Language Identification using Naïve Bayes

Phan Dang Nhan - BI12-336
Duong Hai Nam - BI12-300
Do Huu Nam - BI12-305

February 10, 2025

1 Introduction

Language identification is a fundamental task in Natural Language Processing (NLP). This project aims to build a system that can detect the language of a given text input using the Naïve Bayes classifier. The model is trained using a dataset containing multiple languages and is evaluated based on classification accuracy.

2 Dataset Analysis

The dataset used in this project contains text samples in multiple languages. Key characteristics of the dataset include:

- **Diversity**: The dataset covers a wide range of languages, ensuring robustness in classification.
- **Class Distribution**: Some languages have significantly more samples than others, which might introduce class imbalance.
- **Text Length**: Different languages have varied average word lengths, which might affect vectorization and classification.
- **Noise**: Some text samples contain special characters, numbers, or non-linguistic elements, which can impact model performance.
- **Source**: The dataset is compiled from various sources, ensuring diversity but also introducing inconsistencies in data quality.

3 Methodology

3.1 Preprocessing

The dataset undergoes preprocessing steps including:

- Tokenization
- Stopword removal
- Text vectorization using CountVectorizer
- Data splitting into training and testing sets

3.2 Model Selection

A Naïve Bayes classifier (MultinomialNB) is used for language identification due to its efficiency in handling text classification tasks.

3.3 Training and Evaluation

The model is trained on the preprocessed dataset and evaluated using accuracy, precision, recall, and F1-score.

4 Implementation

The project is implemented using Python with key libraries such as:

- Scikit-learn for machine learning
- Tkinter for GUI development
- Joblib for model persistence

5 Results and Discussion

5.1 Results

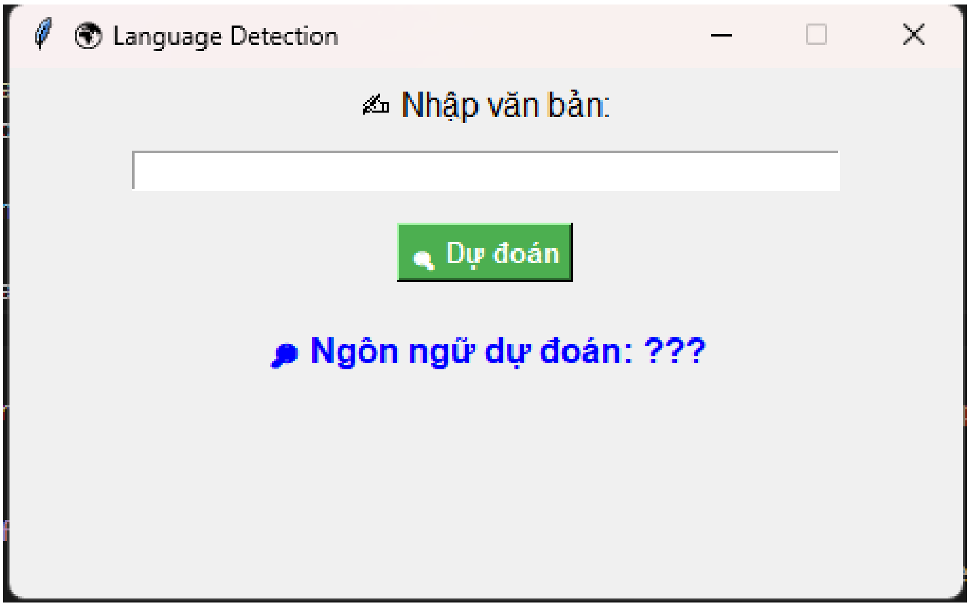


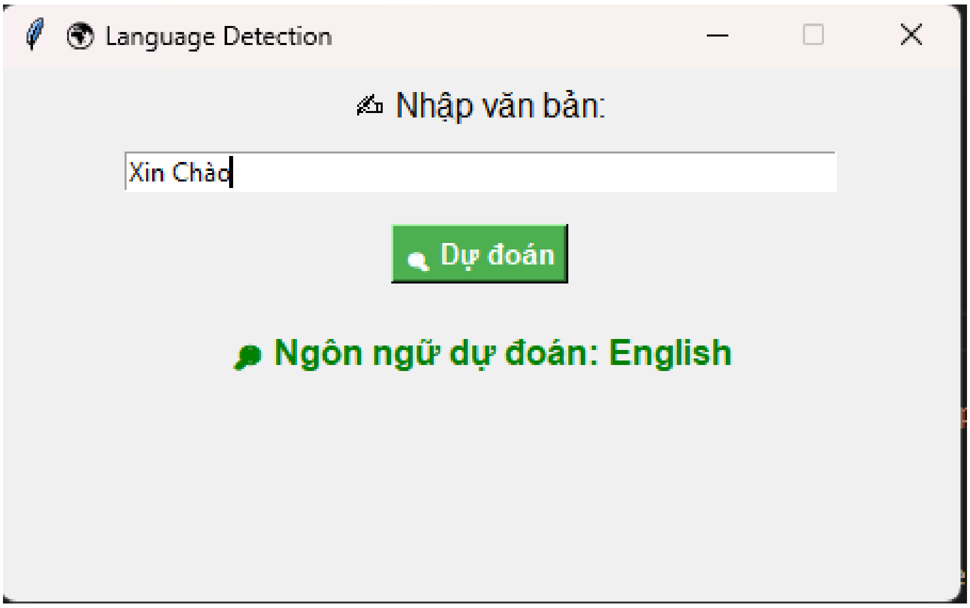
Figure 1: Result1

5.2 Statistic

The trained model achieves satisfactory accuracy on the test dataset. Below is the classification report summarizing the model's performance:

The accuracy of the model is 98.26%. The precision, recall, and F1-score for most languages are very high, indicating strong classification performance. However, we note that:

- Some languages have slightly lower recall scores, which may suggest difficulty in classifying minority classes correctly.
- Class imbalance can impact performance, especially for languages with fewer samples.
- Improving feature extraction techniques, such as TF-IDF or word embeddings, could enhance model robustness.



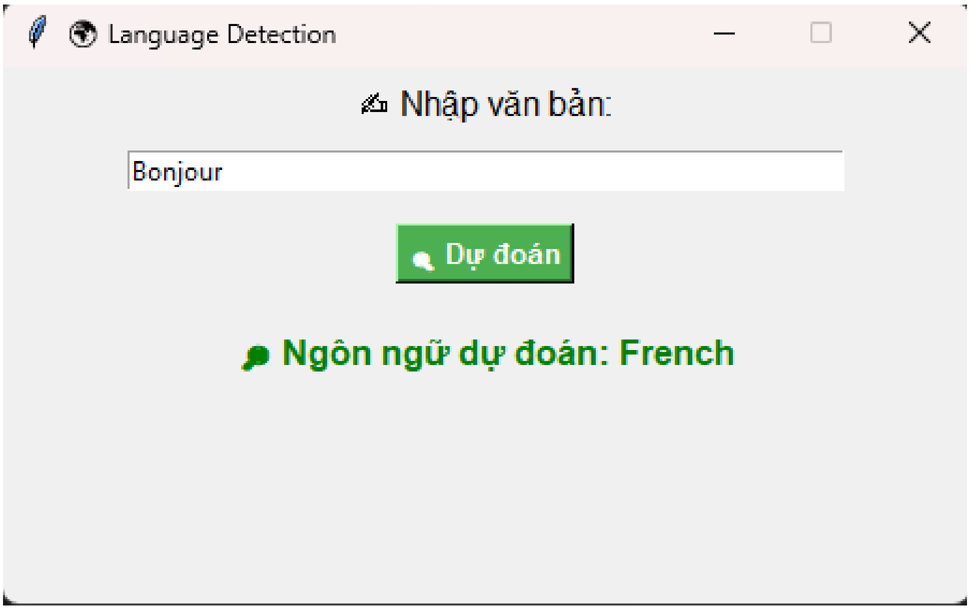
Xin Chào

💬 Ngôn ngữ dự đoán: English

Figure 2: Result2

6 Conclusion

This project successfully implements a language identification system using a Naïve Bayes classifier. The results demonstrate the feasibility of using probabilistic models for text classification tasks. Further enhancements can improve performance, particularly for low-resource languages.

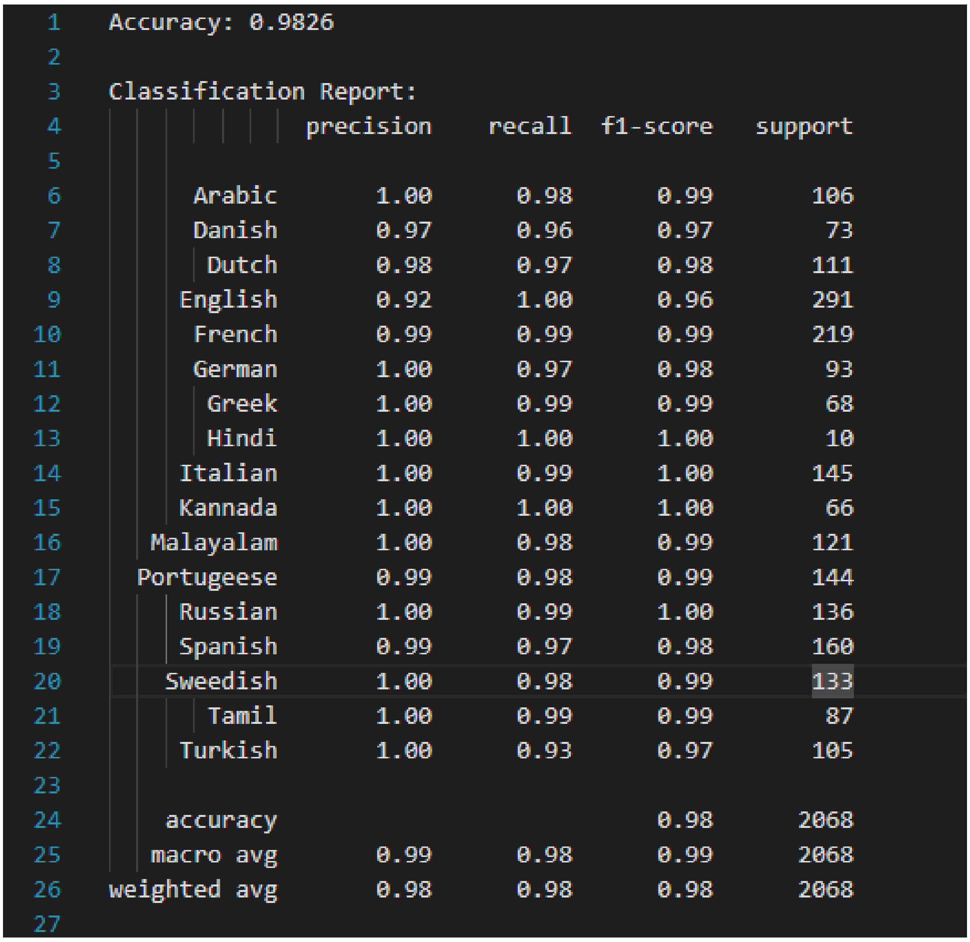


✍ Nhập văn bản:

Bonjour

Ngôn ngữ dự đoán: French

Figure 3: Result3



```
Accuracy: 0.9826

Classification Report:
              precision    recall  f1-score   support

      Arabic       1.00      0.98      0.99       106
      Danish       0.97      0.96      0.97        73
       Dutch       0.98      0.97      0.98       111
     English       0.92      1.00      0.96       291
      French       0.99      0.99      0.99       219
      German       1.00      0.97      0.98        93
       Greek       1.00      0.99      0.99        68
       Hindi       1.00      1.00      1.00        10
     Italian       1.00      0.99      1.00       145
     Kannada       1.00      1.00      1.00        66
   Malayalam       1.00      0.98      0.99       121
  Portugeese       0.99      0.98      0.99       144
     Russian       1.00      0.99      1.00       136
     Spanish       0.99      0.97      0.98       160
    Sweedish       1.00      0.98      0.99       133
       Tamil       1.00      0.99      0.99        87
     Turkish       1.00      0.93      0.97       105

    accuracy                           0.98      2068
   macro avg       0.99      0.98      0.99      2068
weighted avg       0.98      0.98      0.98      2068
```

Figure 4: Classification Report