Improving News Text Classification Using a Hybrid C5.0-KNN Model

Liza Wikarsa [1*], Algy Ngenget [2], Andrew Tumewu [3], Miracle Kenneth Kalempouw [4], Edgard Oley [5]
lwikarsa@unikadelasalle.ac.id [1*] , 22013022@unikadelasalle.ac.id [2], 22013004@unikadelasalle.ac.id [3], 22013041@unikadelasalle.ac.id [4], 22013011@unikadelasalle.ac.id [5]

Article Info

Article history:

Received 2025-10-10
Revised 2025-10-30
Accepted 2025-11-05

Keyword:

News Topic Classification,
C5.0 Decision Tree,
K-Nearest Neighbours,
Low-Resource Languages,
Indonesia.

ABSTRACT

In the digital era, the overwhelming volume of online news far exceeds readers' ability to manually filter information, necessitating automated text classification. However, achieving high classification accuracy remains challenging, especially in low-resource languages like IndonesianThe C5.0 decision tree and K-Nearest Neighbors (KNN) offer complementary strengths but have not yet been jointly utilized for Indonesian news classification; therefore, this study proposes a hybrid C5.0–KNN model designed to enhance news classification performance. A dataset of 1.700 articles was collected from four Indonesian online news, namely CNN Indonesia, Okezone, Tribun Jakarta, and Tribun Jabar, covering five topical categories, namely economy/ekonomi, technology/teknologi, sport/olahraga, entertainment/hiburan, or life style/gaya hidup). The data underwent preprocessing and TF-IDF weighing before classification with the hybrid model. In this approach, C5.0 first generates interpretable decision rules, and KNN then refines borderline cases, combining rule-based and instance-based methods. The findings revealed that the hybrid model achieved a highest accuracy of 0.8847 (using 25% test data and k=5), outperforming standalone C5.0 (0.7426) and KNN (0.8735). Notably, it attained 100% recall for "sport/olahraga" and an F1-score of 0.89 for "entertainment/hiburan". These results demonstrate the model's novelty, efficiency, and strong potential for real-world news classification in low-resource language contexts, offering practical value for journalists, analysts, and media monitoring systems.

I. INTRODUCTION

In the era of digital information, the constant flow and overwhelming volume of real-time online news requires effective tools to sort and interpret large amount of text quickly and accurately. One of the key challenges in this context is text classification that organizes content into meaningful categories based on its core characteristics. A common use of text classification is sorting new articles by topic-like economy, technology, or entertainment-to make them easier to find, more relevant, and more useful for readers [1]. However, existing methods still face limitations in accuracy and interpretability, especially for Indonesian-language news. To address this challenge, many researchers have explored different machine learning algorithms – C5.0 decision tree algorithm valued for its transparent rule-based logic and K-Nearest Neighbors (KNN) performs well at identifying patterns by comparing new data to its closest, most similar neighbors [2] [3]. Individually, these methods have proven their worth: C5.0 attains nearly 80% when distinguishing hoax from factual news [4], while an optimized KNN can lift topic-level accuracy by roughly 20% over a vanilla baseline [5].

To enhance performance further, recent studies have explored ensemble methods, optimization techniques, and hybrid models. For example, Amarta et al. [6] integrated C5.0 with SMOTE and AdaBoosst to classify public sentiment regarding healthcare services, emphasizing the increased complexity introduced by multi-layer boosting, which can enhance accuracy through ensemble methods. While Kasanah et al. [7] implemented SMOTE with KNN to handle class imbalance in online news objectivity classification but found that performance declined when using larger K values and dealing with more complex data. General hybrid models like KNN Tree have performed well on standard datasets, but they have not been tested on news articles or in low-resource languages such as Indonesian [8].

These insights highlight a noticeable gap in current research: while C5.0 and KNN offer complementary advantages, they have yet to be systematically integrated into a dedicated model for classifying news in the Indonesian language. Addressing this gap is considered important and timely, given the growing need for hybrid and interpretable systems that perform well across languages – not just in English [9]. Hence, we propose a hybrid C5.0-KNN in which C5.0 first generates transparent decision rules, providing an interpretable backbone for topic assignment. Furthermore, KNN then refines borderline cases through similarity-based voting, enhancing adaptability to local linguistic variation. The main objective of this study is to develop and evaluate a hybrid C5.0-KNN model for automatic news topic classification in Indonesian. This design yields four important contributions that are: 1) first Indonesian C5.0-KNN news classifier that blends decision-tree rules with neighbor voting for local news, 2) proves NLP (Natural Language Processing) for low-resource language to show one can build clear, high-quality models even outside well-resourced tongues, 3) flexibility to combine C5.0's transparent rules with KNN's nuance to meet the push for explainable AI (Artificial Intelligence) [10], and 4) real-time web development that can accept raw news text and instantly get its topic label (economy/ekonomi, technology/teknologi, sport/olahraga, entertainment/hiburan, or life style/gaya hidup) alongside full performance metrics (accuracy, precision, recall, and F1).

By integrating C5.0 and KNN in a single framework – rare approach in new classification – and transforming into a browser-based tool, this research sets a new standard for hybrid text classification models, especially in low-resource languages like Indonesia as well as provides a simple, reliable tool for journalists, analysts, and educators. This approach paves the way for future models that merge logical rule-based decisions with flexible, data-driven insights to produce a powerful and explainable system ready for real-world use.

II. METODE

This research employs a combination of the C5.0 and KNN algorithms for implementation in the task of news topic classification. The following outlines the research sample and steps in implementing the hybrid models of C5.0-KNN.

A. Research Sample

The data used in this research was obtained from four distinct news APIs (Application Programming Interfaces) - CNN, Okezone, Tribun Jakarta, and Tribun Jabar – resulting in a total of 1.700 news data entries. Each entry includes two key attributes: "topic" and "contentSnippet." The dataset consists of 390 articles on technology (teknologi), 350 on sport (olahraga), and 320 articles each covering economy (ekonomi), entertainment (hiburan), and lifestyle topics (gaya hidup).

Altough the imbalance is minor, a light undersampling strategy is applied to ensure similar representation per class. In this process, the larger classes (Technology and Sport) were randomly undersampled to align more closely with the smaller ones, ensuring a more uniform representation across categories. This balancing method was deliberately conservative to preserve dataset diversity while preventing the classifier from overfitting to majority classes. Also, it is done to ensure that the hybrid C5.0-KNN model evaluated topic category under relatively equal conditions, thereby minimizing bias and improving the reliability of comparative performance metrics such as F1-score.

B. Steps in Implementing the Hybrid C5.0-KNN Model

The hybrid C5.0-KNN pipeline works by which the C5.0 decision tree generates interpretable classification rules based on Information Gain. Instances with low classification (e.g. equal Information Gain) are sent to KNN, which classifies them by measuring Euclidean distances between their TF-IDF vectors and the *k* nearest neighbors. The final decision uses majoruty voting, where C5.0 acts as a rule-based filter and KNN refines uncertain predictions for a balanced, flexible workflow shown in Figure 1.

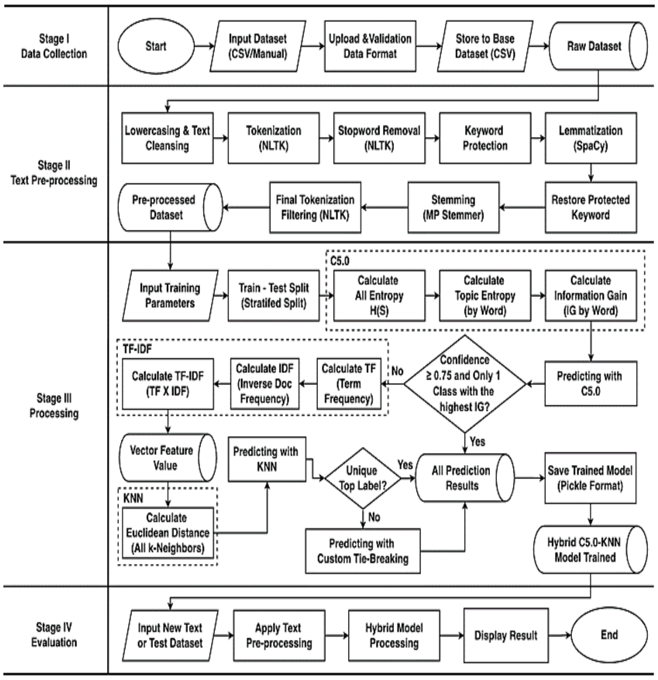

Stage I
Data Collection
Start
Input Dataset (CSV/Manual)
Upload &Validation Data Format
Store to Base Dataset (CSV)
Raw Dataset
Stage II
Text Pre-processing
Lowercasing & Text Cleansing
Tokenization (NLTK)
Stopword Removal (NLTK)
Keyword Protection
Lemmatization (SpaCy)
Restore Protected Keyword
Stemming (MP Stemmer)
Final Tokenization Filtering (NLTK)
Pre-processed Dataset
Stage III
Processing
Input Training Parameters
Train - Test Split (Stratifed Split)
C5.0
Calculate All Entropy H(S)
Calculate Topic Entropy (by Word)
Calculate Information Gain (IG by Word)
Predicting with C5.0
Confidence ≥ 0.75 and Only 1 Class with the highest IG?
No
Yes
TF-IDF
Calculate TF (Term Frequency)
Calculate IDF (Inverse Doc Frequency)
Calculate TF-IDF (TF X IDF)
Vector Feature Value
KNN
Calculate Euclidean Distance (All k-Neighbors)
Predicting with KNN
Unique Top Label?
Yes
No
Predicting with Custom Tie-Breaking
All Prediction Results
Save Trained Model (Pickle Format)
Hybrid C5.0-KNN Model Trained
Stage IV
Evaluation
Input New Text or Test Dataset
Apply Text Pre-processing
Hybrid Model Processing
Display Result
End

Figure 1. Research Framework

There are stages in implementing the hybrid C5.0-KNN model consisting of data collecting, text-preprocessing, processing, and evaluation as follows:

1) Data Collection

This stage begins with collecting and preparing the dataset needed for the classification process. The following section presents five sample news entries that will be used to illustrate how the hybrid C5.0-KNN model works. These samples are taken from the new dataset compiled during the earlier data collection stage.

TABLE 1
NEWS SAMPLES

ID	Content Snippet	Topik
1	Rupiah ditutup di level Rp16.454 per dolar AS pada Kamis (27/2) atau melemah.	Ekonomi (economy)
2	Penyebaran AI di berbagai angka dan masyarakat dengan laju yang lebih cepat.	Teknologi (technology)
3	Berikut 5 film dibintangi sang legendaris Hollywood Gene Hackman.	Hiburan (entertainment)
4	MU berhasil menang 3-1 saat bertanding melawan Chelsea di Old Trafford.	Olahraga (sport)
5	Bagi Anda yang tidak mengonsumsi susu karena intoleransi laktosa, alergi, atau pola makan tertentu, ada banyak makanan alternatif yang tetap kaya kalsium.	Gaya Hidup (life style)

2) Text Pre-processing

After the data collection, the data must go through a preprocessing stage to clean the text, normalize word forms, and minimize word redundancy – making it easire and more effective for the model to learn during training [11]. The preprocessing techniques used are [12]: 1) case folding and text cleansing, 2) tokenization using NTLK, 3) stopword removal, 4) keyword protection, 5) final tokenization filtering using NTLK, 6) stemming using MP Stemmer, 7) restore protected keywords, 8) lemmatizations using SpaCy. Once the dataset has been pre-processed, it will be then used as the input for the processing phase. These techniques are widely used for Indonesian NLP tasks and helps standardize linguistic variations while preserving semantic meaning. Also, this ensures linguistic consistency and improved term weighting during TF-IDF computation.

TABLE 2
PRE-PROCESSED NEWS SAMPLES

ID	Content Snippet	Topic
1	[‘rupiah’, ‘level’, ‘rp’, ‘dolar’, ‘as’, ‘lemah’, ‘persen’, ‘dagang’]	Ekonomi (economy)
2	[‘sebar’, ‘ai’, ‘sektor’, ‘teknologi’]	Teknologi (technology)
3	[‘film’, ‘bintang’, ‘legenda’ ‘hollywood’, ‘gene’, ‘hackman’]	Hiburan (entertainment)
4	[‘mu’, ‘hasil’, ‘menang’, ‘tanding’, ‘lawan’, ‘chelsea’, ‘old’, ‘trafford’]	Olahraga (sport)
5	[‘konsumsi’, ‘susu’, ‘intoleransi’, ‘laktosa’, ‘alergi’, ‘makan’, ‘makan’, ‘kalsium’]	Gaya Hidup (life style)

3) Processing

In this stage, C5.0 generates interpretable classification rules based on Information Gain, TF-IDF is used to assign weights to words and KNN performs instance-based classification by calculating the Euclidean distance between the TF-IDF feature vectors of the test instance and its *k* nearest neighbors.

3.1 C5.0 Algorithm

The following steps outline the implementation of the C5.0 algorithm:

a. Calculate overall entropy using the following formula [13].

$$E(S) = \sum_{I=1}^{c} - p_i \, log_2 \, p_i \qquad (1)$$

Given that the data in Table 2 consists of 1 document for each category: Ekonomi, Teknologi, Olahraga, Hiburan, and Gaya Hidup. Therefore, the initial Entropy is: H(S) = - (1/5 log$_2$ 1/5 + 1/5 log$_2$ 1/5 + 1/5 log$_2$ 1/5 + 1/5 log$_2$ 1/5 + 1/5 log$_2$ 1/5) = 2.321.

b. Calculate topic for every word entropy using the same formula as above.

After obtaining the initial entropy, the next step is to calculate the entropy after word separation. Using test data with keywords ‘sektor’, ‘teknologi’, ‘digital’, ‘kembang’, ‘pesat’, ‘rupiah’, and ‘lemah’, the word distribution can be observed in Table 3.

TABLE 3
UNIQUE WORDS DISTRIBUTION

Word	D1	D2	D3	D4	D5
Sector (sector)	0	**1**	0	0	0
Teknologi (technology)	0	**1**	0	0	0
digital	0	0	0	0	0
Kembang (growing)	0	0	0	0	0
Pesat (rapidly)	0	0	0	0	0
rupiah	**1**	0	0	0	0
Lemah (weakening)	**1**	0	0	0	0

For instance, in this case, as a calculation example, the topic entropy for the word ‘sektor’ will be computed. Based on Table 3, documents containing this word are found in 1 document categorized as Teknologi. Therefore: Entropy (‘sektor’) = - (1/1 $\log_2$ 1/1) = 0.

Meanwhile, documents without the word ‘sektor’ are present in 4 documents: 1 Ekonomi document, 1 Olahraga document, 1 Hiburan document, and 1 Gaya Hidup document. Therefore: Entropy without ‘sektor’ = - (1/4 $\log_2$ 1/4 + 1/4 $\log_2$ 1/4 + 1/4 $\log_2$ 1/4 + 1/4 $\log_2$ 1/4) = 2.

c. Calculate Information Gain using the following formula to determine the weight of each word for the purpose of the C5.0 algorithm [14].

$$Gain(S,A) = E(S) - \sum_{i=1}^{n} \frac{S_i}{S} E(S_i) \qquad (2)$$

Using the word “sektor” from Table 4 as an example: IG(‘sektor’) = 2.321 − (1/5 × 0 + 4/5 × 2) = 0.721.

Based on the data in Table 3, the Information Gain values for each word from the test data that belongs to a specific category are presented in Table 4 below.

TABLE 4
INFORMATION GAIN VALUES

Word	Information Gain	Topic
sektor	0,721	Teknologi (technology)
teknologi	0,721	Teknologi (technology)
rupiah	0,721	Ekonomi (economy)
lemah	0,721	Ekonomi (economy)

After obtaining the overall Information Gain value, if the highest values have equal Information Gain, the news text cannot be directly classified using C5.0 is passed to KNN for proximity-based evalution. This decision branch ensures reliable outcomes by accepting C5.0 predictions with confidence scores ≥ 0.75 and re-evaluating lower-confidence cases through KNN to refine ambiguous classifications.

3.2 Word Weighting with TF-IDF

Before KNN calculations are performed, the data representation of each document must be in the form of feature vectors. In this case, TF-IDF is used to calculate the feature vector values for each document as shown in [15].

$$Wdx = tfdx \times IDF \qquad (3)$$

Each word is given a numerical value by TF-IDF, indicating how relevant a word is to a certain document within a corpus. The results from TF-IDF will be utilized in the KNN algorithm's calculations to determine the proximity between test data and training data, particularly if the data cannot be classified by the C5.0 algorithm.

In this stage, the sentence “Sektor Teknologi Digital Berkembang Pesat, Meskipun Rupiah semakin Melemah.” (The Digital Technology Sector is Rapidly Growing, Despite the Rupiah’s Continued Weakening) will be used as the news data for testing and would be referred to as “DTest”. The preprocessing performed on this sentence yields 7 keywords, namely ‘sektor’ (sector), ‘teknologi’ (technology), ‘digital’, ‘kembang (growing)’, ‘pesat’ (rapidly), ‘rupiah’, and ‘lemah’ (weakining).

For TF calculation, consider the word "sektor" (sector):

- **DTest:** The word "sektor" appears 1 time out of 7 words, so TF = 1/7 = 0.142.
- **D1:** The word "sektor" does not appear, so TF = 0.
- **D2:** The word "sektor" appears 1 time out of 4 words, so TF = 1/4 = 0.25.
- **D3:** The word "sektor" does not appear, so TF = 0.
- **D4:** The word "sektor" does not appear, so TF = 0.
- **D5:** The word "sektor" does not appear, so TF = 0.

For the IDF calculation of the word "sektor": the word "sektor" appears in 2 out of 6 documents, so IDF("sektor") = $\log_2$(6/2) = 0.477.

For the TF-IDF calculation of the word "sektor":

- TF-IDF("sektor", DTest) = 0.142 x 0.477 = 0.068
- TF-IDF("sektor", D1) = 0 x 0.477 = 0
- TF-IDF("sektor", D2) = 0.25 x 0.477 = 0.119
- TF-IDF("sektor", D3) = 0 x 0.477 = 0
- TF-IDF("sektor", D4) = 0 x 0.477 = 0
- TF-IDF("sektor", D5) = 0 x 0.477 = 0

Based on these steps, we calculated the TF-IDF values for each word.

3.3 KNN Algorithm

The vector values of each document will be calculated using the Euclidean Distance formula to determine the proximity between data points, by identifying the lowest resulting value. Euclidean Distance is chosen due to its simple formula and its suitability for feature weights with varying scales. The following is the formula to calculate Euclidean Distance [16].

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (4)$$

Since C5.0 classified the sentence into only Economy and Technology, proximity calculations using Euclidean Distance formula were limited to these categories, as shown in Table 5.

TABLE 5
EUCLIDEAN DISTANCE VALUES

Document	Euclidean Distance	Topic
D1	**0,320**	Ekonomi
D2	0,355	Teknologi

To determine the closest distance, *n* is set to 1, meaning classification relies on the single nearest document-appropriate for the dataset size. As shown in Table 5, the closest match is D1, labeled Economy, so the test data is classified accordingly. If multiple documents share the same minimum Euclidean Distance, additional factors such as average TF-IDF, total TF-IDF, word frequency, or topic count can be used to refine the as presented in Table 6.

TABLE 6
FALLBACK CLASSIFICATION ATRIBUTES

Topic	TF-IDF Average	TF-IDF Total	Word Count	Document Count
Ekonomi (economy)	0,087	0,7	**8**	1
Teknologi (technology)	**0,156**	0,626	4	1
Olahraga (sport)	0,097	**0,776**	**8**	1
Hiburan (entertainment)	0,129	0,774	6	1
Gaya Hidup (life style)	0,11	**0,776**	7	1

Therefore, if there are equal lowest Euclidean Distance values, classification can be performed based on the highest values in the columns of Table 6.

4) Evaluation

This stage begins with collecting and preparing the dataset needed for the classification process. The following section presents five sample news entries that will be used to illustrate how the hybrid C5.0-KNN model works. These samples are taken from the new dataset compiled during the earlier data collection stage.

TABLE 7
CONFUSION MATRIX

Category		Predicted				
		E	GH	H	O	T
Actual	Ekonomi	1	0	0	0	0
	Gaya Hidup	0	1	0	0	0
	Hiburan	0	0	1	0	0
	Olahraga	0	0	0	1	0
	Teknologi	0	0	0	0	1

The confusion matrix results in Table 7 are used to calculate model performance metrics - Accuracy, Precision, Recall, and F1-Score – with the evaluation results summarized in Table 8 [17].

TABLE 8
MODEL PERFORMANCE REPORT

Topic	Precission	Recall	F1-Score
Ekonomi	100%	100%	100%
Teknologi	100%	100%	100%
Olahraga	100%	100%	100%
Hiburan	100%	100%	100%
Gaya Hidup	100%	100%	100%

C. System Design

This section will depict several designs for this news classification system to demonstrate how the system will work.

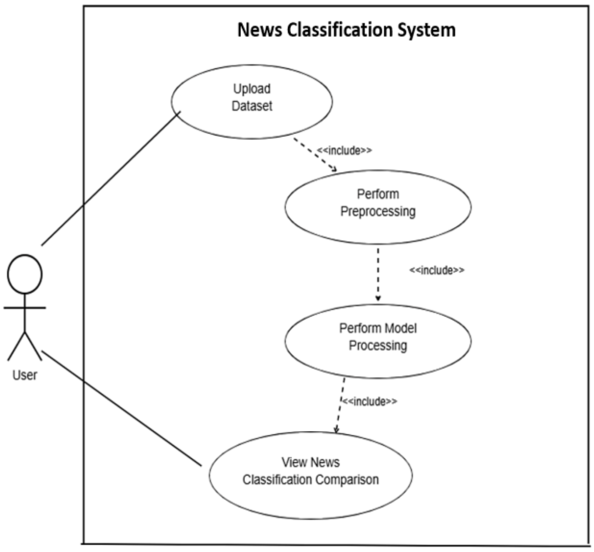


Figure 2. Use Case Diagram

The user initiates the news classification process by uploading a dataset to the system. The system then automatically performs preprocessing on the uploaded data, followed by model processing to analyze and classify the news content. Finally, the user can view a comparison of the news classification results generated by the model.

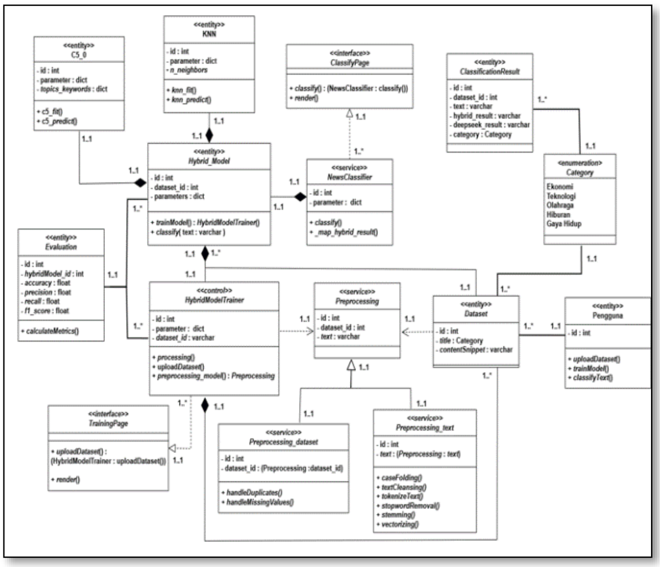


Figure 3. Class Diagram

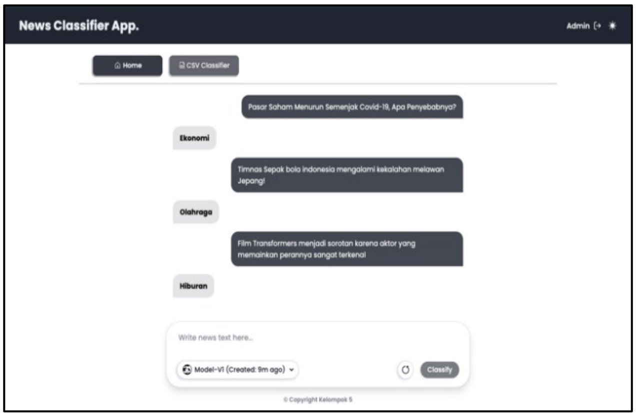
Figure 3 presents the system architecture through a Class Diagram [18], consisting of 16 interconnected classes. The NewsClassifier class depends on the Hybrid_Model class, which in turn relies on C5.0 and KNN classes. A Preprocessing class handles both dataset and text preprocessing, while the User class enables actions such as uploading datasets, training models, and performing classification. Together, these diagrams demonstrate the structure and interactions within the website.

III. Results and Discussion

The primary objective of this research is to develop and implement a novel hybrid model that effectively combines the strengths of the C5.0 decision tree and KNN algorithms. This integrated approach is intended to enhance the accuracy and efficiency of the news topic classification while addressing the individual limitations of each algorithm. In addition to model development, this research also focuses on evaluating its performance through standard metrics - accuracy, precision, recall, and F1-score – to validate its capability in delivering reliable and effective classification outcomes.

A. System Interfaces

This section presents several user interfaces developed for the system.



Pasar Saham Menurun Semenjak Covid-19, Apa Penyebabnya?

Ekonomi

Timnas Sepak bola indonesia mengalami kekalahan melawan Jepang!

Olahraga

Film Transformers menjadi sorotan karena aktor yang memainkan perannya sangat terkenal

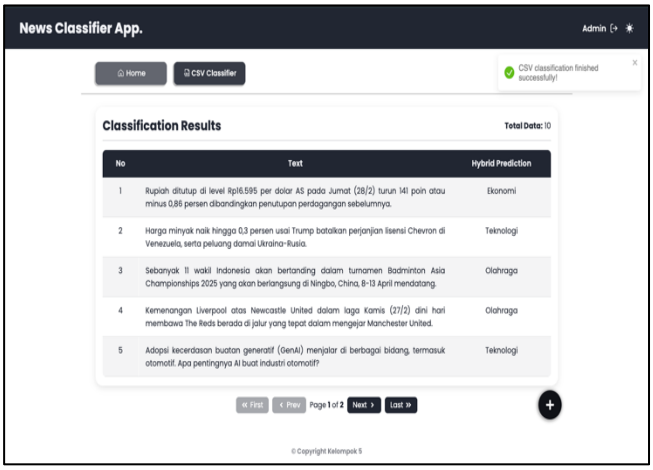
Hiburan

Write news text here...

© Copyright Kelompok 5

Figure 4.Text Classifier

Figure 4 shows the User Homepage Classifier interface after classification, with results displayed in a chat-style format – user input appears on the right, and the Hybrid model's classification appears on the left. New inputs can be entered in the text box at the bottom, with the latest results appearing below.



CSV classification finished successfully!

Classification Results

Total Data: 10

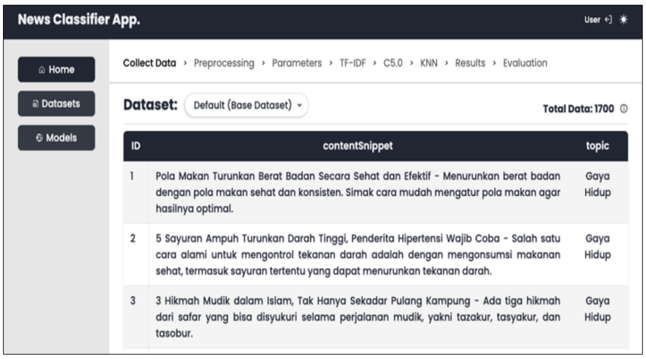
No	Text	Hybrid Prediction
1	Rupiah ditutup di level Rp16.595 per dolar AS pada Jumat (28/2) turun 141 poin atau minus 0,86 persen dibandingkan penutupan perdagangan sebelumnya.	Ekonomi
2	Harga minyak naik hingga 0,3 persen usai Trump batalkan perjanjian lisensi Chevron di Venezuela, serta peluang damai Ukraina-Rusia.	Teknologi
3	Sebanyak 11 wakil Indonesia akan bertanding dalam turnamen Badminton Asia Championships 2025 yang akan berlangsung di Ningbo, China, 8-13 April mendatang.	Olahraga
4	Kemenangan Liverpool atas Newcastle United dalam laga Kamis (27/2) dini hari membawa The Reds berada di jalur yang tepat dalam mengejar Manchester United.	Olahraga
5	Adopsi kecerdasan buatan generatif (GenAI) menjalar di berbagai bidang, termasuk otomotif. Apa pentingnya AI buat industri otomotif?	Teknologi

Page 1 of 2

© Copyright Kelompok 5

Figure 5. CSV Classifier

Figure 5 presents the CSV classifier results page, where classification outcomes are shown in a table format with five entries per page.



News Classifier App.
User
Home
Datasets
Models
Collect Data › Preprocessing › Parameters › TF-IDF › C5.0 › KNN › Results › Evaluation

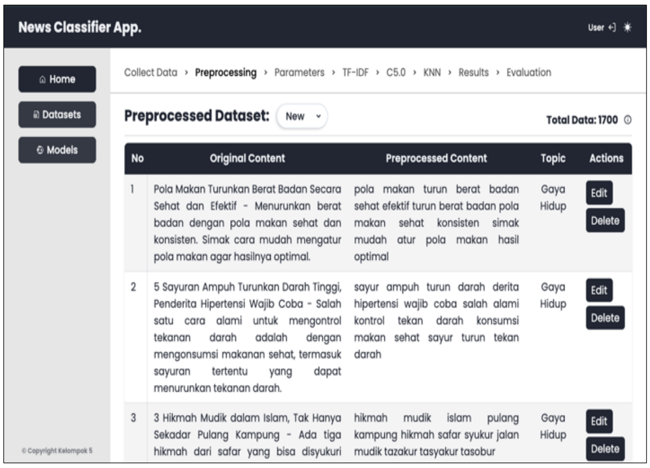
Dataset: Default (Base Dataset)

Total Data: 1700

ID	contentSnippet	topic
1	Pola Makan Turunkan Berat Badan Secara Sehat dan Efektif - Menurunkan berat badan dengan pola makan sehat dan konsisten. Simak cara mudah mengatur pola makan agar hasilnya optimal.	Gaya Hidup
2	5 Sayuran Ampuh Turunkan Darah Tinggi, Penderita Hipertensi Wajib Coba - Salah satu cara alami untuk mengontrol tekanan darah adalah dengan mengonsumsi makanan sehat, termasuk sayuran tertentu yang dapat menurunkan tekanan darah.	Gaya Hidup
3	3 Hikmah Mudik dalam Islam, Tak Hanya Sekadar Pulang Kampung - Ada tiga hikmah dari safar yang bisa disyukuri selama perjalanan mudik, yakni tazakur, tasyakur, dan tasobur.	Gaya Hidup

Figure 6. Data Collection

Figure 6 displays the Data Collection page after a dataset is uploaded or selected, showing the data in a table format with Content Snippets and manually labeled topics. The top-right corner indicates the total number of entries and provides a tooltip with table details.


News Classifier App.
User
Home
Datasets
Models
Collect Data › Preprocessing › Parameters › TF-IDF › C5.0 › KNN › Results › Evaluation
Edit
Delete
© Copyright Kelompok 5

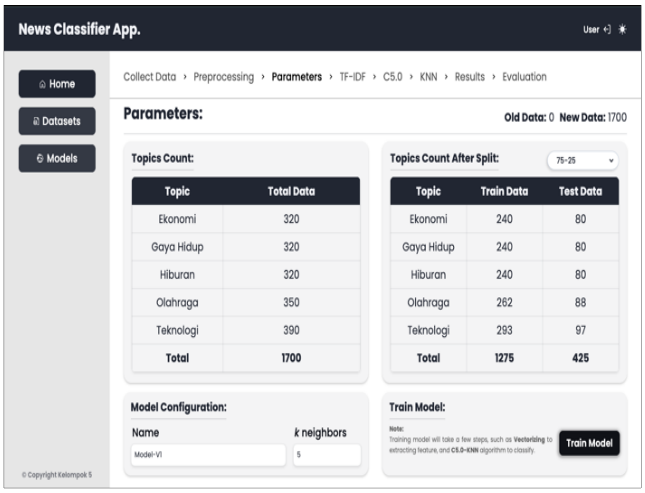
Preprocessed Dataset: New

Total Data: 1700

No	Original Content	Preprocessed Content	Topic	Actions
1	Pola Makan Turunkan Berat Badan Secara Sehat dan Efektif - Menurunkan berat badan dengan pola makan sehat dan konsisten. Simak cara mudah mengatur pola makan agar hasilnya optimal.	pola makan turun berat badan sehat efektif turun berat badan pola makan sehat konsisten simak mudah atur pola makan hasil optimal	Gaya Hidup	Edit Delete
2	5 Sayuran Ampuh Turunkan Darah Tinggi, Penderita Hipertensi Wajib Coba - Salah satu cara alami untuk mengontrol tekanan darah adalah dengan mengonsumsi makanan sehat, termasuk sayuran tertentu yang dapat menurunkan tekanan darah.	sayur ampuh turun darah derita hipertensi wajib coba salah alami kontrol tekan darah konsumsi makan sehat sayur turun tekan darah	Gaya Hidup	Edit Delete
3	3 Hikmah Mudik dalam Islam, Tak Hanya Sekadar Pulang Kampung - Ada tiga hikmah dari safar yang bisa disyukuri	hikmah mudik islam pulang kampung hikmah safar syukur jalan mudik tazakur tasyakur tasobur	Gaya Hidup	Edit Delete

Figure 7. Preprocessing

Figure 7 presents the Preprocessing page, displaying both the originial and pre-processed versions of each Content Snippet.


News Classifier App.
User
Home
Datasets
Models
Collect Data › Preprocessing › Parameters › TF-IDF › C5.0 › KNN › Results › Evaluation
Train Model
© Copyright Kelompok 5

Parameters:

Old Data: 0 New Data: 1700

Topics Count:

Topic	Total Data
Ekonomi	320
Gaya Hidup	320
Hiburan	320
Olahraga	350
Teknologi	390
Total	**1700**

Topics Count After Split: 75-25

Topic	Train Data	Test Data
Ekonomi	240	80
Gaya Hidup	240	80
Hiburan	240	80
Olahraga	262	88
Teknologi	293	97
Total	**1275**	**425**

Model Configuration:

Name: Model-V1

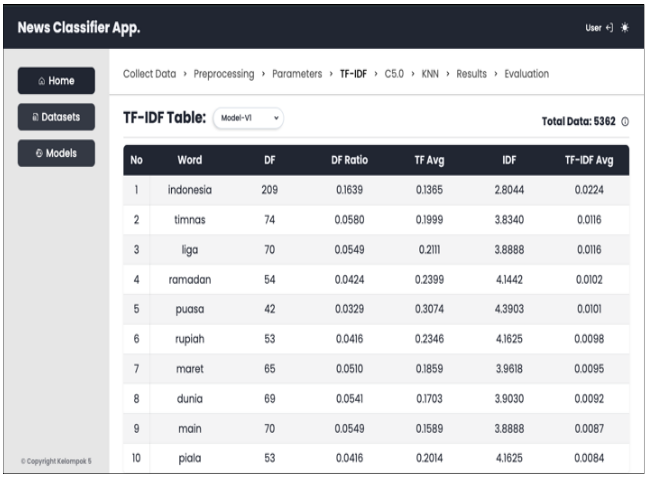
k neighbors: 5

Train Model:

Note:
Training model will take a few steps, such as **Vectorizing** to extracting feature, and **C5.0-KNN** algorithm to classify.

Figure 8. Parameters

Figure 8 shows the Parameters page, where users define training settings, including the number KNN and the train-test split percentage, and total data entries per topic.


News Classifier App.
User
Home
Datasets
Models
Collect Data › Preprocessing › Parameters › TF-IDF › C5.0 › KNN › Results › Evaluation
© Copyright Kelompok 5

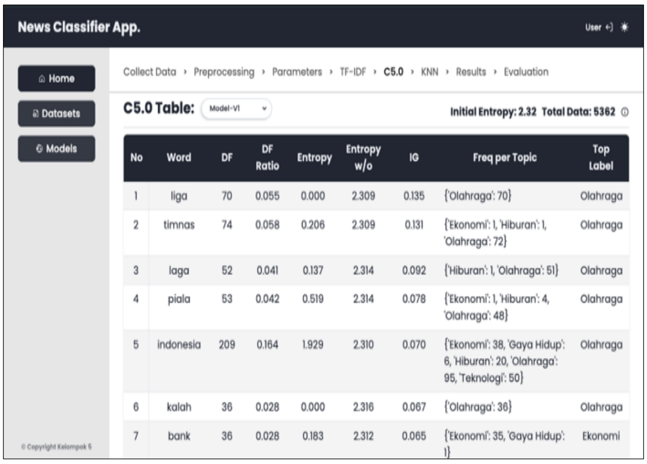
TF-IDF Table: Model-V1

Total Data: 5362

No	Word	DF	DF Ratio	TF Avg	IDF	TF-IDF Avg
1	indonesia	209	0.1639	0.1365	2.8044	0.0224
2	timnas	74	0.0580	0.1999	3.8340	0.0116
3	liga	70	0.0549	0.2111	3.8888	0.0116
4	ramadan	54	0.0424	0.2399	4.1442	0.0102
5	puasa	42	0.0329	0.3074	4.3903	0.0101
6	rupiah	53	0.0416	0.2346	4.1625	0.0098
7	maret	65	0.0510	0.1859	3.9618	0.0095
8	dunia	69	0.0541	0.1703	3.9030	0.0092
9	main	70	0.0549	0.1589	3.8888	0.0087
10	piala	53	0.0416	0.2014	4.1625	0.0084

Figure 9. TF-IDF values

Figure 9 displays the TF-IDF page, presenting word-level TF-IDF values in a table sorted by average score, showing 10 entries per page. The top-right corner includes the total TF-IDF count and a tooltip with table information.


News Classifier App.
User
Home
Datasets
Models
Collect Data › Preprocessing › Parameters › TF-IDF › C5.0 › KNN › Results › Evaluation
© Copyright Kelompok 5

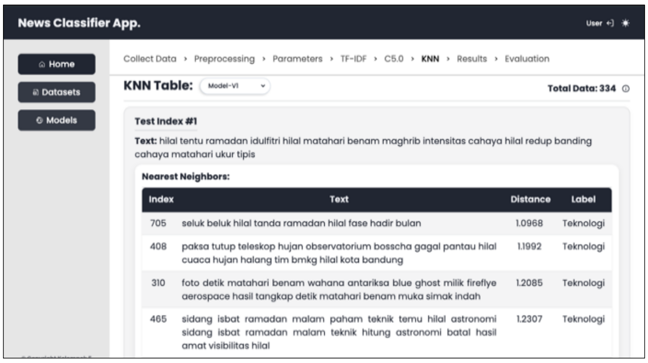
C5.0 Table: Model-V1

Initial Entropy: 2.32 Total Data: 5362

No	Word	DF	DF Ratio	Entropy	Entropy w/o	IG	Freq per Topic	Top Label
1	liga	70	0.055	0.000	2.309	0.135	{'Olahraga': 70}	Olahraga
2	timnas	74	0.058	0.206	2.309	0.131	{'Ekonomi': 1, 'Hiburan': 1, 'Olahraga': 72}	Olahraga
3	laga	52	0.041	0.137	2.314	0.092	{'Hiburan': 1, 'Olahraga': 51}	Olahraga
4	piala	53	0.042	0.519	2.314	0.078	{'Ekonomi': 1, 'Hiburan': 4, 'Olahraga': 48}	Olahraga
5	indonesia	209	0.164	1.929	2.310	0.070	{'Ekonomi': 38, 'Gaya Hidup': 6, 'Hiburan': 20, 'Olahraga': 95, 'Teknologi': 50}	Olahraga
6	kalah	36	0.028	0.000	2.316	0.067	{'Olahraga': 36}	Olahraga
7	bank	36	0.028	0.183	2.312	0.065	{'Ekonomi': 35, 'Gaya Hidup': 1}	Ekonomi

Figure 10. C5.0

Figure 10 shows the C5.0 results page with 10 entries sorted by Information Gain, along with entropy, data count, and table tooltips.


News Classifier App.
User
Home
Datasets
Models
Collect Data › Preprocessing › Parameters › TF-IDF › C5.0 › KNN › Results › Evaluation

KNN Table: Model-V1

Total Data: 334

Test Index #1

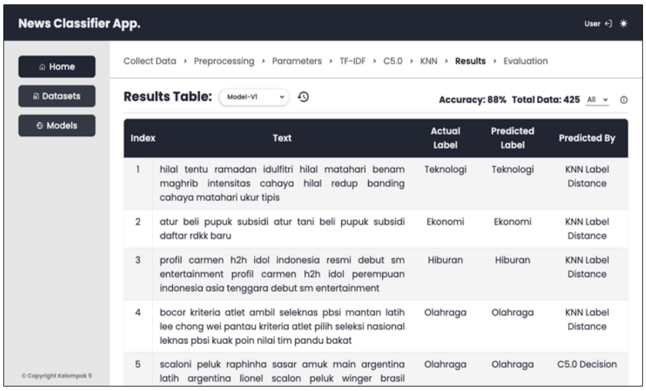
Text: hilal tentu ramadan idulfitri hilal matahari benam maghrib intensitas cahaya hilal redup banding cahaya matahari ukur tipis

Nearest Neighbors:

Index	Text	Distance	Label
705	seluk beluk hilal tanda ramadan hilal fase hadir bulan	1.0968	Teknologi
408	paksa tutup teleskop hujan observatorium bosscha gagal pantau hilal cuaca hujan halang tim bmkg hilal kota bandung	1.1992	Teknologi
310	foto detik matahari benam wahana antariksa blue ghost milik fireflye aerospace hasil tangkap detik matahari benam muka simak indah	1.2085	Teknologi
465	sidang isbat ramadan malam paham teknik temu hilal astronomi sidang isbat ramadan malam teknik hitung astronomi batal hasil amat visibilitas hilal	1.2307	Teknologi

Figure 11. KNN

Figure 11 displays the KNN results page, listing unclassified texts (Test Index) and teir closest matches based on the selectes *k.* The top-right shows the unclassified count and table tooltips.

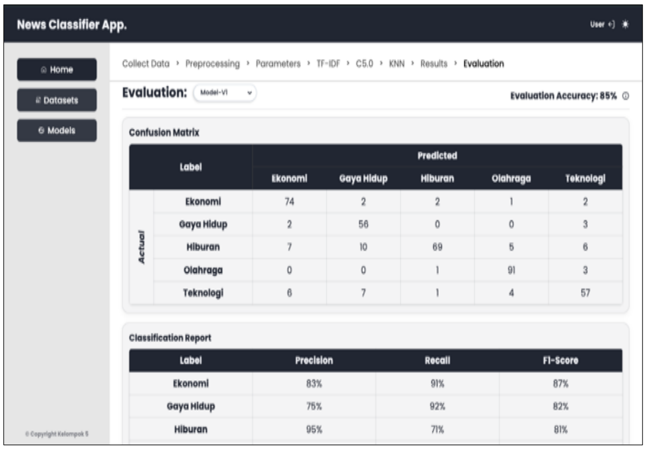

News Classifier App.
User
Home
Datasets
Models
Collect Data › Preprocessing › Parameters › TF-IDF › C5.0 › KNN › Results › Evaluation
Model-VI
All
© Copyright Kelompok 5

Results Table: Accuracy: 88% Total Data: 425

Index	Text	Actual Label	Predicted Label	Predicted By
1	hilal tentu ramadan idulfitri hilal matahari benam maghrib intensitas cahaya hilal redup banding cahaya matahari ukur tipis	Teknologi	Teknologi	KNN Label Distance
2	atur beli pupuk subsidi atur tani beli pupuk subsidi daftar rdkk baru	Ekonomi	Ekonomi	KNN Label Distance
3	profil carmen h2h idol indonesia resmi debut sm entertainment profil carmen h2h idol perempuan indonesia asia tenggara debut sm entertainment	Hiburan	Hiburan	KNN Label Distance
4	bocor kriteria atlet ambil seleknas pbsi mantan latih lee chong wei pantau kriteria atlet pilih seleksi nasional leknas pbsi kuak poin nilai tim pandu bakat	Olahraga	Olahraga	KNN Label Distance
5	scaloni peluk raphinha sasar amuk main argentina latih argentina lionel scalon peluk winger brasil	Olahraga	Olahraga	C5.0 Decision

Figure 12. Test Results

Figure 12 shows the overall classification results table, including actual and predicted labels along with the algorithms used.


News Classifier App.
User
Home
Datasets
Models
Collect Data › Preprocessing › Parameters › TF-IDF › C5.0 › KNN › Results › Evaluation
Model-VI
© Copyright Kelompok 5

Evaluation: Evaluation Accuracy: 85%

Confusion Matrix

	Label	Predicted				
		Ekonomi	Gaya Hidup	Hiburan	Olahraga	Teknologi
Actual	Ekonomi	74	2	2	1	2
	Gaya Hidup	2	56	0	0	3
	Hiburan	7	10	69	5	6
	Olahraga	0	0	1	91	3
	Teknologi	6	7	1	4	57

Classification Report

Label	Precision	Recall	F1-Score
Ekonomi	83%	91%	87%
Gaya Hidup	75%	92%	82%
Hiburan	95%	71%	81%

Figure 13. Evaluation metrics

Figure 13 presents the evaluation page with a Confusion Matrix and Classication Report, and the model's accuracy score shown at the top-right.

B. Testing

There are numerous testing done on the model performance, including 1) using Confusion matrix to calculate accuracy, precision, recall, and F1-score, and 2) statistical significance testing using McNemar's test for model superiority, paired t-Test for comparing accuracy, and mean accuracy and standard deviation across folds.

1) Confusion Matrix

In this part, several tests were conducted to measure and compare model performance:

a. C5.0 algorithm metric – evaluates model accuracy, test size, and training duration.
b. KNN algorithm metric – measures accuracy across varying test sizes to determine the optimal K value.
c. Data Split Ratio vs. Highest Accuracy of C5.0–KNN – to determine the best ratio data split for the hybrid model.
d. Hybrid C5.0-KNN model metric – assess accuracy under different test sizes to identify the best-performing K value for the hybrid model.
e. Best model Confusion Matrix – visualizes the classification outcomes by comparing actual and predicted labels for each news category.
f. Precision, recall, and F1-score per news topic – provides detailed performance metrics for each topic, highlighting the model's balance between accuracy and reliability.

The evaluations conducted with varying test sizes and numbers of neighbors yielded as shown in Table 9.

TABLE 9
C5.0 ALGORITHM METRIC

Rank	Accuracy	Test Size	Train Duration (s)
1.	0,7426	0,4	2,89s
2.	0,7353	0,2	4,72s
3.	0,7341	0,25	4,44s
4.	0,7224	0,5	2,36s
5.	0,7098	0,3	4,07s

Table 9 shows the highest classification accuracy of 0,7426 a 0,4 test sized and completed training in 2,89 seconds.

TABLE 10
KNN ALGORITHM METRIC

Rank	Accuracy	Test Size	N Neighbors
1.	0,8735	0,2	5
2.	0,8659	0,25	5
3.	0,8618	0,2	11
4.	0,8612	0,25	11
5.	0,8569	0,3	5

Table 10 shows the highest classification accuracy of **0,8735** using a 0,2 test size and KNN value of 5.

TABLE 11
DATA SPLIT RATIO VS. HIGHEST ACCURACY OF C5.0–KNN

Test Size	Highest Accuracy	N Neighbors (K)
0.20	0.8765	5
0.25	**0.8847**	5
0.30	0.8471	7
0.35	0.8647	11
0.40	0.8547	5

Table 11 presents the highest accuracy achieved by the hybrid C5.0-KNN model across different data split ratios. The best performance occurred at a 0.25 test size, yielding the best accuracy of 0.8847 with K=5.

TABLE 12
THE HYBRID C5.0-KNN MODEL METRIC

Rank	Accuracy	Test Size	N Neighbors
1.	0,8847	0,25	5
2.	0,8765	0,2	5
3.	0,8729	0,25	11
4.	0,8706	0,2	9
5.	0,8706	0,25	9

Table 12 shows 5-fold cross-validation and the results indicate that the highest classification accuracy of **0,8847** a 0,25 test size and KNN value of 5.

TABLE 13
BEST MODEL CONFUSION MATRIX

Category		Predicted				
		E	GH	H	O	T
Actual	Ekonomi	70	3	2	2	3
	Gaya Hidup	6	66	2	0	6
	Hiburan	1	3	68	5	3
	Olahraga	0	0	0	88	0
	Teknologi	6	2	1	4	84

In Table 13, the Entertainment category shows a high number of true positives with only few misclassifications to other categories, indicating excellent performance for this class. While, categories like Economy exhibit some degree of misclassification into other topics, suggesting areas where the model's discriminative power could be further enhanced.

TABLE 14
PRECISION, RECALL, AND F1-SCORE (%) PER NEWS TOPIC

Topic	Precission	Recall	F1-Score
Ekonomi	84%	88%	86%
Gaya Hidup	89%	82%	86%
Hiburan	93%	85%	89%
Olahraga	89%	100%	94%
Teknologi	88%	87%	87%

In Table 14, the Olahraga category achieved outstanding performance with a perfect recall of 1.00, meaning all actual Sport news were correctly classified, along with a strong precision of 0.89. The Entertainment category also performed well, with a high F1-score of 0.89 and precision of 0.93. In contrast, categories like Economy and Entertainment showed solid F1-socres but revealed some trade-offs between precision and recall, indicating occasional misclassification in these topics.

2) Statistical Significance Testing

a. McNemar’s Test for Model Superiority
This McNemar’s test was applied to evaluate whether the difference between the hybrid model and KNN classifier is statistically significant. The test yielded a chi-square statistic of 0.837 with p-value=0.3602. Since $p<0.05$, the difference is statistically significant, confirming that the hybrid model’s higher accuracy is not due to random chance.

b. Paired t-Test
A paired t-test comparing accuracy between five experimental runs between the hybrid accuracy (0.8759) exceeds that of KNN (0.8639), and p-value<0.05 indicates that the performance improvement is statistically significant. Therefore, the hybrid C5.0-KNN model achieves superior generalization performance with consistent gains across test folds.

c. Mean Accuracy and Standard Deviation Across Folds

The 5-fold cross-validation produced the following results for the hybrid C5.0-KNN model:

- Fold accuracy: 0.8847, 0.8765, 0.8729, 0.8706, 0.8706.
- Mean accuracy: 0.8751
- Standard deviation: 0.0059

These values confirm that the model maintain stable performance across folds, demonstrating its reliability and generalization strength.

C. Error Analysis

An error analysis has been performed, focusing on semantically similarity pairs (‘Entertainment’ vs ‘Lifestyle’)/ The analysis revealed that these two categories were most frequently misclassified, primarily due to overlapping semantic contexts. Articles focusing on topics such as celebrity fitness, fashion trends, or wellness-themed shows often contained linguistic elements that blurred the distinction between these two categories.

The misclassifications highlight several key limitations of the current model:

1) Semantic ambiguity: the contextual boundaries between semantically related topics (e.g. lifestyle vs. entertainment) are often blurred, causing confusion when both categories share similar vocabulary and narrative tone.
2) Lexical redundancy: the reliance on TF-IDF weighting cuases the model to treat words independently, thereby overlooking latent semantic relationships between terms.
3) Limited context modeling: the hybrid framework does not account for word order or long-range dependencies, which reduces its discriminative capability when interpreting complex sentences or nuanced phrasing.

To address these challenges, future research should incorporate contextual embedding techniques such as IndoBERT or fastText to capture deeper semantic meaning and contextual nuance. Additionally, expanding dataset diversity and exploring transformer-based hybrid architectures may enhance both interpretability and contextual understanding, leading to more robust performance across semantically overlapping categories.

IV. Conclusion

This research successfully developed a hybrid C5.0-KNN model to improve Indonesian news topic classification by combining the strengths of both algorithms. C5.0 performs the initial classified, while KNN refines results based on data proximity. The model achieved a peak accuracy of 0.8847, effectively classifying news into categories such as economy, technology, sport, entertainment, and life style. The hybrid model achieved a highest accuracy of 0.8847 (using 25% test data and k=5), outperforming standalone C5.0 (0.7426) and KNN (0.8735). Notably, it attained 100% recall for “sport/olahraga” and an F1-score of 0.89 for “entertainment/hiburan”. These results demonstrate the model’s novelty, efficiency, and strong potential for real-world news classification in low-resource language contexts, offering practical value for journalists, analysts, and media monitoring systems. These findings also indicate that the hybrid C5.0-KNN approach delivers a more accurate and efficient solution compared to using each algorithm independently. Its practical value lies in supporting automated classification systems for use in news aggregators, sentiment analysis, or recommendation engines.

This research offers both theoretical and practical contributions to the field of hybrid classification and explanaible artificial intelligence (XAI). From a thereotical standpoint, the hybrid C5.0-KNN model exemplifies the principles of ensemble learning and meta-classification. The rule-based c5.0 algorithm captures global decision patterns that enhance interpretability, while the instance-based KNN refines ambiguous cases at the local level to improve adapatability. This cooperative two-stage process reflects the ensemble paradigm – combining complementary classifiers to enhance robustness and reduce generalization error. The observed misclassifications highlight both the strengths and limitations of such hybrid approaches, suggesting that performance in semantically overlapping domains could be further improved through contextual embeddings or transformer-based representations.

This integration not only strengthens the theoretical foundation of hybrid modeling but also extends the principles of explainable AI to low-resource language contexts, demonstrating that transparent yet high-performing models can be developed beyond English-dominant datasets. Furthermore, the study’s web-based implementation showcases how hybrid intelligence can be operationalized in real-time, providing a practical demonstration of how interpretable and adaptive Ai systems can support multilingual text analytics and applied informatics research.

Future research should enhance the hybrid C5.0-KNN framework by integrating contextual embedding such as IndoBERT or fastText to capture deeper semantic meaning in Indonesian text. Expanding dataset diversity across domains and dialexts is also recommended to improve generalization. Applying cross-validation and C5.0 pruning can strengthen model reliability, interpretability, and efficiency. Further exploration of transformer-based hybrid architectures, dimensionality reduction methods (e.g. PCA, UMAP), and adaptive distance metrics (e.g. cosine, Mahalabonis) may yield higher accuracy and stability. Finally, developing a real-time web-based classification system would enhance the model’s practical value for automated news and sentiment analysis.

Acknowledgement

We sincerely thank Universitas Katolik De La Salle Manado for their support and commitment throughout this research. Their encouragement, resources, and collaborative spirit have been instrumental in the successful completion of this research.

REFERENCES

[1] S. M. Habib, E. Haerani, S. K. Gusti and S. Ramadhani, “Klasifikasi Berita Menggunakan Metode Naïve Bayes Classifier,” *Jurnal Nasional Komputasi dan Teknologi Informasi,* vol. 5, no. 2, pp. 248-258, 2022.

[2] V. Manurung and A. F. Rozi, “Analisis Perbandingan Algoritma K-Nearest Neighbor dan Decision Tree pada Klasifikasi Tingkat Stress Individu,” *TIN: Terapan Informatika Nusantara,* vol. 5, no. 1, pp. 73-80, 2024.

[3] M. Irfan, W. U. Dewi, K. Nisa and M. Usman, “Implementasi K-Nearest Neighbors, Decision Tree dan Support Vector Machine Pada Data Diabetes,” *JMIK (Jurnal Mahasiswa Ilmu Komputer),* vol. 4, no. 2, pp. 137-150, 2023.

[4] E. B. Santoso, Y. H. Chrisnanto and G. Abdillah, “Identification of Hoax News Using TF-RF and C5.0 Decision-Tree Algorithm,” *Enrichment: Journal of Multidisciplinary Research and Development,* vol. 1, no. 6, pp. 336-348, 2023.

[5] A. Ihsan and E. Rainarli, “ Optimization of K-Nearest Neighbour to Categorize Indonesian’s News Articles,” *Asia-Pacific Journal of Information Technology and Multimedia,* vol. 10, no. 1, pp. 43-51, 2021.

[6] M. R. Amartha, R. Wahyuni and Y. Irawan, “Optimasi Algoritma C5.0 dengan Teknik Ensemble Boosting untuk Peningkatan Akurasi dalam Klasifikasi Ulasan Masyarakat Terhadap Layanan BPJS Kesehatan,” *JEKIN (Jurnal Teknik Informatika),* vol. 5, no. 1, pp. 100-110, 2025.

[7] A. N. Kasanah, M. and U. Pujianto, “Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi),* vol. 3, no. 2, pp. 196-201, 2019.

[8] N. Islam, F. T. Jahra, M. T. Hasan and D. M. Farid, “KNN Tree: A New Method to Ameliorate K-Nearest Neighbour Classification Using Decision Tree,” in *In Proc. 2023 Int. Conf. on Electrical, Computer and Communication Engineering (ECCE)*, Chittagong, Bangladesh, 2023.

[9] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham and K. Akinwolere, “Text Classification: How Machine Learning Is Revolutionizing Text Categorization,” *Information,* vol. 16, no. 2, p. 130, 2025.

[10] Z. Mohammadi-Pirouz, K. Hajian-Tilaki, M. S. Haddat-Zavareh, A. Amoozadeh and S. Bahrami, “Development of Decision-Tree Classification Algorithms in Predicting Mortality of COVID-19 Patients,” *International Journal of Emergency Medicine,* vol. 17, p. 126, 2024.

[11] Y. Wulandari, E. Haerani and S. K. Gusti, “Klasifikasi Berita Menggunakan Algoritma C4.5,” *Jurnal Nasional Komputasi dan Teknologi Informasi,* vol. 5, no. 2, pp. 279-289, 2022.

[12] D. Soyusiawaty, Buku Ajar Pemrosesan Bahasa Alami, Yogyakarta: Universitas Ahmad Dahlan, 2023.

[13] I. Lestari, D. Fitria, Syafriandi and A. Salma, “Comparison of the C5.0 Algorithm and the CART Algorithm in Stroke Classification,” *UNP Journal of Statistics and Data Science,* pp. 90-98, 2024.

[14] N. Tanjung, D. Irmayani and V. Sihombing, “Implementation of C5.0 Algorithm for Prediction of Student Learning Graduation in Computer System Architecture Subjects,” *Sinkron: Jurnal dan Penelitian Teknik Informatika,* pp. 274-280, 2022.

[15] N. D. Bagaskara, “Klasifikasi Sentimen Masyarakat Terhadap Kepolisian Negara Republik Indonesia Menggunakan Naive Bayes Classifier dan Support Vector Machine,” Surabaya, 2022.

[16] L. A. Susanto, “Komparasi Model Support Vector Machine dan K-Nearest Neighbor pada Analisis Sentimen Aplikasi Polri Super App,” *JITET (Jurnal Informatika dan Teknik Elektro Terapan),* vol. 12, no. 2, pp. 1180-1190, 2024.

[17] N. T. Ujianto, G. H. Fadilah, A. P. Fanti, A. D. Saputra and I. G. Ramadhan, “Penerapan Algoritma K-Nearest Neighbors (KNN) untuk Klasifikasi Citra Medis,” *Jurnal Penerapan Teknologi Informasi dan Komunikasi,* vol. 2, no. 2, pp. 33-43, 2023.

[18] S. W. Ramdany, S. A. Kaidar, B. Aguchino, C. A. A. Putri and R. Anggie, “Penerapan UML Class Diagram dalam Perancangan Sistem Informasi Perpustakaan Berbasis Web,” *Journal of Industrial and Engineering System (JIES),* vol. 5, no. 1, pp. 30-41, 2024.

[19] L. P. Sumirat, D. Cahyono, Y. Kristyawan and S. Kacung, Dasar-dasar Rekayasa Perangkat Lunak, Bojonegoro: Madza Media, 2023.