COMP 550 Proposal: Supervised Text Classification Task on Chinese Dataset

Xijuan Sun, Zihan Wang, Jack Wei

1 Introduction

In this proposal, we outline a comprehensive approach for conducting a supervised text classification task on a Chinese dataset, specifically the ”10 Shopping Carts” dataset. Our goal is to explore and compare the performance of text classification models across different language domains: Pinyin-level, Chinese-word-level, and English-level text classification. We also aim to analyze the distinct features and characteristics of each language domain. Additionally, we will evaluate various machine learning models, including LSTM-based models, logistic regression, Naive Bayes, support vector machines (SVM), convolutional neural networks (CNN), and BERT, for text classification on the variant datasets.

2 Objectives

Our primary objectives for this project are as follows: * Implement text classification on the ”10 Shopping Carts” dataset, which is a well-established text classification dataset. * Conduct text classification at three different language levels: Pinyin, Chinese words, and English. * Train LSTM-based models for each language domain and compare their performance. * Analyze the unique linguistic features of each language domain. * Evaluate the performance of various machine learning-based models on the different language domains. * Optionally, explore the possibility of contrastive learning by merging features from the Chinese word, Pinyin, and English word levels to understand the impact of data features on text classification performance.

3 Methodology

Our methodology consists of the following steps: 3.1 Data Preprocessing * Convert Chinese characters to Pinyin and translate Chinese text to English using Python packages such as pinyin and translation APIs. * Prepare the ”10 Shopping Carts” dataset for text classification tasks, ensuring data is cleaned, tokenized, and labeled appropriately. 3.2 Model Training * Train LSTM-based models separately for Pinyin-level, Chinese-word-level, and English-level text classification. Fine-tune these models on the ”10 Shopping Carts” dataset. * Train other machine learning-based models, including logistic regression, Naive Bayes, SVM, CNN, and BERT, on the same datasets for comparison. 3.3 Evaluation and Analysis * Evaluate the performance of all models using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score). * Analyze the features and characteristics of each language domain to gain insights into their strengths and weaknesses. * Compare the models' performance across different language domains to identify which models work best for each domain. 3.4 Optional: Contrastive Learning * If time permits, merge features from the Chinese word, Pinyin, and English word levels and experiment with contrastive learning to examine the impact of data features on text classification performance.

4 Expected Outcomes

We anticipate the following outcomes: * Comparative analysis of text classification model performance across Pinyin, Chinese words, and English levels. * Insights into the unique linguistic features and challenges of each language domain. * A comprehensive comparison of various machine learning models' effectiveness in classifying text in these different domains. * If applicable, insights from the optional contrastive learning experiment, revealing the influence of data features on classification performance.

5 Conclusion

This project aims to provide valuable insights into the challenges and opportunities of text classification in different language domains and the comparative performance of various machine learning models. By undertaking this study, we seek to contribute to the understanding of cross-lingual text classification and improve text classification methods in real-world applications. Please let us know your thoughts and any further details or modifications you would like to discuss before proceeding with this research project.