Bag of words – Input 5000 words – Output CTEC Instructor Score > 5 (1) else (0)

* Unable to run Weka without running out of heap space

Bag of words – Input 2500 words – Output CTEC Instructor Score > 5 (1) else (0)

* ZeroR gives 55% accuracy
* Neural net with 0 hidden layers – Couldn’t get it to run in a reasonable amount of time

Bag of words – Input 1000 words – Output CTEC Instructor Score > 5 (1) else (0)

* Naïve Bayes
* Once again, neural nets simply take too long

Bag of words – Input 100 words – Output CTEC Instructor Score > 5 (1) else (0)

* ZeroR – 55 percent
* Naïve Bayes
* BayesianLogisticRegression did pretty well, with a 70 percent classification instance
  + Interesting coefficients: great (0.452), professor (0.009), fun (0.286), learned (0.190), difficult (-0.239), best (0.599), boring (-0.796), awesome (0.910), could (-0.485), didn (-0.606), amazing (0.940), bad (-0.308), project (-0.460), homework (0.004), subject (0.004)
* LogisticRegression, 71 percent
  + Tells us roughly the same thing

Bag of words – Input 100 words – Output CTEC Instructor Score > 4 (1) else (0)

* Concern is that there are not enough negative variables to make it worthwhile
* ZeroR is around 85 percent, and logistic and bayeslogistic perform around the same (naïve bayes does worse)
  + coefficients don’t really yield any new information

Pretty cool so far. It’s clear the Bayesian Logistic Regression is the best performer here, so we are going to sort of build on top of that, so let’s revisit this:

Bag of words – Input 500 words – Output CTEC Instructor Score > 5 (1) else (0)

* + 75 percent accuracy
  + While our accuracy is just okay, what the algorithm prioritizes tells us quite a lot about what constitutes positive and negative CTECs.

Final Tests: