Shivangi Mundhra
SUID - 842548148

HW 3 – Compare SVM and BERT for Causal Language Detection

Shivangi Mundhra | SUID – 842548148

Best choice of vectorization options for each algorithm and why

I created the below two models –

1. SVM model with TFIDF vectorizer
2. BERT model

The rationale behind these choices is explained below.

SVM model –

I used TFIDF Vectorizer with SVM model because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words. This should increase the accuracy of classification and should help in creating better hyperplanes between different labels for our SVM model.

BERT model –

BERT model includes vectorization using word embedding so we used this type of vectorization and did not have to use separate vectorizer as in case of SVM model. Word embeddings are vector representations of words, and they can capture the context of a word in a document, semantic and syntactic similarity, relation with other words, etc. This makes word embeddings a great vectorizer as it will help in creating better classification/prediction model.

Top 10 word features for each category in the SVM model

The below table shows the top 10 words for each label in the SVM model.

	Rank	1	2	3	4	5	6	7	8	9	10
Labels	No Relationship	needed	studies	research	implications	required	assess	trials	performed	assessment	focus
	Direct Causal	resulted	effective	improved	did	improves	beneficial	reduces	oral	benefits	reduced
	Conditional Causal	improve	reduce	play	appear	result	appeared	responsible	decrease	role	increase
	Correlational	associated	predict	predictor	association	correlated	related	predictors	received	correlation	increased

The top 10 words in each label make sense. For instance, the top 10 words in the label = "Correlational" include words like associated, predict, predicted, association and correlated. These words are synonymous with the label (Correlational) itself! It also makes sense that there is some overlap in the top words for "Direct Causal" and "Conditional Causal" like, improve, reduce, and result. Finally, it is satisfying to see that there is no overlap between "No Relationship" words and words of other labels.

Each model's performance in confusion matrix, precision, recall, and F-measure

SVM model –

The model's overall accuracy is 74% which shows that it predicted 26% features incorrectly. The macro avg of precision, recall and F1-score are 0.64, 0.61 and 0.62 – not great. The precision for Direct Causal and Conditional Causal is almost 50% which shows that for those labels, about 50% of predictions are incorrect. The recall metrics shows that about 73% of Conditional Causal and 45% of Direct Causal features were not predicted as such. Overall, the model's accuracy, precision

Shivangi Mundhra
SUID - 842548148

and recall metrics are not encouraging and show that this model is not perfect. The confusion matrix for test vs prediction is shown below along with each label's precision and recall.

		Prediction Labels				
	`	No Relationship	Direct Causal	Conditional Causal	Correlational	Recall
Test Labels	No Relationship	**249**	21	5	26	0.83
	Direct Causal	21	**47**	4	13	0.55
	Conditional Causal	11	13	**12**	8	0.27
	Correlational	26	11	3	**143**	0.78
	Precision	0.81	0.51	0.50	0.75	

BERT model –

The model's overall macro avg accuracy, F1-score, precision and recall are 97.80%, 0.96, 0.97 and 0.96 respectively, which shows that this is a very good model. The precision and recall for all labels is > 0.90 which also indicates that this is a great model.

		Prediction Labels				
	`	No Relationship	Direct Causal	Conditional Causal	Correlational	Recall
Test Labels	No Relationship	**431**	2	0	4	0.99
	Direct Causal	5	**156**	4	3	0.93
	Conditional Causal	2	1	**64**	1	0.94
	Correlational	7	1	0	**319**	0.98
	Precision	0.97	0.98	0.94	0.98	

Each model's error analysis to identify areas for improvement

SVM model –

For error analysis, we print out different records that were predicted erroneously. From the confusion matrix, there are plenty of errors throughout all labels. From the errors, it is evident that some features are represented in more than one labels, and it would be better if vectorization included the context of words. This would classify features better and perhaps avoid a lot of these erroneous predictions.

BERT model –

2 of the 4 no relationship errors that were predicted as correlational had features that define correlational (like reduced). The word embeddings vectorizer did a really good job of adding context to features which is why we see so less errors in all categories. A way to improve to would be to force more context to word features from all these errors.

A performance comparison of the BERT model and the SVM model

It is clear from the confusion matrix, accuracy, precision, recall and f1-score metrics that the BERT model is far superior as compared to the SVM model. With most metrics > 95%, the BERT model accurately predicts for all 4 labels. The SVM model falls short in all metrics and is not a good model for this prediction. The main differentiating factor that does it for the BERT model is vectorization using word embedding. Since word embedding considers the context of features as well, it increases the model's performance. The TFIDF vectorizer that we used for SVM, although very good on its own, is not sufficient to increase SVM model's performance. SVM model needs to incorporate the contextual vectorization to increase its performance.