*Note: RT.2 and BC.1 – 4 are as per S.3.1 of the main thesis.*

**More Effort**

We noted in S.3.2.2 of the main thesis that RT.2 classes were harder to identify than RT.1 metadata. RT.1 metadata could be identified usually from the first or last pages of each report that consists of usually 9 to 20 pages (or more).

RT.2 classes however, usually required reading whole sections of the report to identify and as such required far more time to do.

**Basic Procedure**

RT.2 datasets for training and testing purposes were difficult to create due the rarity of some BC.1 – 4 classes as per our discussion in S.3.2.2 of the main thesis.

The time allocated to data preparation was three weeks given the time we had to complete the project. We originally targeted for a minimum of a 75% training – 25% test ratio for each RT.2 experiment where N = 300 in each experiment (A common training-test split ratio for relatively small datasets [1, 2]. But as we learnt more about the population proportions of rarer classes we realised that we might come out short.

To make time for more data preparation, we had to be more efficient elsewhere which we found in the Document Retrieval Experiments. If we refer to S.3.4, we note that we were running grid-search cross validation on eight models on a large number of feature-level and model-specific hyperparameters to tune those models to an optimal complexity for our dataset (according to grid-search cross validation’s logic [3–5]). These would require substantial computational time to perform and would involve only having the relevant training sets for RT.2 BC.1 – BC.4 ready. So, when the training sets were at least N = 225[[1]](#footnote-1) (75% of 300) we executed the code for training and tuning models. Then while these computations were running, we continued to code up more and more observations for RT.2 BC.1 – BC.4 *test* sets until we had them to our desired training-test ratio.

As some reports would have multiple classes which would be relevant to more than one RT.2 experiment (e.g. one report may have classifications for BC.1, BC.2 and BC.4), the datasets for each BC ended up with differing numbers as it would impossible to know which reports would have what classifications available and we did not want any additional classifications to go to waste.

**BC.1 Training/ Test Set Sizes**

We will also note that BC.1’s training and test set are far larger than BC.2 – 4. The reason for this was the optimality of System A as described in S.3.3. From the results in S.4.1 we saw that System A could extract MD.1 Appeal Outcome from reports at 100% precision and 100% recall because of how rigid the HTML structure was on the UK Freedom of Information (FOI) tribunal website and it would likely always result in 100% precision and recall in our dataset given the logic we discussed in S.3.3.1 of the main thesis and Appendix C.1.1.

One will note that MD.1 (Appeal Outcome) and BC.1 (The Appeal Outcome) are actually the multi-class problem and binary class problem of the same variable (Appeal Outcome). MD.1 Appeal Outcome has about 6 primary outcomes (“Allowed”, “Dismissed”, “Consent Order”, etc.) possible (among many, many variants) and we only want to know a sub-set for BC.1 being “Allowed” or “Dismissed”.

So, given the perfect precision and recall of System A’s performance on MD.1, we took the minor risk of allowing System A to code all our training and test data for BC.1 experiments for “Allowed” and “Dismissed” and reviewed a sub-set of 350 observations manually to check for the actual appeal outcome with no errors found.

**References**

1. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer New York, New York, NY (2013). https://doi.org/10.1007/978-1-4614-6849-3.

2. Hastie, T., Friedman, J., Tibshirani, R.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, New York (2001).

3. Stone, M.: Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion). Journal of the Royal Statistical Society: Series B (Methodological). 38, 102–102 (1976). https://doi.org/10.1111/j.2517-6161.1976.tb01573.x.

4. Geisser, S.: The Predictive Sample Reuse Method with Applications. Journal of the American Statistical Association. 70, 320–328 (1975). https://doi.org/10.1080/01621459.1975.10479865.

5. Allen, D.M.: The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction. Technometrics. 16, 125–127 (1974). https://doi.org/10.1080/00401706.1974.10489157.

1. No way to know when we would get exactly 225 and each report could have multiple relevant BC.1 – 4 classifications available and we did not want any classifications to go to waste, resulting in minmum 225 but likely more. [↑](#footnote-ref-1)