**Type of contribution:**
Research Paper
Editorial Case
Study Review
Paper Scientific
Data
Technical Application Report

ECSITE

Edutran Computer Science & Information Technology
Vol. 2, No. 1 (2024) pp 42-52
e-issn. 7289-2554

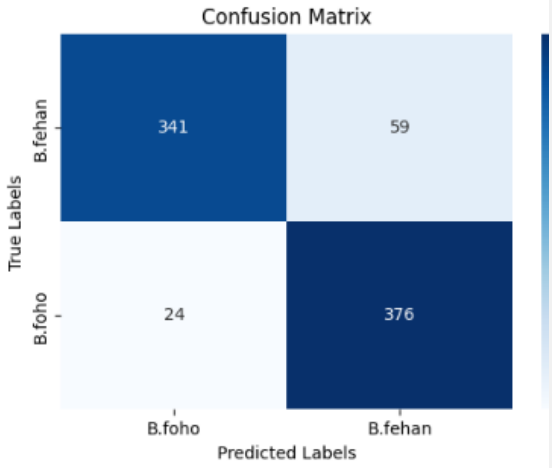
# Classification of Regional Languages of Malaka Foho and Fehan Districts Using the Naïve Bayes Method

**Aprilio Demetrius De Araujo [1], Rangga Pahlevi Putra [2], Istiadi [3]**

[1] Informatics Engineering, Widyagama University, 65128, Indonesia
[2] Informatics Engineering, Widyagama University, 65128, Indonesia
[3] Informatics Engineering, Widyagama University, 65128, Indonesia

emusrteam@gmail.com

This article contributes to:


- **Highlights:**
- High Accuracy: The model achieved an overall accuracy of 90% in classifying the two types of languages foho and fehan.
- Performance of the foho language class 376 sentences were correctly classified as foho language and 24 sentences were incorrectly classified as fehan language.
- There were classification errors that occurred in the fehan language, 59 sentences were incorrectly classified as foho language and 341 were correctly classified as fehan language.
- The highest value is precision 93%, recall 94%, F1-Score 90%.

## Abstract

Regional languages are an important part of cultural identity that must be preserved. Malaka Regency in East Nusa Tenggara has two main dialects, Tetun Foho and Tetun Fehan, which differ in phonology, morphology, and vocabulary. This research aims to develop an automatic classification model that can distinguish between the two languages using the Naïve Bayes algorithm with a TF-IDF feature extraction approach. The dataset consists of 2,000 sentences in Foho and Fehan languages collected through observations, interviews, and social media. The research process included data preprocessing (case folding, tokenizing, and stemming), feature extraction using TF-IDF, model training, and performance evaluation using accuracy, precision, recall, and F1-score metrics. Test results showed that the model was able to classify both languages with 90% accuracy on a 40:60 data split, with high precision and recall, especially in the Foho language class. This research demonstrates the effectiveness of the Naïve Bayes method in regional language classification and has the potential to aid in the preservation and documentation of local languages in the digital age..

**Keywords:** Language Classification, Naïve Bayes, TF-IDF, Foho Language, Fehan Language, Malacca Regency

**Article info**

Submitted: 2025-08-02
Revised: 2025-08-19
Accepted: 2025-08-26



## 1. Introduction

Language is an essential element in human life, not only as a means of communication but also as a reflection of cultural identity. In Malaka Regency, East Nusa Tenggara, there are two main dialect variations known as Foho and Fehan. [1] . These two language varieties represent the geographical and

cultural differences of the Foho people, spoken in the mountainous areas and Fehan in the lowlands [2] . However, the introduction of these two languages faces various obstacles, ranging from a lack of formal documentation, minimal academic research, to limited technology that supports the classification of regional languages [3] .

In several previous studies, such as those conducted by [4] and [5], the text classification method based on the Naïve Bayes algorithm was proven to be able to provide high accuracy in distinguishing regional languages [6] . Naïve Bayes is known for its simplicity and effectiveness in handling linguistic data, especially in feature-based text processing such as TF-IDF [7] .

Based on these problems, this study aims to develop a classification model based on Naïve Bayes [5] , to automatically recognize and differentiate the Foho and Fehan languages [8] . This model is expected to help preserve regional languages through an efficient, accurate, and sustainable technological approach [9] .

## 2. Method

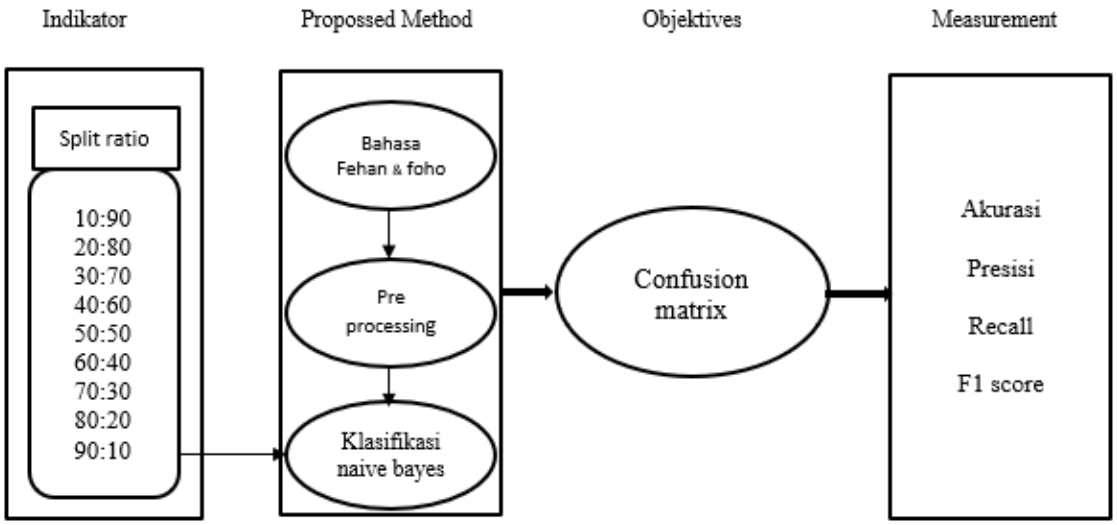



Figure 1 Framework

Figure 1 shows the framework of this research, which begins with inputting text documents in the Foho and Fehan languages. The text documents will then be processed through a pre-processing stage, namely Case Folding and Tokenizing. Next, research and testing are carried out. In the training stage, the data is divided into training data and test data, then features are extracted using Tf-Idf, and a Naïve Bayes model is trained for text classification. In the testing stage, pre-processing of the test data is carried out to ensure that the test data undergoes the same process as the training data, followed by prediction using the model to determine the class of the test data, and then evaluation is carried out. The output of this process is two types of language, namely Foho and Fehan, which are generated from text classification. The overall evaluation uses several matrices, including accuracy, precision, recall, f1-score, and confusion matrix [7] .

### 2.1 Data set

The collected data comes from community observations and interviews with native Foho and Fehan speakers. Each sentence in the data has been labeled according to its respective language to facilitate the classification process. All data was then compiled in an Excel document, as shown in Figure 2. This labeling process was carried out manually to ensure accurate language identification, ensuring that the data used truly represents the characteristics of each dialect. This organized data then served as the basis for training and testing language classification models.

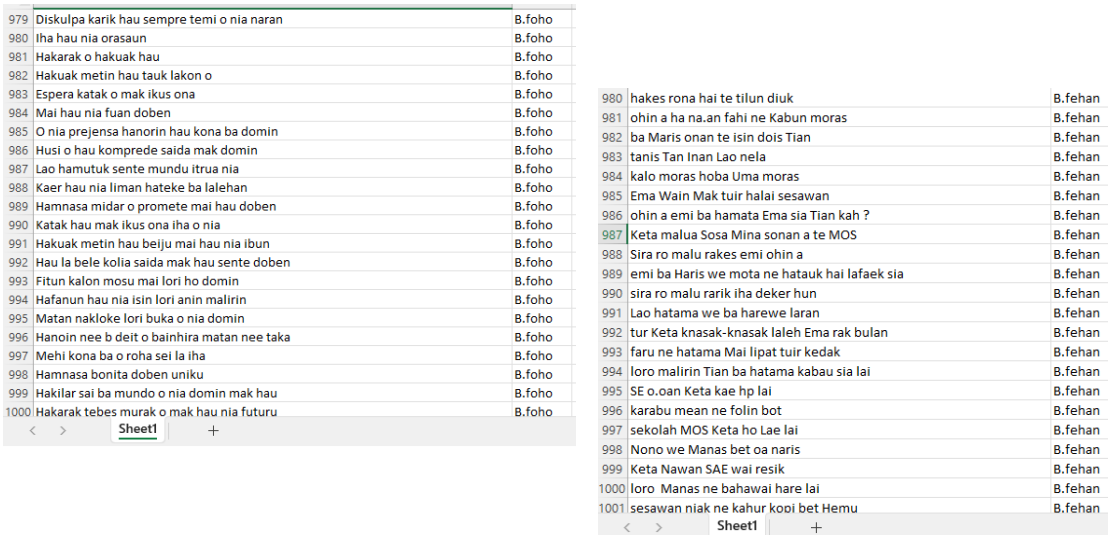
### 2.2 Preprocessing

Preprocessing is a very important initial stage before data is fed into a Naive Bayes model, especially in text classification tasks. The goal is to clean and prepare the data so that the model can work optimally. The preprocessing stage in Naive Bayes usually involves the following steps [3] :

1. Data Collection

The research data was obtained through observations and interviews with native speakers of Foho and Fehan, then manually labeled to ensure classification accuracy. This verified data was compiled in an Excel document in Figure 2 below and became the main basis for the training and testing of the language classification model [10] .

Figure 2
Foho and Fehan
Language Dataset



| | | |
|---|---|---|
| 979 | Diskulpa karik hau sempre temi o nia naran | B.foho |
| 980 | Iha hau nia orasaun | B.foho |
| 981 | Hakarak o hakuak hau | B.foho |
| 982 | Hakuak metin hau tauk lakon o | B.foho |
| 983 | Espera katak o mak ikus ona | B.foho |
| 984 | Mai hau nia fuan doben | B.foho |
| 985 | O nia prejensa hanorin hau kona ba domin | B.foho |
| 986 | Husi o hau komprede saida mak domin | B.foho |
| 987 | Lao hamutuk sente mundu itrua nia | B.foho |
| 988 | Kaer hau nia liman hateke ba lalehan | B.foho |
| 989 | Hamnasa midar o promete mai hau doben | B.foho |
| 990 | Katak hau mak ikus ona iha o nia | B.foho |
| 991 | Hakuak metin hau beiju mai hau nia ibun | B.foho |
| 992 | Hau la bele kolia saida mak hau sente doben | B.foho |
| 993 | Fitun kalon mosu mai lori ho domin | B.foho |
| 994 | Hafanun hau nia isin lori anin malirin | B.foho |
| 995 | Matan nakloke lori buka o nia domin | B.foho |
| 996 | Hanoin nee b deit o bainhira matan nee taka | B.foho |
| 997 | Mehi kona ba o roha sei la iha | B.foho |
| 998 | Hamnasa bonita doben uniku | B.foho |
| 999 | Hakilar sai ba mundo o nia domin mak hau | B.foho |
| 1000 | Hakarak tebes murak o mak hau nia futuru | B.foho |

Sheet1

| | | |
|---|---|---|
| 980 | hakes rona hai te tilun diuk | B.fehan |
| 981 | ohin a ha na.an fahi ne Kabun moras | B.fehan |
| 982 | ba Maris onan te isin dois Tian | B.fehan |
| 983 | tanis Tan Inan Lao nela | B.fehan |
| 984 | kalo moras hoba Uma moras | B.fehan |
| 985 | Ema Wain Mak tuir halai sesawan | B.fehan |
| 986 | ohin a emi ba hamata Ema sia Tian kah ? | B.fehan |
| 987 | Keta malua Sosa Mina sonan a te MOS | B.fehan |
| 988 | Sira ro malu rakes emi ohin a | B.fehan |
| 989 | emi ba Haris we mota ne hatauk hai lafaek sia | B.fehan |
| 990 | sira ro malu rarik iha deker hun | B.fehan |
| 991 | Lao hatama we ba harewe laran | B.fehan |
| 992 | tur Keta knasak-knasak laleh Ema rak bulan | B.fehan |
| 993 | faru ne hatama Mai lipat tuir kedak | B.fehan |
| 994 | loro malirin Tian ba hatama kabau sia lai | B.fehan |
| 995 | SE o.oan Keta kae hp lai | B.fehan |
| 996 | karabu mean ne folin bot | B.fehan |
| 997 | sekolah MOS Keta ho Lae lai | B.fehan |
| 998 | Nono we Manas bet oa naris | B.fehan |
| 999 | Keta Nawan SAE wai resik | B.fehan |
| 1000 | loro  Manas ne bahawai hare lai | B.fehan |
| 1001 | sesawan niak ne kahur kopi bet Hemu | B.fehan |

Sheet1

## 2. Data Sharing

The cleaned, processed, and feature extracted dataset using Tf-IDF was divided into two parts: training data and test data. This division was carried out with a ratio ranging from 10% - 90% to 90% - 10%. The training data was used to train the naive Bayes model, while the test data was used to determine model performance.

### 2.3 TF-IDF Feature Extraction

Feature extraction in this study uses the TF-IDF method to transform text data into a numerical representation that reflects the importance of each word in a document. High weights are given to words that are unique to a document but rarely appear in other documents, thereby increasing classification effectiveness. The resulting TF-IDF representation is then used as the main input in the Naïve Bayes algorithm to accurately classify Foho and Fehan languages.

Figure 3
TF-IDF Feature
Extraction

```
Sample Predictions:
Text: hau rona inan aman sira koalia kona ba festa ita nian
True Label: B.foho, Predicted Label: B.foho

Text: mai saka hau iha kolega nia kost
True Label: B.foho, Predicted Label: B.foho

Text: sira bolu labarik sira atu serbisu iha toos
True Label: B.foho, Predicted Label: B.foho

Text: mai be o ba hasoru malu ho nia
True Label: B.foho, Predicted Label: B.foho

Text: awan hau ba kupang
True Label: B.fehan, Predicted Label: B.foho
```

Figure 3 shows the prediction results of the language classification model for five example sentences. From these results, it can be seen that four sentences originally labeled B.foho were successfully predicted by the model as B.foho. However, there was one prediction error in a sentence that was actually labeled B.fehan but was predicted as B.foho. This indicates that the model has a fairly good performance in recognizing B.foho sentences, but still experiences errors in distinguishing the

characteristics of B.fehan sentences. This error is likely caused by similarities in vocabulary or sentence structure between the two dialects [7] .

### 2.4 Accuracy

Accuracy is a measure used to assess the extent to which a model's prediction or classification results correspond to reality or actual data. In the context of data processing or machine learning, accuracy is calculated as the percentage of the number of correct predictions compared to the total number of predictions made [5] .

$$\text{Curation: } \frac{TP + TN}{TP + TN + FP + FN}$$

### 2.5 Precision

Precision is a measure used to assess the extent to which the positive predictions provided by a model are truly accurate or relevant. In the context of classification, precision is calculated by comparing the number of correct positive predictions (True Positive) with the total number of positive predictions generated, both correct and incorrect (True Positive + False Positive). High precision indicates that the model rarely provides incorrect positive predictions [5] .

$$\textbf{Pre}\text{sis: } \frac{TP}{TP + FP}$$

### 2.6 Recall

Recall is a measure used to assess the extent to which a model can identify all true positive examples. In the context of classification, recall is calculated by comparing the number of correct positive predictions (True Positives) to the total number of actual positive data, namely the sum of True Positives and False Negatives. High recall indicates that the model is able to capture most of the positive examples, although it may miss some [5] .

$$\textbf{Recall}: \frac{TP}{TP + FN}$$

### 2.7 F1-Score

F1 Score is a measure used to evaluate the performance of a classification model by considering both precision and recall. F1 Score is calculated as the harmonic mean between precision and recall, thus providing a more balanced picture of how well the model handles both metrics [11] . F1 Score values range from 0 to 1, where a value of 1 indicates a perfect model in terms of precision and recall, while a value of 0 indicates poor performance [5] .

$$F1 \text{ - } Score: 2\ x\ \frac{Recall\ X\ Precision}{Recall\ +\ Precision}$$

## 3. Evaluation Results

The results and discussion can be made complete, containing research findings and their explanations.

## 3.1. Split Ratio 10% Training Data 90% Test Data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.93 | 0.80 | 0.86 | 98 |
| B.foho | 0.83 | 0.94 | 0.88 | 102 |
| accuracy | | | 0.87 | 200 |
| macro avg | 0.88 | 0.87 | 0.87 | 200 |
| weighted avg | 0.88 | 0.87 | 0.87 | 200 |

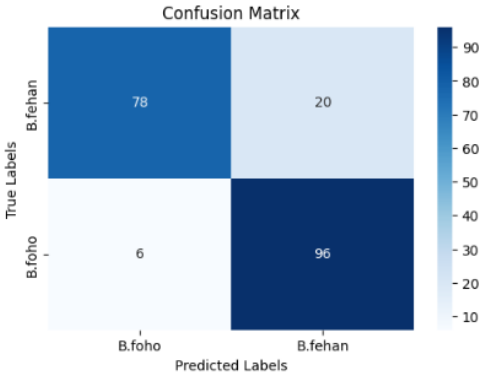



Figure 4 Split Ratio 10:90

The classification model showed good and balanced performance in recognizing the two language labels, B.fehan and B.foho, with an overall accuracy of 87% across 200 test data sets. The high precision of B.fehan (93%) indicates good prediction accuracy, although its recall is lower (80%), indicating that there are still undetected B.fehan sentences. In contrast, B.foho has a high recall (94%) and a precision of 83%, reflecting the model's high sensitivity to this dialect despite some misclassifications. The confusion matrix in Figure 4 supports this evaluation, where out of 98 B.fehan data sets, 78 were correctly classified and 20 were incorrectly classified as B.foho, while out of 102 B.foho data sets, 96 were correctly recognized and only 6 were mispredicted [12] . These results confirm that the model is more effective in recognizing B.foho and generally has stable and consistent performance in distinguishing between the two dialects.

## 3.2. Split Ratio 20% Training Data 80% Test Data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.92 | 0.84 | 0.88 | 193 |
| B.foho | 0.87 | 0.93 | 0.90 | 207 |
| accuracy | | | 0.89 | 400 |
| macro avg | 0.89 | 0.89 | 0.89 | 400 |
| weighted avg | 0.89 | 0.89 | 0.89 | 400 |

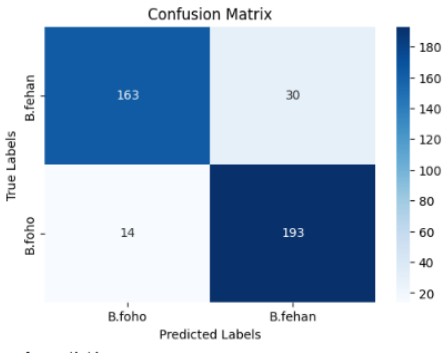



Figure 5 Split Ratio 20:80

The classification model showed quite good and consistent performance in recognizing two language labels, B.fehan and B.foho, with an accuracy of 89% from a total of 400 test data. For the B.fehan label, the precision reached 92% but the recall was slightly lower at 84%, indicating that the model was quite accurate but there were still undetected data. Meanwhile, B.foho had a high recall of 93% and a precision of 87%, indicating the model's strong sensitivity to this class. The balanced F1-score values (88% for B.fehan and 90% for B.foho), as well as the macro and weighted average of 89%, confirmed that the model was able to handle both labels well. This is supported by the confusion matrix [13] in Figure 5, which shows that of the 193 B.fehan data, 163 were correctly classified and 30 were incorrectly classified as B.foho, while of the 207 B.foho data, 193 were correctly recognized and only 14 were misclassified. Overall, these results indicate that the model is very effective, especially in recognizing B.foho, with low misclassification rates for both classes.

### 3.3. Split Ratio 30% Training Data 70% Test Data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.92 | 0.85 | 0.88 | 295 |
| B.foho | 0.87 | 0.93 | 0.90 | 305 |
| accuracy | | | 0.89 | 600 |
| macro avg | 0.89 | 0.89 | 0.89 | 600 |
| weighted avg | 0.89 | 0.89 | 0.89 | 600 |

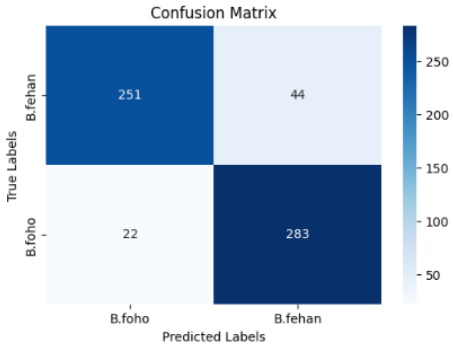



Figure 6 Split Ratio 20:80

The classification model demonstrated strong and balanced performance in classifying two dialects, B.fehan and B.foho, with an accuracy of 89% across 600 test data sets. The high precision of B.fehan (92%) and the highest recall of B.foho (93%) reflect the model's good ability to detect both classes, despite the presence of a small number of missing data sets. The F1-scores of 88% and 90%, respectively, and the consistent macro and weighted average values [4] at 89%, reinforce the model's stable performance. The confusion matrix in Figure 6 shows that out of 295 B.fehan data sets, 251 were correctly classified and 44 incorrectly; while out of 305 B.foho data sets, 283 were correctly recognized and only 22 were incorrectly recognized. These results confirm that the model is quite reliable and effective in distinguishing the two dialects with a low error rate.

### 3.4. Split Ratio 40% Training Data 60% Test Data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.93 | 0.85 | 0.89 | 400 |
| B.foho | 0.86 | 0.94 | 0.90 | 400 |
| accuracy | | | 0.90 | 800 |
| macro avg | 0.90 | 0.90 | 0.90 | 800 |
| weighted avg | 0.90 | 0.90 | 0.90 | 800 |

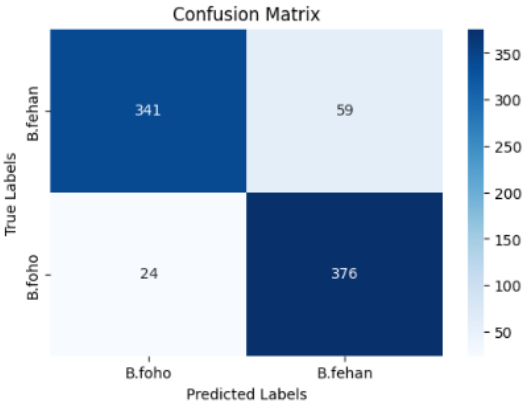



Figure 7 Split Ratio 40:60

The classification model showed excellent and balanced performance in distinguishing between Fehan and Foho, with an overall accuracy of 90%. High precision and recall in both classes—93% and 85% for B.fehan, and 86% and 94% for B.foho—indicate good model accuracy and sensitivity. A balanced F1-score (89% and 90%) and a macro and weighted average of 90% confirm the stability of the model. The confusion matrix shows that out of 400 B.fehan data, 341 were classified correctly and 59 incorrectly, while out of 400 B.foho data, 376 were recognized correctly and only 24 incorrectly [14] . These results indicate that the model is very effective, especially in recognizing B.foho, with low error rates in both classes.

### 3.5. Split Ratio 50% Training Data 50% Test Data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.93 | 0.84 | 0.88 | 493 |
| B.foho | 0.86 | 0.93 | 0.90 | 507 |
| accuracy | | | 0.89 | 1000 |
| macro avg | 0.89 | 0.89 | 0.89 | 1000 |
| weighted avg | 0.89 | 0.89 | 0.89 | 1000 |

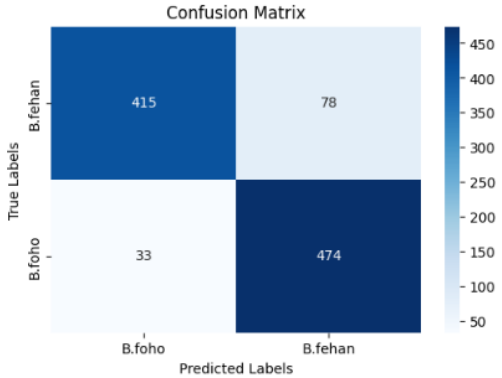


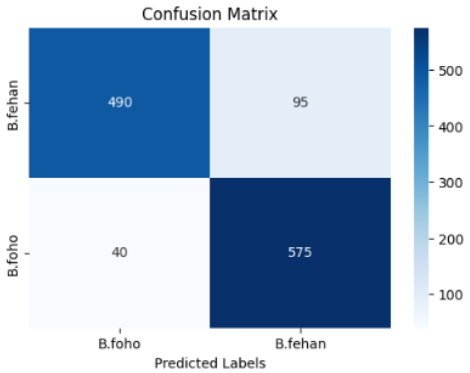

Figure 8 Split Ratio 50:50

The classification model demonstrated reliable and balanced performance with an accuracy of 89% on 1,000 test data sets, indicating that 890 data sets were correctly classified. High precision for B.fehan (93%) indicates accurate predictions, although its recall is lower (84%), while B.foho has the highest recall (93%) and precision of 86%, indicating high sensitivity to this class. F1-scores of 88% and 90%, respectively, and macro and weighted averages of 89%, reflect stable model performance [3] . The confusion matrix shows that out of 493 B.fehan data sets, 415 were correctly classified, while out of 507 B.foho data sets, 474 were correctly recognized. These results confirm that the model is quite accurate and balanced in distinguishing between the two dialects, with a low error rate.

## 3.6. Split Ratio 60% Training Data 40% Test Data

Figure 9 Split Ratio 60:40

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.92 | 0.84 | 0.88 | 585 |
| B.foho | 0.86 | 0.93 | 0.89 | 615 |
| accuracy | | | 0.89 | 1200 |
| macro avg | 0.89 | 0.89 | 0.89 | 1200 |
| weighted avg | 0.89 | 0.89 | 0.89 | 1200 |

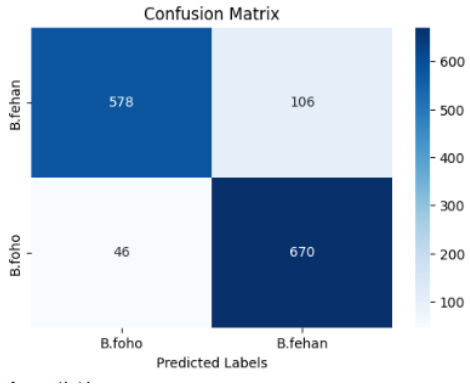



The classification model showed good and stable performance with an accuracy of 89% on 1,200 test data. Fehan language has a high precision (92%) but a lower recall (84%), indicating good prediction accuracy despite missing data. In contrast, Foho language has a high recall (93%) and a precision of 86%, indicating the model is very sensitive to this class [3] . A balanced F1-score (88% for Fehan, 89% for Foho) and a macro and weighted average of 89% both strengthen the stability of the model's performance. The confusion matrix shows that the model is better at recognizing B.foho, with fewer misclassifications than B.fehan, although in general the model is quite reliable in distinguishing the two dialects.

## 3.7. Split Ratio 70% Training Data 30% Test Data

Figure 10 Split Ratio 70:30

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.93 | 0.85 | 0.88 | 684 |
| B.foho | 0.86 | 0.94 | 0.90 | 716 |
| accuracy | | | 0.89 | 1400 |
| macro avg | 0.89 | 0.89 | 0.89 | 1400 |
| weighted avg | 0.89 | 0.89 | 0.89 | 1400 |

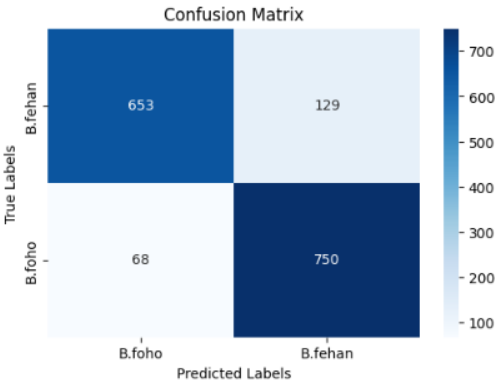



The classification model demonstrated robust and balanced performance with 89% accuracy across 1,400 data sets. Precision for B. fehan was high (93%), but recall was lower (85%), while B. foho had the highest recall (94%) and precision (86%). F1-scores of 88% and 90%, respectively, and macro and weighted average values consistently at 89%, indicated model stability. The confusion matrix supported these results, with 578 of 684 B. fehan data sets and 670 of 716 B. foho data sets correctly classified. These results confirm that the model is more reliable in recognizing B. foho, but still needs improvement in sensitivity to B. fehan data.

## 3.8. Split Ratio 80% Training Data 20% Test Data

Figure 11 Split Ratio 80:20

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.91 | 0.84 | 0.87 | 782 |
| B.foho | 0.85 | 0.92 | 0.88 | 818 |
| accuracy | | | 0.88 | 1600 |
| macro avg | 0.88 | 0.88 | 0.88 | 1600 |
| weighted avg | 0.88 | 0.88 | 0.88 | 1600 |

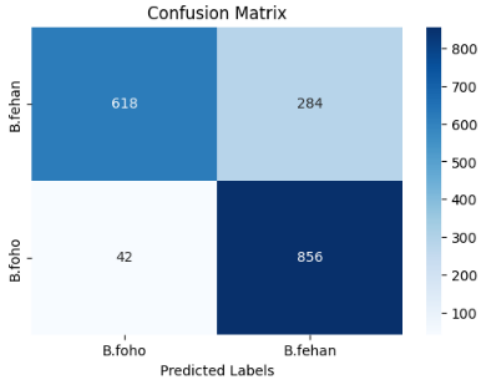



The classification model demonstrated stable performance with an accuracy of 88% across 1,600 test datasets. High precision for B.fehan (91%) indicates good accuracy, although its recall still needs improvement (84%). Conversely, B.foho had the highest recall (92%) and precision of 85%, indicating the model is very good at recognizing this class. The balanced F1-score (87% for B.fehan, 88% for B.foho), and the macro and weighted average were both 88%, reflecting consistent model performance. The confusion matrix supports this result, with a higher number of prediction errors for B.fehan (129 datasets) compared to B.foho (68 datasets), suggesting improvements are needed, particularly in increasing sensitivity to B.fehan datasets.

## 3.9. Split Ratio 90% Training Data 10% Test Data

Figure 12 Split Ratio 90:10

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| B.fehan | 0.94 | 0.69 | 0.79 | 902 |
| B.foho | 0.75 | 0.95 | 0.84 | 898 |
| accuracy | | | 0.82 | 1800 |
| macro avg | 0.84 | 0.82 | 0.82 | 1800 |
| weighted avg | 0.84 | 0.82 | 0.82 | 1800 |




The classification model performed quite well with an accuracy of 82% from 1,800 data sets, but there was an imbalance between the ability to recognize Fehan and Foho. Precision for B.fehan was high (94%), but recall was low (69%), indicating that many Fehan data sets were not detected. Conversely, B.foho had a high recall (95%) but lower precision (75%), indicating that the model was more sensitive to Foho but still often made incorrect predictions. F1-scores of 79 (Fehan) and 84 (Foho) respectively indicated that the model was more optimal in recognizing B.foho. The confusion matrix showed that of the 902 B.fehan data sets, only 618 were correctly recognized, while of the 898 B.foho data sets, 856 were correctly classified. These results emphasize the need to improve accuracy, especially in recognizing B.fehan, to ensure a more balanced model performance [5] .

## 4. Create a Discussion

As part of the evaluation, Precision, Recall, and F1-Score metrics were analyzed across various training-to-test data ratios. The following graphs show the model's performance changes based on these ratios, providing insight into the model's accuracy, sensitivity, and balance in classifying Fehan and Foho.

The precision of Fehan is relatively stable and high across various data ratios, especially when the training data proportion is large, such as at a 90-10 ratio, which reaches 94%. This demonstrates the model's consistency in recognizing the Fehan class. In contrast, the precision of

Foho fluctuates more, with the highest precision at 90% at a 30-70 ratio and the lowest at 75% at a 90-10 ratio. This pattern indicates that the model is more accurate in recognizing Foho when the training and test data ratio is more balanced. Overall, the graph shows that the more balanced the data ratio, the more evenly distributed the precision between classes, and that the right training data proportion can improve the overall model performance.

The model has high and stable recall in recognizing Foho, ranging from 92% to 95% across all data ratios, even when training data is scarce. This reflects the model's strong sensitivity to the Foho class. In contrast, recall for Fehan is lower and more unstable, with a low of 69% at a 90-10 ratio and improving to around 85% at more balanced ratios such as 70-30 or 50-50. This pattern indicates that the model has more difficulty consistently recognizing Fehan, likely due to the more complex structure or vocabulary variations. Overall, the graph highlights the need to improve the model's ability to detect Fehan, to ensure classification performance is less biased toward one class and to ensure fair and representative predictions.

Based on the performance evaluation, the model shows stronger and more stable performance in classifying Foho than Fehan. Fehan's precision tends to be high and consistent across various ratios, but its recall is lower, indicating that the model often misses Fehan data despite being quite accurate in its predictions. In contrast, Foho has very high recall across almost all ratios, indicating that the model is very sensitive to this class, although its precision varies—particularly lower at the initial ratio (90-10) before increasing at more balanced ratios. The F1-score for Foho is consistently higher than Fehan, with a maximum value reaching 90%, while Fehan's F1-score reaches 89%.

Overall, the 50-50 and 40-60 ratios demonstrated optimal model performance, with the balance between precision and recall yielding the highest F1-scores for both classes. This indicates that a balanced proportion of training and testing data significantly impacts the stability and accuracy of model predictions. This comparison also highlights the need to improve sensitivity to Fehan to prevent model bias and to classify both dialects more fairly and representatively in real-world scenarios.

## 5. Conclusion

Based on the research results, it can be concluded that the Naïve Bayes method with TF-IDF architecture is quite effective in classifying the Foho and Fehan languages in Malaka Regency. From a total of 2000 data, consisting of 1000 data for each language, the model showed quite good performance. The highest accuracy achieved was 90% when using 40% of the data for training and 60% of the data for testing. In addition, the model also obtained a precision value of 93% in the Fehan language, 94% Recall in the Foho language and an F1 Score reaching 90% in the Foho language. The model shows that it is able to distinguish the two languages well. Overall, the Naïve Bayes method with TF-IDF architecture can be a useful solution in efforts to preserve and document regional languages automatically, especially in helping identify and recognize languages quickly and efficiently through technology.

## 6. Author's Statement

**Author's contribution and responsibility**

Write each author's contribution here, or mark the following fields.

[√] The authors made substantial contributions to the conception and design of the study.

[√] The authors are responsible for data analysis, interpretation, and discussion of the results.

[√] The author read and approved the final manuscript.

**Funding**

List research funding, if any.

**Availability of data and materials**

[√] All data is available from the author.

**Competing interests**

[√] The author declares no conflict of interest.

**Additional information**

Write down additional information related to this research, if any.

## 7. Confession

The author would like to express his gratitude to God Almighty for His grace and blessings, so that this research entitled "Classification of Regional Languages of Malaka Foho and Fehan Regency Using the Naive Bayes Method" could be successfully completed. The author would also like to express his gratitude to Widyagama University, Malang, especially the Informatics Engineering Study Program, for the support, facilities, and opportunities provided to carry out this research. The author would also like to thank the supervisor who patiently provided guidance, direction, and valuable input during the research process.

The author also expresses his appreciation to his family, friends, and all those who provided moral support, prayers, and assistance in various forms, which contributed to the smooth implementation of this research. He hopes that the results of this study will make a positive contribution to the development of science, particularly in the fields of artificial intelligence and digital image processing, and will serve as a useful reference for further research in the future.

## 8. Reference

[1] CP Ate and STM Ndapa Lawa, “The Shift of Tetun Fehan in the Family Realm in the Belu Community Speech Community in the RI-RDTL Border Region,” SeBaSa, vol. 5, no. 2, pp. 424–437, 2022, doi: 10.29408/sbs.v5i2.6672.

[2] Deni Yosef Nahak Berek and Frysa Wiriantari, ST, MT., “The Construction Process of the Uma Bei Kmeda Traditional House in Lorotolus Village, Malaka Regency - NTT,” J. Anala, vol. 7, no. 1, pp. 10–16, 2020, doi: 10.46650/anala.7.1.997.10-16.

[3] N. Katriani, “Classification of Regional Languages of Toraja, Halmahera, and Kalimantan Using Decision Tree and Gradient Boots Methods,” JATISI (Journal of Information Technology and Information Systems), vol. 9, no. 2, pp. 930–940, 2022, doi: 10.35957/jatisi.v9i2.1670.

[4] GM Momole, “Comparison of Naive Bayes and Random Forest in Regional Language Classification,” JATISI (Journal of Information Technology and Information Systems), vol. 9, no. 2, pp. 855–863, 2022, doi: 10.35957/jatisi.v9i2.1857.

[5] MZ Haq, CS Octiva, A. Ayuliana, UW Nuryanto, and D. Suryadi, “Naive Bayes Algorithm for Identifying Hoaxes on Social Media,” J. Minfo Polgan, vol. 13, no. 1, pp. 1079–1084, 2024, doi: 10.33395/jmp.v13i1.13937.

[6] KB Nahak, A. Rahim, A. Putera, and MD Bano, "Serial Verbs in Tetun Language," vol. 7, pp. 57–67, 2022.

[7] D. Tuhenay and E. Mailoa, “Comparison of Language Classification Using Naive Bayes Classifier (Nbc) and Support Vector Machine (Svm) Methods,” JIKO (Journal of Inform. and Computer), vol. 4, no. 2, pp. 105–111, 2021, doi: 10.33387/jiko.

[8] KB Nahak, "Personal Pronoun Greeting Forms in Tetum, Fehan Dialect," Jubindo J. Educator Science. Idioms and Indonesian Literature, vol. 5, no. 1, pp. 38–49, 2020, doi: 10.32938/jbi.v5i1.484.

[9] IKS Adnyana, "Linguistic Variations of Tetum Fehan Dialect: A Preliminary Study," Masy. Linguist. Indonesia., vol. 36, no. 1, pp. 93–102, 2018.

[10] Aji Priyambodo and Prihati Prihati, “Evaluation of Text Classification Feature Extraction for Classification Accuracy Improvement Using Naive Bayes,” Elkom J. Electron. and Comput., vol. 13, no. 1, pp. 159–175, 2020, doi: 10.51903/elkom.v13i1.277.

[11] D. Ariyanti and K. Iswardani, “Text Mining for Classifying Public Complaints at the Probolinggo City Government Using the Naïve Bayes Algorithm,” J. IKRA-ITH Inform., vol. 4, no. 3, pp. 125–132, 2020.

[12] I. Asyura, R. Dewi, S. Syekh Manshur, JK Raya Labuan, and C. Kadulisung Pandeglang Banten, “ANALYSIS OF MATHEMATICAL ABILITY OF PGSD STUDENTS TOWARDS THE USE OF GEOGEBRA CLASSROOM IN THE ERA AND POST THE COVID-19 PANDEMIC”, [Online]. Available: https://www.geogebra.org/.

[13] Alfath Daryl Alhajir, Yisti Vita Via, and Wahyu Syaifullah Jauharis Saputra, “A Rice Object and Foreign Object Detection System Based on Keras and Google Colab,” J. Inform. and Sist. Inf., vol. 2, no. 3, pp. 580–586, 2021, doi: 10.33005/jifosi.v2i3.369.

[14] N. Khoirunnisaa, K. Nabila Nastiti Kesuma, S. Setiawan, and A. Yunizar Pratama Yusuf, “Classification of Netflix App Review Text on Google Play Store Using Naive Bayes and SVM Algorithms,” SKANIKA Sist. Comput. and Tek. Inform., vol. 7, no. 1, pp. 64–73, 2024, doi: 10.36080/skanika.v7i1.3138.