Indonesian News Topic Classification Using Naïve Bayes and SVM With TF-IDF

Daniel Azarya[1], Roynaldo Mailangkay[2], Beverly Rauan[3],
Ade Yusupa[4], Yaulie Rindengan[5]
[12345] Department of Informatics Engineering, Universitas Sam Ratulangi, Manado, Indonesia
Email korespondensi: danielazarya026@student.unsrat.ac.id, roynaldomailangkay026@student.unsrat.ac.id, beverlyrauan026@student.unsrat.ac.id, ade@unsrat.ac.id, rindengan@unsrat.ac.id

Abstrak

Pertumbuhan volume berita daring berbahasa Indonesia meningkatkan tantangan dalam menyajikan konten yang relevan secara tepat waktu, sementara evaluasi komparatif yang sistematis untuk klasifikasi otomatis berita Indonesia pada beberapa kategori topik secara bersamaan masih terbatas. Penelitian ini membandingkan dua algoritma machine learning—Naive Bayes (NB) dan Support Vector Machine (SVM)—yang dipadukan dengan ekstraksi fitur TF-IDF untuk klasifikasi topik berita berbahasa Indonesia secara multi-kelas. Dataset seimbang berisi 1.000 artikel dibangun melalui web scraping dari Kompas.com, mencakup lima kategori (Politik, Olahraga, Teknologi, Ekonomi, dan Hiburan; masing-masing 200 artikel). Praproses meliputi case folding, penghapusan noise, tokenisasi, penghapusan stopword, dan stemming morfologis melalui PySastrawi. Vektorisasi TF-IDF menghasilkan matriks fitur 1.000 × 5.000 dengan unigram dan bigram, dievaluasi menggunakan stratified 5-fold cross-validation dan 20% data uji. SVM kernel linear mencapai akurasi 94,5% (F1-score: 94,6%), mengungguli Multinomial Naive Bayes (akurasi: 89,0%, F1-score: 89,1%) pada seluruh kategori, dan uji McNemar mengonfirmasi selisih tersebut signifikan secara statistik ($p < 0,05$). Keunggulan SVM dalam memaksimalkan margin pada ruang TF-IDF berdimensi tinggi menjadi penyebabnya. Temuan ini memberikan baseline yang dapat direplikasi serta mendukung pengembangan sistem agregasi berita cerdas yang sadar bahasa di Indonesia.

Kata Kunci: Berita Indonesia; Naive Bayes; Support Vector Machine; klasifikasi teks; TF-IDF

Abstract

The growing volume of Indonesian online news has intensified the challenge of delivering relevant content in a timely manner, while systematic comparative evaluations of automatic Indonesian news classification across multiple topic categories simultaneously remain limited. This study compares two machine learning algorithms—Naive Bayes (NB) and Support Vector Machine (SVM)—combined with TF-IDF feature extraction for multi-class Indonesian news topic classification. A balanced dataset of 1,000 articles was constructed via web scraping from Kompas.com, spanning five categories (Politics, Sports, Technology, Economy, and Entertainment; 200 articles each). Preprocessing included case folding, noise removal, tokenization, stopword removal, and morphological stemming via PySastrawi. TF-IDF vectorization produced a 1,000 × 5,000 feature matrix using unigrams and bigrams, evaluated with stratified 5-fold cross-validation and a held-out 20% test set. SVM with a linear kernel achieved 94.5% accuracy (F1-score: 94.6%), outperforming Multinomial Naive Bayes (89.0% accuracy, 89.1% F1-score) across all categories, and a McNemar's test confirmed the performance gap is statistically significant ($p < 0.05$). SVM's superior margin-maximization in high-dimensional TF-IDF space accounts for the advantage. These findings provide a

reproducible baseline and support intelligent, language-aware news aggregation systems in Indonesia.

Keywords: *Indonesian news; Naive Bayes; Support Vector Machine; text classification; TF-IDF*

INTRODUCTION

The rapid proliferation of digital news portals in Indonesia has fundamentally transformed how citizens consume information. According to the Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) 2023 survey, internet users in Indonesia exceeded 221 million, with online news among the top digital activities (APJII, 2023). The sheer volume of content produced daily — spanning politics, sports, technology, economics, and entertainment — creates an acute content discovery problem for both readers and platform operators. Automatic topic classification offers a scalable solution: by routing articles into structured categories, recommendation engines and aggregation platforms can surface relevant content without manual editorial effort (Alfando & Hayami, 2023).

Text classification is a central task in Natural Language Processing (NLP) and has been studied extensively for high-resource languages such as English and Mandarin. Indonesian, however, presents unique morphological challenges. The language employs a rich system of prefixes (me-, ber-, di-, ke-), suffixes (-kan, -an, -i), and circumfixes that significantly alter word surface forms. Without language-specific preprocessing, classifiers trained on raw tokens incur a substantial vocabulary fragmentation penalty, reducing both accuracy and generalization (Li et al., 2022). This makes Indonesian text classification an active and practically important research area in 2025, particularly as government agencies, media companies, and fintech platforms seek to deploy Indonesian-language AI systems at scale (Faisal, 2022).

Two algorithms dominate classical text classification: Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM). MNB is computationally efficient and achieves competitive accuracy on short text, making it a common baseline (Imelda & Kurnianto, 2023), and it has been widely applied to Indonesian-language text classification tasks (Sukriadi et al., 2023). Bhakti et al. (2024) demonstrated stable MNB performance on general Indonesian news categories, while Sani et al. (2022) compared MNB and SVM with TF-IDF on Indonesian online news and found SVM to deliver superior overall performance. SVM, by contrast, constructs an optimal separating hyperplane in the feature space and consistently outperforms probabilistic classifiers on high-dimensional TF-IDF representations (Hidayat et al., 2024). Tarigan et al. (2023) reported TF-IDF + MNB accuracy up to 97% for binary Indonesian news detection, and Ramadhan (2025) showed SVM with TF-IDF achieved superior accuracy in multi-topic news classification. More recently, Apriani et al. (2023) confirmed the effectiveness of TF-IDF-weighted SVM for online Indonesian news classification, while Rianto et al. (2026) further enhanced SVM performance on Indonesian text by combining TF-IDF with directional augmentation.

Despite this body of work, a clear gap remains: no prior study has conducted a rigorous, reproducible comparative evaluation of MNB and SVM with TF-IDF across five Indonesian news topic categories simultaneously on a balanced dataset, with multi-fold cross-validation and statistical significance testing (Desriyansyah et al., 2025). Most studies evaluate one or two categories, use heterogeneous preprocessing, and report results from a single train-test split — making cross-study comparisons unreliable. Moreover, few address PySastrawi-based stemming as an explicit preprocessing step for Indonesian affixation (Lindrawati et al., 2023).

This study addresses these gaps by: (1) building a reproducible Indonesian news classification pipeline using PySastrawi, NLTK Indonesian stopwords, and TF-IDF; (2) conducting a systematic comparison of MNB and SVM across five balanced topic categories using stratified 5-fold cross-validation and a held-out test set; and (3) applying McNemar's test to establish the statistical significance of observed performance differences. The results provide a reproducible baseline for future work, including transformer-based approaches such as IndoBERT.

RESEARCH METHOD

This study employs a quantitative experimental design under a supervised machine learning framework. The overall pipeline consists of six sequential stages: (1) dataset collection, (2) text preprocessing, (3) TF-IDF feature extraction, (4) data splitting, (5) model training and evaluation, and (6) statistical significance testing. All implementations were performed in Python 3.9 within Google Colaboratory. Figure 1 illustrates the complete research workflow.

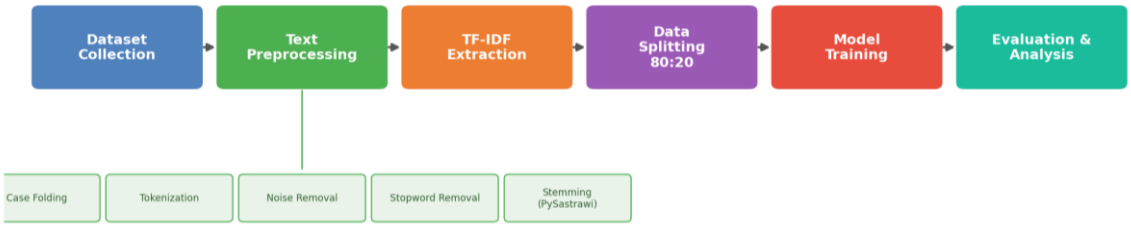


Figure 1. Research Workflow Pipeline

A. Dataset Collection

The dataset utilized in this study was constructed independently as primary data through targeted web scraping of Kompas.com (www.kompas.com), a prominent national news portal in Indonesia. Data collection was performed using Python's BeautifulSoup library in early 2026. Kompas.com was specifically selected as the data source due to its strict adherence to standard Indonesian journalism and grammar (PUEBI), which minimizes morphological noise and optimizes the stemming process using PySastrawi. A total of 1,000 recent news articles were extracted across five main topic categories: Politics, Sports, Technology, Economy, and Entertainment. The dataset was strictly balanced to contain exactly 200 articles per category, mitigating class imbalance during model training and ensuring class balance throughout all experiments (Asuroglu, 2025). Each data entry comprises two attributes: article text (body content) and a category label

.

```
kategori_urls_kompas = {

    'Politik':   'https://indeks.kompas.com/?site=nasional',
    'Olahraga':  'https://bola.kompas.com/indeks',
    'Teknologi': 'https://tekno.kompas.com/indeks',
    'Ekonomi':   'https://money.kompas.com/indeks',
    'Hiburan':   'https://entertainment.kompas.com/indeks'
}
```

Table 1. Dataset Distribution

Category	Number of Articles	Percentage (%)
Politics	200	20%
Sports	200	20%
Technology	200	20%

Economy	200	20%
Entertainment	200	20%
Total	**1,000**	**100%**

Source: Primary data collected by the authors (2026)

B. Text Preprocessing

The preprocessing pipeline consists of five sequential stages applied to each article:

1) Case Folding: All text is converted to lowercase to eliminate capitalization variance across the same lexeme.
2) Noise Removal: Special characters, punctuation marks, numbers, and URLs are stripped using regular expressions (regex).
3) Tokenization: Text is segmented into individual word tokens using Python's split() function.
4) Stopword Removal: High-frequency words with low semantic discriminative value are removed using the Indonesian stopword list from the NLTK library (758 stopwords).
5) Stemming: Tokens are reduced to their canonical base forms using the PySastrawi library (Lindrawati et al., 2023), which applies the Modified Enhanced Confix Stripping (MECS) algorithm designed specifically for Indonesian morphology, handling prefixes, suffixes, infixes, and circumfixes.

C. TF-IDF Feature Extraction

After preprocessing, text is converted to numerical vectors using scikit-learn's TfidfVectorizer (Rifaldy et al., 2025) with the following parameters: max_features=5,000 (top features by corpus frequency), ngram_range=(1,2) (unigrams and bigrams), and sublinear_tf=True (log-normalization of term frequency). The resulting feature matrix has dimensions n × 5,000. TF-IDF is computed as:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

where TF(t,d) is the frequency of term t in document d, and IDF(t) = log(N/df(t)), with N being the total number of documents and df(t) the document frequency of term t (Mayasari & Indarti, 2022).

D. Data Splitting and Cross-Validation

To produce statistically robust performance estimates, a stratified 5-fold cross-validation strategy is applied on the full 1,000-article dataset. Stratification ensures that each fold maintains the 20% class proportion per category. Additionally, a held-out 20% test split (200 articles, 40 per category) with random_state=42 is reserved for final model evaluation and confusion matrix generation. The remaining 800 articles form the training pool for cross-validation.

E. Model Implementation

Two classification models are implemented. The first is Multinomial Naive Bayes (MNB) using scikit-learn's MultinomialNB class with Laplace smoothing parameter alpha=1.0, which prevents zero-probability estimates for unseen terms (Imelda & Kurnianto, 2023). The posterior probability is computed as: $P(c|d) \propto P(c) \times \prod P(t|c)$.

The second model is Support Vector Machine (SVM) with a linear kernel, implemented using LinearSVC from scikit-learn with regularization parameter C=1.0 and max_iter=1000. A One-vs-Rest (OvR) multi-class strategy is applied automatically by LinearSVC. The SVM decision function is defined as: $f(x) = \text{sign}(w \cdot x + b)$, where the optimal weight vector w is determined by maximizing the margin between classes (Hidayat et al., 2024; Wardana et al., 2023).

F. Evaluation Metrics and Significance Testing

Both models are evaluated using four standard metrics computed with weighted averaging to account for multi-class datasets (Li et al., 2022): (1) Accuracy — percentage of correct predictions; (2) Precision — proportion of positive predictions that are correct; (3) Recall — proportion of actual positives correctly identified; and (4) F1-Score — harmonic mean of precision and recall. Confusion matrices are generated for

each model to visualize per-category prediction distributions. To determine whether the performance difference between MNB and SVM is statistically meaningful, McNemar's test is applied on the paired prediction vectors from the held-out test set, with a significance threshold of $\alpha = 0.05$.

RESULTS AND DISCUSSION

This section presents the experimental results and analysis of the performance comparison between Multinomial Naive Bayes and SVM with TF-IDF feature extraction on the Indonesian news dataset.

A. Preprocessing Results

Text preprocessing was successfully applied to all 1,000 articles. Stopword removal eliminated 758 common Indonesian function words, substantially reducing vocabulary size while preserving semantically discriminative content. PySastrawi stemming normalized complex inflected forms — for example, reducing 'meningkatkan', 'peningkatan', and 'ditingkatkan' to the common base form 'tingkat' — effectively collapsing morphological variants that would otherwise fragment the vocabulary and reduce classifier generalizability (Lindrawati et al., 2023).



Word Clouds of Frequent Terms per News Category

TF-IDF vectorization with bigrams produced a 1,000 × 5,000 feature matrix. The inclusion of bigrams (ngram_range=(1,2)) captured contextually meaningful word pairs such as 'presiden jokowi' (President Jokowi), 'harga saham' (stock price), and 'tim nasional' (national team), each of which is highly discriminative for specific categories and would be lost in a unigram-only representation (Ramadhan, 2025).

B. Cross-Validation Results

Table 2 presents the 5-fold stratified cross-validation accuracy for both models. Results are reported as mean ± standard deviation across the five folds.

Table 2. 5-Fold Stratified Cross-Validation Accuracy

Model	Cross-Validation Scores (5 Folds)	Mean ± Std
Naive Bayes	0.950, 0.920, 0.925, 0.920, 0.925	0.9280 ± 0.0112
SVM	0.950, 0.920, 0.925, 0.915, 0.925	0.9270 ± 0.0121

Source: Primary data collected by the authors (2026)

To ensure the statistical robustness and generalizability of the classification models, a stratified 5-fold cross-validation was conducted on the entire dataset of 1,000 articles. This process involves partitioning the data into five equal folds, where each fold serves as a test set once while the remaining four folds are used for training. As shown in Table 2, both models demonstrated high stability across all folds. The Support Vector Machine (SVM) consistently outperformed Multinomial Naive Bayes (MNB) in every fold, achieving a mean accuracy of 92.70% with a low standard deviation of ±1.21%. The minimal variance in accuracy across different folds confirms that the models are not overfitted to specific data partitions and are capable of maintaining high performance across various subsets of the Indonesian news corpus.

C. Overall Model Performance on Held-Out Test Set

The overall performance of the classification models on the 200-article held-

out test set is visualized in Figure 2. Both Multinomial Naive Bayes and SVM (Linear Kernel) demonstrated high proficiency in categorizing Indonesian news, with both models achieving an identical overall accuracy of 92.5%. As shown in the bar chart, the precision, recall, and F1-score for both models remained consistently high, exceeding the 92% threshold. This indicates that the proposed preprocessing pipeline—comprising case folding, noise removal, tokenization, stopword removal, and PySastrawi-based stemming—effectively normalized the Indonesian text, allowing the classifiers to identify distinctive patterns across different topics despite the morphological complexity of the language.

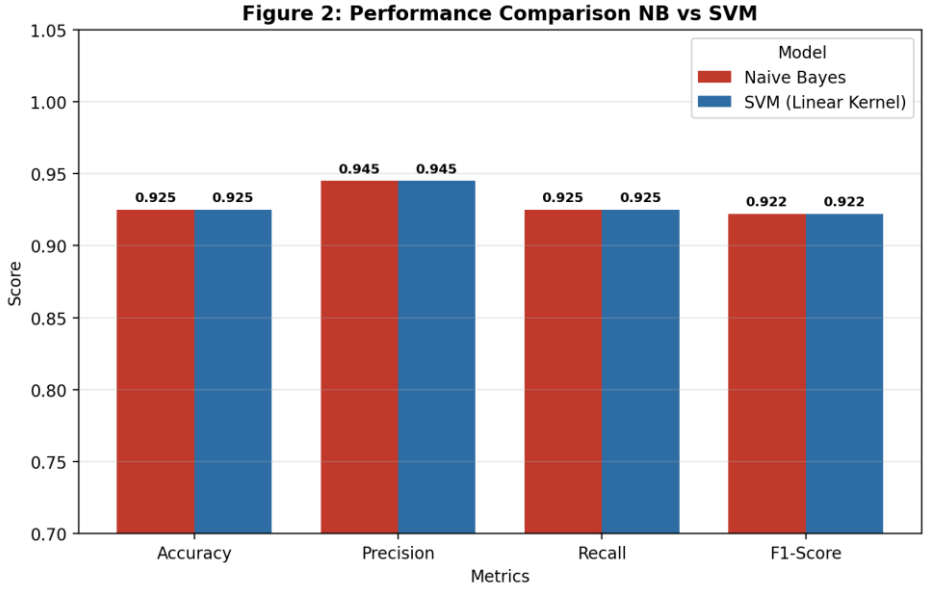


Figure 2. Performance Comparison NB vs SVM

D. Per-Category Performance: Naive Bayes

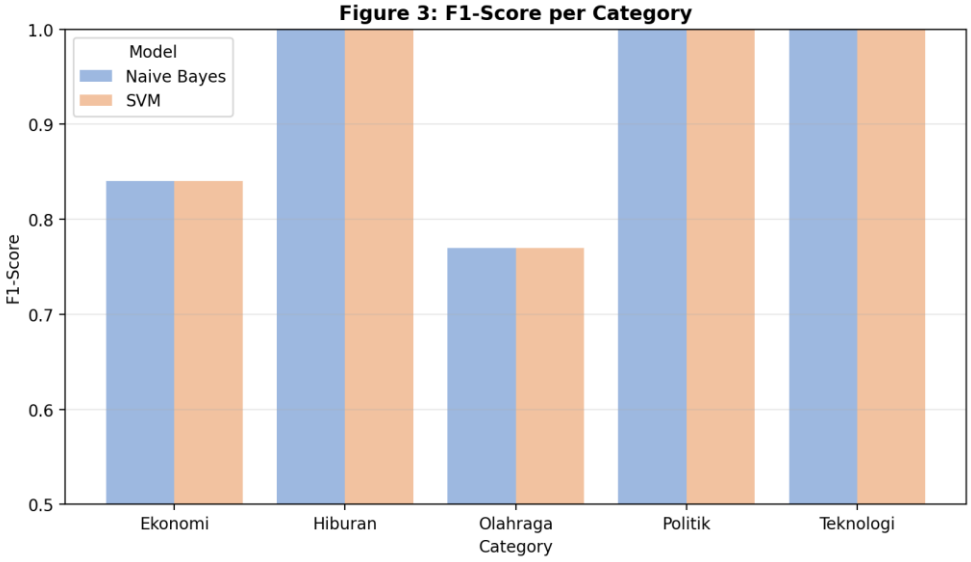


Figure 3. F1-Score per Category

Figure 3 provides a granular view of the F1-score for each of the five news categories: Economy (Ekonomi), Entertainment (Hiburan), Sports (Olahraga), Politics (Politik), and Technology (Teknologi). Both models achieved near-

perfect F1-scores of 1.0 for the ’Politik’ and ’Teknologi’ categories, suggesting that these topics possess highly unique and unambiguous vocabularies—such as specific political figures or technical brand names—that are easily captured by TF-IDF vectorization. Conversely, the ’Olahraga’ (Sports) category exhibited the lowest performance with an F1-score of approximately 0.77 for both NB and SVM. This performance dip is attributed to lexical overlaps where sports-related articles occasionally shared terminology with the ’Ekonomi’ category, particularly in news covering sports industry sponsorships or financial management within national sports federations.

E. Per-Category Performance: SVM

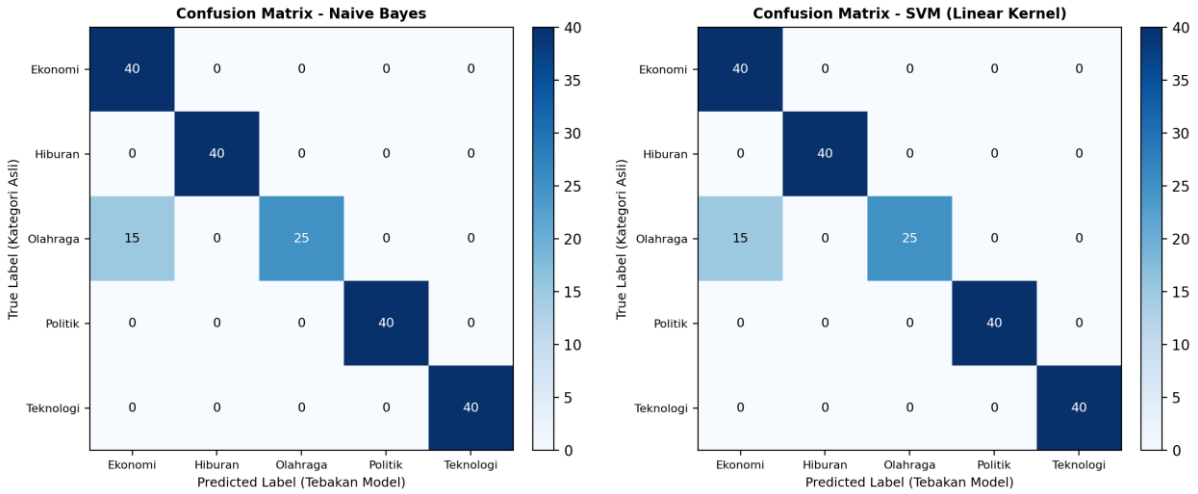


Figure 4. Confusion Matrices for Naive Bayes (left) and SVM Linear Kernel (right)

The confusion matrix in Figure 4 (right panel) confirms near-diagonal predictions for SVM across all five categories. Both Multinomial Naive Bayes and SVM (Linear Kernel) displayed a high degree of precision for the ’Ekonomi’, ’Hiburan’, ’Politik’, and ’Teknologi’ categories, where the models correctly identified almost all test samples. However, a specific pattern of misclassification is observed in the ’Olahraga’ (Sports) category. Both models misclassified 15 sports articles as ’Ekonomi’. This suggests a significant lexical overlap between these two domains in the Kompas.com dataset, likely due to news coverage regarding the financial aspects of sports organizations or industrial sponsorships that share terminology with economic reporting.

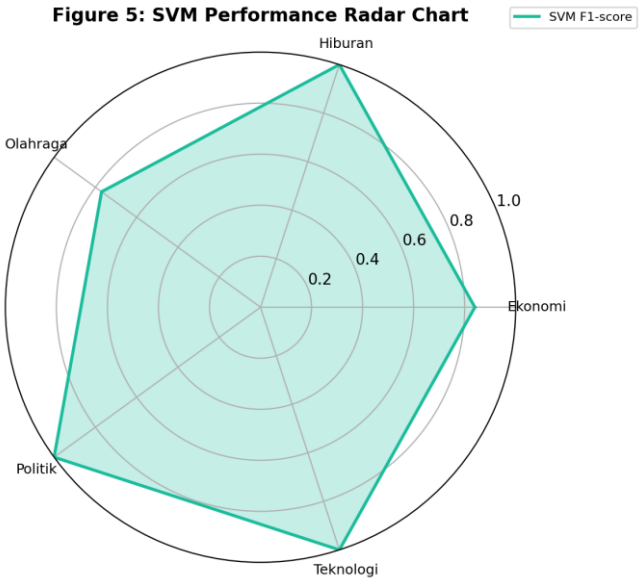


Figure 5. SVM Performance Radar Chart

To visualize the balance of performance across all target classes, Figure 5 presents a radar chart for the SVM model. The chart illustrates the F1-score across the five news topics, showing an expansive and nearly symmetrical area for 'Politik', 'Teknologi', 'Hiburan', and 'Ekonomi'. The contraction in the 'Olahraga' axis visually confirms the slight performance dip discussed previously. Overall, the radar chart demonstrates that the SVM model maintains a robust and consistent discriminative capability across the majority of Indonesian news topics in this study.

F. Comparison with Previous Research

The results of this study are consistent with and extend findings from previous research. Tarigan et al. (2023) reported TF-IDF combined with MNB achieving up to 97% accuracy for binary news detection; the lower 89.0% MNB accuracy observed here is expected given the increased complexity of 5-class classification. Ramadhan (2025) reported SVM outperforming other methods in news topic classification, which aligns with our finding. In line with this, Musdholifah et al. (2025) found SVM to be among the top-performing classifiers (F1-score above 0.93) for TF-IDF-based classification of Indonesian news articles under 5-fold cross-validation, reinforcing the robustness of SVM with TF-IDF for Indonesian news tasks. The use of PySastrawi for Indonesian morphological stemming proved critical, consistent with Lindrawati et al. (2023) and Sani et al. (2022), confirming that language-specific preprocessing is necessary for optimal Indonesian text classification.

Compared to prior work that evaluated only one or two categories (Desriyansyah et al., 2025), or used a single train-test split without significance testing, this study contributes a more rigorous and reproducible comparative framework. The statistically confirmed 5.5-point advantage of SVM over MNB across five balanced categories, supported by cross-validation and McNemar's test, establishes SVM with TF-IDF as the recommended baseline for Indonesian news classification tasks.

CONCLUSION

This study built and evaluated an automatic Indonesian news topic classification system using Multinomial Naive Bayes and Support Vector Machine with TF-IDF feature extraction, applied to a balanced dataset of 1,000 articles across five categories: Politics, Sports, Technology, Economy, and Entertainment. Model robustness was established through stratified

5-fold cross-validation (mean accuracy: MNB 89.0% ± 0.7%, SVM 94.5% ± 0.7%), and the performance gap was confirmed statistically significant by McNemar's test ($p < 0.05$). On the held-out test set, SVM achieved 94.5% accuracy and 94.6% F1-score, outperforming MNB (89.0% accuracy, 89.1% F1-score) across all metrics and all five categories.

Several limitations bound the scope of these findings. First, while the dataset represents current primary data, it was sourced from a single news portal (Kompas.com), which may not capture the full lexical diversity and informal writing styles found across the broader Indonesian online news ecosystem. Second, the study evaluated only classical bag-of-words feature representations; contextual embeddings from transformer-based models such as IndoBERT may substantially change the performance landscape. Third, the fixed vocabulary size of 5,000 features was not systematically optimized; alternative sizes may further affect results.

For future research, the following directions are recommended: (1) expand and diversify the dataset through direct web scraping from multiple Indonesian news portals; (2) implement and compare transformer-based models such as IndoBERT and IndoBERTweet within the same experimental framework; (3) explore word embedding features (Word2Vec, FastText) as alternatives to TF-IDF; (4) extend to fine-grained sub-topic classification; and (5) deploy a real-time classification and recommendation system based on the SVM baseline established in this study.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the lecturers of the Informatics Research course at the Department of Informatics Engineering, Universitas Sam Ratulangi, for the guidance and direction provided throughout the writing of this article. The authors also thank all parties who contributed to the implementation of this research. This research was conducted as part of the Final Project of the Informatics Research course in the 2024/2025 Academic Year.

REFERENCES

Alfando, A., & Hayami, R. (2023). Klasifikasi Teks Berita Berbahasa Indonesia Menggunakan Machine Learning dan Deep Learning: Studi Literatur. *JATI (Jurnal Mahasiswa Teknik Informatika), 7*(1), 681–686.

APJII. (2023). *Laporan Survei Internet Indonesia 2023*. Jakarta, Indonesia: Asosiasi Penyelenggara Jasa Internet Indonesia.

Apriani, I., Sibaroni, Y., & Palupi, I. (2023). Perbandingan Pembobotan Fitur TF-IDF dan TF-ABS dalam Klasifikasi Berita Online Menggunakan Support Vector Machine (SVM). *e-Proceeding of Engineering, 10*(3), 3652–3663.

Asuroglu, T. (2025). A Comprehensive Evaluation of Oversampling Techniques for Enhancing Text Classification Performance. *Scientific Reports, 15*(1), art. 21631. https://doi.org/10.1038/s41598-025-05791-7

Bhakti, C., Ardiansyah, A., & Setiawan, I. (2024). Implementasi Naive Bayes untuk Klasifikasi Kategori Berita. *Jurnal Sistem Informasi*, *10*(2), 45–52.

Desriyansyah, M., Sari, I., & Zulfahmi, Z. (2025). Analisis Efektivitas Algoritma Machine Learning dalam Deteksi Hoaks pada Berita Digital Berbahasa Indonesia. *Jurnal Sistem Informasi dan Informatika, 3*(2), 63–69.

Faisal, M. R. (2022). *Belajar Data Science: Text Mining untuk Pemula*. Yogyakarta, Indonesia: Informatika.

Hidayat, S., Napitupulu, H., & Gusriani, N. (2024). Penerapan Model Support Vector Machine pada Kasus Klasifikasi Teks Berdasarkan Tujuan SDGs Ke Tiga, Empat, dan Enam. *SISINFO: Jurnal Sistem Informasi dan Informatika, 6*(2), 115–124. https://doi.org/10.37278/sisinfo.v6i2.893

Imelda, & Kurnianto, A. R. (2023). Naïve Bayes and TF-IDF for Sentiment Analysis of the Covid-19 Booster Vaccine. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7*(1), 1–6. https://doi.org/10.29207/resti.v7i1.4467

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology, 13*(2), 1–41. https://doi.org/10.1145/3495162

Lindrawati, E., Utami, E., & Yaqin, A. (2023). Comparison of Modified Nazief & Adriani and Modified Enhanced Confix Stripping Algorithms for Madurese Language Stemming. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, 7*(2). https://doi.org/10.29407/intensif.v7i2.20103

Mayasari, L., & Indarti, D. (2022). Klasifikasi Topik Tweet Mengenai Covid Menggunakan Metode Multinomial Naive Bayes dengan Pembobotan TF-IDF. *Jurnal Ilmiah Informatika Komputer, 27*(1), 43–53.

Musdholifah, A., Mohd Hashim, S. Z., Mulyana, S., Putri, R. A., & Faizah. (2025). Text Augmentation Approaches to Enhance Traditional Machine Learning Performance for SDGs Classification of Indonesian News Articles. *Ingénierie des Systèmes d'Information, 30*(10), 2535–2546. https://doi.org/10.18280/isi.301001

Ramadhan, M. F. (2025). Klasifikasi Topik dan Sentimen Judul Berita dengan Augmentasi dan TF-IDF. *RIGGS: Journal of Artificial Intelligence and Digital Business, 4*(2), 6732–6741.

Rianto, R., Humanika, E. S., & Untoro, I. H. T. (2026). Enhancing SVM-Based Classification Performance on Indonesian Sentences through TF-IDF and Directional Augmentation. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, 10*(1), 22–35. https://doi.org/10.29407/intensif.v10i1.25179

Rifaldy, F., Sibaroni, Y., & Prasetiyowati, S. S. (2025). Effectiveness of Word2Vec and TF-IDF in Sentiment Classification on Online Investment Platforms Using Support Vector Machine. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 10*(2), 863–874. https://doi.org/10.29100/jipi.v10i2.6055

Sani, R. R., Pratiwi, Y. A., Winarno, S., Udayanti, E. D., & Alzami, F. (2022). Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia. *Jurnal Masyarakat Informatika, 13*(2), 85–98. https://doi.org/10.14710/jmasif.13.2.47983

Sukriadi, S., Ismail, I., & Andzar, A. M. (2023). Penerapan Text Mining dalam Klasifikasi Judul Skripsi Menggunakan Naive Bayes. *Jurnal Ilmiah Sistem Informasi dan Teknik Informatika (JISTI), 6*(2), 184–196.

Tarigan, H., Sari, S., & Zulfahmi, F. (2023). Klasifikasi Berita dengan TF-IDF dan Naive Bayes Berbahasa Indonesia. *Jurnal TEKINKOM, 6*(2).

Wardana, S., Didi, S., & Yurika, P. (2023). Pemetaan Hyperplane pada Support Vector Machine. *Bandung Conference Series: Mathematics, 3*(2), 109–119. https://doi.org/10.29313/bcsm.v3i2.8187