1. xlmr通用模型，bert只能用在中文里
2. bert和xlmr截取信息时，只截取了前500个，根据词库，bert结果有更少的信息丢失

Limitation

1. 数据没有进行预处理，比如“的”“了”
2. 分类器部分，只用了传统机器学习的方法，比如lstm，cn卷积神经网络，mlp多层感知器

3．调整classifier的参数

Error Analysis

1. 内容比较分散，句子与句子之间的共性比较少，句子与句子之间的主题联系比较少

Since both bert and xlmr only cuts the description in 500 words during the word embedding process, therefore, compare to bert, xlmr is using data set “CommonCrawl” which could be reason causes more information lost. Also, by looking though the real estate’s data, the data is not preprocessed. It contains a lot of useless information in its description, for example, there are a lot of numbers and meaningless words (like “is”, “of”, “to”) in real estate categories compare to other categories. In addition to that, because for both bert and xlmr model, the prediction accuracy is relatively low in real estate categories compare to other categories. It is reasonable to conclude that the content is scattered, there are fewer commonalities between sentences, and fewer topic connections between sentences.

1. To improve the prediction accuracy, we could preprocess the data. For example, remove the meaning less words like “is”, “to”, “of”. Also, we could delete the proper nouns that would confuse the prediction result like movie, street and product names etc. Also, for the classifiers, we use traditional machine learning methods (sklearn) to predict labels. We could improve the prediction result by using advance machine learning methods such as Long Short-Term Memory, Convolutional Neural Network, and Multilayer Perceptron. In addition to that, we are using default hyper-parameters setting in our classifiers. To improve its accuracy, we could adjust the hyper-parameters of the classifiers.