Research on some NLP techniques and application of Vietnamese text classification

Nguyen Thi Thu Hoa

Abstract: This study investigates several natural language processing (NLP) techniques and their application to Vietnamese text classification. Given the unique linguistic characteristics of Vietnamese, including its monosyllabic nature and complex word segmentation challenges, standard NLP methods often require significant adaptation. This research explores and compares multiple approaches, ranging from traditional machine learning models (e.g., Support Vector Machines, Naïve Bayes) with handcrafted features to advanced deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models (e.g., PhoBERT). A key focus is on evaluating the impact of different preprocessing techniques, particularly word segmentation and part-of-speech tagging, on classification performance. Experimental results on benchmark Vietnamese datasets demonstrate that deep learning models, especially those leveraging pre-trained language models like PhoBERT, significantly outperform traditional methods in accuracy and F1-score. The findings provide a comprehensive benchmark for Vietnamese text classification and offer practical insights into selecting optimal techniques for various application domains, including sentiment analysis and topic categorization.

Keywords: NLP, text classification, Vietnamese, techniques