Combination of Rule-Based and Damerau Levenshtein Distance Methods in Stemming Javanese Text

Mufidatul Nabilla Khasanah a), Christian Sri Kusuma Aditya b)

Department of Informatics, Universitas Muhammadiyah Malang, Malang, Indonesia

*a) Corresponding author: mufidatulnabilla@webmail.umm.ac.id  
b)christianskaditya@umm.ac.id*

**Abstract.** Stemming is one of the important stages in Natural Language Processing (NLP) to produce root word forms by removing prefixes and suffixes. In Javanese, the stemming process is a challenge due to the complexity of morphology and the limitations of adequate language processing tools. Javanese has a rich affixation structure, with dialectal and speech variations that do not always follow morphological patterns. In addition, language processing tools such as root word dictionaries, grammatical rules, and corpus that can be used are still very limited. Therefore, a more adaptive approach is needed according to the characteristics of the Javanese language, both rule-based and string similarity, such as Damerau Levenshtein Distance (DLD) to improve accuracy in obtaining valid base words. This research examines the combination of Rule-Based and Damerau Levenshtein Distance (DLD) methods, and evaluates their impact on text classification performance using the Support Vector Machine (SVM) method. The results showed that the proposed approach successfully identified 320 out of 540 root words correctly. In addition, the accuracy of text classification using Support Vector Machine (SVM) reached 92% after the application of this stemming.

**Keywords:** Stemming, Javanese, Rule-Based, Damerau Levenshtein Distance, SVM

# INTRODUCTION

Indonesia is a diverse country, where language has an important role as the main communication tool in everyday life. Language is an arbitrary sound symbol that people use to work, interact and identify themselves [1]. In society, it is not only limited to the use of Indonesian, but also includes regional languages. One of the languages familiar to the Indonesian people is Javanese, which is spoken by residents in Central Java, East Java, and the Special Region of Yogyakarta [2].

The diversity of languages in Indonesia, including Javanese, presents its own challenges in information processing [3]. One of the main challenges is the morphological complexity of the language, where words can have various forms of affixes that have different functions and meanings. Although digital development facilitates access to information in Javanese, it has not been followed by the ability to process it effectively, especially in overcoming language complexity. Therefore, a process is needed to return words to their basic form. This process is known as stemming [4].

Stemming is a technique used to find the root form of a word by removing prefixes, inserts, suffixes, and word combinations [5], [6], and plays an important role as the first step in information retrieval [7] [8]. Although various stemming algorithms have been developed, including for other languages [9]. However, until now the development of stemming algorithms in Javanese has not found stable and effective results (Mohammad Arifin et al., 2020) [10]. In answering the challenges of stemming in Javanese, this research takes a new approach that combines the use of rule-based and Damerau Levenshtein Distance (DLD) methods, which respectively focus on the linguistic rules in Javanese and word similarity through the closest distance.

In an effort to support this concept, a number of previous studies have been conducted to evaluate the development of stemming algorithms in Javanese. Research by Aji Prasetya Wibawa et al, (2021) [11] showed that the use of the Damerau Levenshtein Distance (DLD) algorithm obtained an accuracy of 49.6%. This research focuses on word similarity through the closest distance, in order to get the base word. Meanwhile, other research by Fatkhul Amin et al, (2017) [12] shows that the accuracy of using rule-based and string matching shows 67%. This research combines linguistic rules and word similarities. Both studies can be said to be successful, although the accuracy value is low.

One of the applications of stemming development is Javanese news classification. News classification is done as an effort to help internet users find relevant content [13], [14]. However, given the lack of Javanese news data, this research uses translated data from Indonesian news. In this research, Javanese news classification is done by dividing categories based on the information [15] contained in the content using the Support Vector Machine (SVM) method.

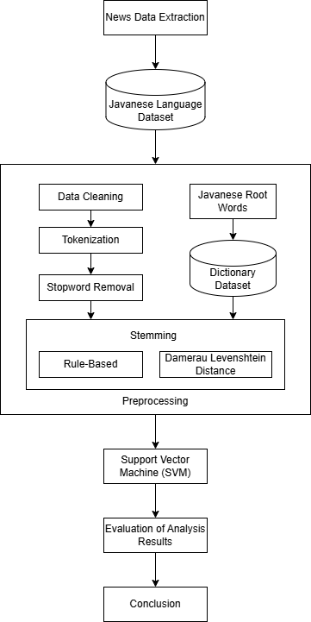
Support Vector Machine (SVM) is one of the data classification methods with high accuracy and can overcome the complexity of nonlinear problems [16]. The basic concept of SVM is to find the optimal hyperplane to separate data into positive and negative classes, and is useful for predicting new data classes [17]. In news classification, SVM analyzes text features to determine the most appropriate category [18]. Thus, SVM is the right choice for Javanese news classification, especially in overcoming the challenges of language variation and sentence structure complexity [19].

Several studies have been conducted in classifying news using the Support Vector Machine (SVM) method. In research conducted by Lalu Gias Irham, et al (2019) [20] classifying Indonesian news using the SVM method and applying the use of Mutual Information, with a research focus on knowing the effect of applying the method. In this study, the best results were achieved with an accuracy of 92.24%. Furthermore, another study was conducted by Sudianto, et al (2022) [21] by applying the SVM and Multi-Layer Perceptron (MLP) algorithms. The study, showed an accuracy value that was not much different, where MLP reached 78% and SVM reached 74%. These studies show the success of SVM in news classification, but the classification of Javanese news certainly has its own challenges.

In light of these challenges, this research aims to address the morphological variations of Javanese through the development of a more comprehensive stemming algorithm. This approach combines the strict linguistic rules of rule-based methods [22] [23] and the flexibility of the Damerau Levenshtein Distance (DLD) algorithm in handling word variations [24], [25], [26]. This combination is expected to produce a stemming algorithm that is more adaptive to various variations of the Javanese language, thus improving accuracy and consistency. The stemming algorithm that has been developed is then applied in Javanese news classification using the Support Vector Machine (SVM) method, to be used in finding news categories based on features extracted from the news text [27]. Thus, this research is expected to contribute to the development of natural language processing technology for Javanese, and support efforts to preserve cultural wealth through regional languages.

# METHODS

In this research, the methods used are explained as shown in **FIGURE 1**.



**FIGURE 1.** Research Method

**FIGURE 1** shows the flow of this research, starting from news data extraction, preparation of datasets in Javanese, pre-processing (supported by basic word dictionary), classification using SVM, to evaluation of results in the form of classification report and confusion matrix.

## Dataset

The data of this study amounted to 315 news articles taken through scraping technique from Kompas.com website. The data was collected from November 1 to 20, 2024 for three news categories: news, money, and bola. The collected data was then translated into Javanese using Google Translate, an automated language translation service that utilizes the Translation API [28]. Consequently, the original categories transformed into *pawarta*, *dhuwit*, and *bola*. **TABLE 1** presents an example of the translated data.

**TABLE 1.** Dataset in Javanese

|  |  |
| --- | --- |
| **Content** | **Category** |
| JAKARTA, KOMPAS.com - Calon gubernur DKI Jakarta nomer 2 Dharma Pongrekun badhe dandosaken cara nyambut damel satuan polisi pamong praja (satpol pp) supados dados langkung beradab dhateng ajengipun. … | *pawarta* |
| JAKARTA, KOMPAS.com - PT ABM Investama Tbk (ABMM) napakasmakaken fasilitas kredit senilai US$395 yuta kaliyan PT Bank Mandiri (Persero) Tbk. … | *dhuwit* |
| KOMPAS.com - Kompetisi Liga Utama Inggris badhe mlebeti ing minggu kaping 10 ing dina Setu (2/11/2024). Minggu iki bakal nampilake pirang-pirang duel sing nyenengake kalebu pertandhingan ageng antawis Manchester United vs Chelsea. … | *bola* |

## Preprocessing

Preprocessing is an important stage in data processing to convert unstructured raw text into a more structured format. This stage is crucial to ensure the data is ready to be used in the next stage of classification or modeling. As shown in **FIGURE 2**, the preprocessing process is done sequentially, starting from data cleaning, tokenization, stopword removal, to stemming.



**FIGURE 2.** Preprocessing Steps

Data cleaning is the process of removing symbols, punctuation marks, and numbers in the original news text. In this process URLs and abbreviations are also removed, with the aim of maintaining the semantic value that exists in each sentence [18]. In this process, case folding is also applied for word uniformity, where letters or words that were originally capitalized (uppercase) become lowercase.

**TABLE 2.** Data Cleaning

|  |  |
| --- | --- |
| **Before Data Cleaning** | **After Data Cleaning** |
| "Yen pancen ora ana, banjur maju, aku uga arep lunga, aku ora pengin metu saka rapat KJP," ujare Deweke. | yen pancen ora ana banjur maju aku uga arep lunga aku ora pengin metu saka rapat ujare deweke |
| Nanging, nasib ala nalika nguatake tim nasional Inggris ngalang-alangi kesempatan kanggo menang penghargaan individu. | nanging nasib ala nalika nguatake tim nasional inggris ngalangalangi kesempatan kanggo menang penghargaan individu |
| Dharmasasampunipun mireng curhatan para pedagang kopi keliling utawi jalak sing ngaku kerep diusir Satpol PP. | dharmasasmpunipun miring curhatan para pedagang kopi keliling utawi jalak sing ngaku karep diusir satpol |

Tokenization is the process of breaking sentences into pieces of words to facilitate subsequent processing [13]. This process is implemented using the NLTK module, which separates tokens based on non-letter characters.

**TABLE 3.** Tokenization

|  |  |
| --- | --- |
| **Before Tokenization** | **After Tokenization** |
| nanging dheweke kuwatir yen salah sijine | [nanging, dheweke, kuwatir, yen, salah, sijine] |
| menang gelar lan amarga kerja tim sing kuwat | [menang, gelar, lan, amarga, kerja, tim, sing, kuwat] |
| wektune mung diwiwiti enem menit liwat | [wektune, mung, diwiwiti, enem, menit, liwat] |

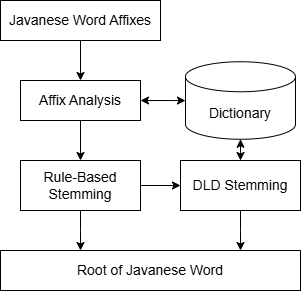
Stopword Removal is a process in text processing to remove common and irrelevant words (such as conjunctions) from a sentence or document [18]. For Javanese texts, the implementation process is done by preparing a special stopword list. This was necessary given the limited availability of language processing tools for Javanese. Some examples of words in this list are *lan*, *ing*, *saka*, *iki*, and *sing*.

**TABLE 4.** Stopword Removal

|  |  |
| --- | --- |
| **Before Stopword** | **After Stopword** |
| [diowahi, maneh, yaiku, sing, diilangi, mung, triliun, bisa, dijupuk, saka, dana] | [diowahi, diilangi, triliun, bisa, dijupuk, dana] |
| [saiki, krasa, ora, kepenak, sing, ndadekake, dheweke, rumangsa, dadi, pemain] | [saiki, krasa, kepenak, ndadekake, dheweke, rumangsa, pemain] |
| [langkah, sing, tepat, kanggo, sing, pengin, nambah, kinerja, ing, manajemen, anyar] | [langkah, tepat, kanggo, pengin, nambah, kinerja, manajemen, anyar] |

Stemming is an important stage in text processing that aims to find the root word by removing prefixes and suffixes [5]. This process utilizes certain algorithms or rules [6]. In Natural Language Processing (NLP), stemming is widely used in many text analysis and information retrieval systems [8]. There are several frequently used stemming algorithms such as Porter, Lovins, Dawson, and Husk, which have also been developed for various languages such as Arabic, Hindi, Uzbek, and Indonesian [11]. However, until now, stemming for Javanese has not been effectively developed.

The stemming process in Javanese is done by utilizing a dictionary. This dictionary is sourced from a dataset containing a list of Javanese root words taken from the Javanese-Indonesian dictionary website [29]. This dictionary serves as the main reference in validating the acquisition of stemming results is the correct base word or not.

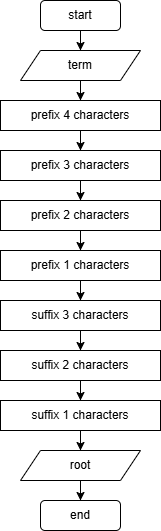


**FIGURE 3.** Diagram of Stemming Process

This root word dictionary is compiled manually, where the spelling of each entry is equalized (from A-Z) and double entries are deleted. The final result of this process is a dictionary containing approximately 36,896 base word entries, in a one-column list called aksara in JSON format. This root word dictionary will be applied in this stemming using rule-based method and Damerau Levenshtein Distance (DLD).

The application of stemming using the rule-based method is based on the morphological approach of the Javanese language, which uses rule-based analysis to find root words [12]. Rule-based is a method that uses rules as a representation of knowledge to be implemented into the system [23]. Its main advantage is simplicity of implementation, but it becomes less effective at high complexity because it relies heavily on human reasoning in processing data and making decisions. As in previous studies, the following flowchart of the stemming process using the rule-based method can be seen in **FIGURE 4**.

The stemming process using rule-based is done sequentially starting from removing prefixes, suffixes, until the root word is found. This affix removal is done sequentially based on the highest number of characters (3 characters, 2 characters, 1 character). Details of the types of prefixes and suffixes are presented in **TABLES 5** and **6** [12].



**FIGURE 4.** Flowchart of The Rule-Based Stemming Method

**TABLE 5.** Types of Prefixes in Javanese

|  |  |  |
| --- | --- | --- |
| **Types of Prefixes** | **Prefix** | **Example** |
| Prefix 4 Characters | kuma- | kuma + wani = *kumawani* |
| kapi- | kapi + lare = *kapilare* |
| Prefix 3 Characters | dak- | dak + rungu = *dakrungu* |
| kok- | kok + jupuk = *kokjupuk* |
| pan- | pan + jaluk = *panjaluk* |
| pra- | pra + lambang = *pralambang* |
| tar- | tar + buka = *tarbuka* |
| tak- | tak + jupuk = *takjupuk* |
| tok- | tok + simpen = *toksimpen* |
| Prefix 2 Characters | di- | di + tuku = *dituku* |
|  | ka- | ka + wujud = *kawujud* |
|  | ke- | ke + pungkur = *kepungkur* |
|  | ma- | ma +gawe = *magawe* |
|  | pa- | pa + warto = *pawarto* |
|  | pi- | pi + wulang = *piwulang* |
|  | sa- | sa + cedhak = *sacedhak* |
| Prefix 1 Characters | a- | a + marga = *amarga* |

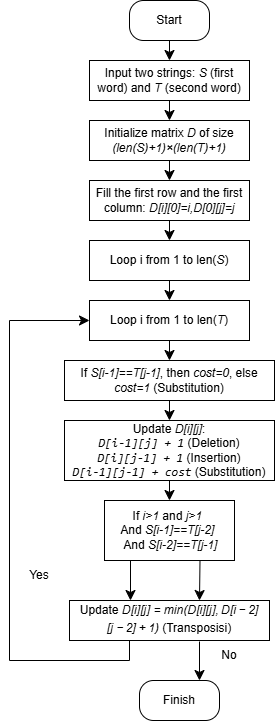
**TABLE 6.** Types of Suffixes in Javanese

|  |  |  |
| --- | --- | --- |
| **Types of Suffixes** | **Suffix** | **Example** |
| Suffix 3 Characters | -ana | silih + ana = *silihana* |
| -ane | dukung + ana = *dukungana* |
| Suffix 2 Characters | -ake | menang + ake = *menangake* |
| -an | wiwit + an = *wiwitan* |
| -na | gambar + na = *gambarna* |
| -ne | suwe + ne = *suwene* |
| -ku | montor + ku = *montorku* |
| -ke | dhewe + ke = *dheweke* |
| Suffix 1 Characters | -a | tuku + a = *tukua* |
|  | -e | omah + e = *omahe* |
|  | -i | tandur + I = *tanduri* |

### Stemming using Damerau Levenshtein Distance (DLD)

Damerau Levenshtein Distance (DLD) is an algorithm that applies insertion, deletion, substitution, and transposition operations to check spelling against word separation errors [11]. This algorithm is a development of Levenshtein Distance, which works to measure the level of similarity between strings [25]. The addition of the transposition operation is what gives DLD an edge, especially in handling spelling correction or word separation errors [26]. DLD works by calculating the closest distance based on the words contained in the base word dictionary, starting from the first to the last index in each column.

Based on the pseudocode by Aji Prasetya Wibawa [11], the stages of the Damerau Levenshtein Distance (DLD) algorithm are summarized in **FIGURE 5**.



**FIGURE 5.** Flowchart of The DLD Process

## Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification method that aims to find the optimal hyperplane to separate between two classes, with the aim of maximizing the margin, through the farthest distance to the nearest data point from each class [27]. In inputting SVM classification, it is necessary to ensure that the data is in numerical form. Term Frequency - Inverse Document Frequency (TF-IDF) is a feature extraction technique used to convert text into numeric vectors and calculate the weighted value of each feature based on word frequency. The selection of this method is based on the ability to overcome high-dimensional problems to optimize the separation between classes in text data processing.

In data classification [21], the optimal separation hyperplane is denoted by ℎ, determined by the value of 𝑥 where the decision function 𝐷(𝑥) is equal to zero. With a training dataset (𝑥I;𝑦i) from the input space 𝑥, the decision function is defined as follows:

(1)

To find the optimal value of ∝\_i^\*, SVM solves the following quadratic optimization problem:

(2)

With the subject to be restricted to:

(3)

## Evaluation

The final stage of this research is an evaluation aimed at determining the accuracy of the developed model [21] on Javanese text data that has undergone stemming. This evaluation was conducted indirectly, by observing the impact of the stemming results on classification performance using the SVM method. Thus, the success of the stemming method is measured by its contribution to improving classification accuracy, rather than by linguistic validity.

The evaluation metrics used include Precision, Recall, F1-Score, and Accuracy. Precision, as described in equation (4), is the proportion of positive data that is correctly predicted out of all data classified as positive. Recall, as formulated in equation (5), is the proportion of positive data successfully identified from all actual positive data. F1-Score, in (6), is the average obtained from the precision and Recall values. Meanwhile, Accuracy, as in equation (7), indicates the proportion of data correctly predicted from the total data tested.

The Confusion Matrix is used to improve model performance evaluation. This matrix provides a visual representation of the classification process prediction performance [30]. This method provides a detailed description of the model's classification performance by identifying four key values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP is the number of positive data predicted correctly. TN is the number of negative data predicted incorrectly. FP is the number of positive data with incorrect values. FN is the number of negative data with correct values.

*Precision =* (5)

*Recall =* (6)

*F1-Score =* (7)

*Accuracy =* (8)

# RESULTS AND DISCUSSION

The results of this study are useful for assessing the effectiveness of the application of the Javanese stemming method, before then evaluating its impact on the performance of text classification using the SVM method. It is important to understand that the dataset to be used is translated data. For this reason, it is necessary to conduct an initial exploration using preprocessing to gain an understanding of the characteristics and representation of the dataset. This stage is useful to ensure that the data is in a clean and structured form, so that it can be optimally applied in the stemming process.

The implementation of the stemming method in Javanese text aims to return words to their root form. The stemming testing process was conducted on three approaches, namely the Rule-Based method, the Damerau Levenshtein Distance (DLD) algorithm, and a combination of both.

The implementation process of both stemming method approaches relies heavily on the use of dictionaries. In the rule-based stemming method, each preprocessed word will be matched with a dictionary to find the root word. If not found, the affix removal process will be performed. Meanwhile, in the Damerau Levenshtein Distance (DLD) stemming method, the base word search is also done by dictionary matching, but with a distance tolerance of no more than 2 characters.

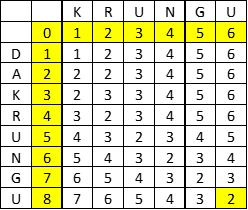
In the combination of rule-based and Damerau Levenshtein Distance (DLD) method, the application of stemming is done sequentially. The process starts with a rule-based approach to find the base word based on Javanese morphology. If a word fails to be stemmed by rule-based and is identified as not a base word, then DLD is used. In this combination, word matching in DLD is limited to the closest distance of 1 character. The results of this stemming method approach are presented in **TABLE 7**.

**TABLE 7.** Comparison of Stemming Result Words

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Initial Word** | **Rule-Based Result** | **DLD Result** | **Rule-Based and DLD Combination Result** |
| 1. | *nutup* | *nutup* | *tutup* | *tutup* |
| 2. | *diowahi* | *owah* | *dhowah* | *owah* |
| 3. | *dakrungu* | *rungu* | *krungu* | *rungu* |
| 4. | *wiwitan* | *wiwit* | *wiwit* | *wiwit* |
| 5. | *ngganti* | *ngganti* | *ganti* | *ganti* |

In **TABLE 7**, it shows each stemming approach results in different base word retrievals. Rule-based methods are accurate for words with affixes that follow morphological rules, e.g. “diowahi” becomes ‘owah’, but is less flexible for variations such as “ngganti”. In contrast, the Damerau Levenshtein Distance (DLD) excels at finding similarities across word variations, such as “ngganti” successfully stemming to “ganti”, although sometimes the results have different meanings, for example in ‘dhowah’ derived from “diowahi”. The combination of these two methods utilizes the strengths of each, resulting in better performance, as evidenced by the success of “diowahi” becoming “owah” and ‘ngganti’ becoming “ganti”. This demonstrates that while DLD might yield high scores, it doesn't always guarantee accurate root word retrieval, whereas the more rigid rule-based approach, despite its limitations, often provides better root word identification. The combined approach, by integrating the strengths of both, ultimately offered a more robust solution.

To identify the word search process using DLD, a letter distance matrix needs to be created. Damerau Levenshtein Distance is a distance matrix that compares a series of strings to detect typing errors [26]. For example, the word *dakrungu* stemmed using DLD can produce *krungu*, even though krungu is not the base word but has the closest distance of 2.



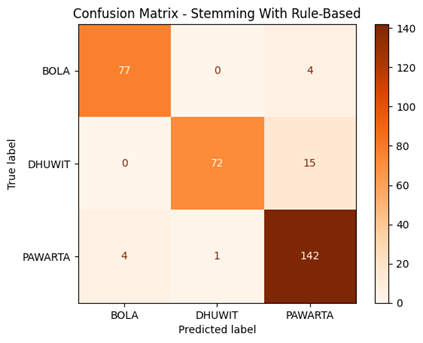
**FIGURE 6.** Stemming Process using DLD

Based on the results obtained from each stemming approach, the data is then evaluated using Support Vector Machine (SVM) for category grouping in Javanese text classification. Each stemming result must be ensured in numerical form. The process of converting text data into numerical representation uses TF-IDF (Term Frequency - Inverse Document Frequency), a feature extraction technique that aims to give weighted values based on word frequency. Once the data is ready, the next step is to train and evaluate the overall performance of the SVM method.

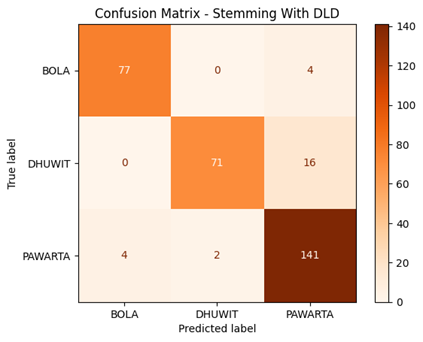
In this research, cross-validation is used as an evaluation technique to measure model performance. This cross-validation technique divides the dataset into several parts (folds). At each iteration, part of the data is used as test data as well as validation data, to measure the performance of the model that has been trained on the training data. The model used is LinearSVC, which is an SVM implementation with a linear kernel. The choice of this kernel type is based on its lighter and more efficient implementation.

With the model and data that has been prepared, the text classification process using SVM aims to measure the effectiveness of each stemming method in the three previous approach scenarios, namely using the rule-based method, Damerau Levenshtein Distance, and a combination of both. Classification through all three approaches resulted in an accuracy of 92%. This relatively equal result is due to the identical words detected by the classification model despite the difference in stemming results. In addition, most of the words are already in their base word form or have not undergone significant changes, so stemming variations do not have a significant impact on the features used in the classification process.

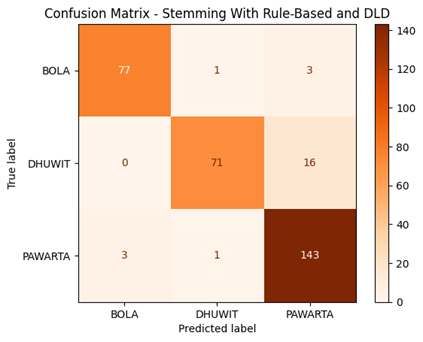
To further analyze the classification performance of each stemming approach, Figure 7, Figure 8, and Figure 9 respectively present the confusion matrix of the Javanese text classification results using the combination of Rule-Based and DLD, DLD only, and Rule-Based stemming methods. The confusion matrix provides a detailed overview of the number of True Positives, True Negatives, False Positives, and False Negatives for each category (*bola*, *dhuwit*, and *pawarta*), enabling a comprehensive evaluation of the accuracy and types of errors made by the model in each scenario.



**FIGURE 7.** Confusion Matrix Stemming with Rule-Based



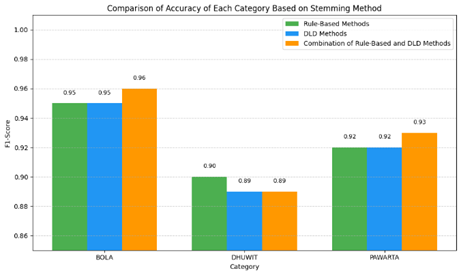
**FIGURE 8.** Confusion Matrix Stemming with DLD



**FIGURE 9.** Confusion Matrix Stemming with Rule-Based and DLD

The confusion matrix analysis of the three stemming approaches showed interesting patterns of classification performance in each category, with nuanced differences in the error distribution. subject the *bola* category showed very high classification consistency and accuracy across all scenarios, with a stable number of True Positives and minimal False Positives (i.e., 77 TPs and very few FPs). In contrast, the *dhuwit* category was consistently the biggest challenge, with the most significant number of False Negatives (misclassified as other categories, especially *pawarta*) across all three stemming methods. Meanwhile, the *pawarta* category also showed strong classification performance with a high number of correct predictions (TP) and relatively low errors. The small differences in the specific numbers between methods in each category (e.g., a difference of 1-2 errors) show that although the stemming characteristics of each method slightly affect the error distribution, the effect is not large enough to drastically change the accuracy.

To present more specific quantitative results, the classification accuracy of the three stemming approaches can be seen in **FIGURE 10**.



**FIGURE 10.** Comparison of Stemming Approach Categories

Further analysis of the accuracy values shows the potential advantages of the rule-based method and the combination method, especially in understanding and interpreting word context and identifying root words in complex word variations. As shown in **TABLE 7**, stemming results tend to be better when the rule-based and DLD methods are used together. This is also reflected in **FIGURE 10**, where the combined method produced the highest F1-score in the *bola* (0.96) and *pawarta* (0.93) categories, although it experienced a slight decrease in the *dhuwit* category. This finding suggests that the effectiveness of stemming is strongly influenced by the morphological and vocabulary characteristics of each news category, not just by the level of word similarity.

To gain a deeper understanding of the performance of the combination method, it is crucial to analyze the number of words that are successfully stemmed and those that are not. As shown in Table 8, the combination method was able to identify 320 correct stemmed words out of a total of 540 words.

**TABLE 8.** Number of Successful Stemming Words

|  |  |  |
| --- | --- | --- |
| **Content** | **Word Count** | **Number of Correct Stemming** |
| mratelakake pihake ngaturake rembug babagan putusan babagan pemilihan kepala daerah pilkada taun ditindakake penyelenggara … | 100 | 66 |
| babak regional dianakake enem kutha tim melu dening wilayah diwakili dening diwakili dening ana wilayah perwakilan wilayah … | 99 | 60 |
| ngumumake pembagian dividen final tunai senilai milyar dolar sekitar triliun ngacu kurs dolar keputusan disepakati deviden gedhe … | 107 | 62 |
| pengamanan pelaksanaan di tiga provinsi yaitu dan personel kalebu personel ... | 91 | 46 |
| emas sansaya populer amarga nduweni macemmacem kaluwihan ndadekake pilihan akeh wong .. | 143 | 86 |

Overall, this research successfully explored and analyzed the effectiveness of various stemming approaches in the context of Javanese for text classification. Although the classification accuracy generally shows similar results among rule-based, Damerau Levenshtein Distance (DLD), and combination methods, an in-depth analysis of the confusion matrix and the number of successfully stemmed words (as shown in **TABLE 8**) indicates that the combination method has more potential in handling the morphological complexity of Javanese. The effectiveness of stemming is highly influenced by the characteristics of the data. In the future, it is necessary to develop more adaptive stemming methods, by integrating machine learning or enriching representative datasets, so that it will be crucial to significantly improve the performance of Javanese NLP.

# CONCLUSION

This research aims to assess the effectiveness of Javanese stemming method implementation by applying rule-based approach and Damerau Levenshtein Distance (DLD), and evaluate its impact on text classification performance using Support Vector Machine (SVM) method. The application was carried out by implementing three scenarios, through the rule-based approach, DLD, and a combination of both. The evaluation results showed that the combined rule-based and DLD approaches successfully stemmed 320 out of a total of 540 words. However, in terms of classification performance, this combination approach did not show any significant improvement, with a consistent accuracy of 92% across all three scenarios. This indicates that the combination of rule-based and DLD has not been able to consistently improve the overall classification quality.

Nonetheless, the application of the SVM algorithm proved to be able to handle local language text classification well, supported by an accurate preprocessing stage, especially in stemming. The results of this study underline that the quality of stemming affects the performance of classification methods. Thus, this study concludes that the combination of rule-based and DLD can be a potential alternative approach to improve the comprehensiveness of Javanese text stemming and classification. Thus, this research is expected to provide an important contribution in efforts to strengthen Natural Language Processing for Javanese, especially in the development of regional language-based information systems.

# References

1. V. C. M, R. Rudy, and D. S. Naga, "Fast and accurate spelling correction using trie and Damerau-levenshtein distance bigram," *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol. 16, no. 2, pp. 827-833, 2018, doi: 10.12928/TELKOMNIKA.v16i2.6890.

2. D. Soyusiawaty, A. H. S. Jones, and N. L. Lestariw, "The stemming application on affixed javanese words by using nazief and adriani algorithm," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 771, no. 1: IOP Publishing, p. 012026, doi: 10.1088/1757-899X/771/1/012026.

3. R. A. Pratama, A. A. Suryani, and W. Maharani, "Part of Speech Tagging for Javanese Ngoko Language with Hidden Markov Model," J-Cosine(Journal of Computer Science and Informatics Engineering), vol. 4, doi: <https://doi.org/10.29303/jcosine.v4i1.346>.

4. D. E. Cahyani, L. M. T. Utami, and H. Setiadi, "Clustering of Javanese News in Krama Alus Level with Javanese Stemming," in 2019 International Conference on Information and Communications Technology (ICOIACT), 2019: IEEE, pp. 462-467.

5. N. Hidayatullah, A. P. Wibawa, and H. A. Rosyid, "Penerapan ECS Stemmer untuk Modifikasi Nazief & Adriani Berbahasa Jawa," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi),* vol. 3, no. 3, pp. 343-348, 2019, doi: 10.29207/resti.v3i3.994.

6. S. I. Melia, J. Sholihah, D. Nisak, I. S. Juniaristha, and A. T. Ni'mah, "The Ngoko Javanese Stemmer uses the Enhanced Confix Stripping Stemmer Method," *Rekayasa,* vol. 16, no. 1, pp. 107-112, 2023.

7. A. P. Wibawa, F. A. Dwiyanto, I. A. E. Zaeni, R. K. Nurrohman, and A. Afandi, "Stemming javanese affix words using nazief and adriani modifications," *J. Inform,* vol. 14, no. 1, p. 36, 2020, doi: 10.26555/jifo.v14i1.a17106.

8. Z. Abidin and A. Junaidi, "Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review," *Journal of Information Systems Engineering and Business Intelligence,* vol. 10, no. 2, pp. 217-231, 2024, doi: <http://dx.doi.org/10.20473/jisebi.10.2.217-231>.

9. D. Mustikasari, I. Widaningrum, R. Arifin, and W. H. E. Putri, "Comparison of effectiveness of stemming algorithms in indonesian documents," in 2nd Borobudur International Symposium on Science and Technology (BIS-STE 2020), 2021: Atlantis Press, pp. 154-158.

10. M. A. Nq, L. P. Manik, and D. Widiyatmoko, "Stemming Javanese: Another Adaptation of the Nazief-Adriani Algorithm," in 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020: IEEE, pp. 627-631.

11. A. P. Wibawa and M. N. m. Hakim, "Stemming Bahasa Jawa Menggunakan Damerau Levenshtein Distance (Dld)," *Jurnal Teknik Informatika,* vol. 14, no. 1, pp. 22-27, 2021, doi: 10.15408/jti.v14i1.15010.

12. F. Amin, W. HADIKURNIAWATI, S. WIBISONO, H. FEBRUARIYANTI, and J. S. WIBOWO, "A HYBRID METHOD OF RULE-BASED AND STRING MATCHING STEMMER FOR JAVANESE LANGUAGE," *Journal of Theoretical & Applied Information Technology,* vol. 95, no. 19, 2017.

13. I. A. Ropikoh, R. Abdulhakim, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) untuk Klasifikasi Berita Hoax Covid-19," *Journal of Applied Informatics and Computing,* vol. 5, no. 1, pp. 64-73, 2021.

14. J. Ahmed and M. Ahmed, "Online news classification using machine learning techniques," IIUM Engineering Journal, vol. 22, no. 2, pp. 210-225, 2021.

15. H. B. Dogru, S. Tilki, A. Jamil, and A. A. Hameed, "Deep learning-based classification of news texts using doc2vec model," in 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), 2021: IEEE, pp. 91-96.

16. T. S. Sabrila, Y. Azhar, and C. S. K. Aditya, "Analisis Sentimen Tweet Tentang UU Cipta Kerja Menggunakan Algoritma SVM Berbasis PSO," JISKA (Jurnal Informatika Sunan Kalijaga), vol. 7, no. 1, pp. 10-19, 2022.

17. S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification," in 2019 International conference on intelligent sustainable systems (ICISS), 2019: IEEE, pp. 24-28.

18. L. Triyono, R. Gernowo, M. Rahaman, and T. R. Yudantoro, "Indonesian Fake News Detection Using Various Machine Learning Technique," International Journal on Informatics Visualization, vol. 7, no. 3, pp. 726-732, 2023.

19. N. Kalcheva, M. Karova, and I. Penev, "Comparison of the accuracy of SVM kemel functions in text classification," in 2020 International Conference on Biomedical Innovations and Applications (BIA), 2020: IEEE, pp. 141-145.

20. L. Irham, A. Adiwijaya, and U. Wisesty, "Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine. JURNAL MEDIA INFORMATIKA BUDIDARMA, 3 (4), 284–292," ed, 2019.

21. S. Sudianto, A. D. Sripamuji, I. R. Ramadhanti, R. R. Amalia, J. Saputra, and B. Prihatnowo, "Penerapan Algoritma Support Vector Machine dan Multi-Layer Perceptron pada Klasifikasi Topik Berita," *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI,* vol. 11, no. 2, pp. 84-91, 2022.

22. A. Fahda and A. Purwarianti, "A statistical and rule-based spelling and grammar checker for Indonesian text," in 2017 international conference on data and software engineering (icodse), 2017: IEEE, pp. 1-6.

23. M. A. P. Subali and C. Fatichah, "KOMBINASI METODE RULE-BASED DAN N-GRAM STEMMING UNTUK MENGENAL STEMMER BAHASA BALI," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK),* vol. 6, no. 2, 2019.

24. S. Yuliani, S. Sahib, M. Abdollah, Y. Wijaya, and N. Yusoff, "Hoax news validation using similarity algorithms," in Journal of Physics: Conference Series, 2020, vol. 1524, no. 1: IOP Publishing, p. 012035.

25. P. Santoso, P. Yuliawati, R. Shalahuddin, and A. P. Wibawa, "Damerau Levenshtein Distance for Indonesian Spelling Correction," *Jurnal Informatika Ahmad Dahlan,* vol. 13, no. 2, pp. 11-15, 2019.

26. B. V. Indriyono, "Kombinasi damerau levenshtein dan jaro-winkler distance untuk koreksi kata bahasa inggris," Jurnal Teknik Informatika Dan Sistem Informasi, vol. 6, no. 2, 2020.

27. D. Tuhenay, "Perbandingan Klasifikasi Bahasa Menggunakan Metode Naïve Bayes Classifier (NBC) Dan Support Vector Machine (SVM)," JIKO (Jurnal Informatika dan Komputer), vol. 4, no. 2, pp. 105-111, 2021.

28. M. Maryamah, A. Z. Arifin, R. Sarno, and A. M. Hasan, "Adapting Google translate using dictionary and word embedding for Arabic-Indonesian cross-lingual information retrieval," in *2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 2021: IEEE, pp. 205-209.

29. Balai Bahasa Provinsi Daerah Istimewa Yogyakarta, "Kamus Bahasa Jawa-Indonesia (KBJI)," Kementerian Pendidikan Dasar dan Menengah, [*Online*]. Tersedia:  [https://kbji.kemendikdasmen.go.id/](https://kbji.kemdikbud.go.id/).

30. L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: a review with examples from ecology," Ecological Monographs, vol. 93, no. 1, p. e1557, 2023.