**IST – 736 Text Mining**

Comparison of MNB and SVM for Causal Language Detection

Performed analysis on the Biomedical and health Research paper sentences using MNB and LinearSVC which are annotated as “No Relationship”, “Direct Causal”, “Conditional causal” and “Correlational”.

From the given CSV file “Pubmed casual language use”. I have considered the independent variable as “Sentence” and dependent variable as “Label”. For hold-out test, I have split the entire data into 70 percent train data for fitting the models and 30 percent train test for testing the predictions and comparing the model performances.

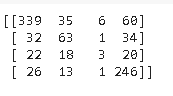
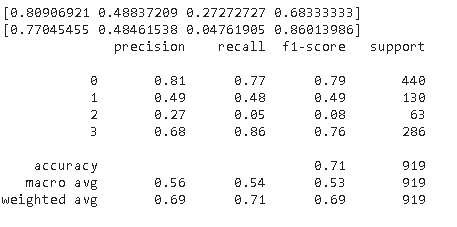
**Multinomial Naïve Bayes:**

I have used both CountVectorizer and TFIDF Vectorizer by considering the minimum document frequency as “2” and removing the stop words .For Multinomial Naïve Bayes, it is observed that the on Hold out test, the accuracy is more for the CountVectorizer. However, this can vary for different split. Hence, I have considered the 5 fold cross validation to compare the vectorization performances. It is clear, from the results, that the average score of all the categories is more for CountVectorizer **(68.76)** compared to TFIDF Vectorizer **(65.24).** Also, it is observed that the label “Conditional casual” is not predicted when using TFIDF Vectorizer.

***Top 10 Features for four categories using Count Vectorizer:***

**No relationship** (-5.33, 'high'), (-5.333581085290859, 'results'), (-5.3158815, 'trial'), (-5.29, 'treatment'), (-5.15, 'needed'), (-5.0, 'clinical'), (-4.912, 'risk'), (-4.78, 'studies'), (-4.77, 'study'), (-4.21, 'patients')]**Direct Casual** [(-5.61, 'significantly'), (-5.58, 'effective'), (-5.50, 'did'), (-5.50, 'improved'), (-5.47, 'effect'), (-5.43, 'treatment'), (-5.28, 'study'), (-5.14, 'weight'), (-4.98, 'risk'), (-4.53, 'patients')]**Conditional Casual** [(-5.98, 'therapy'), (-5.89, 'disease'), (-5.89, 'results'), (-5.8, 'reduce'), (-5.812, 'role'), (-5.745, 'cancer'), (-5.676, 'increase'), (-5.61, 'improve'), (-5.01, 'patients'), (-4.85, 'risk')]**Correlational** [(-5.2, 'cancer'), (-5.26, 'diabetes'), (-5.209, 'women'), (-5.095, 'higher'), (-5.079, 'study'), (-5.00, 'levels'), (-4.9791870367252, 'increased'), (-4.2860, 'patients'), (-4.038, 'risk'), (-3.91, 'associated')]

**Model Performance: (Confusion Matrix, Precision, Recall, FScore)**

**Error Analysis:**On comparing the “No relation” label predicted as “Correlated”, it is observed that there are 50 errors. For eg: “*However, most had playground equipment, courts, and outdoor play areas*” predicted as correlated, this can be due the words play and playground in the sentence

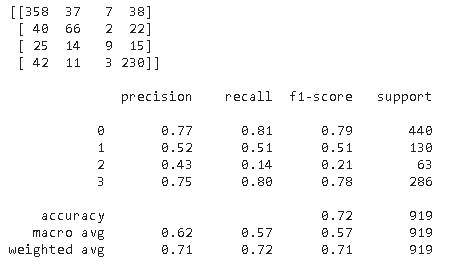
**Linear SVC:**

For Linear SVC, on comparing the accuracy for both hold-out test/train split (70 percent train and 30 percent test) and 5 Fold cross validation, it is observed that the accuracy is more for TFIDF Vectorizer. . With 5 fold cross validation, the accuracy for CountVectorizer is 68.9 percent and TFIDF Vectorizer is 72 percent. Hence, I choose TFIDF, for selecting the top features in each category and the for error analysis .

***Top 10 Features for four categories using TFIDF Vectorizer:***

**No relationship**[(1.374, 'performed'), (1.403, 'obtained'), (1.436, 'need'), (1.458, 'safety'), (1.58, 'implications'), (1.69, 'research'), (1.70, 'appropriate'), (1.736, 'required'), (2.1021671184605255, 'studies'), (2.475, 'needed')]**Direct\_casual**[(1.49, 'achieved'), (1.45, 'does'), (1.60, 'oral'), (1.723, 'benefits'), (1.78, 'improves'), (1.805, 'effect'), (1.883, 'did'), (2.19978, 'effective'), (2.4053, 'improved'), (2.798, 'resulted')]**Conditional Causal**[(1.1655, 'useful'), (1.189, 'increase'), (1.2015, 'role'), (1.2, 'contribute'), (1.201, 'protective'), (1.27, 'decrease'), (1.30, 'reduce'), (1.529, 'play'), (1.58, 'appear'), (1.60, 'improve')]**Correlational** [(1.6, 'higher'), (1.801, 'correlation'), (1.8595, 'received'), (1.94, 'correlated'), (2.06, 'related'), (2.126, 'predictor'), (2.22, 'predictors'), (2.494, 'predict'), (2.628, 'association'), (5.1637, 'associated')]

**Model Performance: (Confusion Matrix, Precision, Recall, FScore)**



**Error Analysis:**

On comparing the “No relation” label predicted as “Correlation”, it is observed that there are 38 errors. For eg: “*It appears important for pharmacists to show their daily involvement in the quality of medical care*” predicted as Direct relation, this can be due the words important and Daily involvement in the sentence

**Performance comparison of MNB and Linear SVC:**

On comparing both the models, it is observed that the errors are more for MNB compared to Linear SVC. i.e., the labels are not properly predicted. Further analysis can be performed using trigrams and bigrams to check the predictions for both the models.

**Bonus-point question (3 points): Use BERT to build a prediction model.**

Using BERT Classifier with 5 fold cross validation, the accuracy is more which is around 80 percent with less number of errors compared to bot MNB and LinearSVC Models and is around 88 percent for holdout test

**Model Performance using BERT**:

For 5 fold cross validation:



For Holdout test:

