CMPUT 497 Assignment 5 Report

Yonael Bekele
University of Alberta
Edmonton, Canada
yonael@ualberta.ca

Michael Lin
University of Alberta
Edmonton, Canada
michael.lin@ualberta.ca

OVERVIEW

In assignment 5, we implemented a Naive Bayes Text Classifier to identify texts as belonging to one of the 5 classes: business, entertainment, politics, sport, or tech.

DATA

We are provided with 3 datasets, see Fig. 1 for detail

Train	Test	Evaluate
1704	668	71

Fig 1. *Datasets*

Pre-processing

We use the NLTK built-in tokenizer and predefined stop words by NLTK to preprocess the document. We originally used the tokenizer from the spaCy library, but found that NLTK's tokenizer performed better and faster. We chose to remove the stop words to see if there was a difference, because in the text, *Speech and Language Processing,* it states that the improvement in the effectiveness of classification by removing stop words varies (Jurafsky and Martin, 2019). While we will describe our findings in further detail in the Discussion portion of this paper, we went forward with removing stop words in the pre-processing stage. The training dataset was also randomized for each document to remove any pre-existing order it may have been provided in.

METHODS

We implement k-fold cross-validation to train the Naive Bayes (NB) Classifier. We divided the training dataset into *k* chunks as evenly as possible, chunks are obtained sequentially instead of random sampling since the dataset has been shuffle during preprocessing. We train *k* NB classifier, in each iteration, we pick one chunk as validation dataset and the rest as the training dataset. After each iteration, we will record the accuracy of the classifier on the validation dataset. Each chunk will only be used as the validation dataset once in the whole training process. The classifier that has the highest validation accuracy is used to perform testing and evaluation later on.

ERROR ANALYSIS

We found that the aggregated pooled micro-average precision was 0.99251 and macro-average was 0.99228.

	sport	business	politics	entertainment	tech
sport	162	1	0	0	0
business	0	156	1	0	0
politics	0	1	123	0	0
entertainment	0	0	0	116	1
tech	0	0	0	1	106

Fig 2. *Confusion Matrix for a 5-class categorization task, showing for each pair of classes in a document from the datasets were (in)correctly assigned to another class. Row = Predicted; Column = Actual.*

The classifier performed very well, as shown in the results of the analysis displayed in Fig 2. For the sport class, the precision was 1.00, recall was 0.993865 and F-score was 0.996923. It performed the best out of all classes. For the business class, the precision was 0.987342, recall was 0.993631 and

F-score was 0.990476. The business class performed the worst out of all classes. For the politics class, the precision was 0.991935, recall was 0.991935 and F-score was 0.991935. For the entertainment class, the precision was 0.991453, recall was 0.991453, F-score was 0.991453. For the tech class, precision was 0.990654, recall was 0.990654 and F-score was 0.990654.

DISCUSSION

We found that the classifier is prone to errors when evaluating text which serves as a hybrid between two classes. For example, the classifier's only error in misclassifying the sport class came when discussing a financial takeover of an English soccer club with financial details. One could argue that this piece of text could technically fall under the business class as well. This is a common theme in the errors where the text is misclassified under the business class, the text discusses financial details which furthers our argument that we might consider this a hybrid piece of text. Therefore, the classifier's logic seems like a reasonable classification because one can see how a human would make the same inference. Similar cases of hybrid text are found to be the main cause in the other misclassifications of class.

We found that removing stop words marginally increased the aggregated micro-average precision by 0.003 and the macro-average by 0.004. In our case, removing the stop words made our classifier slightly more accurate

TEAM COLLABORATION

We spoke with team Daniella and Delaney about high accuracy results of the classifier, and particularly about the very high accuracy of the sports class. We also spoke with team Helen about how NB works.

REFERENCES

[1] spaCy documentation

[2] NLTK documentation

[3] Jurafsky and Martin. 2019. Speech and Language Processing (3rd Edition).