Online Version Available at: https://jurnal.ubd.ac.id/index.php/te

JOURNAL OF TECH-E

| 2581-1916 (ONLINE) | 2598-7585 (PRINTED) |


Article

# Enhancing Sundanese News Articles Classification: A Comparative Study of Models and Feature Extraction Techniques

*Yadhi A. Permana[1], Irwan Setiawan[2], Fitri Diani[3], Suprihanto[4]*

*[1,2,3,4] Politeknik Negeri Bandung, Department of Computer Engineering and Informatics, West Java, Indonesia*

SUBMISSION TRACK

Received: 08, 07, 2024
Final Revision: 01, 07, 2025
Available Online: 02, 03, 2025

KEYWORD

Sundanese Language Processing, Naive Bayes, Text Classification, TF-IDF, Logistic Regression

CORRESPONDENCE

E-mail:
irwan@jtk.polban.ac.id

**A B S T R A C T**

This paper presents a comprehensive investigation into the classification of Sundanese news articles, focusing on the evaluation of various classification models and feature extraction methods. Using a dataset obtained from Sundanese news websites, this study conducts a systematic comparison of Naive Bayes and Logistic Regression classifiers combined with TF-IDF and Bag-of-Words feature extraction methods. The research process involves critical steps such as data preprocessing, model training, hyperparameter optimization, and performance assessment based on standard metrics, including accuracy, precision, recall, and F1-score. Results demonstrate high accuracy across all combinations, with the Logistic Regression model using Bag-of-Words feature extraction achieving the highest accuracy of 98.20%. Beyond model evaluation, the research delves into qualitative data analysis. Word clouds and TF-IDF weighting are employed to uncover prominent themes and topics within the news articles, highlighting recurring patterns in the Sundanese language. The study identifies key challenges, including the scarcity of annotated datasets for low-resource languages like Sundanese and the limitations of traditional models in capturing complex linguistic structures. Future opportunities are highlighted, such as leveraging deep learning models, including transformers, to enhance classification performance and address current limitations. Additionally, ensemble methods and domain-specific adaptations could further improve accuracy. Overall, this research contributes to advancing Sundanese language processing and provides a roadmap for future innovations in text classification and natural language processing applications.


DOI: 10.31253/te.v8i2.3212

## I. INTRODUCTION

In the era of information overload, efficient organization and categorization of textual data have become imperative for various applications, ranging from information retrieval to sentiment analysis. Text classification, a core aspect of natural language processing (NLP), is pivotal in automatically assigning predefined tags or categories to textual data based on its content. One significant application of text classification is in news articles, which aids in content recommendation, topic analysis, and information filtering.

Text classification, a fundamental task in NLP, has garnered significant attention in both academic research and industrial applications. Numerous studies have explored various techniques and algorithms for text classification across different languages and domains. However, research focusing on text classification in regional languages, such as Sundanese, still needs to be completed.

Given Indonesia's linguistic diversity, efforts to develop NLP tools and resources for regional languages have gained momentum in recent years. As one of the major regional languages spoken by millions of people, Sundanese presents unique challenges and opportunities for NLP research. This study contributes to the growing research in Sundanese language processing by exploring text classification in Sundanese news articles. It lays the groundwork for future investigations in this domain.

This paper addresses this gap by presenting a comprehensive study on Sundanese news article classification. The primary objective is to explore and compare different machine learning models and feature extraction techniques for accurately categorizing Sundanese news articles into predefined topics or classes. To this end, we utilize a dataset from www.wasunda.com, a prominent Sundanese news and media platform.

To systematically examine the efficacy of various models and feature extraction techniques for Sundanese news item classification, we created a comparative study with four unique combinations. First, we used the TF-IDF representation with the Naive Bayes classifier. A popular method for extracting text features is TF-IDF, which determines a term's significance in a document concerning a corpus of documents. The Bayes theorem-based probabilistic classifier Naive Bayes is well-known for its ease of use and effectiveness in text classification problems.

We applied the Bag-of-Words representation to the Naive Bayes classifier in the second combination. Using Bag-of-Words, documents are represented as vectors, where each dimension represents a distinct word from the lexicon, and the value of each dimension shows how frequently the word occurs in the document. Despite its simplicity, Bag-of-Words has been commonly utilized in text categorization problems because it can capture the overall distribution of words in a document.

We used the TF-IDF representation with the Logistic Regression classifier for the third combination. A linear classification model called logistic regression calculates the likelihood that a sample will fall into a specific class. We combined Logistic Regression with TF-IDF to assess the effectiveness of an alternative classification algorithm while maintaining the benefits of TF-IDF in capturing the significance of terms.

Finally, we investigated how well the Bag-of-Words representation worked with the Logistic Regression classifier. This combination allowed us to evaluate how the feature extraction method selection affects the classification outcomes when used with an alternative classifier. Logistic regression, which can model intricate correlations between features, is a good option for examining bag-of-words representations.

Standard measures, including accuracy, precision, recall, and F1-score, were used to analyze each combination, guaranteeing a thorough evaluation of classification performance across various models and feature extraction methods. Additionally, we used cross-validation approaches to get reliable estimates of model performance to lessen the impact of random variability in model training and evaluation.



DOI: 10.31253/te.v8i2.3212

By systematically comparing these four combinations, we aimed to identify the most effective approach for classifying Sundanese news articles. This provided valuable insights for future research and practical applications in Sundanese language processing and text classification tasks. We hypothesize that the choice of the classification model and the feature extraction method significantly impacts the classification performance, particularly in the context of Sundanese text, which may exhibit unique linguistic characteristics compared to standard Indonesian or English.

This study addresses the knowledge gap in NLP for local languages by systematically analyzing the effectiveness of widely used machine learning models and feature extraction techniques in the context of Sundanese, a low-resource language. By utilizing a real-world dataset from a Sundanese news platform, this research evaluates existing methods and highlights their adaptability to Sundanese's unique linguistic and contextual characteristics. Additionally, the comparative analysis serves as a benchmark for future studies, encouraging the development of NLP tools tailored for local languages with similar resource constraints.

Through this study, we aim to provide insights into the effectiveness of various approaches for classifying Sundanese news articles. We strive to contribute to advancing NLP research in regional languages and facilitate the development of practical applications for Sundanese-speaking communities. Moreover, the findings of this research can serve as a foundation for future investigations in text classification tasks across other regional languages in Indonesia and beyond.

## II. LITERATURES REVIEW

Several studies have investigated the effectiveness of different machine learning algorithms and feature extraction methods in the broader context of text classification. For instance, Li et al. [1] conducted a comprehensive survey of text classification algorithms, including Naive Bayes [2], Support Vector Machines (SVM) [3]–[5], and Decision Trees [6], highlighting their strengths and limitations in different scenarios. Similarly, Hartmann [7] proposed using SVM with varying kernel functions for SVM tasks, demonstrating competitive performance compared to traditional approaches.

TF-IDF (Term Frequency-Inverse Document Frequency) is a feature extraction technique in text classification tasks. Salton [8] originally introduced TF-IDF for information retrieval, which measures the importance of a term within a document relative to a corpus of documents. Subsequently, TF-IDF has been adopted in various text classification studies due to its effectiveness in capturing the discriminative features of documents.

Bag-of-words representation, another prevalent feature extraction method, represents documents as vectors of word frequencies. Despite its simplicity, Bag-of-Words has demonstrated robust performance in text classification tasks. Qader [9] and Walkowiak [10] explored the use of Bag-of-Words representations with different classification algorithms, highlighting their versatility and effectiveness across various domains.

In the specific context of regional language processing, there is a growing interest in developing NLP tools and resources for languages other than English. However, research focusing on Sundanese language processing is relatively scarce [11]–[16]. Putra et al. [13] studied Sundanese emotion analysis, employing machine learning techniques to classify Sundanese text into emotion categories.

## III. FRAMEWORK

The research on Sundanese News Articles Classification (Figure 1) begins with the crucial step of dataset collection. This involves gathering diverse and representative Sundanese news articles from various reliable sources. The dataset must cover a range of topics and writing styles to ensure the robustness and generalizability of the classification models. The collected articles are then compiled into a structured dataset, the foundation for subsequent preprocessing and analysis.



DOI: 10.31253/te.v8i2.3212

Next, the preprocessing steps are applied to clean and prepare the dataset for feature extraction. This includes removing irrelevant content such as advertisements or metadata, normalizing text by converting it to lowercase, and eliminating stop words and punctuation. Additionally, stemming or lemmatization is performed to reduce words to their root forms, enhancing the consistency of the textual data. These preprocessing steps are essential to transform the raw dataset into a more uniform and analyzable format, setting the stage for effective feature extraction.

Feature extraction entails translating the preprocessed text into numerical representations that machine learning algorithms can employ. Two approaches are used: TF-IDF (Term Frequency-Inverse Document Frequency) and Bag-of-Words. TF-IDF captures the importance of words in the context of the complete dataset, while Bag-of-Words concentrates on the presence or absence of words in each document. These features are then used to train two types of models: Naive Bayes and Logistic Regression. Combining each model with feature extraction methods generates four distinct configurations: Naive Bayes with TF-IDF, Naive Bayes with Bag-of-Words, Logistic Regression with TF-IDF, and Logistic Regression with Bag-of-Words. To identify the best strategy for categorizing Sundanese news articles, these models are trained and assessed, offering insightful information about the advantages and disadvantages of each technique.

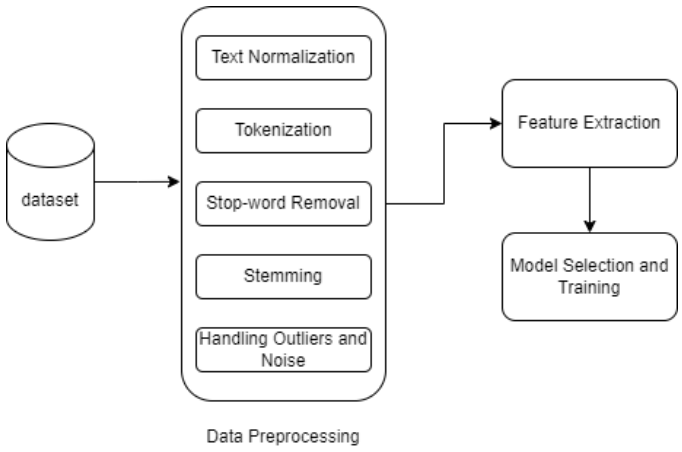


**Figure 1. Sundanese News Articles Classification Research Framework**

## IV. METHODS

The research method employed in this study aims to systematically evaluate the efficacy of various classification models and feature extraction techniques for classifying Sundanese news articles. By comparing different combinations of models and feature extraction methods, we seek to identify the most effective approach for accurately categorizing Sundanese news articles into predefined topics or classes. This section outlines the dataset collection, preprocessing steps, feature extraction, model selection and training, and evaluation metrics used to assess the performance of the classification models.

Dataset Collection and Preprocessing

The dataset utilized in this study was obtained from www.wasunda.com, a prominent online platform that aggregates Sundanese news articles and media content. Wasunda.com is a Sundanese-language news portal managed by PT Beritau Indonesia Prosperous, which was launched on July 7, 2020, in Bandung. This portal has several categories of articles, such as news, sports, culinary, tourism, and stories, thereby providing a representative sample of Sundanese textual data for classification purposes. In this study, we only took articles in the news and sports categories.

To provide a more detailed overview of the data collection process in this study, Table 1 Summarizes the key steps, methods employed, and outputs generated at each stage. It aims to help readers understand the workflow from source identification to storing a dataset ready for further analysis.



DOI: 10.31253/te.v8i2.3212

**Table 1. Data collection Steps**

| Step | Description | Tools/Methods | Output |
|---|---|---|---|
| Data Source Identification | Identify www.wasunda.com as a reliable source for Sundanese articles. | Manual Selection | A website with target categories |
| Category Selection | Select specific categories: news and sports. | Manual Filtering | Relevant URLs for scraping |
| Web Scraping | Extract articles from selected categories. | Python (BeautifulSoup) | Raw textual data |
| Preprocessing | Remove duplicates, HTML tags, and irrelevant content; ensure UTF-8 encoding. | Text Cleaning Scripts | Cleaned text files |
| Data Storage | Store the cleaned data for further analysis. | CSV/Database Storage | Structured dataset |

The raw textual data was subjected to several preparation procedures to improve its quality and eligibility for analysis before the classification task. The preprocessing methods listed below were used:

- Text Normalization [17]: Variations in word forms and spelling are common in Sundanese texts. Thus, text normalization strategies such as eliminating diacritical marks, standardizing spellings, and changing text to lowercase were used to guarantee document uniformity.
- Tokenization [18]To enable additional analysis, the text was tokenized into discrete words or tokens. Due to the language's distinctive linguistic features, tokenization techniques that consider word boundaries and compound words common in Sundanese were given particular consideration.
- Stop-word Removal [19]: To reduce noise and boost the effectiveness of the classification models, frequently used terms with less semantic meaning (such as articles and prepositions) were eliminated from the text. For this, Sundanese-specific stop-word lists were used.
- Stemming [12]: To reduce word inflections and variations, stemming or lemmatization techniques were applied to normalize words to their base forms. This helps consolidate semantically similar words and enhance the generalization capability of the classification models.
- Handling Outliers and Noise [20]: Any outliers, anomalies, or noisy data points were identified and either corrected or removed from the dataset to ensure the integrity and reliability of the analysis.

The dataset was divided into three subsets following preprocessing to simplify the classification models' training, validation, and evaluation phases. The classification models were trained using the first subset, the training set. This collection serves as the baseline data that the models use to identify trends and connections between textual characteristics and their associated labels. The training set allows the models to effectively generalize to unseen cases and capture the underlying structure of the data by exposing them to a wide variety of Sundanese news articles.

Simultaneously, the validation set, the second subset, was essential for adjusting the classification models' hyperparameters and evaluating how well they performed during training—providing an independent sample for model evaluation established safeguards against overfitting. By iteratively adjusting hyperparameters based on the validation set's performance metrics, such as accuracy or F1-score, the models can attain optimal generalization to unseen data while avoiding the dangers of excessive complexity or simplicity.

ROC-AUC is another commonly used metric for binary or multi-class classification. Still, it was not prioritized in this study because it focuses on the trade-off between true positive and false positive rates across thresholds, which is more suited for tasks emphasizing ranking or probabilistic outputs rather than hard classification decisions. As the primary goal of this study is to evaluate the ability of models to assign accurate labels to Sundanese news articles, the F1 score was chosen as



DOI: 10.31253/te.v8i2.3212

the primary evaluation metric because it provides a balanced measure of a model's precision and recall, which is particularly important in situations where the class distribution is unbalanced.

Finally, the third subset, referred to as the test set, remained untouched during the training and validation phases and served as the ultimate benchmark for evaluating the final performance of the trained models on unseen data. This set provides an unbiased assessment of the model's ability to generalize to real-world Sundanese news articles, ensuring that the reported performance metrics accurately reflect the models' predictive capabilities. Moreover, the test set allows for an objective comparison of different models and feature extraction methods, free from the influence of hyperparameter tuning or model selection biases.

Crucially, the dataset-splitting process was meticulously executed to maintain class balance across the subsets, thereby mitigating the risk of bias in model evaluation. Ensuring that each subset contains representative samples from all classes of Sundanese news articles enhances the reliability and fairness of the evaluation process, enabling robust conclusions about the relative performance of the classification models and feature extraction methods. The systematic partitioning of the dataset into training, validation, and test sets lays the foundation for rigorous experimentation and credible results in Sundanese news article classification.

Feature Extraction

Text categorization relies heavily on feature extraction, which converts unstructured textual data into a format that machine learning systems can use. This investigation used two standard feature extraction methods: TF-IDF and Bag-of-Words representation.

TF-IDF is a crucial statistical metric in natural language processing, especially for text classification. It confirms the importance of specific terms in a text concerning a larger body of documents. This metric has two essential elements that effectively capture terms' discriminative potential in textual data.

Term Frequency (TF), the first component, measures how frequently a particular phrase appears in a document. To do this, divide the total number of times a term appears in a document by the entire number of terms in that document. In essence, TF captures a term's relative importance within a particular document; higher values signify a term's greater prominence and relevance to the document's content. The foundation for further analysis and classification tasks is laid by the TF-IDF measure's ability to identify the saliency of specific phrases within each document by utilizing TF.

On the other hand, the second element, Inverse Document Frequency (IDF), enhances TF by considering how common phrases are throughout the whole document corpus. IDF measures how uncommon a term is by calculating the logarithm of the ratio between the total number of documents in the corpus and the number of documents that contain the phrase. While down-weighting terms are commonly found throughout documents, IDF gives larger weights to relatively uncommon terms throughout the corpus. IDF improves the classification models' capacity to differentiate between various document categories or themes by highlighting the discriminative strength of rare terms. This allows the TF-IDF measure to capture the uniqueness and specificity of terms inside particular documents. Combining TF and IDF, TF-IDF captures the distinctive features of each text by giving larger weights to phrases common in a given document but uncommon throughout the corpus.

The Bag-of-Words (BoW) representation, often called the "vector space model," encodes documents as vectors, with each dimension representing a unique word in the vocabulary and its value indicating the word's frequency within the document. This model ignores words' order and syntactic structure, focusing solely on their presence and frequency. This representation process unfolds through several sequential steps, each contributing to creating a comprehensive and informative representation of the textual corpus.

The process begins with Vocabulary Construction, wherein a vocabulary comprising unique words across the entire corpus is compiled. This step ensures the creation of a standardized set of



DOI: 10.31253/te.v8i2.3212

terms encompassing the linguistic diversity inherent in the corpus. Constructing a cohesive vocabulary, the Bag-of-Words representation lays the groundwork for subsequent operations, enabling the systematic analysis and processing of textual data.

Subsequently, Tokenization emerges as a pivotal step in the representation process, wherein each document undergoes segmentation into individual words or tokens. This process decomposes the textual content into its constituent elements, facilitating granular analysis and manipulation of language constructs. Tokenization delineates the boundaries between words, punctuation, and other linguistic units, thereby enabling precise counting and characterizing textual features.

Following Tokenization, the Counting phase ensues, wherein the frequency of each word within the constructed vocabulary is tallied for every document in the corpus. This step involves enumerating the occurrences of each term within individual documents, thereby quantifying the prominence and prevalence of specific words across the corpus. By capturing the frequency distribution of terms, the Counting phase encapsulates the underlying structure and content of the textual data, paving the way for subsequent feature extraction and analysis.

Finally, the Vectorization step culminates transforming each document into a numerical vector representation. In this phase, the frequency counts obtained during the Counting step are leveraged to construct a vector for each document, wherein the values correspond to the frequency of words in the vocabulary. This vectorization process encodes textual data into a structured numerical format amenable to machine learning algorithms, facilitating subsequent classification, clustering, or regression tasks.

Despite its simplicity, the Bag-of-Words representation effectively captures the overall word distribution in a document and has been widely used in text classification tasks.

Model Selection and Training

In this study, we evaluated two popular classification algorithms, Naive Bayes and Logistic Regression, each paired with two different feature extraction methods: TF-IDF and Bag-of-Words representation.

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with an assumption of conditional independence between features. Despite its simplicity, Naive Bayes often performs well in text classification tasks and is computationally efficient. We employed the Multinomial Naive Bayes variant, suitable for discrete features such as word counts.

Logistic Regression is a linear classification algorithm that estimates the probability of a sample belonging to a particular class. Despite its name, Logistic Regression is used for binary classification tasks and can be extended to handle multi-class classification using techniques such as one-vs-rest or softmax regression.

In each combination of classification algorithm and feature extraction method, a series of standardized steps were meticulously executed to ensure systematic model training and evaluation. The initial phase involved data preparation, where the preprocessed dataset underwent partitioning into training, validation, and test sets. The training set facilitated model training, while the validation set served as a crucial component for hyperparameter tuning and performance monitoring during the training process. Meanwhile, the test set remained untouched until the final evaluation, ensuring an unbiased assessment of model performance on unseen data. Subsequently, the classification models, including Naive Bayes and Logistic Regression, were initialized with default parameters, laying the foundation for subsequent training and evaluation steps.

Following model initialization, feature extraction techniques tailored to each chosen method, namely TF-IDF or Bag-of-Words, were applied to transform the training data into feature vectors. This step enabled the representation of textual information in a numerical format that is amenable to computational analysis and model training. With feature extraction completed, the transformed feature vectors served as inputs for training the classification models. Throughout the training process, hyperparameters such as Naive Bayes' smoothing parameters and Logistic Regression's regularization strength were fine-tuned using grid or random search techniques, optimizing model



DOI: 10.31253/te.v8i2.3212

performance. Finally, the performance of each trained model was rigorously evaluated using standard metrics such as accuracy, precision, recall, and F1-score on the validation set. Cross-validation techniques were employed to obtain robust estimates of model performance, mitigating the impact of random variability in model training and ensuring the reliability of the results. Collectively, these systematic steps laid the groundwork for a thorough and reliable assessment of classification model performance across different feature extraction methods.

After training and evaluation, the performance of each model was compared across different combinations of classification algorithms and feature extraction methods. The models were assessed based on their ability to accurately classify Sundanese news articles into predefined topics or classes.

## V. RESULT

The Results and Discussions section presents the outcomes of our experiments on classifying Sundanese news articles and the ensuing analysis and interpretation. This section provides a comprehensive overview of the performance metrics obtained from evaluating different classification models and feature extraction methods.

Experimental Setup

The dataset from www.wasunda.com was partitioned into three subsets: training, validation, and test sets. We employed a stratified sampling strategy to ensure that each subgroup maintains the same class distribution as the original dataset, thus preventing bias in model evaluation. The dataset was partitioned into training, validation, and test sets using the following ratios: 70% for training, 15% for validation, and 15% for testing.

Before training the models, the raw textual data underwent preprocessing, including text normalization, tokenization, stop-word removal, and stemming or lemmatization. These preprocessing techniques were specifically tailored to accommodate the linguistic nuances of Sundanese text and ensure the data's consistency and quality.

Two feature extraction techniques were employed: TF-IDF and Bag-of-Words representation. For TF-IDF, the TfidfVectorizer class from the scikit-learn library was used to transform the text data into TF-IDF feature vectors. Similarly, the CountVectorizer class from scikit-learn generated Bag-of-Words feature vectors.

A systematic training and evaluation process was conducted separately for each pairing to comprehensively assess the performance of each combination of classification algorithm and feature extraction method. This involved standardized steps to ensure consistency and rigor throughout the experimentation process. Initially, the classification model was initialized with default parameters, providing a baseline configuration to begin training. Subsequently, the preprocessed text data underwent feature extraction, which was transformed into numerical feature vectors utilizing the selected feature extraction method. This crucial step enabled the representation of textual information in a format suitable for computational analysis and model training.

Following feature extraction, the classification model was trained on the transformed feature vectors using the designated training set. Throughout the training process, hyperparameters of the classification models were fine-tuned to optimize performance. Techniques such as grid or random search were employed to systematically explore the hyperparameter space and identify configurations that yielded optimal results. Additionally, a validation set was utilized to evaluate the models' performance during hyperparameter tuning, ensuring robustness and generalization capability. Finally, the trained models were rigorously assessed on the validation set utilizing typical evaluation metrics, including accuracy, precision, recall, and F1-score. Cross-validation methods were applied to obtain reliable estimates of model performance, mitigating the impact of randomness and variability in the training process. Collectively, these systematic steps enabled the thorough assessment and comparison of classification models across different feature extraction methods, laying the foundation for informed decision-making and model selection. After training



DOI: 10.31253/te.v8i2.3212

and evaluation, the performance of each model was compared across various combinations of classification algorithms and feature extraction methods. The models were assessed based on their capacity to classify Sundanese news articles into predefined topics or classes accurately.

Experimental Design

The experiments were designed to systematically evaluate the performance of four different combinations of classification models and feature extraction methods for classifying Sundanese news articles. Each combination was trained and evaluated using the same dataset and experimental setup described in the previous section.

Performance Metrics

Each model's performance was evaluated using standard assessment measures, such as F1-score, recall, accuracy, and precision. Precision quantifies the percentage of correctly categorized occurrences among those projected to be positive, whereas accuracy assesses the classification's total correctness. The percentage of accurately categorized positive cases among all positive instances is known as recall. A balanced classifier performance indicator is the F1-score, the harmonic mean of precision and recall.

Results

As previously stated, the data in this study were taken from a Sundanese-language news portal. We collected 8010 articles from 7 July 2020 to 14 December 2021. The dataset consists of 6567 news articles and 1443 sports articles. Table 2 Shows examples of data collection results for news and sports categories.

**Table 2. Data collection results example**

| ID | Category | Text Article |
|---|---|---|
| 1 | Warta (news) | Saurang budak nu aranna Akmal (30), kalayan euweuh rosa kamanusaan lantaran manéhna mengas pati nu jadi bapa, indung, jeung dulur tegesna. Kajadian éta diiberkeun aya di Kelurahan Ereng-ereng,….. |
| 2 | Olahraga (sport) | Dua wawakil Indonesia tina katégori nomor ganda putra jeung ganda putri bakal nyanghareupan lawan kuat di babak semifinal World Tour Finals …. |

In addition to evaluating the performance of classification models, we employed word clouds to visually depict the most frequently occurring words within each category of news articles. Word clouds serve as an intuitive data visualization technique, where the size of each word corresponds to its frequency or importance within the text data. In Figure 2, we present the word cloud representing news article-related words, highlighting prominent terms and themes across the corpus of news articles. Similarly, Figure 3 showcases the word cloud for sports article-related words, offering insights into sports-related content's distinctive vocabulary and topics. Through these visualizations, we gain a nuanced understanding of the prevalent themes and subjects within each category, providing valuable context for interpreting the classification results and identifying key features contributing to the classification task's success.


**Figure 2. News Articles Related Words**


**Figure 3. Sports Articles Related Words**

Figure 4 presents the keywords with the most excellent TF-IDF scores within the News class. As depicted in the figure, the word "vaksinasi" (vaccination) occupies the top position based on the TF-IDF score for the news category. This observation indicates that from July 7, 2020, to December 14, 2021, there was a significant emphasis on vaccination-related news articles. An intriguing aspect of the TF-IDF results for the news category is the absence of the word "bandung" in the top 10 TF-IDF list, despite Bandung being the capital city of West Java province. Instead, words like "bogor" and "garut" emerge as the second and third highest-ranking terms. as close as possible to the text they refer to and aligned center.

Figure 5 showcases the keywords with the most excellent TF-IDF scores within the Sports class. Here, the words "persib" (referring to a football club) and "gol" (goal) occupy the first and second positions, respectively. These findings shed light on the dominant themes and topics within the sports-related news articles, highlighting the prominence of specific terms indicative of sports events, teams, and activities. Analyzing TF-IDF scores gives us valuable insights into the key terms and subjects that characterize news articles across different categories. This provides a deeper context for understanding the content and themes prevalent within each classification class.

Table 2 comprehensively compares the categorization model results, showing the performance metrics for each combination of the classification model and feature extraction method. Across all configurations, the models exhibited high accuracy, precision, recall, and F1-score levels, revealing robust performance in classifying Sundanese news articles into predefined categories.

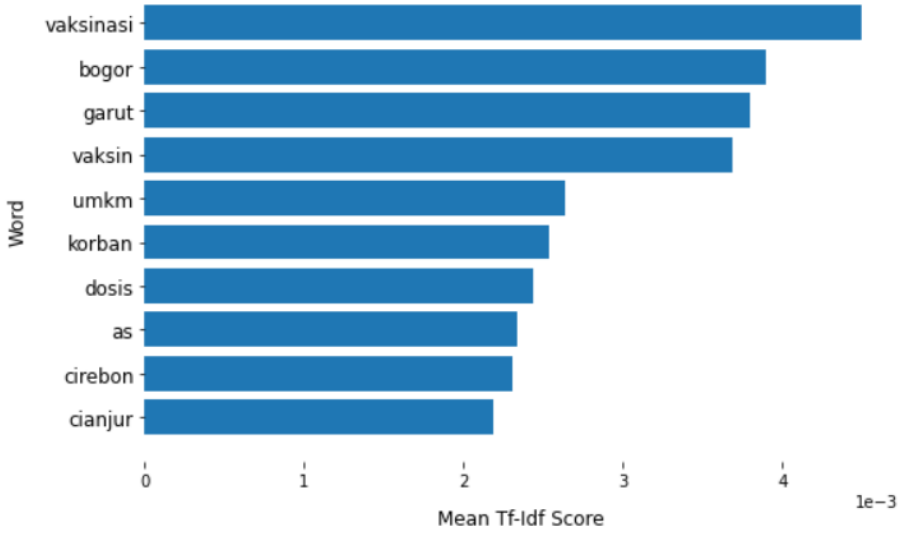


**Figure 4. Best 10 keywords with the most excellent TF-IDF score in News class**



DOI: 10.31253/te.v8i2.3212

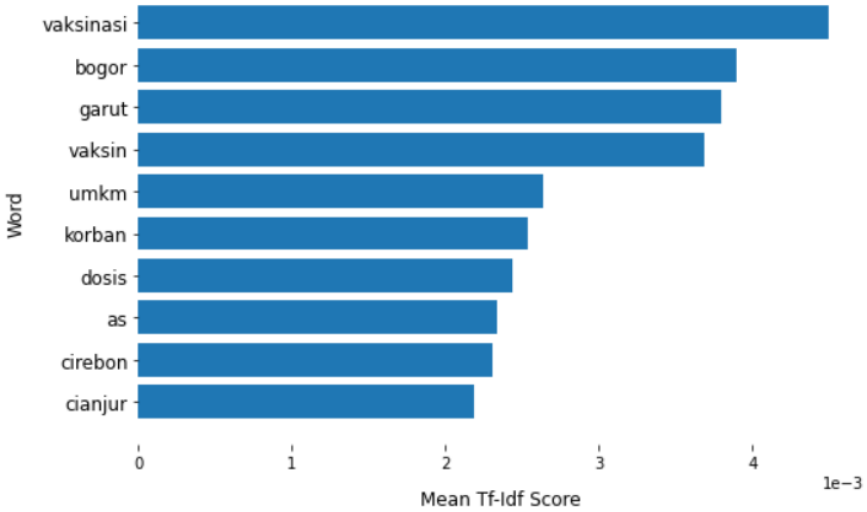


**Figure 5. Best 10 keywords with the most excellent TF-IDF score in Sport class**

**Table 2. Comparison of Categorization Model Result**

| Model | Feature Extraction | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Naive Bayes | TF-IDF | 0.9732 | 0.9742 | 0.9732 | 0.9735 |
| Naive Bayes | Bag-of-Words | 0.9801 | 0.98 | 0.9801 | 0.9799 |
| Logistic Regression | TF-IDF | 0.9757 | 0.9765 | 0.9757 | 0.976 |
| Logistic Regression | Bag-of-Words | 0.982 | 0.9819 | 0.982 | 0.9819 |

Figure 6 visualizes the execution metrics for every classification model paired with different feature extraction methods. The rows correspond to the model-feature combinations, while the columns represent the respective performance metrics. The color's intensity draws attention to the size of the values; deeper hues indicate higher performance.

The heatmap shows that Logistic Regression with Bag-of-Words consistently achieved the highest performance across all metrics, followed closely by Naive Bayes with Bag-of-Words. The TF-IDF feature extraction method yielded slightly lower but comparable results for both models. This visualization underscores the robustness of the Bag-of-Words method, particularly when paired with Logistic Regression, in capturing the nuanced patterns of Sundanese news articles.

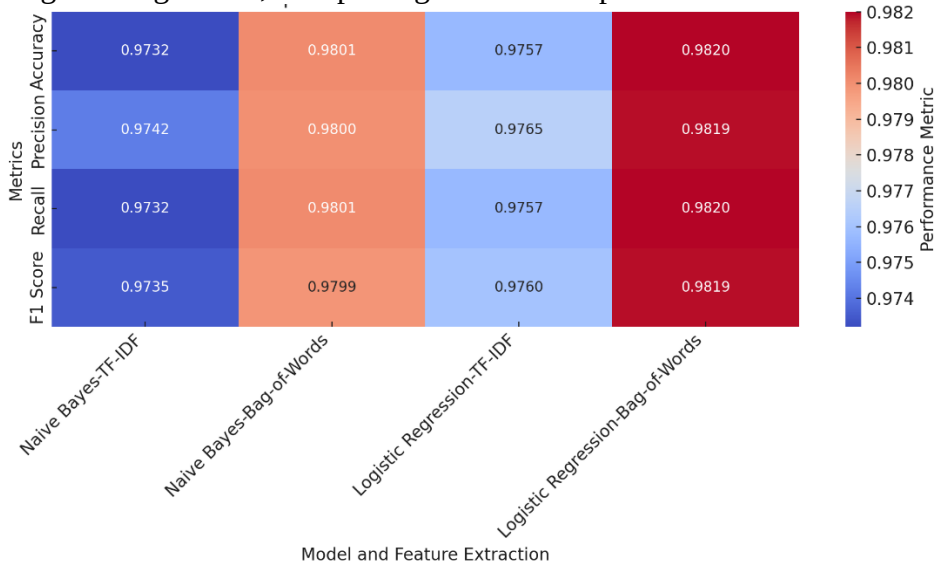


**Figure 6. Comparative Heatmap of Model Performance Metrics across Feature Extraction Techniques**



DOI: 10.31253/te.v8i2.3212

Figure 7 illustrates the relationship between Precision and F1 Score for the evaluated models and feature extraction methods. Each data point corresponds to a model-feature combination color-coded based on the feature extraction method. The proximity of points to the diagonal line (Precision ≈ F1 Score) indicates a strong balance between the two metrics, reinforcing the models' overall reliability.

The plot reveals that Logistic Regression with Bag-of-Words achieved the highest Precision and F1 Score, as reflected by its placement near the top-right corner of the plot. Naive Bayes with Bag-of-Words also demonstrated competitive performance, albeit slightly below Logistic Regression. The compact clustering of points within the high-performance region further illustrates the consistency of all model-feature combinations. This scatter plot effectively conveys the comparative advantages of each approach while emphasizing the relationship between these critical performance metrics.

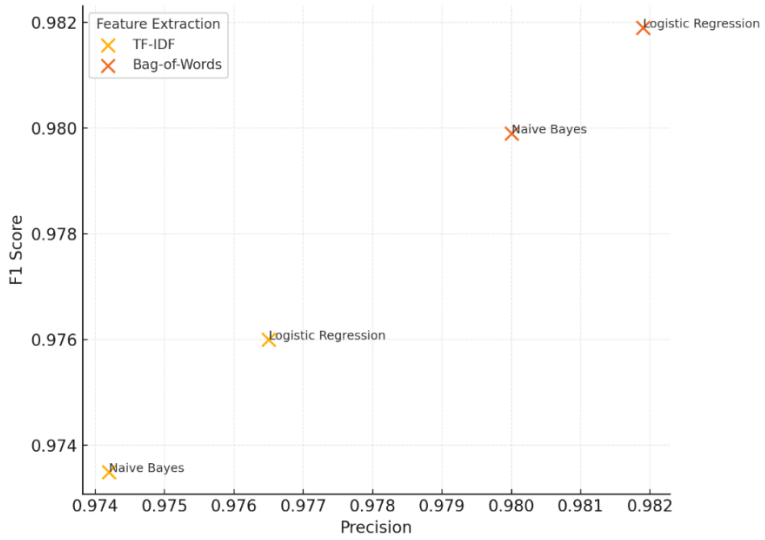


**Figure 7. Precision vs. F1 Score for Classification Models and Feature Extraction Methods**

## VI. DISCUSSION

When considering the Naive Bayes classifier, it is evident that both TF-IDF and Bag-of-Words feature extraction methods yielded impressive results. The model achieved an accuracy of 97.32% with TF-IDF and slightly higher, 98.01%, with Bag-of-Words. Precision, recall, and F1-score metrics were also consistently high for both feature extraction methods, demonstrating the effectiveness of Naïve Bayes in capturing the underlying patterns within the textual data.

Similarly, the Logistic Regression classifier exhibited commendable performance across TF-IDF and Bag-of-Words feature extraction methods. With TF-IDF, the model achieved an accuracy of 97.57%, while with Bag-of-Words, it reached an even superior accuracy of 98.20%. These outcomes emphasize the robustness and versatility of Logistic Regression in effectively discerning between different categories of Sundanese news articles.

A deeper examination of why Logistic Regression combined with Bag-of-Words outperformed other combinations points to both the algorithm's strengths and the Sundanese language's characteristics. Logistic Regression is a linear model that can effectively capture relationships between features while accommodating high-dimensional data, such as that generated by Bag-of-Words. The simplicity of Bag-of-Words, which represents documents as vectors of word frequencies, aligns well with the linguistic structure of Sundanese, where the frequency and co-occurrence of key terms play a significant role in distinguishing between topics.

As a regional language, Sundanese often employs contextually rich vocabulary and repetitive word structures to convey meaning. Bag-of-Words inherently captures these frequency patterns, providing an advantage over TF-IDF, which focuses on term uniqueness across documents.



DOI: 10.31253/te.v8i2.3212

Additionally, Sundanese's relatively minor vocabulary and morphological structure may reduce the sparsity issues commonly associated with Bag-of-Words representations in other languages. This makes Bag-of-Words an effective feature extraction method for this dataset.

Logistic Regression's ability to assign weights to features also complements the Bag-of-Words approach by highlighting the most influential words for classification. Unlike Naive Bayes, which assumes feature independence, Logistic Regression can model word dependencies, better leveraging Sundanese's contextual and co-occurrence patterns.

Overall, these findings suggest that the combination of Logistic Regression and Bag-of-Words is particularly well-suited to Sundanese news article classification. This combination leverages the language's linguistic features and the algorithm's ability to model complex relationships in high-dimensional data, offering insights for future research and practical applications in regional language processing.

## VII. CONCLUSION

In conclusion, this study comprehensively investigated classification models and feature extraction methods for classifying Sundanese news articles. Through systematic experimentation and analysis, we have identified logistic regression with TF-IDF as the preferred approach, achieving the highest performance among the evaluated combinations.

Our results highlight the importance of considering feature extraction methods and algorithm selection to get accurate and trustworthy classification results. Additionally, this study advances our understanding of Sundanese language processing and establishes the groundwork for further regional language text classification research.

We advocate for continued research efforts in Sundanese language processing, encompassing linguistic analysis, cultural considerations, and practical applications in text classification tasks. By leveraging advancements in machine learning and NLP techniques, we aim to foster the development of inclusive and accessible technologies for Sundanese-speaking communities.

In addition to recommending deep learning techniques to enhance performance, future research can explore collaborative efforts across regional languages, such as Javanese, Balinese, and other Austronesian languages, to create multi-lingual or cross-lingual classification systems. This collaboration would enable the sharing of linguistic insights and resources, fostering a deeper understanding of regional languages' structural similarities and differences. Such efforts could also improve classification performance by utilizing transfer learning techniques where models trained on larger datasets in one language inform models for less-resourced languages.

Another promising direction is the development of more affluent and diverse datasets that capture various genres, contexts, and linguistic styles within Sundanese and other regional languages. Expanding datasets to include conversational text, traditional narratives, cultural expressions, and even code-mixed data would enhance the applicability of machine learning models across different domains and use cases. This could also involve curating audio transcriptions, annotated data, and parallel corpora, facilitating advancements in tasks like speech recognition, machine translation, and sentiment analysis for Sundanese.

Additionally, there may be ways to increase the accuracy and robustness of Sundanese news item classification systems by utilizing ensemble or hybrid approaches incorporating several classifiers and feature extraction techniques. These methods could integrate the strengths of traditional machine learning algorithms and modern deep learning architectures to create a more comprehensive and adaptive system.

Overall, this study represents a step forward in Sundanese language processing while highlighting opportunities to advance the field through inter-regional collaboration and innovative dataset development. By continuing to explore Indonesia's linguistic diversity and beyond, we remain committed to advancing the frontiers of regional language processing and contributing to the broader landscape of NLP research and innovation.



DOI: 10.31253/te.v8i2.3212

## REFERENCES


[1] Q. Li *et al.*, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, Apr. 2022, doi: 10.1145/3495162.

[2] M. E. Maron, "Automatic Indexing: An Experimental Inquiry," *J. ACM*, vol. 8, no. 3, pp. 404–417, Jul. 1961, doi: 10.1145/321075.321084.

[3] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, C. Nédellec and C. Rouveirol, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142.

[4] M. S. Simanjuntak, N. Damanik, and Allwine, "Performance Analysis Of Support Vector Machine In Identifying Comments And Ratings On E-Commerce," *Int. J. Basic Appl. Sci.*, vol. 11, no. 1, pp. 37–46, Jul. 2022, doi: 10.35335/IJOBAS.V11I1.79.

[5] Y. A. Singgalen, "Comparative analysis of decision tree and support vector machine algorithm in sentiment classification for birds of paradise content," *Int. J. Basic Appl. Sci.*, vol. 12, no. 3, pp. 100–109, Dec. 2023, doi: 10.35335/IJOBAS.V12I3.298.

[6] P. Vateekul and M. Kubat, "Fast Induction of Multiple Decision Trees in Text Categorization from Large Scale, Imbalanced, and Multi-label Data," in *2009 IEEE International Conference on Data Mining Workshops*, Dec. 2009, pp. 320–325. doi: 10.1109/ICDMW.2009.94.

[7] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Mark.*, vol. 36, no. 1, pp. 20–38, 2019, doi: https://doi.org/10.1016/j.ijresmar.2018.09.009.

[8] G. Salton, "Recent trends in automatic information retrieval," in *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '86. New York, NY, USA: Association for Computing Machinery, 1986, pp. 1–10. doi: 10.1145/253168.253171.

[9] W. A. Qader, M. M. Ameen, and B. I. Ahmed, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," in *2019 International Engineering Conference (IEC)*, Jun. 2019, pp. 200–204. doi: 10.1109/IEC47844.2019.8950616.

[10] T. Walkowiak, S. Datko, and H. Maciejewski, "Bag-of-Words, Bag-of-Topics and Word-to-Vec Based Subject Classification of Text Documents in Polish - A Comparative Study," in *Contemporary Complex Systems and Their Dependability*, W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, and J. Kacprzyk, Eds., Cham: Springer International Publishing, 2019, pp. 526–535.

[11] A. A. Suryani, D. H. Widyantoro, A. Purwarianti, and Y. Sudaryat, "The Rule-Based Sundanese Stemmer," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 4, pp. 1–28, Aug. 2018, doi: 10.1145/3195634.

[12] I. Setiawan and H.-Y. Kao, "SUSTEM: An Improved Rule-Based Sundanese Stemmer," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Apr. 2024, doi: 10.1145/3656342.

[13] O. V. Putra, F. M. Wasmanson, T. Harmini, and S. N. Utama, "Sundanese Twitter Dataset for Emotion Classification," *CENIM 2020 - Proceeding Int. Conf. Comput. Eng. Network, Intell. Multimed. 2020*, no. Cenim, pp. 391–395, 2020, doi: 10.1109/CENIM51130.2020.9297929.

[14] Y. Sudaryat, A. Prawirasumantri, and K. Yudibrata, *Tata Basa Sunda Kiwari (Sundanese Grammar Today)*. Bandung: Yrama Widya, 2013.

[15] R. H. Robins, *Sistem dan Struktur Bahasa Sunda (System and Structure ofSundanese Language)*. Jakarta: DJAMBATAN, 1983.

[16] F. Djajasudarma, *Tata Bahasa Acuan Bahasa Sunda (Sundanese Reference Grammar)*. Pusat Pembinaan dan Pengembangan Bahasa, 1994.

[17] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 11, pp. 1457–1466,




DOI: 10.31253/te.v8i2.3212

Nov. 2006, doi: 10.1109/TKDE.2006.180.

[18] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast WordPiece Tokenization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 2089–2103. doi: 10.18653/v1/2021.emnlp-main.160.

[19] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2020, pp. 466–472. doi: 10.1109/ICACCS48705.2020.9074166.

[20] L. V. Subramaniam, S. Roy, T. A. Faruquie, and S. Negi, "A survey of types of text noise and techniques to handle noisy text," in *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, in AND '09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 115–122. doi: 10.1145/1568296.1568315.



DOI: 10.31253/te.v8i2.3212

## BIOGRAPHY

**Yadhi A. Permana** obtained his Bachelor's degree in Computer Science in 2005 and his Master's in Socio Informatics in 2008. He began his career as a software developer from 2000 to 2008, then transitioned to an academic role at Politeknik Negeri Bandung in 2008 at the Department of Computer Engineering and Informatics. He has secured several research grants. His current research interests include software engineering, artificial intelligence, data analysis, and the development of intelligent systems, aiming to bridge the gap between academic research and real-world applications.

**Irwan Setiawan** obtained his Bachelor's degree in Computer Science in 2004 and his Master's in Computer Science in 2008. He began his career as a software developer from 2001 to 2005, then transitioned to an academic role at Politeknik Negeri Bandung in 2005 at the Department of Computer Engineering and Informatics. He has secured several research grants. He has published over 16 papers in reputable journals and international conferences. His current research interests include software engineering, artificial intelligence, data analysis, and the development of intelligent systems, aiming to bridge the gap between academic research and real-world applications.

**Fitri Diani** obtained her Bachelor's degree in Computer Science in 2005 and her Master's in Computer Science in 2008. She began his career as an IT consultant from 2001 to 2009, then transitioned to an academic role at Politeknik Negeri Bandung in 2009 at the Department of Computer Engineering and Informatics. She has secured several research grants. Her research interests include artificial intelligence and data analysis, aiming to bridge the gap between academic research and real-world applications.

**Suprihanto** obtained his Bachelor's degree in Computer Science in 1990 and his Master's in Computer Science in 2016. He began his career at Politeknik Negeri Bandung from 1995 at the Department of Computer Engineering and Informatics.



DOI: 10.31253/te.v8i2.3212