**Homework 4: Causal Language Detection**

IST736: Text Mining

Oliver Posewitz

Overview:

There are clear disconnects of information in the context of research papers and people’s interpretation, especially when consumed by a general populace that have a less specific background. Scientific, statistical, or field specific research papers could include language that describe relationships in an accurate and concise manner, however, these relationship characteristics fail to convey to a wider audience.

This analysis attempts to identify key words, or word pairs, from research paper conclusion statements that describe either no relationship, direct causal, conditional causal, or correlational relationships using statistical methods.

Initial Analysis:

The data set is comprised of about 3k documents that are labeled in one of the four categories previously mentioned. The categories are skewed, as depicted below:

|  |  |
| --- | --- |
| No Relation | 44% |
| Direct Causal | 33% |
| Conditional Causal | 16% |
| Correlational | 07% |

The sentence length variations are relatively high, with the mean length of all sentences 131 characters with a standard deviation of 53 characters.

Methods:

Since the sentence lengths have relatively high variance, the first vectorization method that comes to mind is the term-frequency inverse document frequency, as it has potential to reduce noise from longer sentences although count vectorization is still being included.

Prior to vectorization, a hold-out test dataset was created from 20% of the data, maintaining equal label proportions between the test and training partitions. Cross validation with a k-fold of 5 will be used when training the models as there are limited samples and further partitioning may result in mischaracterization of patterns.

The estimation models are Multinomial Naïve Bayes (MNB) and C-Support Vector Classification (SVC). Both models reduce computational cost when working with a large set of features and offer interpretable coefficients preceding their predictions.

Four pipelines were made for each vectorization method and classifying model combination using the sklearn pipeline tool. Each pipeline was trained multiple times with different parameters using the gridsearch tool to find the best performing model. Some of the hyperparameters for count vectorization included ranges of n-gram range, min and max DF, and whether to include stop words. All the same hyperparameter ranges that were included in count vectorization are also included in TFIDF, with an additional normalization parameter where either the sum of squares of vector elements is one, or the sum of absolute values of vector elements is one. The MNB model hyperparameters include a range of additive smoothing, and the SVC model hyperparameter included a regularization parameter range.

Each of the pipelines were iterated through all combinations of hyperparameter values, and the average cross validation score was used to determine the best combination of vectorization method, model, and hyperparameter values. If the best performing pipeline contained a hyperparameter value at the end of the range, the range was recentered to ensure optimization. Surprisingly, MNB performed best with count vectorization that included English stop words with an average accuracy of 70%, however, the best performing model was TFIDF vectors with a SVC estimator. The most important words for each category can be found in the output of cell 18 in the hw4.ipynb file. The SVM model performed better overall, having greater f1-measures on the test data. This can be intuitively seen in the confusion matrix by observing that the numbers going diagonal from top left to bottom right are higher for SVM as they are for the MNB model.

As depicted in the confusion matrices attached, both models struggled to correctly label the sentences that indicate no relationship and would predict these no relation sentences as the other extreme, correlational relationship. Zooming in on some of these cases, the sentences are hard to distinguish intuitively as having no relationship or correlational. This result could be an indication of poor execution of analysis or could be indicative of a more overarching issue of the academic community not making their papers accessible. Although in this case there is no evidence supporting a relationship between statistical analysis and conclusion sentence effectiveness. I’m not saying these papers are not effective because the models employed underperformed. Although, the difficulties of gleaning characteristic relationships in conclusion sentences, either intuitively or statistically, could be suggestive of a need for the academic community to make their research more accessible to the general population and less susceptible to manipulation of the main ideas.