Detecting Vietnamese Language and Lao Language

Nguyen Viet Bac

Abstract

This report details the implementation and evaluation of models for detecting Vietnamese and Lao languages. The dataset consists of 10,000 Vietnamese sentences and 10,000 Lao sentences. Several classification algorithms were employed, including Naive Bayes, Logistic Regression, Linear Classification, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The performance of these models is compared based on accuracy, F1 score, recall, and precision.

1 Introduction

Language detection is a fundamental task in natural language processing (NLP) that involves identifying the language of a given text. In this project, we focus on detecting Vietnamese and Lao languages. This task is crucial for various applications such as machine translation, sentiment analysis, and information retrieval.

2 Methods

To detect Vietnamese and Lao languages, we used several machine learning algorithms: Naive Bayes, Logistic Regression, Linear Classification, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). These models were selected due to their varying strengths in handling different types of data distributions and complexities.

The preprocessing steps involved:

- Removing symbols and numbers from the text.
- Converting text to lowercase to maintain uniformity.
- Using CountVectorizer to create a bag-of-words representation of the text data.

After obtaining the bag-of-words representation using CountVectorizer, the data was fed into each machine learning model for training and evaluation. The model training and evaluation process was carried out using the scikit-learn library in Python.

3 Experiments

3.1 Data

The dataset was obtained from the Leipzig Corpora Collection at wortschatz.uni-leipzig.de. It consists of two text files, one containing 10,000 Vietnamese sentences and the other containing 11,000 Lao sentences. To balance the dataset, we used 10,000 sentences from each file, resulting in a combined dataset of 20,000 sentences. Each sentence was labeled as either Vietnamese or Lao.

3.2 Settings

We split the data into training (80%) and testing (20%) sets. The training set was used to train the models, while the testing set was used to evaluate their performance.

3.3 Results

The performance of each model is summarized in Table 1. The Naive Bayes and Logistic Regression models achieved perfect scores, indicating their strong performance on this dataset.

Model	**Accuracy**	**F1 Score**	**Recall**	**Precision**
Naive Bayes	1.0	1.0	1.0	1.0
Logistic Regression	1.0	1.0	1.0	1.0
Linear Classification	0.99975	0.99975	0.99975	0.99975
KNN	0.84375	0.8396	0.84375	0.8807
SVM	0.9985	0.9985	0.9985	0.9985

Table 1: Performance of various models on the Vietnamese-Lao language detection task.

3.4 Analyses

The Naive Bayes and Logistic Regression models performed exceptionally well, likely due to their effectiveness in handling the bag-of-words features. Linear Classification and SVM also performed well, indicating that the linear separability of the data was high. KNN showed relatively lower performance, which may be due to the high dimensionality of the bag-of-words representation affecting the distance calculations.

3.5 Discussions

The high performance of most models suggests that the Vietnamese and Lao sentences are distinguishable based on the features extracted. This can be attributed to the distinct linguistic characteristics of the two languages. However, the near-perfect scores also raise questions about the potential overfitting of the models, which could be investigated further by using cross-validation or more complex datasets.

3.6 Limitations

One limitation of this study is the use of a bag-of-words model, which does not capture the semantic meaning of the text. Future work could explore more advanced text representations such as TF-IDF or word embeddings. Additionally, the dataset size is relatively small, which might not represent the full complexity of the languages.

3.7 Future Work

Future work could involve expanding the dataset to include more sentences and a wider variety of topics. Additionally, exploring more sophisticated text representations, such as TF-IDF or word embeddings, could enhance the models' ability to capture semantic meaning and context. Implementing and comparing the performance of advanced machine learning and deep learning techniques, such as ensemble methods and neural networks, may also provide further insights into improving classification accuracy and robustness.

4 Conclusion

In this project, we successfully implemented and evaluated several models for detecting Vietnamese and Lao languages. The Naive Bayes and Logistic Regression models achieved perfect performance on the test set, demonstrating the effectiveness of these methods for this task. Future work will aim to address the limitations and further improve the robustness of the models.

References

[1] Jose Maria Gomez Hidalgo: Language Identification as Text Classification with WEKA.