SI 630 HW4 Report

Problem 1: See jupyter notebook

Note: To select my “best” prompt model, I re-ran all of the models 10 times on randomly selected data, and picked the model which had the highest f1 score most number of times out of 10.

Problem 2:
The best prompt was as follows:

'{"placeholder":"text_a"}. On a scale of 1-5, how rude was that sentence? {"mask"}'

With scores of 4 and 5 being verbalized to toxic, and 1-3 being verbalized to non-toxic.

On a random samples of the dev data, this model had F1 scores as high as 0.6, but on average it was much lower (~0.17)

Problem 3:
The best zero-shot prompt was as follows:

If your waiter said this, how much would you tip them between 0 and 20%:

{}

Answer with a number between 0 and 20.

With values below 10 being verbalized to toxic, and the rest verbalized to non-toxic.

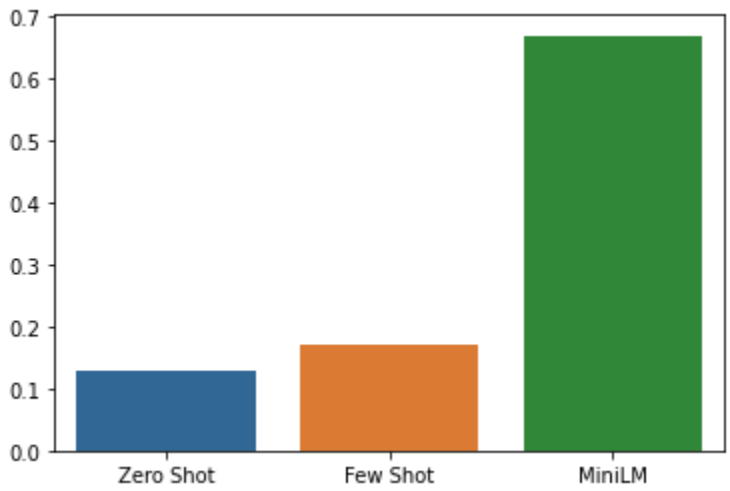
This model performed reasonably well on samples of the dev data (F1: ~0.66), but only had an F1 score of around 0.17 on the test data (see Kaggle)

Problem 4: The MiniLM model had an F1 score of 0.67 on the dev dataset.

Problem 5:

The Few-Shot model performance did not seem to change when increasing the training instances from 10 to 500. It stayed the same at around 0.17

Problem 6:



The Zero Shot and Few Shot perform significantly worse than the MiniLM, and the Few Shot model did not seem to increase in performance with more training examples (hence the bar plot which does not account for training instances).

The MiniLM was trained on all the ~160,000 training instances, the few shot model was trained on 10 - 500 instances (with performance remaining same)

I'm unsure on how to make a guess for the number of training instances required for the prompt based models to match the MiniLM, but it definitely is more than 500-1000 instances.

I did notice that in some randomized runs, the F1 scores did reach close to 0.6 so it is possible that with more prompt engineering the models could perform better and match the MiniLM.

Problem 7:
Kaggle Username: Chittaranjan Velambur Rajan