*International Journal Software Engineering and Computer Science (IJSECS)*
*4 (1)*, 2024, 40-48
Published Online April 2024 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN : 2776-4869, E-ISSN : 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v4i1.2068.

RESEARCH ARTICLE Open Access

# Classification of Hoax News Using the Naïve Bayes Method

**Rama Qubra Putra** *
Universitas Halu Oleo, Kendari City, Southeast Sulawesi, Indonesia.
Corresponding Email: ramaqubra.15@gmail.com.

**Rizal Adi Saputra**
Universitas Halu Oleo, Kendari City, Southeast Sulawesi, Indonesia.
Email: rizaladisaputra@uho.ac.id.

*Received: January 9, 2024; Accepted: March 10, 2024; Published: April 1, 2024.*

**Abstract**: The rampant dissemination of false and unsourced information, commonly known as hoaxes, has become pervasive in internet media. In the digital age, spreading false and unverified information has become a critical concern within internet media. Hoax news can influence elections, sway public opinion, and create political instability. The rapid evolution of information technology has contributed to the uncontrollable proliferation of hoax content, necessitating the development of intelligent systems for adequate classification. This research focuses on implementing a robust classification system for identifying hoax news circulating through internet media. The method used in this program is the Naive Bayes method, specifically the Naive Bayes Multinomial, which works with the assumption that each feature (word) is considered independent from the others. Text vectorization using CountVectorizer converts text into a numeric vector, which classification algorithms can use. This program uses a trained model to predict testing data and calculate evaluation metrics such as accuracy, confusion matrix, and classification reports. By leveraging these methodologies, the study aims to enhance the accuracy and efficiency of distinguishing genuine news from deceptive hoaxes. The highest accuracy value obtained in this research was 94.73%, with a division of 20% test data and 80% training data. True Negative (TN): 4555, False Positive (FP): 178 False Negative (FN): 295, True Positive (TP): 3952. The 94.73% accuracy value underscores the potential of the classification system to significantly contribute to the identification and containment of hoaxes in Internet media, thereby fostering a more reliable and secure information ecosystem.

**Keywords**: Classification; Hoax News; Internet Media; Information Technology; Naive Bayes Algorithm.

## 1. Introduction

The Internet has exponentially grown as a popular information medium, encompassing various domains such as news, product reviews, public services, movies, and more. These diverse sources include social media, news articles, blogs [1], and other platforms. In the contemporary digital age, the pervasive spread of false and unverified information, commonly labeled as hoaxes, poses a significant threat within Internet media [2]. Hoax news can influence elections, manipulate public opinion, and sow seeds of political instability [3]. A hoax is a piece of information or news that contains things that have not been identified or are not facts that happened. To gain profit and achieve personal goals, hoaxes are often deliberately created and shared to spread more quickly. Information obtained from hoaxes can certainly influence society because it creates doubts and confusion about the information received and can damage the image of individuals and related groups. Headlines used by hoax information are often sensational and provocative. It is deliberately made to attract the interest and curiosity of readers. At least 30% to nearly 60% of Indonesians are exposed to hoaxes when accessing and communicating through cyberspace. Meanwhile, only 21% to 36% can recognize or detect hoaxes. Most hoaxes were related to political, health, and education issues [4].

The rapid evolution of information technology has propelled the rampant proliferation of deceptive content, necessitating the development of advanced and intelligent systems for precise and effective classification. This research is ardently dedicated to confronting this multifaceted challenge by strategically implementing a robust and state-of-the-art classification system meticulously designed to discern and identify hoax news circulating through the intricate web of internet media channels. At the core of this classification process lies a meticulous preprocessing stage, and among the classification algorithms utilized, Naïve Bayes stands out for its simplicity and high accuracy [5][6].

The Naïve Bayes method is a classification method in machine learning that excels in using training data samples to rapidly estimate parameters involved in the classification process, resulting in high accuracy. Related research on Naïve Bayes classification has proposed various methods and algorithms. For instance, Muhabatin H *et al*. achieved a Naïve Bayes accuracy rate of 91.82%, while Dinesh T *et al*. achieved 95.38%; on the other hand, other research by Marissa Audina *et al* "Classification of Covid-19 Hoax News Using a Combination of K-Nearest Neighbor and Information Gain Methods" managed to achieve K-Nearest Neighbor model without feature selection with a value of k=5 obtained precision, recall, F1-Score, and accuracy performance of 87.5%, 96.5%, 91.8%, and 91.6%, respectively. While the K-Nearest Neighbor model with a combination of 0.5% Information Gain threshold feature selection with a value of k=3 obtained precision, recall, F1-Score, and accuracy performance of 93.3%, 96.6%, 95%, and 95% [7]. Sagita R *et al*., "Clickbait News Classification Using K-Nearest Neighbor (KNN)," managed to achieve the best results at k = 11 using scenario one on data division with a total of 800 data and 200 test data, which resulted in an accuracy of 71%, precision of 72%, and recall of 71% [8]. Ramadhan N *et al*., "Fake News Detection Using Random Forest and Logistic Regression Methods," achieved results using the random forest model of 84%, which is higher than the logistic regression model of 77% [9].

As we delve into the intricate nuances of hoax news classification, this groundbreaking research significantly contributes to the evolving technological landscape. It bears broader implications for advancing media literacy in the digital era. By skillfully navigating the intricate complexities of information dissemination, the profound insights derived from this study hold the promise of being instrumental in innovating tools that empower individuals to distinguish between trustworthy news sources and potentially misleading content adeptly. In this transformative era of continually evolving information landscapes, this pioneering research diligently aspires to fortify our collective ability to combat misinformation, thereby fostering and advancing a more enlightened and discerning society.

## 2. Research Method

The method used in this program is the Naive Bayes method, specifically the Naive Bayes Multinomial, which works with the assumption that each feature (word) is considered independent from the others. Text vectorization using CountVectorizer converts text into a numeric vector, which classification algorithms can use. This program uses a trained model to predict testing data and calculate evaluation metrics such as accuracy, confusion matrix, and classification reports [10]. The choice of the Naïve Bayes Multinomial method is driven by its efficiency and simplicity, particularly in handling large datasets standard in internet media. Its assumption of feature independence aligns well with text classification tasks. CountVectorizer is selected for text vectorization, converting text into a bag-of-words representation. This technique captures word frequencies, which is crucial for distinguishing between genuine and hoax news. CountVectorizer's numeric vector output seamlessly integrates with the Naïve Bayes Multinomial method. Together, these choices form a robust framework for efficiently classifying news articles in the challenging landscape of internet media misinformation.

 

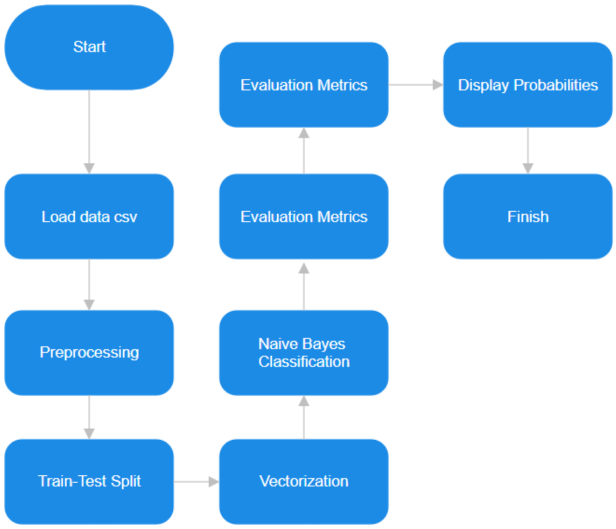


Figure 1. Research Stages

### 2.1 Load Dataset

The initial step in the program involves loading the datasets containing both fake and accurate news. This is facilitated by leveraging the Pandas library, a versatile tool for data manipulation and analysis in Python. Two separate CSV (Comma-Separated Values) files serve as the sources of these datasets, each representing a distinct news category. The utilization of Pandas ensures seamless handling of structured data and provides functionalities for efficient data exploration [11]. Labels are introduced to distinguish between the two categories during subsequent analysis. A label of "0" is assigned to news articles categorized as fake, while a "1" is assigned to those classified as true. This labeling process is fundamental for supervised machine learning, enabling the model to learn and generalize patterns based on known outcomes. Once labeled, the two datasets are merged into a unified data frame. Merging the datasets ensures a consolidated and coherent dataset for comprehensive analysis. The resulting data frame now contains all the pertinent information, including the news content, labels, and any additional attributes in the original datasets. This meticulous process of loading, labeling, and merging datasets sets the stage for subsequent stages in the program, laying the groundwork for adequate data preprocessing and model training. The structured data frame becomes the focal point for further exploration and manipulation as the program progresses.

### 2.2 Preprocessing

In the preprocessing phase, missing values are handled by filling them with an empty string (''). This proactive strategy ensures a seamless flow in subsequent processing steps, eliminating potential issues caused by null values in the dataset. Additionally, all news headlines are converted to lowercase. This standardization simplifies subsequent analyses by treating identical words uniformly, regardless of their original letter casing. The aim is to enhance consistency in feature representation during vectorization and model training, contributing to a more accurate and robust textual content analysis [12]. By addressing missing values and harmonizing letter casing, the program establishes a clean and standardized dataset, laying the foundation for subsequent stages in the machine learning pipeline.

### 2.3 Train-Test Split

In the train-test split stage, the dataset is partitioned into two subsets: a training set, constituting 80% of the entire data, and a testing set, comprising the remaining 20%. This division ensures a robust evaluation of the model's performance, as it is trained on a substantial portion of the data and tested on a separate, unseen portion. By employing an 80-20 split and presenting a detailed overview of the dataset's class distribution, the program establishes a strong foundation for subsequent training and testing phases in the machine learning workflow. The division of the dataset into training and testing sets is a fundamental practice in machine learning for model evaluation and validation. By allocating 80% of the data to the training set, the model learns patterns and relationships within the dataset. The remaining 20%, designated as the testing set, is an independent dataset to assess the model's ability to generalize to unseen data. This split helps detect overfitting, where a model memorizes the training data but performs poorly on new data. The 80-20 division strikes a balance between training sufficiency and the need for a diverse evaluation set, ensuring a reliable assessment of the model's performance and its potential to make accurate predictions in real-world scenarios.

### 2.4 Vectorization

In the text vectorization phase, news article titles transform to allow machine learning algorithms to process and analyze text information. CountVectorizer, a feature extraction technique in natural language processing, facilitates this transformation. CountVectorizer operates by converting each news article title into a numeric vector. This process allows a mathematical representation of the text, where each word in the title is calculated and attributed as a vector feature. The result is a data representation ready to be processed by machine learning algorithms for fake news detection and classification. Thus, the text vectorization phase is a critical step in data preparation for further fake news detection research analysis.

### 2.5 Naive Bayes Classification

The basic formula for Naive Bayes in binary classification scenarios, such as the case of distinguishing between fake news and true news, can be explained using Bayes' Theorem [13]. Let's use the following notations:

$P$(Fake|X) : The probability that the news is fake given features X.
$P$(True|X) : The probability that the news is true given features X.
$P$(X|Fake) : The probability of features X given that the news is fake.
$P$(X|True) : The probability of features X given that the news is true.
$P$(Fake) : The prior probability that the news is fake.
$P$(True) : The prior probability that the news is true.

Using Bayes' Theorem, we can write:

$$P(\,Fake \mid X\,) = \frac{P(X|Fake)\cdot P(Fake)}{P(X)} \tag{1}$$

$$P(True \mid X) = \frac{P(X|True)\cdot P(True)}{P(X)} \tag{2}$$

Assuming that the features (X) are independent of each other (Naive Bayes Assumption), we can simplify $P$(X|Fake)$P$(X|Fake) and $P$(X|True)$P$(X|True) as the multiplication of individual feature probabilities:

$$P(X \mid Fake) = P(word1 \mid Fake)\cdot P(word2 \mid Fake)\cdot \ldots \cdot P(wordn \mid Fake) \tag{3}$$

$$P(X \mid True) = P(word1 \mid True)\cdot P(word2 \mid True)\cdot \ldots \cdot P(wordn \mid True) \tag{4}$$

By approximating these feature probabilities using the training data, we can predict whether the news is fake or true based on the calculated probabilities [14].

### 2.6 Evaluation Metrics

Upon training the Naive Bayes classifier and making predictions on the testing set, the program evaluates the model's performance using three key metrics: accuracy, confusion matrix, and a classification report. Accuracy provides an overall measure of the model's correctness in classifying fake and accurate news. It is calculated as the ratio of correctly predicted instances to the total number of cases in the testing set. A higher accuracy indicates a more effective classification model. The confusion matrix delves deeper into the model's performance by breaking down the predictions into four categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This matrix offers insights into the model's ability to distinguish between fake and accurate news, revealing areas of strength and areas that may require improvement. The classification report provides a comprehensive overview of the model's precision, recall, and F1 score for each class (fake and genuine news). Precision measures the accuracy of optimistic predictions, recall assesses the model's ability to capture all positive instances, and the F1-score represents the harmonic mean of precision and recall. This detailed report aids in understanding the model's performance on a class-specific level. Together, these evaluation metrics furnish a thorough assessment of the Naive Bayes classifier, offering valuable insights into its strengths and areas for refinement in discerning between fake and authentic news articles.

### 2.7 Count Fake and True News

The program comprehensively examines the dataset's composition, aiming to show the distribution of fake and genuine news articles. This analysis is crucial in delineating the equilibrium between the two classes, laying the groundwork for subsequent model evaluation. Upon execution, the computation unveils significant insights into the dataset. Specifically, it discloses 23,503 instances of fake news articles and 21,418 cases of true news articles. This information assumes a pivotal role as it not only characterizes the dataset but also substantially impacts the model's learning dynamics. The emphasis on a balanced dataset becomes evident, ensuring the model encounters an ample representation of both classes. This balance, in turn, enhances the model's capacity to generalize and make accurate predictions when confronted with unseen data. Understanding the class distribution is fundamental for interpreting the model's performance accurately. If there were a significant imbalance between the number of fake and genuine news articles, it could impact the model's ability to discern patterns effectively and necessitate additional techniques such as resampling or adjusting class weights during training. In this case, the relatively balanced dataset contributes to a robust evaluation of the Naive Bayes classifier [15].

### 2.8 Display Probabilities

In the presented program, the classification probabilities for the first five samples in the testing data are computed and displayed. This analysis provides valuable insights into the model's confidence in assigning each news article to either the fake or accurate category. The model generates a probability distribution for each sample, indicating the likelihood of the news article belonging to each class. The program prints these probabilities alongside the corresponding news article titles. The classification probabilities are represented as numeric values between 0 and 1, where higher values signify greater confidence in the assigned class.

## 3. Result and Discussion

### 3.1 Results

The model achieves an impressive accuracy of 94.73%, showcasing its effectiveness in distinguishing between fake and authentic news articles during testing. This high accuracy underscores the model's reliability in real-world scenarios, suggesting its potential in mitigating the impact of misinformation. In practical terms, 94.73% accuracy holds significant promise for addressing the challenges posed by hoaxes in everyday life and the dynamic internet media environment. With such a reliable classification system, individuals navigating the vast online information landscape are empowered to discern between trustworthy and deceptive news sources. In the evolving realm of internet media, this model contributes to promoting a more informed and resilient society.

3.1.1 Confusion Matrix

Table 1. Confusion Matrix

| True Negative (TN) | False Positive (FP) | False Negative (FN) | True Positive (TP) |
|---|---|---|---|
| 4555 | 178 | 295 | 3952 |

1) True Negative (TN)
   The number of instances correctly predicted as fake news (class 0). In our case: 4555 news articles were accurately classified as fake.
2) False Positive (FP):
   The number of instances incorrectly predicted as true news (class 1). In our case: 178 fake news articles were wrongly classified as true.
3) False Negative (FN):
   The number of instances incorrectly predicted as fake news (class 0). In our case: 295 true news articles were wrongly classified as fake.
4) True Positive (TP):
   The number of instances correctly predicted as true news (class 1). In our case: 3952 news articles were accurately classified as true.

3.1.2 Classification Report

Table 2. Classification Report

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.94 | 0.96 | 0.95 | 4733 |
| 1 | 0.96 | 0.93 | 0.94 | 4247 |
| Accuracy | | | 0.95 | 8980 |
| Macro avg | 0.95 | 0.95 | 0.95 | 8980 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 8980 |

1) Precision
Precision is the ratio of correctly predicted positive observations to the total predicted positives. A high precision relates to a low false positive rate. In your case, the precision for class 0 (fake news) is 0.94, and for class 1 (true news) is 0.96.
2) Recall
Recall is the ratio of correctly predicted positive observations to the all observations in the actual class. It is also known as Sensitivity or True Positive Rate. In your case, the recall for class 0 is 0.96, and for class 1 is 0.93.
3) F1-Score
F1-Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is a good way to show that a classifier has a good value for both recall and precision. In your case, the F1-score is approximately 0.95 for both classes.

**3.2 Discussion**

The program exhibits a notable accuracy of 94.73%, effectively discerning between fake and authentic news even in a dataset imbalance with more instances of fake news (23,503) than accurate news (21,418). The confusion matrix provides a comprehensive view of the model's performance, highlighting areas for improvement, particularly in false positives and negatives. Introducing classification probabilities for the initial five test samples enhances interpretability, showcasing the model's confidence in its predictions. Tokenization, employed for text-to-feature vector conversion, streamlines the processing of textual data, while the 80% training and 20% testing data split ensures robust evaluation and generalization. Future enhancements may involve addressing class imbalance and exploring advanced natural language processing techniques to refine the model's news classification capabilities further.

## 4. Related Work

In recent years, efforts to address the problem of the spread of fake news or hoaxes have become the focus of significant research in various fields, including computer science, information science, and data analysis. Different approaches and techniques have been developed to detect and classify fake news with high accuracy. One commonly used approach is a classification algorithm, such as the Naïve Bayes Classifier. Research conducted by Pratiwi *et al.* (2018) explored the use of the Naïve Bayes Classifier in hoax detection in Indonesian. Their research results show the great potential of this approach in the context of the Indonesian language [1]. Furthermore, research by Santoso *et al.* (2020) optimized hoax news classification and sentiment analysis in Indonesia using Naïve Bayes. This research highlights the importance of developing efficient hoax detection methods in different language contexts [2]. Muhabatin *et al.* (2021) explored using the Naïve Bayes algorithm based on Particle Swarm Optimization (PSO) for hoax news classification. Their research results show a significant improvement in hoax detection performance through this approach [3]. In addition, several studies have compared the performance of Naïve Bayes with other algorithms, such as Support Vector Machine (SVM), as done by Febriyanty *et al.* (2023). Experimental results highlight the superiority of Naïve Bayes in some fake news detection [4]. Apart from that, recent research has also expanded the application of the Naïve Bayes algorithm to other fields, such as sentiment analysis of application user reviews, as carried out by Pasaribu & Sriani (2023) [5]. Related research highlights essential developments in using Naïve Bayes algorithms in fake news detection and classification. It shows the potential and challenges associated with this approach in different cultural and linguistic contexts.

The Naive Bayes classifier has demonstrated a commendable accuracy rate of 94.73% [16], making it a valuable tool in the fight against fake news. It consistently maintains high F1 scores, typically between 0.94 and 0.95 across categories, demonstrating its ability to balance precision and recall effectively [17]. This balance is critical in mitigating the impact of false positives and false negatives, especially in the challenging field of misinformation detection [18]. Moreover, the Naive Bayes model's strong performance confirms its current effectiveness and highlights its great potential for practical applications [19]. As technology advances, further improvements to Naive Bayes models are promising. They could potentially enable innovative applications such as real-time monitoring of news sources to quickly identify and counter the spread of misinformation [20]. The foundation of the Naive Bayes classifier in Bayes' theorem, which computes posterior probabilities while assuming attribute independence, underscores its efficacy in classification tasks [21]. This algorithm is widely used in various studies because of its efficiency, especially with large data sets [22]. Its simplicity and strong probabilistic basis make it a favorite choice in machine learning and data mining applications [23]. Naive Bayes classifiers are emerging as an essential tool in the fight against fake news, offering high levels of accuracy, a balanced approach to precision and recall, and significant potential to advance misinformation detection and prevention.

## 5. Conclusion

The Naive Bayes classifier's remarkable accuracy, reaching an outstanding 94.73%, positions it as a pivotal weapon in the battle against fake news. The F1 scores, consistently high between 0.94 and 0.95 across diverse categories, underscore the model's exceptional ability to strike a harmonious balance between precision and recall. This equilibrium is critical for minimizing the impact of false positives and false negatives, particularly in the intricate landscape of misinformation detection. The model's robust performance affirms its current effectiveness and illuminates its vast potential for real-world applications. Looking ahead and envisioning the trajectory of this technology, the continued refinement of Naive Bayes models holds immense promise. Its evolution may lead to groundbreaking applications, including real-time monitoring of news sources, providing platforms with the capability to identify and curb misinformation's dissemination swiftly. The Naive Bayes model doesn't just address the challenges of today; it emerges as a catalyst for shaping resilient, misinformation-resistant societies in the dynamic realm of Internet media.

## References

[1] Pratiwi, I. Y. R., Asmara, R. A., & Rahutomo, F. (2018). Study of hoax news detection using naïve bayes classifier in Indonesian language. *Proceedings of the 11th International Conference on Information and Communication Technology and System, ICTS 2017*, 73–78. https://doi.org/10.1109/ICTS.2017.8265649

[2] Santoso, H. A., Rachmawanto, E. H., Nugraha, A., Nugroho, A. A., Setiadi, D. R. I. M., & Basuki, R. S. (2020). Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization. *Telkomnika (Telecommunication Computing Electronics and Control)*, 18(2), 799–806. https://doi.org/10.12928/TELKOMNIKA.V18I2.14744

[3] Muhabatin, H., Prabowo, C., Ali, I., Lukman Rohmat, C., Rizki Amalia, D., sitasi, C., & Rizki, D. (2021). Klasifikasi Berita Hoax Menggunakan Algoritma Naïve Bayes Berbasis PSO. *Informatics for Educators and Professionals*, 5(2), 156–165. https://turnbackhoax.id/

[4] Febriyanty, N. E., Hariyadi, M. A., & Crysdian, C. (2023). Hoax Detection News Using Naïve Bayes and Support Vector Machine Algorithm. *International Journal of Advances in Data and Information Systems*, 4(2), 191–200. https://doi.org/10.25008/ijadis.v4i2.1306

[5] Pasaribu, N. A., & Sriani. (2023). The Shopee Application User Reviews Sentiment Analysis Employing Naïve Bayes Algorithm. *International Journal Software Engineering and Computer Science (IJSECS)*, 3(3), 194–204. https://doi.org/10.35870/ijsecs.v3i3.1699

[6] Mustofa, H., & Mahfudh, A. A. (2019). Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes. *Walisongo Journal of Information Technology*, 1(1), 1–12. https://doi.org/10.21580/wjit.2019.1.1.3915

[7] Audina, M., Karyawati, A. E., Supriana, I. W., Suhartana, I. K. G., Astawa, I. G. S., & Santiyasa, I. W. (2022). Klasifikasi Berita Hoaks Covid-19 Menggunakan Kombinasi Metode K-Nearest Neighbor dan Information Gain. *Jurnal Elektronik Ilmu Komputer Udayana*, 10(4), 319–327. https://cekfakta.com

[8] Ramadhan, N. G., Adhinata, F. D., Segara, A. J. T., & Rakhmadani, D. P. (2022). Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression. *JURIKOM (Jurnal Riset Komputer)*, 9(2), 251. https://doi.org/10.30865/jurikom.v9i2.3979

[9] Sagita, R., Enri, U., & Primajaya, A. (2020). Klasifikasi Berita Clickbait Menggunakan K-Nearest Neighbor (KNN). *JOINS (Journal of Information System)*, 5(2), 230–239. https://doi.org/10.33633/joins.v5i2.3705

[10] Dinesh, T., & Rajendran, T. (2021). Higher Classification of Fake Political News Using Decision Tree Algorithm Over Naive Bayes Algorithm. *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS, 2-11*, 1084–1096. https://doi.org/10.03.2021

[11] Siti Nurul Hassanah, W., Puji Lestari, Y., & Adi Saputra, R. (2023). Digital Image Processing to Detect Cracks in Buildings Using Naïve Bayes Algorithm (Case Study: Faculty of Engineering, Halu Oleo University). *Jurnal Informatika Dan Teknologi Informasi*, 20(1), 1–14. https://doi.org/10.31515/telematika.v20i1.8925

[12] Zaman, B., Justitia, A., Sani, K. N., & Purwanti, E. (2020). An Indonesian Hoax News Detection System Using Reader Feedback and Naïve Bayes Algorithm. *Cybernetics and Information Technologies*, 20(1), 82–94. https://doi.org/10.2478/cait-2020-0006

[13] Kesumawati, A., & Thalib, A. K. (2018). Hoax Classification with Term Frequency-Inverse Document Frequency Using Non-Linear SVM and Naïve Bayes. *Int. J. Advance Soft Compu. Appl*, 10(3).

[14] Ogilvie, T. L., Sharma, M., & Huu, Z. X. (2022). Fake news detection using naive bayes classifier and forward selection in the digital era. *Machine Learning, 29*(2/3), 131-163. https://doi.org/10.1023/a:1007465528199

[15] Muhammad Habib, S., Haerani, E., Kurnia Gusti, S., Ramadhani, S., & Informatika UIN Sultan Syarif Kasim Riau Jl Soebrantas, T. H. (2022). Klasifikasi Berita Menggunakan Metode Naïve Bayes Classifier. *Jurnal Nasional Komputasi Dan Teknologi Informasi, 5*(2), 248–258.

[16] Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. M. (1998). Using em to classify text from labeled and unlabeled documents. https://doi.org/10.21236/ada350490

[17] Carvalho, V. D. H. d., Nepomuceno, T. C. C., Poleto, T., Turet, J. G., & Costa, A. P. C. S. (2022). Mining public opinions on covid-19 vaccination: a temporal analysis to support combating misinformation. *Tropical Medicine and Infectious Disease, 7*(10), 256. https://doi.org/10.3390/tropicalmed7100256

[18] Senhadji, S., & Ahmed, R. A. S. (2022). Fake news detection using naïve bayes and long short term memory algorithms. *IAES International Journal of Artificial Intelligence (IJ-AI), 11*(2), 746. https://doi.org/10.11591/ijai.v11.i2.pp746-752

[19] Valentina, S., & Songpan, W. (2022). Features extraction based on probability weighting for fake news classification on social media. *Current Applied Science and Technology, 23*(2). https://doi.org/10.55003/cast.2022.02.23.014

[20] Davardoost, F., Sangar, A. B., & Majidzadeh, K. (2022). An innovative model for extracting olap cubes from nosql database based on scalable naïve bayes classifier. *Mathematical Problems in Engineering, 2022*, 1-11. https://doi.org/10.1155/2022/2860735

[21] Muhamad, H., Prasojo, C. A., Sugianto, N. A., Surtiningsih, L., & Cholissodin, I. (2017). Optimasi naïve bayes classifier dengan menggunakan particle swarm optimization pada data iris. *Jurnal Teknologi Informasi Dan Ilmu Komputer, 4*(3), 180. https://doi.org/10.25126/jtiik.201743251

[22] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Untitled. *Machine Learning, 29*(2/3), 131-163. https://doi.org/10.1023/a:1007465528199

[23] Taheri, S., Yearwood, J., Mammadov, M., & Seifollahi, S. (2013). Attribute weighted naive bayes classifier using a local optimization. *Neural Computing and Applications, 24*(5), 995-1002. https://doi.org/10.1007/s00521-012-1329-z.