*J. Intell. Comput. Health Inform. Vol. 5, No. 2, 2024*
ISSN: 2715-6923 e-ISSN: 2721-9186, pp.57~64 | DOI: 10.26714/jichi.v5i2.15420

# NAKNN: An Efficient Classification of Indonesian News Texts with Nazief-Adriani and KNN

Basirudin Ansor*, Aditya Putra Ramdani, Nova Christina Sari, Muhammad Zainudin Al Amin, Achmad Solichan, Kilala Mahadewi

Department of Information Technology, Universitas Muhammadiyah Semarang, Jl. Kedungmundu Raya No. 18, Semarang 50273, Indonesia
*Corresponding author: basirudinansor@unimus.ac.id

Article history :
Received: 11 AUG 2024
Accepted: 23 SEP 2024
Avalaible online: 30 SEP 2024

*Research article*

**Abstract:** Internet usage in Indonesia has seen a significant increase, reaching 215.63 million users in 2022-2023, or 78.19% of the population. With the ease of internet access, digital news portals like Narasi TV have become a primary source of information for many people. However, the large number of news articles makes manual categorizing challenging. This study aims to classify Indonesian-language news documents from Narasi TV using the Nazief-Adriani algorithm for stemming and the K-Nearest Neighbor (KNN) method for classification. The text mining process begins with preprocessing, which includes case folding, tokenizing, stop-word filtering, and stemming. Using a dataset of 500 news documents, the study demonstrated that with a 90:10 data split, the average accuracy reached 93%, with the highest value being 100%. For the 80:20 data split, the average accuracy was 89%, with the highest value being 93%, and for a 70:30 data split, the average accuracy was 87%, with the highest value being 89%. In conclusion, the combination of the Nazief-Adriani algorithm and the KNN method with optimal k selection and random states obtained high accuracy, obtaining an average accuracy of 93%) in classifying Indonesian-language news documents. These results demonstrate the significant potential of text mining and classification techniques to manage digital news.

**Keywords:** NEWS CLASSIFICATION; NAZIEF-ADRIANI; K-NEAREST NEIGHBOR; TEXT MINING



## 1. Introduction

The rapid advancement of technology has significantly impacted various aspects of life, including how people access information. According to a survey by the Indonesian Internet Service Providers Association (APJII) for the 2022-2023 period, internet users in Indonesia comprised 215.63 million people. This accounts for 78.19% of Indonesia's total population of 275.77 million people (IndonesiaBaik.id, 2023). One of the main activities of the digital community is reading news. News portals that provide rich content from various mass media sources contribute positively by offering valuable information to the public. Data from the DNR 2022 report reveal that nearly a quarter of digital news readers still rely on news portals as their primary source of information (Krisdamarjati, 2022).

News content in digital media is typically categorized into several sections, such as sports, education, culture, and economy. For example, Narasi TV offers young people space to exchange ideas and share their views through various programs (Narasi People, (2024). However, the growing number of news articles distributed through digital media presents a challenge for editors to manually categorize them (Septian, Susanto, & Shidik, 2017). Therefore, an efficient and timely method to organize news content is needed. One approach to address this problem is the application of text mining and classification algorithms, which can automatically group news articles based on the text they contain.

Text mining and classification algorithms are pivotal in several fields, including text retrieval, information summarization, and question-answering. Typically, the data used in the classification are diverse, originating from various sources such as news articles, discussion groups, advertisements, and reviews. Given the wide range of vocabulary, format, and writing styles, automatic text classification has become increasingly important. With the massive amount of text stored electronically, it is essential to understand and analyze this data to extract useful insights that can inform decision-making processes (Ahmed & Ahmed, 2021).

The application of automatic text classification, particularly in the news domain, has significant potential for enhancing information management. As the volume of digital information grows, manual processing becomes impractical and inefficient. By employing machine learning techniques like text mining, organizations can better manage, categorize, and retrieve information, leading to improved decision-making and faster access to relevant content. This study's focus on Indonesian-language news documents not only addresses this growing demand but also contributes to the broader field of information management by exploring the application of specific algorithms like nazief adriani (NA) and k-nearest neighbors (KNN) algorithms (Septian et al., 2017; Sinaga & Nainggolan, 2023).

Previous studies have demonstrated that text mining and classification methods such as naive bayes (NB) and KNN can improve news categorization, they still face limitations in terms of accuracy and efficiency. For example, although the NB algorithm has demonstrated an accuracy rate of 94% in news classification (Septian et al., 2017), it is less effective compared to KNN in handling more complex data structures. Studies comparing these two algorithms found that KNN achieved higher accuracy (86%) than Naive Bayes (51%) when tested on 100 news articles (Tejawati, Septiarini, Rismawati, & Puspitasari, 2023). However, both approaches are limited by their reliance on simple preprocessing techniques, which may not capture the full complexity of the data. This study aims to address these limitations by implementing the Nazief-Adriani algorithm for stemming in the preprocessing stage, followed by the KNN method for classification.

The news text data will undergo preprocessing to enhance their quality and clarity. Several subprocesses are involved in this stage, including case folding, tokenizing, stop-word filtering, and stemming (Sudrajat, Wulandari, & Syafwan, 2022). The NA algorithm is known for its high accuracy and fast stemming process (Sinaga & Nainggolan, 2023). Combined with the KNN classification algorithm, the proposed method offers an effective solution for classifying news articles, particularly on the Narasi TV platform. This combination improves both the accuracy and efficiency of news classification (Septian et al., 2017).

Based on the description above, this study focuses on classifying Indonesian-language news documents from the Narasi TV website using the NA algorithm in the preprocessing stage and the KNN method for classification. By leveraging these advanced techniques, this research seeks to overcome the limitations of previous methods and contribute to the broader field of text mining and digital information management.

## 2. Research Methodology

The text classification process involves several steps to group texts into specific categories or classes based on their content and characteristics. This study focused on classifying Indonesian-language news using the nazief-adriani (NA) algorithm for stemming and the k-nearest neighbor (KNN) algorithm for classification.

### 2.1 Nazief-adriani algorithm

The NA algorithm was chosen for stemming because of its deep understanding of the Indonesian morphological structure. This algorithm efficiently identifies root words by removing affixes based on extensive morphological rules specific to the Indonesian language (Wahyudi et al., 2017). One of the key advantages of this algorithm is its ability to handle complex affix combinations with high accuracy and processing speed, which makes it ideal for text mining tasks involving the Indonesian language. Studies have shown that the NA algorithm outperforms other stemming algorithms in terms of both accuracy and time efficiency (Lakonawa, Mola, & Fanggidae, 2021).

The rationale for using the NA algorithm stems from its suitability for processing large volumes of Indonesian language text. Unlike other stemming algorithms that may lack specificity in handling unique Indonesian affixes, NA provides a tailored approach to Indonesian language processing, improving both the accuracy and relevance of the resulting root words (Septian et al., 2017).

### 2.2 K-nearest neighbor (KNN)

The KNN algorithm was selected for classification due to its simplicity and effectiveness in handling text data. The KNN algorithm is a nonparametric algorithm that classifies documents based on the majority class of their nearest neighbors. One of its main advantages is its ability to handle high-dimensional data without making assumptions about the data distribution (Rozi, Sukmana, & Adani, 2021). This flexibility makes the proposed KNN suitable for text classification tasks in which document vectors often involve thousands of features, such as word frequencies or TF-IDF scores.

KNN's performance, particularly when used in conjunction with the Nazief-Adriani stemming process, demonstrates high classification accuracy. In this study, different k values (1, 3, 5, 7, 9) were tested to determine the optimal number of neighbors for classification. The Euclidean distance metric was employed to measure the similarity between documents, ensuring that documents with similar content were grouped into the same category.

### 2.3 Rationale for algorithm selection

The combination of Nazief-Adriani for preprocessing and the KNN method for classification provides a robust solution for Indonesian-language news classification. The Nazief-Adriani algorithm provides an effective stemming process tailored to Indonesian text, and the KNN algorithm offers flexibility and classification simplicity, making it an ideal choice for this task. Other algorithms, such as Naive Bayes or SVM, could have been used, but KNN's ability to achieve high accuracy with minimal computational complexity and straightforward implementation made it preferable for this study (Tejawati, Septiarini, Rismawati, & Puspitasari, 2023). The KNN's adaptability to handling high-dimensional data further supports its selection.

### 2.4 Flowchart of preprocessing and classification process

To enhance clarity, a flowchart summarizing the preprocessing and classification steps is visualised in Fig. 1.

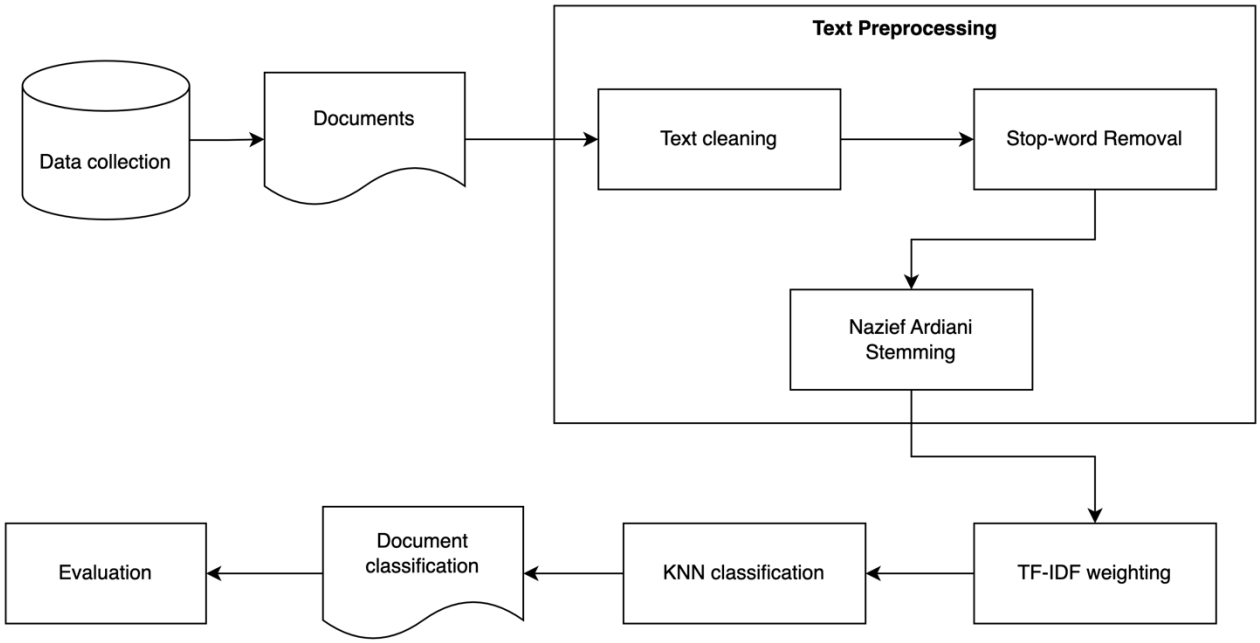



Fig 1. Research framework

*a. Data collection*

This task gathered news articles from the Narasi TV platform.

*b. Preprocessing*

- Text cleaning: Remove unwanted characters, numbers, and punctuation.
- Stop-word removal: We filter out common words (stopwords) using a predefined stop-word list.
- Tokenization: The text is broken into individual words (tokens).
- Stemming: Apply the nazief-adriani algorithm to reduce words to their root forms.

*c. Feature extraction*

The TF-IDF scores for each word in the preprocessed documents are calculated.

*d. Classification*

Use the KNN algorithm with varying values of k (1, 3, 5, 7, 9) to classify documents.

*e. Evaluation*

Classification model accuracy was evaluated.

## 3. Results and Discussion

### 3.1 Dataset

In this study, the data used were news documents on the Narasi TV website, totaling 500 data with 5 categories: financial, legal, political, international and public policy, with each category consisting of 100 documents. Data is taken based on the category and then entered into the database as training and testing data. Data is divided into 90:10, 80:20, and 70:30 to determine the accuracy obtained if the data separation is performed on a different scale. Table 1 presents an example document sentence used for training and testing the proposed model.

Table 1. Example of news document data.

| DOKUMEN | LABEL |
|---|---|
| Menerapkan *financial freedom* membuat seseorang tidak perlu khawatir akan kondisi finansialnya. Hal ini tentu kebalikan dari *financial insecurity* atau ketakutan finansial… | Finansial |
| Ketua IPW Sugeng Teguh Santoso mengatakan bahwa laporan itu atas dugaan penerimaan *cashback* dari perusahaan asuransi. Nilai dugaan gratifikasi atau suap itu mencapai lebih dari Rp100 miliar… | Hukum |
| Pasca penetapan hasil Pemilu 2024 untuk kursi legislatif oleh KPU, muncul ketegangan antara PDIP dan Golkar terkait posisi Ketua DPR… | Politik |
| Utusan Palestina di PBB Riyad Mansour mengungkap kekejaman Israel terhadap warga Gaza. Mansour menyebut Israel telah membuat warga Gaza kelaparan… | Internasional |
| Menteri Ketenagakerjaan (Menaker) Ida Fauziyah, baru-baru ini mengeluarkan Surat Edaran (SE) yang berkaitan dengan prosedur pemberian Tunjangan Hari Raya (THR) Keagamaan tahun 2024... | Kebijakan Publik |

### 3.2 Preprocessing

*a. Text cleaning*

Data preprocessing involves cleaning data to prepare them for processing using techniques applied in machine

learning (see Table 2). The methods typically used in this preprocessing stage include the following:

| |
|---|
| Text Cleaning: This step involves cleaning sentences if there are characters such as (!?<'") and others. |

Table 2. Text cleaning.

| TEXT DOCUMENT | TEXT CLEANING |
|---|---|
| Menteri ketenagakerjaan (Menaker) Ida Fauziyah, baru-baru ini mengeluarkan Surat Edaran (SE) yang berkaitan dengan prosedur pemberian Tunjangan Hari Raya (THR) Keagamaan tahun 2024... | menteri ketenagakerjaan menaker ida fauziyah barubaru ini mengeluarkan surat edaran se yang berkaitan dengan prosedur pemberian tunjangan hari raya thr keagamaan tahun. |

*b. Stop-word removal*

This step involves removing unimportant words such as "di," "yang," "dari," "ke," "dan," "ini," etc. In the example data with the category of public policy mentioned above, there are several stopwords that need to be eliminated.

Table 3. Stop-word removal.

| TEXT DOCUMENT | TEXT CLEANING |
|---|---|
| menteri ketenagakerjaan menaker ida fauziyah barubaru ini mengeluarkan surat edaran se yang berkaitan dengan prosedur pemberian tunjangan hari raya thr keagamaan tahun. | menteri ketenagakerjaan menaker ida fauziyah barubaru mengeluarkan surat edaran berkaitan prosedur pemberian tunjangan raya thr keagamaan |

*c. Nazief-adriani for stemming*

Removing affixed words based on the rules provided by the Nazief-Adriani algorithm is as follows in Table 4.

Table 4. Nazief-adriani for stemming.

| TOKEN | STEMMING RESULT |
|---|---|
| menteri | menteri |
| ketenagakerjaan | ketenagakerjaan |
| menaker | menaker |
| ida | ida |
| fauziyah | fauziyah |
| barubaru | barubaru |
| mengeluarkan | keluar |
| surat | surat |
| edaran | edar |
| berkaitan | kait |
| prosedur | prosedur |
| pemberian | beri |
| tunjangan | tunjang |
| raya | raya |
| thr | thr |
| keagamaan | agama |

### 3.3 Classification

Dataset was split into training and test sets using scikit-learn's train_test_split method.

1) Pipeline and Grid Search
   The pipeline is created using TF-IDF Vectorizer for text vectorization and KNeighborsClassifier for the k-NN model. Grid search with different k parameters ([1, 3, 5, 7, 9]) is conducted to find the best k value using cross-validation
2) Experiments with Random State
   - Experiments are conducted with random_state values ranging from 1 to 100.
   - For each random state, 5 trials are performed to calculate the average accuracy.
   - The best k value obtained from each trial is recorded for further analysis.

Table 5. Best *k* value accuracy.

| DATASET | RANDOM STATE | BEST K | ACCURACY |
|---|---|---|---|
| 90:10 | 42 | 7 | 0.9286 |
| | 101 | 9 | 0.8333 |
| | 202 | 9 | 0.881 |
| | 303 | 5 | 0.881 |
| | 404 | 9 | 0.8095 |
| 80:20 | 42 | 9 | 0.881 |
| | 101 | 7 | 0.8571 |
| | 202 | 9 | 0.8571 |
| | 303 | 7 | 0.869 |
| | 404 | 9 | 0.7976 |
| 70:30 | 42 | 7 | 0.864 |
| | 101 | 7 | 0.872 |
| | 202 | 7 | 0.832 |
| | 303 | 7 | 0.872 |
| | 404 | 9 | 0.808 |

## 4. Results

This study aimed to test the effectiveness of combination Nazief-Adriani and K-Nearest Neighbor (NA+KNN) algorithms in classifying Indonesian news documents. The performance of the algorithms was assessed based on different splits (90:10, 80:20, and 70:30) and random states to ensure consistent results. The performance analysis results are shown in Table 6, which summarizes the random states with the highest accuracy for each data proportion.

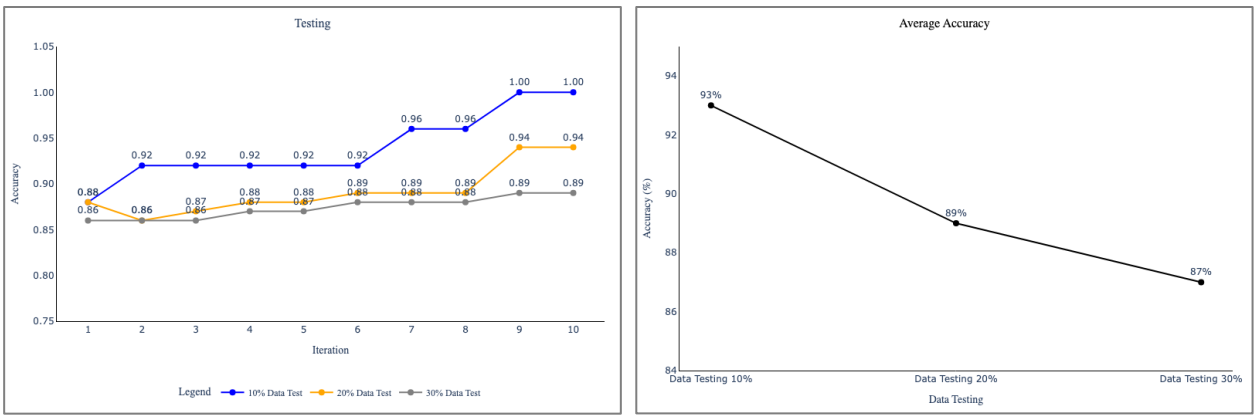
As shown in Table 6, we found that the best accuracy was achieved when a proportion of 10% of the test data was used, and the highest accuracy was 97.62% for random states 40. This shows that the larger the training data, the better the performance of the model in accurately classifying the data. At 20% of the test data, random states 60 and 69 yielded the highest accuracy of 91.67%, indicating a moderate decrease in performance as the size of the test data increases. Meanwhile, at 30% of the test data, the maximum accuracy of 90.4% was achieved at random state 60, confirming that increasing the amount of test data can reduce the model's ability to predict correctly. Overall, the model performed consistently across the various random state configurations, with relatively small differences in accuracy between the best random state values for each data proportion.

Table 6. Random state with the best average accuracy.

| 10% TEST DATA | | 20% TEST DATA | | 30% TEST DATA | |
|---|---|---|---|---|---|
| *Random State* | *Average Accuracy* | *Random State* | *Average Accuracy* | *Random State* | *Average Accuracy* |
| 40 | 0.9762 | 60 | 0.9167 | 60 | 0.904 |
| 2 | 0.9524 | 69 | 0.9167 | 25 | 0.896 |
| 37 | 0.9524 | 2 | 0.9048 | 31 | 0.896 |
| 57 | 0.9524 | 11 | 0.9048 | 53 | 0.896 |
| 60 | 0.9524 | 40 | 0.9048 | 2 | 0.888 |
| 8 | 0.9286 | 45 | 0.9048 | 9 | 0.888 |
| 15 | 0.9286 | 97 | 0.9048 | 58 | 0.888 |
| 25 | 0.9286 | 21 | 0.8929 | 62 | 0.888 |
| 26 | 0.9286 | 31 | 0.8929 | 34 | 0.88 |
| 42 | 0.9286 | 52 | 0.8929 | 40 | 0.88 |

Table 7. Optimal *k* value and accuracy.

| TESTING | 10% TEST DATA | 20% TEST DATA | 30% TEST DATA |
|---|---|---|---|
| $k = 1$ | 0.88 | 0.88 | 0.86 |
| $k = 2$ | 0.92 | 0.88 | 0.86 |
| $k = 3$ | 0.92 | 0.88 | 0.86 |
| $k = 4$ | 0.92 | 0.88 | 0.87 |
| $k = 5$ | 0.92 | 0.88 | 0.87 |
| $k = 6$ | 0.92 | 0.88 | 0.88 |
| $k = 7$ | 0.92 | 0.88 | 0.88 |
| $k = 8$ | 0.96 | 0.90 | 0.88 |
| $k = 9$ | 0.96 | 0.92 | 0.89 |
| $k = 10$ | 1.00 | 0.94 | 0.89 |




(a) (b)

Fig 2. (a) Accuracy testing, and (b) average accuracy

### 4.1 Optimization of parameter *k* in knn

For optimization KNN classification we need to select of parameter k-fold to ensure optimal performance. Table 7 and Fig. 2 shows the average accuracy for various *k*-fold values.

Based on the test results shown in Table 7, Fig. 2 and 3, it can be seen that the k value affects the accuracy of the model significantly at each proportion of test data. For 10% of the test data, the model achieved a maximum accuracy of 100% at k = 10, while for 20% of the test data, the best accuracy was 94% at the same k. Although k = 10 gave the highest accuracy, the values of k = 7 and k = 8 showed almost similar performance with stable results in each data split, achieving 92% and 96% accuracy for 10% of the test data, respectively. Overall, the model showed a trend that the higher the value of k, the more optimal the accuracy obtained, with the best performance seen when the proportion of training data is larger.

### 4.2 Accuracy analysis based on data proportion

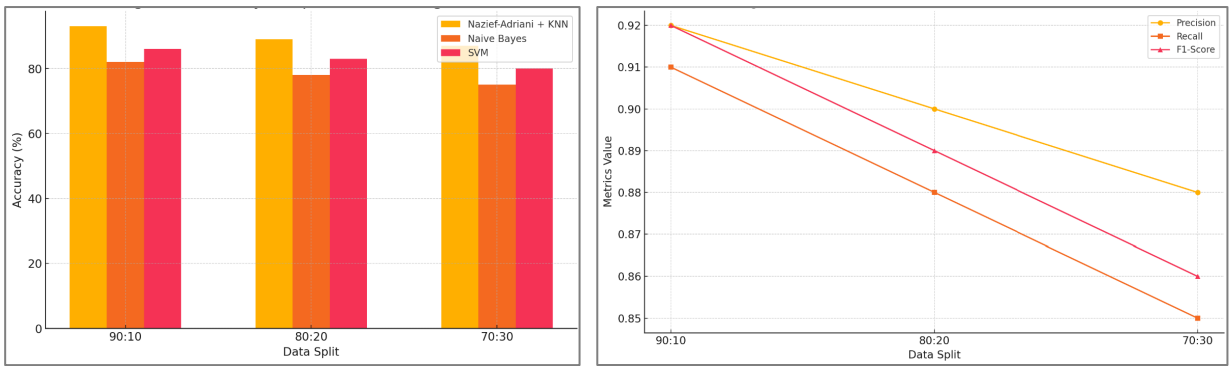
Experimental results with various proportions of data show that the larger the training data, the better the performance of the model. Table 8 summarizes the

average accuracy based on the proportion of test data.
Table 8. Performance comparison of algorithms.

| ALGORITHM | DATA SPLIT | ACCURACY | PRECISION | RECALL | F1-SCORE |
|---|---|---|---|---|---|
| Nazief-Adriani + KNN | 90:10 | 93% | 0.92 | 0.91 | 0.92 |
| Nazief-Adriani + KNN | 80:20 | 89% | 0.90 | 0.88 | 0.89 |
| Nazief-Adriani + KNN | 70:30 | 87% | 0.88 | 0.85 | 0.86 |
| Naive Bayes | 90:10 | 82% | 0.85 | 0.80 | 0.82 |
| SVM | 90:10 | 86% | 0.87 | 0.85 | 0.86 |

Table 9 Summarizes the results for each method.

| ALGORITHM | ACCURACY | PRECISION | RECALL | F1-SCORE |
|---|---|---|---|---|
| Nazief-Adriani + KNN | 89% | 0.90 | 0.88 | 0.89 |
| Naive Bayes | 82% | 0.85 | 0.80 | 0.82 |
| SVM | 86% | 0.87 | 0.84 | 0.85 |




(a) (b)

Fig 3. (a) Accuracy comparison across algorithms, and (b) performance metrics for NA+KNN

As shown in Table 8, the NA+KNN combination shows the best performance in terms of accuracy, precision, recall, and F1-score compared to the naive bayes and support vector machine algorithms. At a data proportion of 90:10, the NA+KNN combination achieved the highest accuracy of 93%, with an F1-score of 0.92, outperforming Naive Bayes and SVM which only achieved 82% and 86% accuracy respectively. As the proportion of training data decreased to 80:20 and 70:30, the accuracy of NA+KNN decreased to 89% and 87%, but still maintained a higher performance metric compared to Naive Bayes and SVM at the same split. This indicates that the NA+KNN combination is more effective in handling data with high dimensionality and variation, and maintains consistent performance even when the amount of test data increases.

### 4.3 Accuracy analysis based on best split using matrix evaluation

The performance results of the NA+KNN compared naive bayes and support vector machine using metrics evaluation, such as accuracy, precision, recall, and F1-score shown in Table 9 and Fig. 3.

As shown in Table 9 and Fig. 3, the combination of NA+KNN produces the best performance compared to naive bayes and support vector machine in all evaluation metrics. The Nazief-Adriani + KNN algorithm achieved 89% accuracy with an F1-score of 0.89, showing a good balance between precision (0.90) and recall (0.88). Naive Bayes, although simpler, showed a lower accuracy of 82% with an F1-score of 0.82, indicating that this method is less effective in handling data complexity. SVM, although more competitive, only achieved 86% accuracy with F1-score 0.85, which is still below the performance of NA+KNN. These results emphasize that the combination of NA+KNN algorithm is superior in news classification, especially in capturing patterns more accurately and consistently.

## 5. Discussions

### 5.1 Algorithm performance analysis

Based on the results, the NA+KNN combination proved superior in all evaluation metrics compared to Naive Bayes and Support Vector Machine (SVM). At 90:10 data split, the combination achieved the highest accuracy of 93% and maintained consistent metrics on precision (0.92), recall (0.91), and F1-score (0.92). The decrease in accuracy to 89% and 87% for the 80:20 and 70:30 data proportions shows that a smaller amount of training data affects the model's ability to capture patterns, but the NA+KNN combination still performs better than the other algorithms.

The decrease in accuracy that occurs with increasing test data is in line with the characteristics of the KNN algorithm, which relies on the amount of training data to produce accurate classifications. Thus, optimal

performance is achieved when more data is used for training, such as in the 90:10 split.

### 5.2 Comparison with other algorithms

Compared to naive bayes, which only achieved 82% accuracy and F1-score of 0.82, the combination of NA+KNN excels in handling data complexity. This can be explained by the limitation of Naive Bayes that assumes independence between features, making it less effective in processing data that has complex relationships such as news text. Meanwhile, SVM provides better accuracy than Naive Bayes with 86% and F1-score 0.85, but still lags behind the NA+KNN combination. This confirms that KNN has an advantage in handling high-dimensional data with the support of the stemming process of the NA algorithm, which improves the consistency of the base word in Indonesian.

### 5.3 Practical implications

The results of this study have significant implications for digital news management. By combining NA+KNN, the news classification process can be automated with a high level of accuracy, reducing manual workload and speeding up publication time. This implementation can be applied to various news platforms, such as Narasi TV, to improve the quality of content management and provide more relevant information to readers.

### 5.4 Research limitations

Although the combination of NA+KNN shows superior performance, there are some limitations in this study. Firstly, KNN has a disadvantage in terms of computation time, especially when the dataset becomes larger, as the algorithm requires nearest neighbor search in each prediction process. In addition, the results show that the model is sensitive to the proportion of training data, where the performance decreases when the training data decreases.

### 5.5 Future research directions

Future research could focus on applying deep learning-based algorithms such as LSTM or BERT, which have the potential to further improve accuracy and efficiency. In addition, testing on a multilingual news corpus will help assess the generalizability of this model beyond Bahasa Indonesia. Thus, this model can be extended to various applications in digital information management, not just limited to news platforms.

## 4. Conclusion

This research shows that the combination of Nazief-Adriani+K-Nearest Neighbor (KNN) is an effective solution for Indonesian news classification, with consistently superior performance compared to Naive Bayes (NB) and Support Vector Machine (SVM) algorithms. The model achieved the highest accuracy of 93% at 90:10 data proportion, with precision, recall, and F1-score of 0.92 each. Despite a drop in accuracy at 80:20 and 70:30 proportions, the model still performed competitively, with accuracies of 89% and 87% respectively. This confirms that the larger the amount of training data, the better the model's ability to capture relevant patterns.

The performance of Nazief-Adriani + KNN also proved superior to Naive Bayes, which only achieved 82% accuracy, and SVM, which produced 86% accuracy. This is mainly due to the combination of Nazief-Adriani's effective stemming process and KNN's ability to handle high-dimensional data without assuming data distribution.

Practically, these findings are relevant for digital news platforms such as Narasi TV, where automating news classification can reduce manual workload and speed up publication. The system allows for more efficient and accurate information management, and provides a better experience for users. However, there are some limitations to this research, particularly regarding KNN computation time on larger datasets and performance degradation when training data is reduced.

Therefore, future research can explore the use of deep learning models such as BERT or LSTM to further improve accuracy and efficiency. In addition, testing this model on a multilingual corpus will open up opportunities for wider generalization in various digital information management contexts.

## Acknowledgement

We would like to thank the Department of Information Technology, Faculty of Engineering and Computer Science, Universitas Muhammadiyah Semarang for its support and encouragement throughout the process of conducting this study.

## Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

Ahmed, J., & Ahmed, M. (2021). Online News Classification Using Machine Learning Techniques. *IIUM Engineering Journal*, *22*(2), 210–225. https://doi.org/10.31436/iiumej.v22i2.1662

Apriani, H., Jaman, J. H., & Adam, R. I. (2022). Optimasi SVM menggunakan algoritme grid search untuk identifikasi citra biji kopi robusta berdasarkan circularity dan eccentricity. *Jurnal Teknologi Dan Sistem Komputer*.

Herlingga, A. C., Prismana, I. P. E., Prehanto, D. R., & Dermawan, D. A. (2020). Algoritma Stemming Nazief & Adriani dengan Metode Cosine Similarity untuk Chatbot Telegram Terintegrasi dengan E-layanan. *Journal of Informatics and Computer Science (JINACS)*, *2*(01), 19–26. https://doi.org/10.26740/jinacs.v2n01.p19-26

IndonesiaBaik.id. (2023). Pengguna Internet Indonesia Paling banyak Usia Berapa? Retrieved February 6, 2024, from Indonesia.go.id website: https://indonesia.go.id/mediapublik/detail/2093#:~:text=Berdasarkan hasil survei Asosiasi Penyelenggara,orang pada periode 2022-2023.

Krisdamarjati, Y. A. (2022). Membaca Menjadi Pilihan Mengakses Berita Digital. Retrieved February 6, 2024, from kompas.id website: https://www.kompas.id/baca/riset/2022/07/07/membaca-

menjadi-pilihan-mengakses-berita-digital
Lakonawa, K. N., Mola, S. A. S., & Fanggidae, A. (2021). Nazief-Adriani Stemmer Dengan Imbuhan Tak Baku Pada Normalisasi Bahasa Percakapan Di Media Sosial. *Jurnal Komputer Dan Informatika*, *9*(1), 65–73. https://doi.org/10.35508/jicon.v9i1.3749
Narasi People. (2024). TENTANG NARASI. Retrieved February 16, 2024, from Narasi.tv website: https://narasi.tv/about-us?menu=informasi-perusahaan
Rozi, F., Sukmana, F., & Adani, M. N. (2021). Pengelompokkan Judul Buku dengan Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Term Frequency – Inverse Document Frequency (TF-IDF). *JIMP: Jurnal Informatika Merdeka Pasuruan*, *6*(3), 1–5.
Septian, G., Susanto, A., & Shidik, G. F. (2017). Indonesian news classification based on NaBaNA. *Proceedings - 2017 International Seminar on Application for Technology of Information and Communication: Empowering Technology for a Better Human Life, ISemantic 2017*, *2018-Janua*, 175–180. https://doi.org/10.1109/ISEMANTIC.2017.8251865
Sinaga, A., & Nainggolan, S. P. (2023). Analisis Perbandingan Akurasi Dan Waktu Proses Algoritma Stemming Arifin-Setiono Dan Nazief-Adriani Pada Dokumen Teks Bahasa Indonesia. *Sebatik*, *27*(1), 63–69. https://doi.org/10.46984/sebatik.v27i1.2072
Sudrajat, A., Wulandari, R. R., & Syafwan, E. (2022). Indonesian Language Hoax News Classification Basedn on Naïve Bayes. *Journal of Applied Intelligent System*, *7*(1), 70–79. https://doi.org/10.33633/jais.v7i1.5985
Tejawati, A., Septiarini, A., Rismawati, R., & Puspitasari, N. (2023). Perbandingan Metode K-Nearest Neighbor dan Naive Bayes untuk Klasifikasi Konten Berita. *Jurnal Teknik Informatika (Jutif)*, *4*(2), 401–412.
Wahyudi, D., Susyanto, T., Nugroho, D., Studi Teknik Informatika, P., Sinar Nusantara Surakarta, S., & Studi Sistem Informasi, P. (2017). Implementasi dan Analisis Algoritma Stemming Nazief & Adriani dan Porter pada Dokumen Berbahasa Indonesia. *Jurnal Ilmiah SINUS STIMIK Sinar Nusantara Surakarta*, *15*(2).