INTENSIF, Vol.10 No.1 2026
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

Enhancing SVM-Based Classification Performance on Indonesian Sentences through TF-IDF and Directional Augmentation

Received:
8 June 2025
Accepted:
9 October 2025
Published:
31 January 2026

[1*]Rianto, [2]Eko Setyo Humanika, [3]Iwan Hartadi Tri Untoro
[1]Sains Data, Universitas Teknologi Yogyakarta
[2]Sastra Inggris, Universitas Teknologi Yogyakarta
[3]Sistem Informasi, Universitas Teknologi Yogyakarta
E-mail: [1]rianto@uty.ac.id, [2]eko.humanika@uty.ac.id, [3]iwan.hartadi@uty.ac.id
*Corresponding Author

Abstract— **Background:** The distinction between standard and non-standard Indonesian sentences is traditionally well-defined, yet the ubiquity of digital communication has increasingly blurred these boundaries. This convergence introduces significant lexical ambiguity in formal contexts, complicating the performance of automated text classification systems. **Objective:** This study aims to enhance the robustness of Support Vector Machine (SVM) classification by addressing these linguistic irregularities through TF-IDF vectorization and a targeted directional augmentation strategy. **Methods:** A corpus comprising 5,394 labeled sentences was processed under a strict anti-leak grouping strategy to rigorously prevent semantic leakage between training, validation, and testing sets. To resolve decision boundary overlaps often missed by the baseline model, manual directional augmentation was applied, specifically targeting ambiguous sentence structures to enrich the training distribution and linguistic diversity. **Results:** The experiments demonstrated that directional augmentation significantly refined the model's decision margins. While the baseline model achieved a test accuracy of 94.39%, the augmented approach substantially improved generalization capabilities across unseen groups, elevating validation accuracy from 96.11% to 97.39% and test accuracy to 96.16%. **Conclusion:** These findings substantiate that structurally enriching the dataset effectively mitigates overfitting and improves sensitivity. However, given the scalability constraints of manual intervention, future research should prioritize automated augmentation techniques and contextual embeddings to handle deep linguistic nuances further.

Keywords— Directional Augmentation; Indonesian Sentences; SVM; Text Classification; TF-IDF

Corresponding Author:

Rianto,
Department of Sains Data,
Universitas Teknologi Yogyakarta,
Email: rianto@uty.ac.id
Orchid ID: https://orcid.org/0000-0002-5058-4580

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

I. INTRODUCTION

Text classification, a core task in Natural Language Processing (NLP), is driven by the growing adoption of intelligent systems for interpreting human language [1]. One of its key applications is distinguishing between standard and non-standard language styles, a particularly relevant task in the Indonesian context where speakers often blend formal and informal styles within a single discourse [2]. The expansion of digital platforms, while amplifying linguistic fluidity, has also introduced new complexities to the classification process [3].

Standard Indonesian sentences—typically found in formal domains such as government or academic communication—are increasingly mixed with non-standard expressions that dominate daily interactions [4]. This is evident in short expressions like *"Kesabaran adalah kunci"* (Patience is the key) or *"Cinta itu sederhana"* (Love is simple), which are structurally standard but commonly used in informal settings, posing a challenge for machine learning models that lack deep semantic understanding, particularly in detecting stylistic ambiguity [5].

Existing approaches to text classification include rule-based systems, statistical representations like CountVectorizer, and machine learning techniques such as Support Vector Machines (SVMs) [6-8]. SVMs are particularly effective in high-dimensional spaces but heavily rely on the quality of input features [9]. CountVectorizer, while proficient at capturing word frequency, often misses contextual relevance, especially in short texts [10]. However, TF-IDF representations, with their ability to highlight more informative terms, offer a promising solution for improving feature quality in short-text classification [11-12].

Data augmentation is essential for improving model generalization, especially when labeled data is limited or imbalanced. Studies by Bayer et al. [13-14] and Kesgin et al. [15] have demonstrated the effectiveness of augmentation techniques such as synonym replacement and paraphrasing in enhancing model robustness, predominantly in English-language settings. However, research on Indonesian data augmentation remains scarce, particularly in addressing classification challenges involving poetic or stylistically ambiguous expressions—highlighting an urgent gap in current literature. To address this, the present study introduces a context-aware manual augmentation strategy that incorporates pragmatic usage patterns and discourse-level variation in Indonesian texts.

This study introduces a method combining TF-IDF with manual data augmentation to enhance SVM performance in detecting standard sentences with subtle semantics or contextual ambiguity [16]. The augmentation emphasizes expressive language such as wise sayings, motivational quotes, and reflective phrases that frequently challenge traditional classifiers [17].

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

The key contributions of this study are: (1) a context-aware, manual augmentation strategy tailored to the unique characteristics of Indonesian texts, and (2) improved classification of short, ambiguous standard sentences—providing a practical basis for more effective text classification in low-resource Indonesian language settings.

II. RESEARCH METHOD

This study uses a quantitative approach with an experimental method to evaluate the performance of the Support Vector Machine (SVM) algorithm in classifying standard and non-standard sentences in Indonesian. The primary focus is on increasing the model's sensitivity to short poetic and philosophical sentences that often cause misclassification. The Overview of Research Methodologies is shown in **Error! Reference source not found.**.

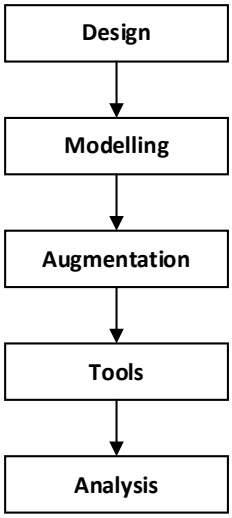

Fig 1. Overview of Research Methodology

The diagram in **Error! Reference source not found.** provides an overview of the research methodology. Detailed descriptions of each component are presented in the subsequent sections.

A. Research Design

The data used in this study consists of 5,394 Indonesian sentences, each labeled as standard and non-standard. The standard sentences were collected from formal sources such as academic documents, news, and government texts. In contrast, non-standard sentences came from social media, online forums, and informal conversations that reflect everyday language varieties.

All data underwent a preprocessing process that included text normalization (changing capital letters, removing special characters, and refining spelling) and removing punctuation and

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

numbers that were irrelevant to the classification task [18-19]. The data was divided into three parts: 70% for training, 15% for validation, and 15% for testing [20-21] . This division aims to maintain a balanced distribution between classes and avoid overfitting during training [22].

B. Modelling and Text Representation

The model used in this study is SVM with a linear kernel. The selection of SVM is based on its ability to produce an optimal hyperplane that can separate two classes with a maximum margin and its stability in handling high-dimensional data. Text representation is done using two approaches that are compared in stages:

1. Baseline: Using CountVectorizer to convert sentences into vectors based on word frequency [23].
2. TF-IDF-based Model: Using TF-IDF Vectorizer [24] to give more weight to more informative and less frequent words so that the model can be more sensitive to the difference between standard and non-standard sentences.

C. Manual Data Augmentation

To address data limitations and better handle ambiguous patterns, a series of manual data augmentation strategies [25] —chosen to precisely control linguistic variations and ensure contextual relevance—were applied, including

1. Replacing words with synonyms [26] contextually without changing the core meaning of the sentence,
2. Rearranging phrase structures to create variations of expressions that remain valid,
3. Adding new sentences with a poetic or philosophical style that represent sentence types that are often misclassified.

The following examples, as presented in **Error! Reference source not found.**, illustrate the implementation of each manual augmentation technique described previously. These instances highlight the importance of contextual relevance and stylistic variation in addressing classification challenges associated with semantically subtle or pragmatically ambiguous sentences.

Table 1. Examples of manual data augmentation techniques

	Original Sentence	**Augmented Sentence**	**English**
Synonym Replacement	*Kesabaran adalah kunci*	*Kesabaran merupakan kunci utama*	Patience is the key
Phrase Reordering	*Cinta itu sederhana*	*Sederhana itu cinta*	Love is simple
Style Injection		*Harapan sejati menerangi kehidupan*	True hope illuminates life

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

While the classification task addresses both standard and non-standard sentences, the augmentation specifically targets short standard sentences. This focus was based on an initial analysis, which showed that such sentences were frequently misclassified due to their structural similarity to informal expressions, thereby impacting model performance.

D. Tools and Experiment Environment

All experiments were conducted in Python using libraries such as Scikit-learn (for SVM implementation and evaluation) [27], NLTK and SpaCy [28] (for text preprocessing and augmentation), NumPy and Pandas [29] (for data manipulation), and Matplotlib [30] (for result visualization). Experiments were executed on Google Colaboratory, a cloud-based platform that enables efficient notebook execution without reliance on high local computing power.

E. Data Analysis

The performance evaluation was carried out using standard quantitative metrics [31], including accuracy, precision, recall, and F1-score, to comprehensively assess the model's classification capability. Furthermore, confusion matrix analysis was employed to examine the distribution of true positives, false positives, true negatives, and false negatives, offering more profound insights into model behavior [32]. Comparative analysis was performed between the baseline and the improved models, with particular attention given to persistent error patterns after applying manual augmentation techniques.

III. RESULT AND DISCUSSION

A. Performance Comparison Before and After Augmentation

Prior to model evaluation, a dataset of 5,394 sentences was meticulously prepared, comprising 2,725 standard and 2,669 non-standard Indonesian sentences to ensure balanced class representation. Although relatively modest, this dataset reflects the typical challenges of low-resource settings, where high-quality annotated linguistic data in Indonesian remains scarce.

This study consistently employed the TF-IDF Vectorizer as the feature extraction method across all experimental conditions. The core comparison examined the impact of manual data augmentation by contrasting the baseline model with the augmented variant. Results indicated a validation accuracy improvement from 96.11% to 97.39%, while the test accuracy improved from 94.39% to 96.16%, confirming that augmentation enhanced intra-split generalization without affecting overall robustness. These findings underscore the effectiveness of the augmentation strategy in addressing structurally challenging cases. The data in Table 2 summarizes the model's accuracy before and after manual augmentation.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

Table 2. Model Accuracy before and after Manual Augmentation

Experimental Setting	Accuracy (%)
SVM + TF-IDF Vectorizer (before augmentation)	94,39 (test) / 96,11 (validation)
SVM + TF-IDF Vectorizer + Manual Augmentation	96,16 (test) / 97,39 (validation)

To highlight the improvement in classification performance, the confusion matrices before and after augmentation are presented in Fig 2 and Fig 3, respectively. The detailed confusion matrix data is shown in Table 3.

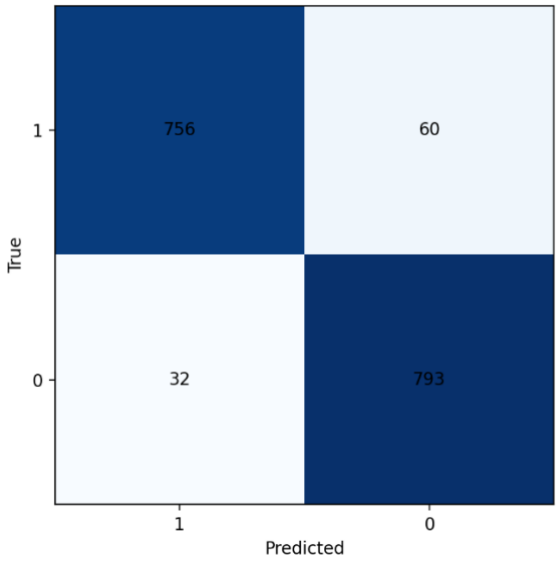

Fig 2. Classification Results before Augmentation

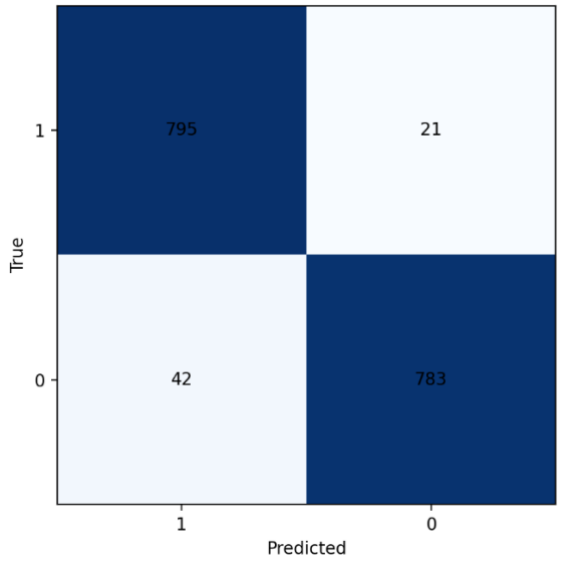

Fig 3. Classification Results after Augmentation

INTENSIF, Vol.10 No.1 2026
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

Table 3. Comparison of Augmentation Results (Based on Test Set Confusion Matrix)

	TP	FN	TN	FP	Errors
Before Augmentation	756	60	793	32	92
After Augmentation	795	21	783	42	63

The improvement observed in the confusion matrices indicates that manual data augmentation enhanced the model's precision in identifying standard sentences, as evidenced by the reduction of False Negatives from 60 to 21 and total errors from 92 to 63. The results demonstrate stronger discrimination between sentence types. The TF-IDF Vectorizer further reinforced this effect by highlighting contextually informative words. Furthermore, the augmented data broadened the model's linguistic coverage by introducing a wider variety of patterns, particularly short and ambiguous standard sentences that were underrepresented in the baseline. As a result, the model exhibited better generalization to previously challenging sentence structures.

B. Error Analysis of Ambiguous Sentences

The decrease in False Negatives from 60 to 21 indicates that ambiguous patterns in standard sentences—such as poetic expressions or informal stylistic elements—remain challenging for the model. This suggests that manual data augmentation has yet to capture these complex linguistic variations fully. In real-world applications, minimizing False Positives from 32 to 42 is crucial for ensuring data integrity, particularly in tasks such as official document analysis. Conversely, False Negatives warrant increased attention in contexts like chatbot systems, where maintaining response quality is essential.

Compared to previous approaches, the results highlight the effectiveness of TF-IDF combined with manual augmentation. However, the scalability of this method remains a limitation that should be addressed in future research. The remaining prediction errors predominantly occur in short or ambiguous standard sentences, typically characterized by simple structures and everyday vocabulary. To further illustrate the classification challenges posed by such instances, several representative examples are presented in ***Error! Reference source not found.***.

Table 4. Examples of Standard Sentences

	Indonesian	English
1	*Waktu tidak pernah menunggu siapa pun*	Time never waits for anyone
2	*Hidup adalah perjalanan untuk belajar*	Life is a journey to learn
3	*Kejujuran adalah dasar dari kepercayaan*	Honesty is the foundation of trust
4	*Setiap usaha membawa hasil yang sepadan*	Every effort brings a fair result
5	*Kedamaian dimulai dari hati yang tenang*	Peace begins with a calm heart

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

Sentences such as *"Waktu tidak pernah menunggu siapa pun"* and *"Hidup adalah perjalanan untuk belajar"* possess a poetic and philosophical tone that makes them difficult for machine learning models to classify. This complexity arises from the abstract meaning of words such as *"waktu"* and *"belajar"*, which do not explicitly indicate whether the sentence belongs to a standard or non-standard category. Machine learning models such as Support Vector Machines (SVMs) rely on concrete lexical and syntactic patterns for classification. However, in the case of these sentences, the extracted textual features are often ambiguous because similar constructions may appear in both standard and non-standard contexts.

Furthermore, the concise structure of sentences like *"Kejujuran adalah dasar dari kepercayaan"* and *"Kedamaian dimulai dari hati yang tenang"* provides limited linguistic information for the model to infer contextual nuances, leading to potential misclassification. Ambiguity in semantic usage also plays an important role. For instance, *"Setiap usaha membawa hasil yang sepadan"* can appear in formal or informal discourse depending on the situation. In casual conversation, such a sentence may be perceived as expressive or motivational, while it would still be considered grammatically standard in a formal document. This semantic overlap complicates the model's ability to draw a clear boundary between the two classes without incorporating deeper contextual or pragmatic features.

C. Effectiveness of Manual Augmentation

Furthermore, the words used in these sentences are neutral and frequently appear across standard and non-standard expressions. For instance, terms such as *"waktu"*, *"usaha"*, or *"hati"* carry general meanings that do not explicitly define the formality of a sentence. As a result, distinguishing between stylistic and contextual usage becomes challenging when relying solely on lexical features.

These limited semantic cues indicate that simple feature-based approaches, such as CountVectorizer or TF-IDF Vectorizer, cannot handle such nuanced cases. Consequently, more targeted manual augmentation—particularly through the inclusion of short, standard sentences with ambiguous or philosophical tones like *"Kedamaian dimulai dari hati yang tenang"* or *"Setiap usaha membawa hasil yang sepadan"*—is essential to expose the model to subtle linguistic variations. This augmentation strategy enables the classifier to recognize better contextual signals that differentiate standard from non-standard constructions. Furthermore, integrating context-aware modeling techniques can enhance the model's sensitivity to these ambiguous linguistic patterns.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

To address this challenge, data augmentation was performed manually with human intervention to maintain the quality and relevance of the text. A total of 100 ambiguous standard sentences were selected and augmented using the following approach:

1. Synonym Replacement, in which common words in a sentence are replaced with their synonyms to maintain the original meaning while slightly altering the structure. This technique was applied to a subset of the 100 augmented sentences. Examples of this process are:
    a. before: *Hidup adalah perjalanan untuk belajar* (Life is a journey to learn).
    b. After: *Hidup merupakan perjalanan untuk memahami makna kehidupan* (Life is a journey to understand the meaning of life).
2. Reordering phrases, where the sentence structure is rearranged to increase syntactic diversity and enrich the dataset variation. This transformation was also performed on several of the 100 manually augmented sentences. Examples of this process are:
    a. before: *Kejujuran adalah dasar dari kepercayaan* (Honesty is the foundation of trust).
    b. after: *Dasar dari kepercayaan adalah kejujuran* (The foundation of trust is honesty).

The results of this augmentation help the model understand more complex variations of standard sentences, especially ambiguous or poetic sentences that tend to be misclassified. Although small, the increase in accuracy from 94.39% before augmentation to 96.16% after augmentation (and from 96.11% to 97.39% on the validation set) shows significant improvement in overcoming False Negatives. Although Deep Learning-based models are often the first choice in natural language processing, the results of this study show that SVM can still be utilized effectively, especially in the following conditions:

1. Limited Dataset: SVM performs well on small to medium-sized datasets, such as those used in this study, because its computational complexity is lower than deep learning models.
2. Simple Feature Representation: Using TF-IDF, SVM can capture text patterns well without requiring deeper context representations like in transformer models [33].
3. Controlled Text Ambiguity: In the case of ambiguous sentences, simple data augmentation can help improve the performance of SVM at a lower computational cost.

D. Limitations and Future Work

This improvement is mainly due to manual augmentation that provides variations to the ambiguous standard sentence patterns. However, this manual augmentation is also a limitation in the study because it is impractical when applied to larger data scales. In such scenarios, NLP-

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

based automatic augmentation approaches, such as transformer models or generative algorithms, are potential solutions for the future.

The findings confirm that combining a simple feature-based approach with data augmentation can provide optimal results in classifying standard and non-standard sentences. This work offers practical contributions to applications such as official document analysis, text clustering, and classification systems that are sensitive to variations in language style in Indonesia. Future research directions, which are crucial for advancing the field, include exploring deep learning-based models to capture deeper contextual understanding and adopting automated augmentation strategies to improve efficiency at a larger scale.

While the study results are promising, it is important to note the limitations. The manual data augmentation strategy, which relies heavily on human intervention, may not be practical for large-scale or dynamic automated systems. Additionally, the TF-IDF-based text representation used in this work may not fully capture contextual meaning and pragmatic intent, particularly in short, poetic, or expressive Indonesian sentences. Furthermore, the SVM model's inability to interpret semantic relationships between words and account for more complex contextual or syntactic dimensions is a significant limitation when dealing with the flexibility inherent in Indonesian language varieties.

The flexible characteristics of Indonesian in shifting sentence functions from standard to non-standard (and vice versa) make this classification task challenging if it only relies on word frequency-based representation. Therefore, an advanced approach is needed that integrates context-based models and automatic augmentation strategies tailored to Indonesia's linguistic structure and cultural variations.

IV. CONCLUSION

This study contributes a manually crafted augmentation approach tailored to structural ambiguities in Indonesian, which has rarely been explored in previous works. The approach enables a Support Vector Machine (SVM), combined with a TF-IDF Vectorizer, to effectively classify standard and non-standard Indonesian sentences. The model achieved a high accuracy of 96.16%, with notable improvements in handling short and ambiguous standard sentences and reducing false negatives. These findings highlight the potential of traditional machine learning models in resource-constrained settings, particularly when supported by well-targeted data augmentation, as evidenced by the observed improvement in classification accuracy following manual sentence expansion. However, limitations in semantic understanding and the scalability of manual augmentation remain challenges. Future research should explore context-aware models, such as those based on contextual embeddings, to better capture the nuances of meaning

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

in Indonesian. Additionally, scalable and linguistically informed automatic augmentation methods are needed to support wider applications in dynamic and large-scale environments.

Author Contributions: *Rianto Rianto*: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision. *Eko Setyo Humanika*: Investigation, Data Curation. *Iwan Hartadi Tri Untoro*: Software, Investigation, Data Curation, Writing - Original Draft.

All authors have read and agreed to the published version of the manusripts

Funding: This research was conducted with the institutional support of Universitas Teknologi Yogyakarta. While no specific external grant was received, the authors gratefully acknowledge the resources provided by the university.

Acknowledgments: The authors would like to thank the Department of Data Science at Universitas Teknologi Yogyakarta for their administrative support. We also express our sincere gratitude to the anonymous reviewers and the editor for their constructive comments, which have significantly improved the quality of this manuscript.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this article.

Data Availability: The data used in this study comprises a compiled corpus derived from primary sources. Books, magazines, and articles were used as sources for standard sentences, while native speaker elicitation was employed to capture daily conversational usage for non-standard sentences. The compiled dataset is available from the corresponding author upon request.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Rianto Rianto: https://orcid.org/0000-0002-5058-4580
Eko Setyo Humanika: https://orcid.org/0000-0001-9789-0764
Iwan Hartadi Tri Untoro: https://orcid.org/0009-0000-3041-3381

REFERENCES

[1] S. U. Hassan, J. Ahamed, and K. Ahmad, “Analytics of machine learning-based algorithms for text classification,” *Sustainable Operations and Computers*, vol. 3, pp. 238–248, Jan. 2022, doi: 10.1016/J.SUSOC.2022.03.001.

[2] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, “Word Level Language Identification in Indonesian-Javanese-English Code-Mixed Text,” *Procedia Comput Sci*, vol. 244, pp. 105–112, Jan. 2024, doi: 10.1016/J.PROCS.2024.10.183.

[3] D. R. Febryanti, I. Hamad, and U. Rusadi, “Pemetaan Wacana Berbasis Korpus di Media Sosial,” *Jurnal ILMU KOMUNIKASI*, vol. 21, no. 1, pp. 1–18, Jun. 2024, doi: 10.24002/JIK.V21I1.6452.

[4] N. S. Nurliza, N. Hidayah, S. V. Azzahra, and B. Ginanjar, "Afiks Ng- pada Bahasa Gaul di Media Sosial Beserta Padanan Formalnya: Kajian Morfologi," *Linguistik Indonesia*, vol. 43, no. 1, pp. 81–98, Feb. 2025, doi: 10.26499/li.v43i1.673.

[5] H. Murfi, S. Theresia Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," *Appl Soft Comput*, vol. 151, p. 111112, Jan. 2024, doi: 10.1016/J.ASOC.2023.111112.

[6] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A Survey on Text Classification Algorithms: From Text to Predictions," *Information 2022*, vol. 13, no. 2, p. 83, Feb. 2022, doi: 10.3390/INFO13020083.

[7] W. F. Satrya, R. Aprilliyani, and E. H. Yossy, "Sentiment analysis of Indonesian police chief using multi-level ensemble model," *Procedia Comput Sci*, vol. 216, pp. 620–629, Jan. 2023, doi: 10.1016/J.PROCS.2022.12.177.

[8] I. S. M. Fadhil, M. H. M. Yusof, I. A. Khalid, S. H. Teoh, and A. A. Almohammedi, "Sentiment analysis comparisons across selected ml models: application on Malaysia online banking twitter data," *Procedia Comput Sci*, vol. 245, no. C, pp. 979–988, Jan. 2024, doi: 10.1016/J.PROCS.2024.10.326.

[9] K. S. B. Kharthik et al., "Transfer learned deep feature based crack detection using support vector machine: a comparative study," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–19, Jun. 2024, doi: 10.1038/s41598-024-63767-5.

[10] E. Mitreva, V. Georgiev, and A. Nikolova, "Classification of Short Noisy Text," *ACM International Conference Proceeding Series*, pp. 227–231, Jun. 2024, doi: 10.1145/3674912.3674935.

[11] B. Bakiyev, "Method for Determining the Similarity of Text Documents for the Kazakh language, Taking Into Account Synonyms: Extension to TF-IDF," *SIST 2022 - 2022 International Conference on Smart Information Systems and Technologies*, Proceedings, 2022, doi: 10.1109/SIST54437.2022.9945747.

[12] Y. Bilgen and M. Kaya, "EGMA: Ensemble Learning-Based Hybrid Model Approach for Spam Detection," *Applied Sciences 2024*, Vol. 14, Page 9669, vol. 14, no. 21, p. 9669, Oct. 2024, doi: 10.3390/APP14219669.\

[13] M. Bayer, M. A. Kaufhold, and C. Reuter, "A Survey on Data Augmentation for Text Classification," *ACM Comput Surv*, vol. 55, no. 7, Jul. 2022, doi: 10.1145/3544558/SUPPL_FILE/3544558.SUPP.PDF.

[14] M. Bayer, M. A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 1, pp. 135–150, Jan. 2023, doi: 10.1007/S13042-022-01553-3/TABLES/5.

[15] H. T. Kesgin and M. F. Amasyali, "Advancing NLP models with strategic text augmentation: A comprehensive study of augmentation methods and curriculum strategies," *Natural Language Processing Journal*, vol. 7, p. 100071, Jun. 2024, doi: 10.1016/J.NLP.2024.100071.

[16] I. Putu Widiarta Nandana Githa, A. Syananda, R. Faustine, I. S. Edbert, and D. Suhartono, "Hate Speech Classification in Indonesian Tweets Using TF-IDF and Data Augmentation," *2024 International Conference on Green Energy, Computing and Sustainable Technology*, GECOST 2024, pp. 61–65, 2024, doi: 10.1109/GECOST60902.2024.10474781.

[17] Y. C. Zhou, Z. Zheng, J. R. Lin, and X. Z. Lu, "Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking," *Comput Ind*, vol. 142, p. 103746, Nov. 2022, doi: 10.1016/J.COMPIND.2022.103746.

[18] S. Li et al., "Preprocessing of natural language process variables using a data-driven method improves the association with suicide risk in a large veterans affairs population,"

Comput Biol Med, vol. 189, p. 109939, May 2025, doi: 10.1016/J.COMPBIOMED.2025.109939.

[19] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf Syst*, vol. 121, p. 102342, Mar. 2024, doi: 10.1016/J.IS.2023.102342.

[20] S. Raza and V. Chatrath, "HarmonyNet: Navigating hate speech detection," *Natural Language Processing Journal*, vol. 8, p. 100098, Sep. 2024, doi: 10.1016/J.NLP.2024.100098.

[21] C. Bulut and E. Arslan, "Comparison of the impact of dimensionality reduction and data splitting on classification performance in credit risk assessment," *Artif Intell Rev*, vol. 57, no. 9, pp. 1–23, Sep. 2024, doi: 10.1007/S10462-024-10904-1/TABLES/9.

[22] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics*, vol. 64, no. 2, pp. 166–176, 2022, doi: 10.1080/00401706.2021.1921037.

[23] T. Turki and S. S. Roy, "Novel Hate Speech Detection Using Word Cloud Visualization and Ensemble Learning Coupled with Count Vectorizer," *Applied Sciences 2022*, Vol. 12, Page 6611, vol. 12, no. 13, p. 6611, Jun. 2022, doi: 10.3390/APP12136611.

[24] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, M. B. Ghaffar, and W. Khan, "Sentiment analysis of movie reviews based on NB approaches using TF–IDF and count vectorizer," *Soc Netw Anal Min*, vol. 14, no. 1, pp. 1–15, Dec. 2024, doi: 10.1007/S13278-024-01250-9/METRICS.

[25] M. Bayer, M. A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 1, pp. 135–150, Jan. 2023, doi: 10.1007/S13042-022-01553-3/TABLES/5.

[26] G. Chao, J. Liu, M. Wang, and D. Chu, "Data augmentation for sentiment classification with semantic preservation and diversity," *Knowl Based Syst*, vol. 280, p. 111038, Nov. 2023, doi: 10.1016/J.KNOSYS.2023.111038.

[27] S. S. Sikarwar, C. Kumar Babubhai Patel, P. U. Dhanjibhai, V. Vir Singh, A. T. Ravi, and L. Mohan Sakya, "Enhancing False News Detection through Supervised Machine Learning and NLP Techniques: A Comparative Study of Feature Extraction and Selection Methods Using Python Scikit-Learn," *Proceedings - IEEE 2024 1st International Conference on Advances in Computing, Communication and Networking, ICAC2N 2024*, pp. 1361–1366, 2024, doi: 10.1109/ICAC2N63387.2024.10895797.

[28] R. Dumbre, P. Ankalwar, S. Bhagwat, D. Pandit, M. Bhutada, and S. Gund, "SpaCy and NLTK NLP Techniques for Text Summarization: *A Comprehensive Comparison," Lecture Notes in Networks and Systems*, vol. 1149, pp. 55–64, 2025, doi: 10.1007/978-981-97-8160-7_5.

[29] C. Lompa and P. Luczynski, "Analysis and Reproducibility of 'Productivity, Portability, Performance: Data-Centric Python,'" *IEEE Transactions on Parallel and Distributed Systems*, 2024, doi: 10.1109/TPDS.2024.3366571.

[30] P. Gupta and A. Bagchi, "Data Visualization with Python," *In: Essentials of Python for Artificial Intelligence and Machine Learning. Synthesis Lectures on Engineering, Science, and Technology*, pp. 237–282, 2024, doi: 10.1007/978-3-031-43725-0_7.

[31] G. Phillips et al., "Setting nutrient boundaries to protect aquatic communities: The importance of comparing observed and predicted classifications using measures derived from a confusion matrix," *Science of The Total Environment*, vol. 912, p. 168872, Feb. 2024, doi: 10.1016/J.SCITOTENV.2023.168872.

[32] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix," *Comput Stat*, vol. 39, no. 1, pp. 363–383, Feb. 2024, doi: 10.1007/S00180-022-01301-9/TABLES/5.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v10i1.25179

[33] N. C. A. Agustina, R. Novita, Mustakim, and N. E. Rozanda, "The Implementation of TF-IDF and Word2Vec on Booster Vaccine Sentiment Analysis Using Support Vector Machine Algorithm," *Procedia Comput Sci*, vol. 234, pp. 156–163, Jan. 2024, doi: 10.1016/J.PROCS.2024.02.162.