Proposal of modeling approach:

Our objective is to use consumer complaint dataset released by Consumer Financial Protection Bureau (CFPB) to classify complaints to different issues We specifically focus on mortgage complaints in this project. Roughly 35 percent of customers provided consent to release their complaint narrative. These complaints can provide other customers and companies meaningful information such as: issues occurring in certain company, how a company treats their customers and what and when an issue is likely to occur.

For this project we will rely on pre-trained large language models such as ELMo, BERT, RoBERTa, ULMFiT and XLNET. These language models have been trained on very large datasets and perform very good on benchmark tasks such as GLUE. We will tune these models with our consumer complaint. We will split the data into training-validation-test sets, and apply cross validation in the training to not overfit our model.

Most of these models tend to have 512 token length. In our EDA we found that median length of our documents was around 1000 tokens. There are models that can handle longer documents, but they are computationally more expensive. We will just use first 512 tokens in our documents. We will tokenize our text and use word embeddings as inputs in our models. We don’t plan to perform extensive pre-processing since these language models are not affected much by pre-processing steps.