RQ 2.1: Summarize the results of your experiments for Problem 2. Are your results “good” or “bad”, and importantly, how do you determine that?

From my perspective, the result is not good. Since the “macro avg” and “weighted avg” are 0.24 and 0.34 respectively. As we all know the more F1 score tends to 1, the better the model performs. Apparently, it is not the case.

RQ 2.2: In Problem 4, you implemented undersampling. How did your results change compared to Problem 2? How would “oversampling” have looked like for this task?

The result of using undersmapling is better, since the “macro avg” and “weighted avg” are 0.39 and 0.41 respectively.

If in this task “oversampling” is used, many new instances would be created for those parties with small numbers. This would first increase the training time, and second add some noises to the training_dateset. The efficiency and accuracy would be both affected.

RQ 2.3: Why is it important to do a hyperparameter search before drawing conclusions about the performance of a model? Why do you think it is often not done, anyway? Why should you never tune hyperparameters on the test set?

Without doing a hyperparameter search, it is possible we do not leverage the real power of the model, when the inappropriate parameters are picked. Maybe the model we choose is not the optimal one and lose the real good one.

The reason for not being used is it is time consuming, especially when we have to use a large dataset and many parameters we want to try.

The only role the test set plays is to show us how good the prediction performance is when facing new situations. If we use this dataset to tune hyperparameters, it is not the data the model has never met, so the performance on the dataset can be regarded as the real ability of the model.