Type of contribution:
Editorial
▶ Research Paper
Case Study
Review Paper
Scientific Data
Report of Tech. Application

ECSITE

Edutran Computer Science & Information Technology
Vol. 3, No. 1 (2025) pp 1-10
e-issn. 7289-2554

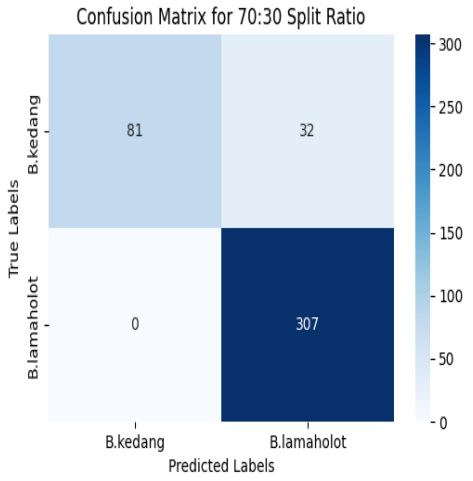
CLASSIFICATION OF REGIONAL LANGUAGE TEXTS IN LEMBATA REGENCY USING NAÏVE BAYES WITH TF-IDF FEATURES

Bartolomeus Wadan Ladopurab[1*], Aviv Yuniar Rahman[2], Rangga Pahlevi[3]

[1] Teknik Informatika, University Widyagama, 65128, Indonesia
[2] Teknik Informatika, University Widyagama, 65128, Indonesia
[3] Teknik Informatika, University Widyagama, 65128, Indonesia

bertopurab@gmail.com

This article contributes to:



Highlights:

- To identify the Kedang and Lamaholot languages using the Naïve Bayes method with TF-IDF features.
- The model achieved 93% accuracy in classifying both languages.
- The data is divided into 80% for training and 20% for testing.

Abstract

Language has an important role in human life. With language, humans can communicate and exchange ideas with each other. However, the diversity of ethnic groups in Indonesia means that Indonesia has a variety of regional languages, therefore regional languages can make it difficult to convey information and communication. This research aims to identify the Kedang language and the Lamaholot language in text form. Identification is carried out to find out the language of each region using computerized technology. This identification uses a classification technique using a method, namely NAIVE BAYES WITH TF-IDF FEATURES. This method is used to identify the language according to the text that has been entered and then calculate the accuracy value. The identified data is 2000 sentences, so it can be seen which methods are effective and can be used to identify language. The research results found that this method was quite effective in identifying Kedang language with an accuracy value of 0.93 or 93%. And Lamaholot language with an accuracy value of 0.93 or 93%. And there are 63 examples of Kedang language and 58 examples of Lamaholot language

Keywords: Regional Languange Classification, TF-IDF, Naïve Baye

Article info
Submitted: 2025-04-21
Revised: 2025-05-06
Accepted: 2025-05-07

Publisher
Edutran Academic Publisher

1. Introduction

Language serves as a vital component of regional culture, encapsulating the identity and distinctiveness of each ethnic group or community. In Indonesia, renowned for its rich ethnic and cultural diversity, over 700 regional languages are spoken across various regions. Among these, Kedang and Lamaholot are significant regional languages utilized by communities in the Flores Islands, located in East Nusa Tenggara. However, these languages remain relatively obscure outside their native-speaking areas, underscoring the necessity for enhanced efforts in their documentation and preservation[1], [2]. The preservation of such languages is crucial not only for maintaining cultural heritage but also for fostering inclusive communication within Indonesian society.

Natural Language Processing (NLP) emerges as a pivotal discipline in the contemporary era, particularly concerning the preservation and promotion of less commonly spoken languages. Text classification is a primary application within NLP, designed to categorize written texts into predefined categories or labels. This technique possesses broad utility, ranging from language detection and document clustering to sentiment analysis[3]. As methods in text classification continue to advance, it becomes increasingly evident that applications specific to regional languages, like Kedang and Lamaholot, are essential to leverage digital technology for language preservation. The integration of NLP technologies can enhance the accessibility and recognition of these regional languages in digital spaces that dominate contemporary communication[4], [5].

The Naive Bayes algorithm serves as an effective method for text classification within the NLP realm. Known for its probabilistic approach, Naive Bayes is celebrated for its combination of efficiency and accuracy across various classification tasks, particularly in text data scenarios[6]. The algorithm evaluates the probability of a given document belonging to a specific category based on the presence of certain features, making it conducive for tasks involving linguistic data. By applying the Naive Bayes algorithm, researchers can facilitate the identification and classification of the Kedang and Lamaholot languages, thus sustaining linguistic diversity and contributing to the larger body of NLP literature that presently bears limited studies on regional languages[7], [8].

One foundational aspect of the Naive Bayes algorithm's effectiveness in text classification lies in feature extraction methods, specifically TF-IDF (Term Frequency-Inverse Document Frequency). The TF-IDF technique quantifies the importance of a word in a document relative to a collection of documents, gauging both the frequency of the term and its inverse frequency across the corpus[9]. This statistical measure is instrumental in distinguishing salient features from noise within text, thus enhancing the performance of classification models. When coupled with the Naive Bayes algorithm, the TF-IDF approach has yielded promising results across a myriad of classification applications, evidencing its robustness within diverse contexts[10], [11]. By harnessing these combined methodologies, researchers can create a sophisticated classification model capable of effectively discerning between texts written in Kedang and Lamaholot.

In examining the potential outcomes of applying the Naive Bayes algorithm alongside TF-IDF for classifying Kedang and Lamaholot languages, the implications extend beyond merely creating a linguistic model. The anticipated classification model, rooted in well-researched methodologies, is aimed at preserving these regional languages while simultaneously lending support for the development of language-centric technologies tailored to their respective speaking communities. Enabling such technologies not only fosters communication but also enriches the cultural fabric of the region through the preservation of linguistic diversity. Furthermore, this endeavor contributes significantly to the academic discourse in NLP, illuminating the potential and challenges involved in working with underrepresented languages in the field of computational linguistics[12], [13].

2. Methods

Figure 1. Research Flow Chart

This research was conducted through several main stages. The first stage is data collection, which consists of Kedang and Lamaholot language texts gathered from various sources such as books, articles, and interviews. Next, data preprocessing is performed to clean the text and prepare it for analysis. Then, feature extraction is carried out using the TF-IDF method, which aims to represent the text in a numeric form so that it can be processed by the model. After the features are extracted, the Naïve Bayes algorithm is applied to perform language classification[14], [15]. The final stage is model evaluation, where the system's performance is assessed using various evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The research stages used in this study are shown in Figure 1 below.

2.1 Data Collection

Data collection in this study involved obtaining texts in Kedang and Lamaholot languages from various sources, such as books, articles, and interviews. Each document in the dataset was labeled according to the language used to facilitate the classification process. Overall, the dataset used consists of 1000 Kedang language documents and 1000 Lamaholot language documents. The data was then divided into two subsets: 80% for training data and 20% for test data. This division aims to ensure that the model can be trained with sufficient data and tested with representative data to assess its performance. Before being used in the analysis, the data was also reformatted to ensure consistency, including the removal of irrelevant symbols, elimination of extra spaces, and normalization of letters to lowercase to facilitate the processing by the classification algorithm[16].

2.1.1. Description of the Kedang and Lamaholot Language Dataset:

The dataset used in this study consists of texts in Kedang and Lamaholot languages collected from various sources such as books, articles, and interviews. Each document is labeled according to the language used to facilitate the classification process. Overall, the dataset consists of 1000 documents in Kedang and 1000 documents in Lamaholot. The data is distributed into two subsets: 80% for the training data and 20% for the test data, allowing the model to be trained and tested optimally[17].

2.1.2 Text Data Processing

The collected text data was reformatted to ensure consistency before being used in the analysis. This process includes the removal of irrelevant symbols such as punctuation marks, numbers, and other special characters. Additionally, double spaces that could cause inconsistencies in the data were removed. Finally, letter normalization was performed by converting all letters to lowercase to ensure the data has a uniform format, making it easier for the classification algorithms to process[18].

2.2 Preprocessing

Data preprocessing in this study was conducted to ensure that the text used in the analysis is clean and ready for further processing. The text cleaning process involves three main steps. The first step is the removal of special characters such as punctuation marks, numbers, and irrelevant symbols, aiming to eliminate elements that do not carry meaningful information in the context of classification. The second step is the removal of common words, or stopwords, that do not contribute significantly to the analysis, such as "and," "or," and "with." Lastly, stemming is applied to convert words into their root form, for example, the word "berlari" is changed to "lari." After

the cleaning process is complete, feature extraction is performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF represents the text in numerical form by calculating the weight of each word based on its frequency in a specific document and how unique it is across the entire corpus[19]. This process gives higher weights to words that frequently appear in a particular document but are rare across other documents, making them more informative for the classification process. The resulting TF-IDF numerical representation is then used as input for the Naive Bayes algorithm to perform language classification.

2.2.1 Text Cleaning Process

The text cleaning process involves several key steps to ensure that the data is ready for analysis. The first step is the removal of special characters such as punctuation marks, numbers, and irrelevant symbols, which aims to eliminate elements that do not carry meaningful information in the context of classification. Next, common words, or stopwords, which do not contribute significantly to the analysis, such as "and," "or," and "with," are removed. Finally, stemming is applied to convert words into their root form, for example, the word "berlari" is changed to "lari." With these steps, the text data becomes cleaner and more informative for use in the classification model.

2.2.2 Feature Extraction Using TF-IDF

Feature extraction in this study is performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method to represent the text in a numerical form. The TF-IDF method calculates the weight of each word based on its frequency of occurrence in a specific document and how unique that word is across the entire corpus. This process assigns higher weights to words that frequently appear in a particular document but are rare in other documents, making them more informative for the classification process. The resulting numerical representation from TF-IDF is then used as input for the Naïve Bayes algorithm to perform language classification.

2.3 Naïve Bayes Algorithm

The application of the Naïve Bayes algorithm in this study is performed to classify texts in Kedang and Lamaholot languages. The process begins by splitting the dataset into two parts: 80% of the total dataset for training data and 20% for test data. The training data is used to train the model to understand patterns within the text, while the test data is used to evaluate the model's performance. After splitting the dataset, the extracted text features using the TF-IDF method are input into the Naïve Bayes algorithm. The model is trained by calculating the probability of words in each document based on the given language labels. Once the training process is complete, the model is tested using the test data to measure its performance in classifying text into Kedang and Lamaholot languages. This evaluation aims to ensure that the model can make accurate predictions on new, previously unseen data. The Naïve Bayes algorithm is a probabilistic classification method that uses Bayes' theorem, assuming that each feature is independent of each other. In this study, the Naïve Bayes algorithm is applied to classify text in Kedang and Lamaholot languages. The process involves the following steps: dataset splitting, feature extraction using the TF-IDF method, model training, and performance evaluation using test data[20].

Bayes' theorem is used as the foundation for probability calculations in the Naïve Bayes algorithm. The formula for Bayes' theorem is as follows:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

3. Results and Discussion (b)

Results and discussion can be made as a whole that contains research findings and explanations.

3.1. Presenting the Results (b)

3.1.1 Desired Model Accuracy

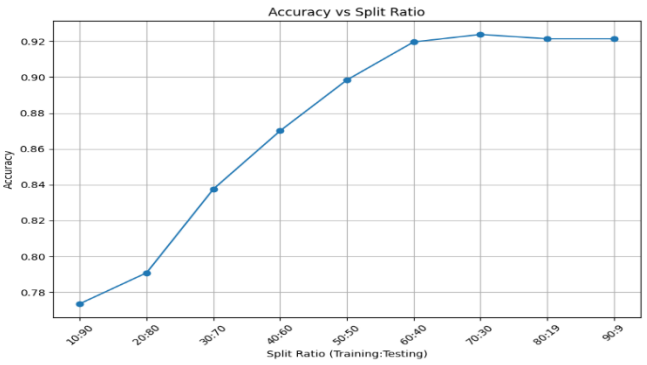


Figure 2. Accuracy_Vs-Split_Ratio

(Figure 1) illustrates the relationship between the data split ratio used for training and testing and the classification model's accuracy. In the graph, the X-axis (horizontal) represents the data split ratios for training and testing, ranging from 10:90 to 90:10. This means that the higher the percentage of data used for training, the lower the percentage of data available for testing[21]. Meanwhile, the Y-axis (vertical) displays the model's accuracy values corresponding to each split ratio. Based on the displayed results, it is evident that the model's accuracy increases as the proportion of training data grows. At the 10:90 split ratio, the model's accuracy is relatively low, around 0.77. However, as the amount of training data increases, the accuracy also improves significantly. A notable increase is observed up to the 70:30 ratio, where accuracy approaches 92%. Beyond this point, the accuracy growth slows down and tends to stabilize at around 92% for the 80:20 to 90:10 ratios. This pattern indicates that having more training data significantly contributes to improving model performance, especially in the early stages. However, after a certain point, adding more training data yields diminishing returns in terms of accuracy improvement[22]

3.1.2 Model Performance Based on Precision, Recall, and F1-Score

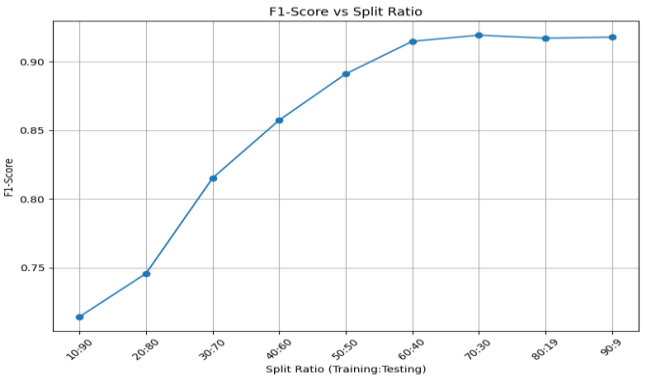


Figure 3. F1_Score_Vs_Split_Ratio

(Figure 2) illustrates the relationship between the data split ratio (training:testing) and the F1-Score in a machine learning model. In this graph, the X-axis (horizontal) represents the data split ratio between training and testing, while the Y-axis (vertical) displays the F1-Score, which serves as a performance evaluation metric for classification models[23].

Based on the graph, it can be seen that when the proportion of training data is low, the F1-Score also tends to be low. This indicates that the model is unable to learn effectively due to the limited amount of training data. As the proportion of training data increases, the F1-Score significantly improves, indicating better model performance in terms of classification, with a more balanced precision and recall.

However, after reaching a certain training ratio, the improvement in F1-Score begins to slow down and eventually stabilizes. This phenomenon suggests that adding a large amount of training data does not always result in a significant improvement in model performance.

From this pattern, it can be concluded that there is an optimal point in the data split between training and testing. If too little data is used for training, the model's performance will be poor. Conversely, if too much data is allocated for training, the additional benefits gained become increasingly marginal and disproportionate to the improvement in model performance.

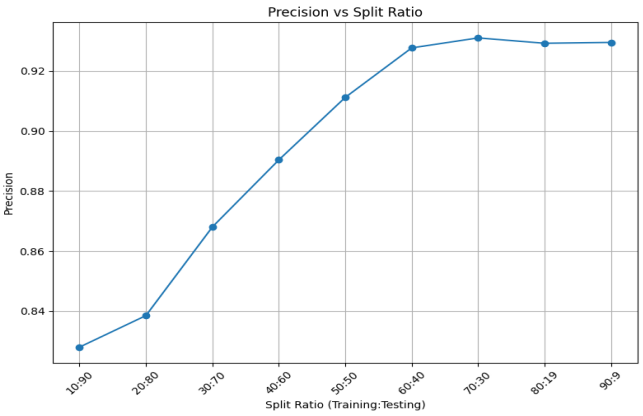


Figure 4. Precision_Vs-Split_Ratio

(Figure 3) ilustrates the relationship between the data split ratio (training: testing) and the precision score of a machine learning model. The graph reveals that with a smaller proportion of training data, the model struggles to make accurate positive predictions, resulting in low precision scores. As the amount of training data increases, the precision score improves significantly, indicating the model's growing ability to recognize data patterns and make accurate classifications. However, after reaching a training ratio of approximately 70:30, the improvement in precision begins to slow down and eventually stabilizes. This indicates that beyond a certain point, adding more training data no longer provides substantial benefits to model performance[24]. Therefore, an optimal split ratio exists where the model learns effectively without unnecessary use of additional training data.

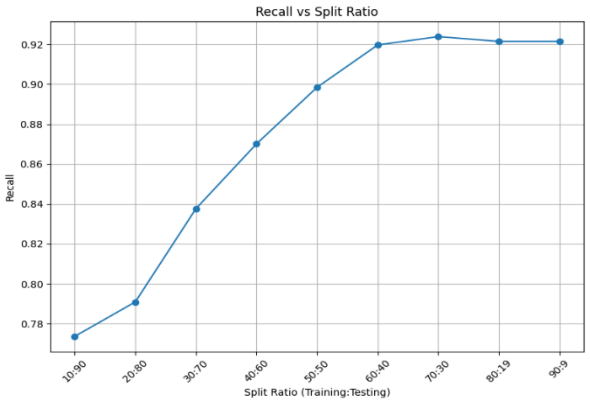


Figure 5. Recall_Vs-Split_Ratio

(Figure 4 illustrates the relationship between the training-to-testing data split and the recall score of a machine learning model. The recall score improves as the proportion of training data increases, indicating that the model becomes more effective at identifying relevant instances. However, after reaching a split of around 70:30, the improvement in recall slows down and eventually plateaus[25]. This suggests that adding more training data beyond this point yields minimal benefits. Therefore, finding an optimal balance in the data split is essential—too little training data leads to underperformance, while too much offers only limited additional value. This insight is important for making efficient use of available data during model development[26].

3.2. Create a Discussion (b)

This study aims to develop a text classification model to distinguish between Lamaholot and Kedang languages using the Naive Bayes algorithm and TF-IDF feature extraction. With a total of 2000 data samples, consisting of 1000 texts from each language, the model is expected to achieve a minimum accuracy of 90% as an indicator of good performance, particularly considering the balanced class distribution and linguistic complexity of both languages. The experimental results show that the model's accuracy increases as the proportion of training data increases. At a training-to-testing ratio of 10:90, the accuracy is still relatively low, around 0.77. However, as the training portion increases, the accuracy steadily improves, reaching approximately 92% at a 70:30 ratio. Beyond this point, accuracy gains begin to slow down and eventually plateau, indicating that adding more training data does not always result in significant performance improvement[27].

When compared to previous studies using similar algorithms, the results obtained in this study are quite competitive. For example, the study by [28], which applied Naive Bayes and TF-IDF for sentiment analysis on YouTube comments, reported an average accuracy of 91.8%. Meanwhile, studies by [29] and [30] showed lower accuracy rates of 79% and 61%, respectively.

This comparison suggests that model effectiveness is highly influenced by the data context and preprocessing techniques used. The present study demonstrates that with a sufficient amount of data and appropriate feature representation, a Naive Bayes-based model can perform very well, even in the context of local languages that share similar structures and vocabulary.

Despite the high level of accuracy achieved, several challenges need to be addressed. One key issue is misclassification caused by linguistic similarities between Lamaholot and Kedang, which makes it difficult for the model to distinguish between some texts with similar sentence structures or contexts. This challenge presents opportunities for future model development, such as incorporating semantic-based feature representations like word embeddings or applying deep learning classification techniques to capture more complex linguistic nuances[31].

The overall performance of the model is considered stable based on other evaluation metrics such as precision, recall, and F1-score, which indicate that the model maintains a balanced ability to recognize both classes. These findings reinforce the effectiveness of the Naive Bayes and TF-IDF approach for text classification, including in the context of regional languages that have been underexplored in natural language processing (NLP) research. Furthermore, this study opens avenues for further development, whether in terms of methodological enhancements, dataset expansion, or broader application to other dialects or languages with similar characteristics. These findings provide a valuable initial contribution to the advancement of NLP technologies for local languages in Indonesia.

4. Conclusion

Based on the available dataset, which consists of 2000 data points, with 1000 data points in Lamaholot language and 1000 data points in Kedang language, this study expects the model to achieve a sufficiently high accuracy level to demonstrate its ability to distinguish between Kedang and Lamaholot texts. The desired accuracy is at least 90%, considering the dataset's complexity and the class distribution used. This indicates the relationship between the split ratio (Training: Testing) and the accuracy of a machine learning model. The X-axis (Horizontal) represents the data split ratio for training and testing, for example, 10:90, 20:80, up to 90:10. This indicates what percentage of the data is used for training and testing.

The Y-axis (Vertical) shows the model's accuracy across various split ratios. As the ratio increases, the model's performance improves. The pattern shows that accuracy increases as the proportion of training data increases. At a 10:90 ratio, the accuracy is still low (~0.77), but the accuracy continues to rise until around a 70:30 ratio, approaching 92%. After that, the improvement becomes slower and eventually stabilizes at around 92%.

Based on the results of the study, the Naive Bayes algorithm with TF-IDF features was able to classify texts in Kedang and Lamaholot languages with a satisfactory level of accuracy. The model showed stable performance based on evaluation metrics such as precision, recall, and F1-score. However, some challenges were identified, such as prediction errors caused by the linguistic similarities between the two languages. This conclusion serves as the foundation for model improvements and better development strategies in future research.

5. Authors' Declaration (b)

Authors' contributions and responsibilities

Write the contribution of each author here, or mark the following column.

√	The authors made substantial contributions to the conception and design of the study.
√	The authors took responsibility for data analysis, interpretation and discussion of results.
√	The authors read and approved the final manuscript.

Funding

Write down the research funding, if any.

Availability of data and materials

√ All data are available from the authors.

Competing interests

√ The authors declare no competing interest.

Additional information

Write additional information related to this research, if any.

6. Acknowledgement

We express our gratitude to God Almighty for all His blessings and grace, which has allowed us to prepare this research proposal report entitled "CLASSIFICATION OF REGIONAL LANGUAGE TEXTS IN LEMBATA REGENCY USING NAÏVE BAYES WITH TF-IDF FEATURES." This research is motivated by the importance of classifying regional language texts, particularly in our country, which is rich in biodiversity. With the advancement of technology, the researcher aims to utilize image classification methods to automatically classify regional languages.

For the moral and material support provided in the preparation of this report, the author would like to express gratitude to:

1. Dr. Anwar Cengkeng, S.H., M., Hum, Rector of Widyagama University Malang
2. Ir. Candra Aditya, S.T., M.T., Dean of the Faculty of Engineering, Widyagama University Malang
3. Aviv Yuniar Rahman, S.T., M.T., Supervisor I and Head of the Informatics Engineering Study Program, Widyagama University Malang
4. Rangga Pahlevi Putra, S.Pd., M.T., Supervisor II and Secretary of the Faculty of Engineering
5. Our beloved parents and family, who have always prayed for me and provided full support and encouragement for this success.
6. All lecturers and staff of the Informatics Engineering Program at Widyagama University Malang
7. And my friends who have always provided extraordinary support.

In the preparation of this thesis, the researcher has made efforts to systematically explain the background, objectives, methodology, and benefits of the research to be conducted. The researcher acknowledges that this research would not have come to fruition without the support and guidance from various parties, including the supervising lecturers and colleagues. The researcher hopes that this proposal report can provide a positive contribution to the development of image recognition technology in the field of language, as well as serve as a reference for future research. Thanks to all parties who have assisted in the preparation of this proposal. Finally, the researcher hopes that this research will yield useful results.

7. References

[1] G. Setiawan and I. Adnyana, “Improving helpdesk chatbot performance with term frequency-inverse document frequency (tf-idf) and cosine similarity models,” *Journal of Applied Informatics and Computing*, vol. 7, no. 2, pp. 252–257, 2023, doi: https://doi.org/10.30871/jaic.v7i2.6527.

[2] M. Afif, M. Ula, L. Rosnita, and R. Rizal, “Applying tf-idf and k-nn for clickbait detection in indonesian online news headlines,” *Jo. Adv. Comp. Know. Algo*, vol. 1,

no. 2, pp. 38–41, 2024, doi: https://doi.org/10.29103/jacka.v1i2.15810.
[3] "An atn based framework for arabic text analysis," *International Research Journal of Modernization in Engineering Technology and Science*, 2024, doi: https://doi.org/10.56726/irjmets64774.
[4] J. Xue, "Research on korean literature corpus processing based on computer system improved tf-idf algorithm," *Intelligent Decision Technologies*, vol. 18, no. 4, pp. 3011–3024, 2024, doi: https://doi.org/10.3233/idt-230772.
[5] A. Purpura, D. Giorgianni, G. Orrù, G. Melis, and G. Sartori, "Identifying single-item faked responses in personality tests: a new tf-idf-based method," *PLoS One*, vol. 17, no. 8, p. e0272970, 2022, doi: https://doi.org/10.1371/journal.pone.0272970.
[6] L. Xiang, "Application of an improved tf-idf method in literary text classification," *Advances in Multimedia*, vol. 2022, pp. 1–10, 2022, doi: https://doi.org/10.1155/2022/9285324.
[7] H. Jadia, "Comparative analysis of sentiment analysis techniques: svm, logistic regression, and tf-idf feature extraction," *International Research Journal of Modernization in Engineering Technology and Science*, 2023, doi: https://doi.org/10.56726/irjmets45265.
[8] R. Putranto, M. Purbolaksono, and W. Astuti, "Sentiment analysis of practo application reviews using naïve bayes and tf-idf methods," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, p. 1070, 2023, doi: https://doi.org/10.30865/mib.v7i3.6311.
[9] E. Heikel and L. Espinosa-Leal, "Indoor scene recognition via object detection and tf-idf," *J Imaging*, vol. 8, no. 8, p. 209, 2022, doi: https://doi.org/10.3390/jimaging8080209.
[10] J. Abbas, C. Zhang, and B. Luo, "Bet-bilstm model: a robust solution for automated requirements classification," *Journal of Software Evolution and Process*, vol. 37, no. 3, 2025, doi: https://doi.org/10.1002/smr.70012.
[11] M. Dhiyaulhaq and P. Gunawan, "Sentiment analysis of the jakarta - bandung fast train project using the svm method," *Jurnal Media Informatika Budidarma*, vol. 7, no. 4, p. 2128, 2023, doi: https://doi.org/10.30865/mib.v7i4.6855.
[12] T. Swalar, "Deiksis persona bahasa lamaholot dialek lamaholot tengah," *MAJU*, vol. 1, no. 3, pp. 83–93, 2024, doi: https://doi.org/10.62335/t792n796.
[13] Y. Demon, "Morphophonemics in the lamalera dialect of lamaholot," *Randwick International of Education and Linguistics Science Journal*, vol. 3, no. 1, pp. 112–127, 2022, doi: https://doi.org/10.47175/rielsj.v3i1.414.
[14] K. Austad and B. W. Jack, "Linguistic and Cultural Competence at Hospital Discharge," *Journal of Healthcare Management Standards*, 2023, doi: 10.4018/jhms.330644.
[15] S. V Kusnoor *et al.*, "Design and Implementation of a Massive Open Online Course on Enhancing the Recruitment of Minorities in Clinical Trials – Faster Together," *BMC Med Res Methodol*, 2021, doi: 10.1186/s12874-021-01240-x.
[16] L. Cayón and T. C. Chacon, "Diversity, Multilingualism and Inter-Ethnic Relations in the Long-Term History of the Upper Rio Negro Region of the Amazon," *Interface Focus*, 2022, doi: 10.1098/rsfs.2022.0050.
[17] L. T. Kim Ha, T. Van Le, L. T. Phan, L. T. Bich Nguyen, and A. T. Van Dam, "Perspectives of Vietnamese Students and Teachers Regarding the Preservation of Languages of Ethnic Minorities," *Revista De Gestão Social E Ambiental*, 2024, doi: 10.24857/rgsa.v18n9-026.
[18] S. Budiono and T. Jaya, "Evaluation of Local Language Learning in the Limola Language Revitalization," *Journal of Applied Studies in Language*, 2024, doi: 10.31940/jasl.v8i1.20-30.
[19] J. Jupri, A. Aprianoto, and E. Firman, "The Application of Linguistic Landscape in Mataram City Kota Madya Mataram, West Nusa Tenggara Province, Indonesia," *Jurnal Ilmiah Mandala Education*, 2022, doi: 10.58258/jime.v8i3.3761.
[20] B. Rivaldo, "Interest of Youth of the Batak Karo Protestant Church (GBKP) Cikarang in Using Regional Language Communication," *Adv*, 2024, doi: 10.46799/adv.v2i10.291.

[21] B. Sinclair *et al.*, "Machine Learning Approaches for Imaging-based Prognostication of the Outcome of Surgery for Mesial Temporal Lobe Epilepsy," *Epilepsia*, 2022, doi: 10.1111/epi.17217.

[22] O. Yossofzai *et al.*, "Development and Validation of Machine Learning Models for Prediction of Seizure Outcome After Pediatric Epilepsy Surgery," *Epilepsia*, 2022, doi: 10.1111/epi.17320.

[23] G. O. Ghosheh, L. Thwaites, and T. Zhu, "Synthesizing Electronic Health Records for Predictive Models in Low-Middle-Income Countries (LMICs)," *Biomedicines*, 2023, doi: 10.3390/biomedicines11061749.

[24] R. Muralidhar, M. L. Demory, and M. M. Kesselman, "Exploring the Impact of Batch Size on Deep Learning Artificial Intelligence Models for Malaria Detection," *Cureus*, 2024, doi: 10.7759/cureus.60224.

[25] S. T. Arasteh, C. Kühl, M.-J. Saehn, P. Isfort, D. Truhn, and S. Nebelung, "Enhancing Domain Generalization in the AI-based Analysis of Chest Radiographs With Federated Learning," *Sci Rep*, 2023, doi: 10.1038/s41598-023-49956-8.

[26] M. Aliyari and Y. Z. Ayele, "Application of Artificial Neural Networks for Power Load Prediction in Critical Infrastructure: A Comparative Case Study," *Applied System Innovation*, 2023, doi: 10.3390/asi6060115.

[27] A. B. Nugraha and A. Romadhony, "Identification of 10 Regional Indonesian Languages Using Machine Learning," *Sinkron*, 2023, doi: 10.33395/sinkron.v8i4.12989.

[28] D. Farah Zhafira, B. Rahayudi, and P. Korespondensi, "ANALISIS SENTIMEN KEBIJAKAN KAMPUS MERDEKA MENGGUNAKAN NAIVE BAYES DAN PEMBOBOTAN TF-IDF BERDASARKAN KOMENTAR PADA YOUTUBE," 2021.

[29] G. Mandar, A. H. Muhamamd, and S. Sudin, "Klasifikasi Berita Indonesia Menggunakan Naïve Bayes dengan Porter Stemmer," *Jurnal Teknik Informatika (J-Tifa)*, vol. 3, no. 2, pp. 17–22, Sep. 2020, doi: 10.52046/j-tifa.v3i2.1121.

[30] L. Mayasari and D. Indarti, "KLASIFIKASI TOPIK TWEET MENGENAI COVID MENGGUNAKAN METODE MULTINOMIAL NAÏVE BAYES DENGAN PEMBOBOTAN TF-IDF," *Jurnal Ilmiah Informatika Komputer*, vol. 27, no. 1, pp. 43–53, 2022, doi: 10.35760/ik.2022.v27i1.6184.

[31] N. Abinaya, P. Jayadharshini, S. Priyanka, S. Keerthika, and S. Santhiya, "Identification of Language From Multi-Lingual Dataset Using Classification Algorithms," *J Phys Conf Ser*, 2023, doi: 10.1088/1742-6596/2664/1/012009.