Text Analysis for Social Scientists and Leaders

Assignment 3

You were introduced to some functions for calculating word embeddings. Next, you will apply it to a dataset of consumer complaints to the CFPB. Your primary outcome of interest is whether the company's response to the claim is disputed or not (binary, 0 or 1).

Before you start, filter the data to focus on only the "Credit Reporting" products. Then randomly split the CFPB data into a training set (80% of documents) and a test set (20% of documents). Also, please download the full 6GB fastext word embedding file, don't just use the small one shared in class.

1. Train a LASSO model using ngrams from the narrative text box to predict whether each claim is disputed. Train another LASSO model using word2vec embeddings from the same text as features. Finally, train a LASSO model that uses both the ngram and embedding features together. Calculate the accuracy of these three models. (hint - the training set size is small, so you may need to set s="lambda.min" when you make predictions or extract coefficients from the glmnet model)

2. Choose two other benchmarks you would like to use to compare to the trained models. Create a plot showing the average accuracy scores (with 95% confidence intervals) of the models and benchmarks from questions 1 and 2. Give a description of what you found.

3. Use the Distributed Dictionary Representation technique to compute similarity of all the narratives to a dictionary - the Loughran & McDonald uncertainty dictionary - to create an "uncertainty score" for each narrative. Now notice there is a "sub-issue" column, indicating different types of of issues within the credit-reporting product. Calculate the average similarity for each sub issue (mean and standard error) and put these all on a plot. Make sure the sub-issues are listed in order, from highest to lowest average uncertainty. Do these differences make sense?

4. Use these uncertainty scores to predict if the complaint is disputed or not. Then use dfm_lookup to apply the dictionary list the traditional way (i.e. without the vector representation). Make a plot comparing these two accuracy scores.

[note: if you can't get spacy working, then you can run this function on posit.cloud easily, if you sign up for a free account]

5. Use the politeness package to extract politeness features from the training data. Make sure to use spacy to extract grammar-aware features. Then, use those features to make a politeness plot to show which features are more or less common among narratives that are disputed (set middle_out = .05 to filter marginal features). What do the features suggest about the complaints language?

6. Filter the data down to something much smaller, that only includes companies that have at least twenty complaints in the data. Then select 20 complaints at random from each one of those companies. Process this dataset through the spacy engine.

a) Make a list of 20 words that are processed differently by stemming and lemmatisation algorithm.

b) What are the 30 most common noun phrases in the narratives in your data?