IST736 Text Mining

Professor Bei Yu

Spring 2021

HW\_4 Compare MNB and SVM for Causal Language Detection

Yu et al. (2019) annotated some conclusion sentences in biomedical and health research papers to identify their claim strength. Each sentence was annotated as belonging to one of the four categories:

Label = 0 : No relationship (1356 cases)

Label = 1 : Direct causal (494 cases)

Label = 2 : Conditional causal (213 cases)

Label = 3 : Correlational (998 cases)

You are going to use both MNB and SVMs algorithm to build prediction models to predict the claim strength.

For each algorithm, try different vectorization options to build the model, and evaluate them using 5-fold cross validation method.

Deliverable –

1. Your Python script.
2. A word document (Up to two pages) to describe
   1. Your best choice of vectorization options for each algorithm and why
   2. Top 10 word features for each category in the SVM model
   3. Each model’s performance in confusion matrix, precision, recall, and F-measure
   4. Each model’s error analysis to identify areas for improvement
   5. A performance comparison of the MNB model and the SVM model.

Grading criteria:

* Accuracy and reproducibility in analysis
* Clarity in explaining the analytical process and result

Bonus-point question (3 points):

Use BERT to build a prediction model. Compare this model’s performance with the naïve Bayes and SVM models. Do error analysis to identify areas for improvement.

Reference:

Yu, B., Li, Y., & Wang, J. (2019, November). Detecting Causal Language Use in Science Findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4656-4666).