CSE 3521 HOMEWORK 6

BRENDAN WHITAKER

1. Our awesome new model, trained on a random sample of 80% of the given dataset will make different and better classifications than our rivals' model when both are run on the remaining 20% of the dataset.

2. First, you need to control for the language that the tweet is in as it appears when it is parsed by our model. In other words, we must control for the proportion of English, French, Arabic, and Mandarin tweets used as input during testing for both the training and testing data. We need to do this because either our model or our competitors' may work particularly well on a certain language, but not so well on the others, for example, and if our test favored (proportion-wise) that language, the model which worked better for it would appear better overall, when in reality it is only better for that specific language. Second, since our data consists of some tweets which have been translated and some which have not, we must control for whether or not the tweet was translated from another language or not. The grammatical and thematic structures of French and English may not be preserved during translation, and one of the models may work better on content from a tweet written in the language in which it was originally composed, and for this reason we must control for the translation status as well.

3. (a) Since there are less Arabic tweets than Mandarin, we first take a random sample of 16,500 Mandarin tweets, plus a random sample of 16,500 each of the total 25,000 original English tweets and the total 28,500 original French tweets, so we have 4 total strata, one for each language. Then we will take a simple random sample of 80% of each of these strata and make that our training data. This way we get an equal proportion of each language, and an equal proportion of translated vs non-translated tweets.

   (b) The remaining 20% of each of our equal sized strata will compose the testing data.

   (c) We will choose to train both models on the selected training data, so that we control for differences between our rivals' proprietary training data and our own.

   (d) We will evaluate both models on same testing data generated in part (b), so that we know that there are no differences in difficulty in classifying tweets between the datasets that each model was tested on.

4. The process by which we organized and sampled our training and testing data means that we already know that each of the variables we wish to control for is being given equal weight in each experiment. Therefore we will run a number, say 20, of experiments using this same sampling strategy, but taking a new random sample from each strata population with each new experiment, so that, apart from the smallest group, the Arabic tweets at 16,500 in total, we have randomly incorporated as much of our data as possible to get the best possible prediction of how our model will perform on the population of all tweets in all languages. These experiments will train both models on the same generated training data, and test them both on the test data, and using a statistical test, say a 2-sample t-test for the mean number of correctly classified languages, determine whether the difference between the two models' results is meaningful.

Date: AU17.

5. The results from this experiment will not give enough evidence to support/reject the given claim. This is because the claim is very general and specifies no language. Thus, the comparison between the models must be valid for all languages, and so we need a lot more datasets, enough to comprise a random sample of all languages, perhaps weighted by what proportion of people tweet in them worldwide, in order to make a valid conclusion on this claim.