Comp 550 Project Proposal

Zihan Wang 260825558

Xijuan Sun 260896211

Jack Wei 260837048

Chinese NLP tasks are challenging due to the absence of natural delimiters like spaces, which are common in languages like English. For this project, our team aims to explore and evaluate different word segmentation schemes for a Chinese online-shopping review dataset to improve the performance of LSTM on text classification and sentiment analysis tasks. Specifically, we investigate various tokenization schemes, such as pinyin-based, character-level tokenization as well as word/subword segmentation. Through this study, we want to answer the following questions:

1. Which granularity of Chinese feature tokens is optimal for sentiment analysis?
2. Which granularity is best suited for general text classification?

1 Feature Extraction

The focus of our study is on three different ways to tokenize chinese words. These include: pinyin-based tokenization, character-level tokenization, word-level segmentation. We aim to perform these tokenization as follows

1. **Pinyin-based**: Pinyin is a pronunciation-based Romanization system for Chinese. For this feature, we convert Chinese characters into pinyin without tonal information. This can be done through the Python pinyin package.
2. **Character-level**: This is the most straightforward method for Chinese language tokenization.
3. **Word-level Segmentation**: Word-level segmentation tokenizes multi-character words or phrases. Various tools, like Jieba and the Stanford Word Segmenter, can be employed to execute segmentation for Chinese.

2 Dataset

At this preliminary stage, we have chosen one of the classification dataset from a Chinese NLP Corpus [1], ”online_shopping_10_cats”, a collection of online reviews encompassing 10 product/service categories, including books, tablets, phones, hotels, shampoos, and more. Each one of the 60000 reviews is labeled with its respective sentiment (positive or negative), facilitating both classification and sentiment analysis tasks. If the dataset is shown to be too easy for sentiment analysis/text classification task then we are open to selecting other datasets available in the same repository.

3 Model

Our base model for the classification/sentiment is LSTM. An embedding layer will be employed to convert the processed tokens into dense vectors. For the training process, we will employ categorical cross-entropy for multi-class classification and binary cross-entropy for the binary sentiment classification. Moreover, we will consider early stopping based on validation loss to ensure the most optimal model is retained.

4 Evaluation

The quality of the word segmentation schemes will be evaluated by the performance of the classifier models on both tasks. Since the classes/labels in the datasets are well-balanced, we will primarily use accuracy as the evaluation metric, though we’re open to considering other metrics like F1-score and AUC-ROC.

5 Optional Experimentation Options

If the scale of our project is limited and if time permits, we may expand our project to include experimentation on some more advanced Chinese segmentation methods in the literature. Moreover, we may expand our models to include linear models with bag-of-words and n-grams features.

[1] https://github.com/SophonPlus/ChineseNlpCorpus