocessing*, *feature extraction using TF-IDF*, and *classification using the K-Nearest Neighbor (KNN)* method. The final output of the system is a decision on whether the news is a hoax or valid news, which is used as the basis for research analysis.
4. System Implementation
The implementation phase is the process of translating the entire system design and methods into a concrete, functional application or program. In this study, implementation was carried out by integrating the *K-Nearest Neighbor* (KNN) algorithm into the program/system, then inputting the collected news text data into the program. Through this phase, the system will execute the data processing and classification (hoax/fact) in a structured manner, resulting in a fully functional system ready for accuracy testing.

2.1. System schematic

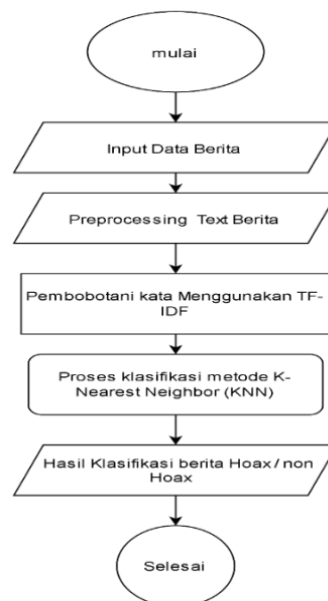


Fig. 1 : System schematic

The system scheme of the hoax news classification process using the *K-Nearest Neighbor* (KNN) algorithm is as follows:

1. Start
The process begins with the system receiving input data.
2. News data input
News data is entered into the system for further processing.
3. *Pre-Processing*

The data that has been entered will go through a *pre-processing stage*, which usually includes data cleaning, removing unimportant words (*stop words*), text normalization, and tokenization. [6] .

4. TF-IDF word weighting

After pre-processing, word weighting is carried out using the *Term Frequency - Inverse Document Frequency* (TF-IDF) method to determine the level of importance of each word in the document. [7] .

5. Calculation with the *K-Nearest Neighbor* (KNN) algorithm

The test data and training data will be calculated using the *K-Nearest Neighbor* (KNN) algorithm based on the proximity between data in the feature space to determine the classification based on its nearest neighbors. [8] .

6. News classification results

The system produces output based on the classification results of the algorithm used.

7. Finished

The system process is completed after producing the classification output.

This scheme illustrates the process of text analysis and classification with a combination of two machine learning algorithms commonly used in *Natural Language Processing* (NLP).

2.1. Use case diagram

Use case diagram is a *Unified Modeling Language model diagram* used to describe an interaction between a user and the system being designed [9] .

The use case diagram for this research can be seen in Figure 2. This system allows *users* to input news, process the classification, and view the results. As the news classification process progresses, the system automatically calculates data similarity using the K - NN process . *Euclidean Distance* until it successfully determines the hoax / fact class as the final result for the user.

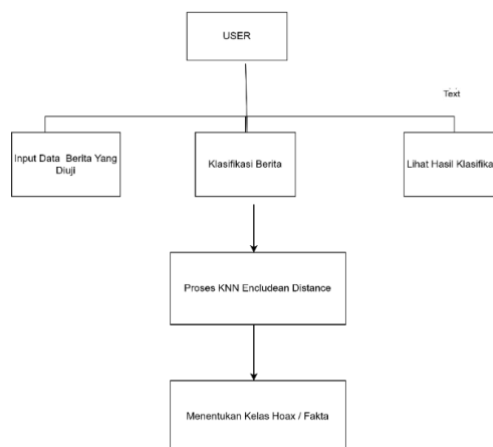


Fig. 2 : Use Case Diagram

3. Results and Discussion

This section presents the results of testing the *K-Nearest Neighbor* (K-NN) method in classifying news into two categories, namely Hoax and Fact. The dataset used in this study is a collection of news obtained from the News Media portal and verified using references from fact-checking sites such as *TurnBackHoax* and other verification sources. The total dataset used is 1600 training data and 400 testing data that have gone through a text *preprocessing process*, including: *case folding* , *tokenizing* , *stopword removal* , and *stemming* to ensure that the data is ready to be used in the classification process.

The dataset used in this study has several criteria and sub-criteria within each criterion. To facilitate calculations, each sub-criterion value was weighted.

Table 1: Criteria Data

Criteria Code	Criteria
X1	Clarity of news sources
X2	Validity of News Content
X3	Sensational Title
X4	Content Not in Accordance with Facts
X5	Types of distribution media
X6	Clarity of time of incident
X7	Author credibility

X8	The presence of clickbait elements
X9	Conformity to official facts

3.1. Manual calculation of K-NN

The classification process using the *K-Nearest Neighbor* (K-NN) algorithm consists of two steps: training and testing. In the training stage, K-NN utilizing sample data Which includes variables and classes The target is taken from the number of classification classes as input. Meanwhile, in the testing phase, K-NN uses the distance calculation value for the attributes of each test data against all attributes in the training data using the *Euclidean distance formula*. [10]. The results of the distance calculation will be compared and sorted based on the smallest *Euclidean distance* (*ascending*). Then the top data is grouped according to the number of specified parameter values k. After that, a number of values of k nearest neighbors are generated, where the classification of the newly obtained test data is based on the majority of classes in the k nearest neighbors.

The following are the stages of working on the K-NN algorithm:

1. Determining the value of k, in solving this case the value of k=3 is determined.
2. Calculate the distance between test data (data to be classified) and all training data using the *Euclidean Distance approach*.

Table 2: *Euclidean Distance*
Euclidean Distance

$d_1 = \sqrt{(1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2}$ $d_1 = \sqrt{0+0+0+0+0+0+0}$ $d_1 = 0$
$d_2 = \sqrt{(2-1)^2 + (2-1)^2 + (2-1)^2 + (2-1)^2}$ $d_2 = \sqrt{1+1+1+1}$ $d_2 = \sqrt{4}$ $d_2 = 2$
$d_3 = \sqrt{(2-1)^2 + (2-2)^2 + (2-1)^2 + (2-1)^2}$ $No. 3 = \sqrt{1+0+1+1}$ $No. 3 = \sqrt{3}$ $No. 3 = 1.73205$
$d_4 = \sqrt{(2-1)^2 + (1-1)^2 + (2-1)^2 + (2-1)^2}$ $d_4 = \sqrt{1+1+1+1}$ $d_4 = \sqrt{4}$ $d_4 = 2$
$d_5 = \sqrt{(2-1)^2 + (2-2)^2 + (2-1)^2 + (2-1)^2}$ $d_5 = \sqrt{1+0+1+1}$ $d_5 = \sqrt{3}$ $d_5 = 1.73205$

Table 2. above is an example of a manual calculation of the *Euclidean distance* from test data 1 to 5 training data sets. Repeat the same method to calculate the distance to all remaining training data sets, resulting in the following distance calculation results:

Table 3: Results Calculation Distance <i>Euclidean Distance</i>		
No	Distance	Classification
1	0	Hoax
2	0	Hoax
3	0	Hoax
8	3	Fact
15	2,828	Fact
16	2,828	Fact
28	1	Hoax
...		

3. Sort the results of the distance calculation from the smallest to the largest distance (*ascending*).

4. Create data groups as many as the previously determined k values.
5. Select the most dominant category or the one that appears the most as a result of the classification.

3.2. System testing

System testing was carried out using several evaluation methods, namely train-test split, cross validation, confusion matrix, and visualization of accuracy comparison.

3.2.1 . Train-test split accuracy results

The test used a *train-test split* with a total of 2,000 data sets, divided into **80%** training data (1,600 data sets) and **20%** testing data (400 data sets). Based on the test results on the 400 test data sets, the K-Nearest Neighbor (KNN) model achieved an accuracy of 0.7625, or 76.25%. This value indicates that the model is capable of classifying hoax and non-hoax news with an accuracy rate of 76.25% on data not used during the training process.

3.2.2. Cross Validation Accuracy

The image below shows the accuracy results for each fold in the cross validation process.

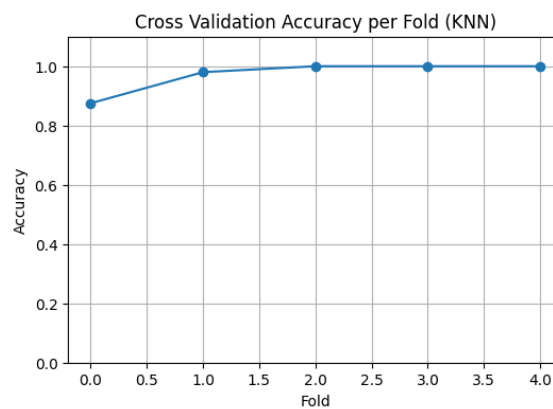


Fig. 3: Cross Validation Accuracy per Fold (KNN)

Testing using *5-Fold Cross Validation* resulted in an accuracy of 88% on *Fold 1*, 99% on *Fold 2*, and 100% on *Folds 3, 4, and 5*. This achievement resulted in an average accuracy of 97.4%, which indicates that the model has excellent and consistent performance in distinguishing between hoax and factual news based on TF-IDF text features.

3.2.3. Confusion matrix

The results of the confusion matrix test on the test data are shown in the following image:

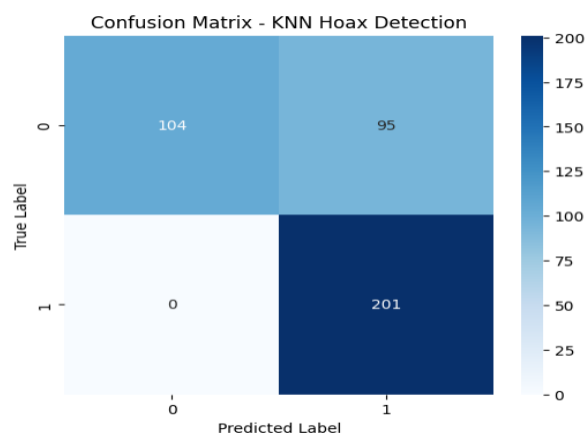


Fig. 4: Confusion Matrix-KNN Detection

Confusion matrix evaluation on 400 test data sets showed that the K-NN model successfully predicted 201 factual news items and 104 hoax news items correctly, with an accuracy of 76.25%, a precision of 67.91%, a recall of 100%, and an *F1-score* of 80.89%. The perfect recall rate demonstrates the model's superiority in identifying all factual news items without error. However, hoax news detection still needs to be optimized as 95 hoax items were incorrectly classified as factual news items.

3.3. Implementation

1. Home page

For can login to system, admin must fill in *username* And *password* match with Which stored in database.

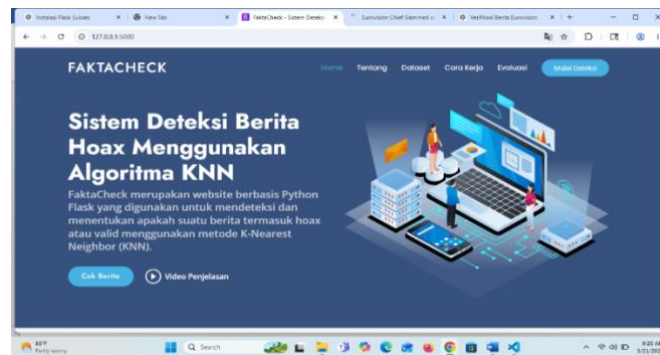


Fig. 5: Home Page

2. System information page

This page displays information about the system.

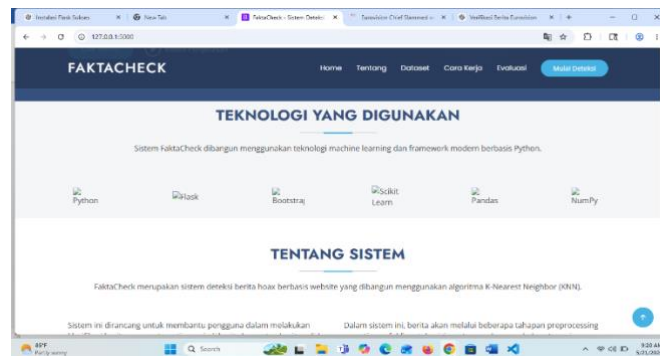


Fig. 6: System Information Page

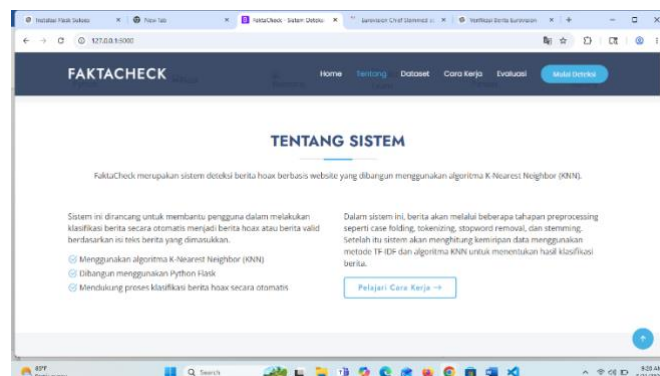


Fig. 7: System Information Page

3. Dataset page

This page displays the research dataset of the hoax news detection system.

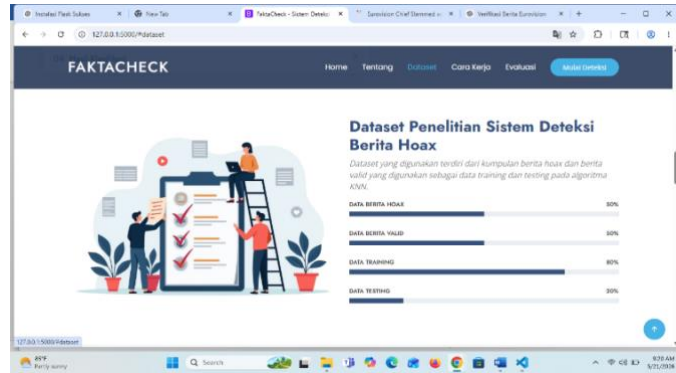


Fig. 8: Dataset Page

4. How it works page

This page displays the process flow of the hoax news detection system using K-NN.

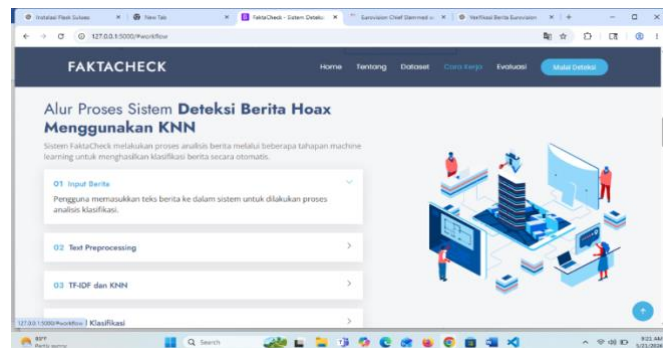


Fig. 9: How It Works Page

5. Evaluation data page

The following page displays the results of the performance evaluation of the K-Nearest Neighbor (K-NN) algorithm on the hoax news detection system.

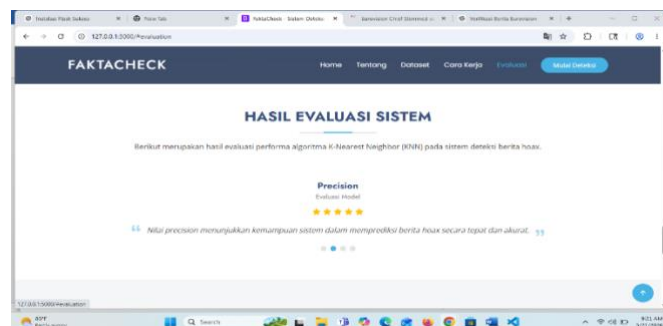


Fig. 10: Evaluation Data Page

6. System detection page

This page shows how the hoax news detection system works.

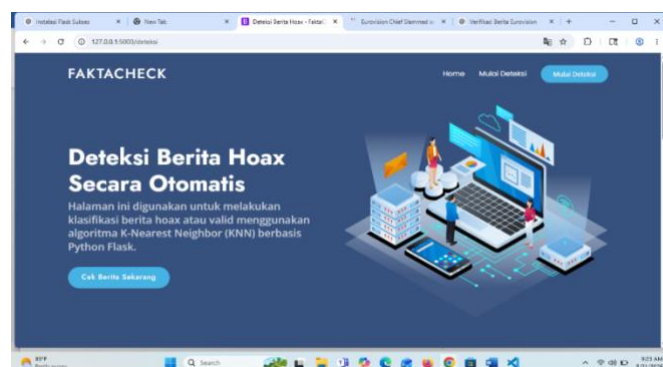


Fig. 11: System Detection Page

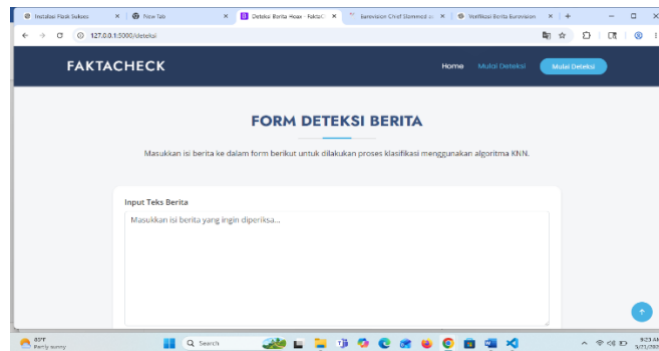


Fig. 12: System Detection Page

4. Conclusion

The conclusions that can be drawn from the results of this research are as follows:

1. This hoax news classification system was created using the *K-Nearest Neighbor* (K-NN) method and uses 9 criteria, namely clarity of news sources, validation of news content, sensational titles, content that does not match the facts, type of distribution media, clarity of time of incident, credibility of the author, presence of clickbait elements, conformity with official facts.
2. The output results from the system are in the form of a classification of “Hoax” and “Not Hoax/Fact” news data.
3. This study uses the K-Nearest Neighbor (KNN) method with the Euclidean Distance approach and a parameter value of $k = 3$. The total dataset used in this study is 2,000 news data that have been collected and verified using references from official fact-checking sites. The dataset consists of hoax news and non-hoax/fact news, then divided using the *train-test split method* with a ratio of 80:20, resulting in 1,600 data as training data and 400 data as testing data. Training data is used to build a classification model, while testing data is used to evaluate model performance based on *accuracy*, *precision*, and *recall values*. Before the training and testing process is carried out, all data first goes through a *preprocessing stage* which includes *case folding*, *tokenizing*, *stopword removal*, and *stemming* to produce cleaner data and is ready to be used in the classification process.

Reference

- [1] M. Audina, AE Karyawati, IW Supriana, IKG Suhartana, IGS Astawa, and IW Santiyasa, “Classification of Covid-19 Hoax News Using a Combination of K-Nearest Neighbor and Information Gain Methods,” *JELIKU (Journal of Electron. Comput Science. Udayana)*, vol. 10, no. 4, p. 319, 2022, doi: 10.24843/jlk.2022.v10.i04.p02.
- [2] Nur Halimah and Purwanti, “Implementing Critical Thinking in Daily Life: A Case Study of Handling Hoax News,” *J. Media Commun. | E-ISSN 3063-9581*, vol. 1, no. 3, pp. 40–44, 2025, doi: 10.62379/jmc.v1i3.158.
- [3] MD Hendriyanto and BN Sari, “Application of the K-Nearest Neighbor Algorithm in Classifying Hoax News Titles,” *J. Ilm. Inform.*, vol. 10, no. 02, pp. 80–84, 2022, doi: 10.33884/jif.v10i02.5477.
- [4] F. Putra, HF Tahiyat, RM Ihsan, R. Rahmadden, and L. Efrizoni, “Application of K-Nearest Neighbor Algorithm Using Wrapper as Preprocessing for Determining Human Body Weight Information,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.
- [5] S. Ernawati and R. Wati, “Application of the K-Nearest Neighbors Algorithm in Travel Agent Review Sentiment Analysis,” *J. Khatulistiwa Inform.*, vol. 6, no. 1, pp. 64–69, 2018, [Online]. Available: <https://www.trustpilot.com/categories/tr>
- [6] L. Rhomaningtias, A. Khairunisa, S. Shella May Wara, and K. Maulida Hindrayani, “Sentiment Analysis of Smile Indonesia Application Reviews Using Naive Bayes Method and Support Vector Machine (SVM),” *HOAQ (High Educ. Organ. Arch. Qual. J. Teknol. Inf.)*, vol. 16, no. 1, pp. 79–91, 2025, doi: 10.52972/hoaq.vol16no1.p79-91.
- [7] C. Saputra, W. Wilcent, H. Irsyad, and A. Rahman, “Query Processing and Ranking of News Titles Related to the Governor of West Java Using TF-IDF and Cosine Similarity,” *Appl. Inf. Technol. Comput. Sci.*, vol. 4, no. 1, pp. 25–32, 2025, doi: 10.58466/aicoms.v4i1.1799.
- [8] AR Putri and RK Dinata, “Comparison of the K-Nearest Neighbor and Random Forest Methods in Classifying the Best Selling Medicines at Khan Pharmacy Matang Glumpang Dua,” vol. 3, no. 2, pp. 90–95, 2026.
- [9] R. Fauzan, “A Different Approach on Automated Use Case Diagram Semantic Assessment,” vol. 14, no. 1, pp. 496–505, 2021, doi: 10.22266/ijies2021.0228.46.
- [10] B. Christopher, “Comparative Analysis of K-Nearest Neighbor and Naïve Bayes Classifier Methods Based on Randomized Search Optimization in Hoax News Classification,” *JATISI (Jurnal of Information Technology and Information Systems)*, vol. 12, no. 2, pp. 272–283, 2025, doi: 10.35957/jatisi.v12i2.11371.