IST736 HW5

Yicun Deng

Prof. Bei Yu

**Bert and SVM and MNB**

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

A screenshot of a computer

Description automatically generated with low confidenceTable

Description automatically generated

Confusion matrix: SVM: MNB :

Text, letter

Description automatically generatedText, letter

Description automatically generated

Bert:

A picture containing text, receipt

Description automatically generated

[[592 41 17 55]

[ 16 198 6 14]

[ 6 6 91 3]

[ 28 12 3 443]]

D) Errores

SVM error: MNB error: Bert error:

errors: 505 errors: 474 errors: 207

too related errors: 251 too related errors: 194 too related errors: 71

poor related errors: 254 poor related errors: 280 poor related errors: 136

Conclusion:

Bert is outperforming SVC and MNB with their best match of convector. Bert’s weighted accuracy is 87% compared to SVC’s 69% or MNB’s 68%. Also, Bert made less mistakes. Bert both make less predicting a text to be too related or too poor related to the topic compared to the SVC and MNB models (more than 50% reduction, this is meaningful because it shows that Bert not only predicts better, but also less over-fit than SVC and MNB). After browsing Bert’s errors, I found that SVC and MNB are marking these texts wrong too. Because of that Bert doesn’t support word feature, we can’t really investigate its word feature’s relationship to the errors. So I did a eye-balling test and vectorized these errors. I found that BERT is not doing well when the text is embedded of abbreviation word like MRI, BMI. Bert don’t have knowledges of these MRI. I don’t have a good solution for improving Bert’s knowledge of abbreviations but perhaps we can supply the knowledge of abbreviation by change all the abbreviation word to its original versions in the text on data cleaning phase.

Errors:

“The absence of two mtDNA mutations in ND1 gene rules out the possibility of involvement of these mutations in early onset diabetes in Pakistani population.

Multiple and immediate access to the web-based education program at home may prove useful as a source of reference for women with GDM.”