Group 3 Project 2 Proposal

Limitations of Translation: How much translation affect the analysis of Chinese text in different models? (Natural language processing with Chinese character)

MATH 4995 (2021 Fall)

# The dataset description

The dataset we will be using would be one of the dataset from “weibo\_senti\_100k” (<https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo_senti_100k/intro.ipynb>). It is a dataset containing 100k reviews in Chinese 50k positive and 50k negative, which all of them comes from Sina Weibo, and contains either positive or negative review (also known as label).

From the dataset, the label would be treated as response variable, and the review would be used as predictor variable.

# Measure of accuracy

We would first separate the dataset into 2 parts, with 20% of the data being the testing set and 80% of the data being the training set. Moreover, as the dataset has equal amount of positive and negative review, the amount of positive and negative review in both testing and training set would be equal. During the training phase, only the data in training set would be used. The scoring would depend on the testing set.

As the response is balanced, the formula for scoring would be the accuracy of the result. (i.e. )

# Method

We would be using multiple method in this project.

First, there are some pre-existed method in public (<https://github.com/ymcui/Chinese-BERT-wwm/tree/master/data>). We would be using some of the method as a way of training. Moreover, we would also be using some other method to train and test the model. Some potential method includes CNN, RNN, Bayes Techniques.

Moreover, we would be analyzing how much better it is if the text is analyzed directly in Chinese when compare to text is analyzed after it is translated to English. We would be using the same method as above, but before feeding the data to the model, we would translate the data into English first. The scoring will also be using the method stated in measure of accuracy. We hypothesize that there exist some performance lost on analysis if the text is translated.