IST736 HW3

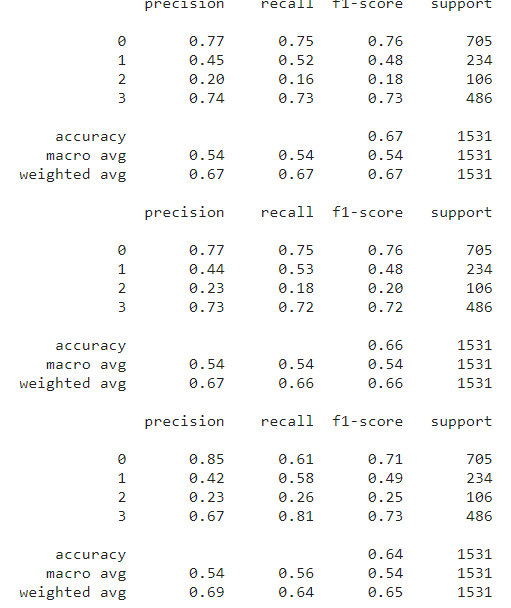
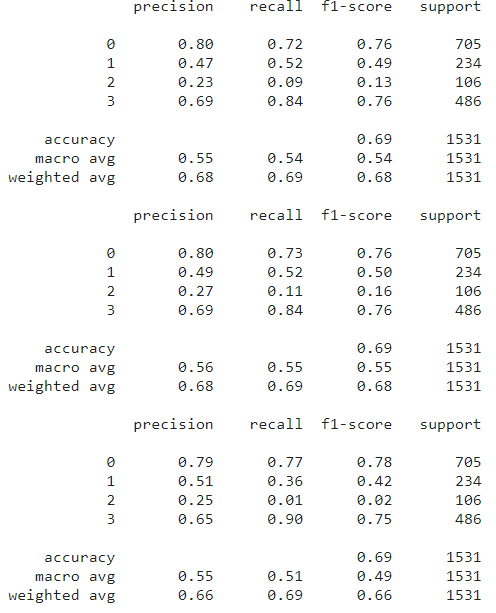
Yicun Deng

Prof. Bei Yu

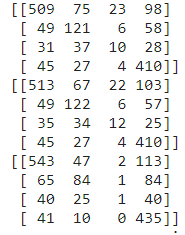
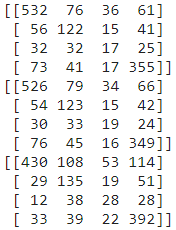
**SVM or MNB for Causal Language Detection**

A and C) I had run countvectorizer, booleanvectorizer, and tfidfvectorizer for both SVM and MNB (count, Boolean, tfidf from top to bottom). (confusion matrix is in the ipynb file)

The result shows that, SVM: countvectorizer MNB: booleanvectorizer



Confusion matrix: SVM: MNB :



B)

The top ten words features for each category of SVM are: no relation (esophageal, proposed, specifically, meal, bias, obtained, policy, superior, needed, european), direct causal (did, resulted, determinants, administered, successful, highly, tolerated, effective, contributed, theory), conditional causal (appeared, improve, mg, cad, influence, helpful, mediated, problem, useful, length), correlation (variant, dysfunction, correlation, older, significant, correlated, rygb, predictors, predicted, associated)

The top ten words features for each category of MNB are: no relation (result, trial, high, treatment, needed, clinical, risk, study, studies, patients) direct causal (woman, effect, significantly, effective, cancer, study, weight, treatment, risk, patients), conditional causal (role, breast, early, effective, reduce, cancer, increase, improve, patients, risk), correlation (woman, cancer, diabetes, higher, levels, study, increased, patien0ts, risk, associated).

D)

SVM error: MNB error:

errors: 505 errors: 474

too related errors: 251 too related errors: 194

poor related errors: 254 poor related errors: 280

Combining with our knowledge of top10 features, We can clearly see that SVM’s feature is very literate: ‘correlation’, ‘correlated’ are marking correlation. Most of the errors contains the top10 features for SVM’s correlation and direct causal categories feature. To improve the performance, we should create lexical entry for words like ‘correlation’ and ‘correlated’. Also, word like ‘did’ should be removed from the feature. For the MNB, the performance is indeed better. But we can still remove words like ’risk’ which appeared twice or ‘woman’ which doesn’t make sense to improve the performance.

E)

Overall, the performance of MNB is better than SVM. Part of the reasons why is that: SVM is a simpler regressive algorism that it is natural not good for a task like categorizing relations which is more complicated semantically and thus mathematically. Also, Boolean vectorizer seems to best fit with MNB. This is interesting. I believe that this might be caused by the nature of Bayes theorem that it is more looking into a status instead of counts.