Name: ADITI PALA

IST736 Text Mining HW3

Health Claim Analysis

INTRODUCTION:

This report aims to leverage NLP & Text mining techniques to detect exaggerated causal claims within health-related content, to various datasets and tasks, specifically cross-domain classification, zero-shot classification, and clustering, with the goal of achieving several key outcomes.

The datasets annotated_pubmed.csv, annotated_eureka.csv, and Eureka_headlines_10000 contain 3061 rows, 2076 rows and 10000 rows respectively

TASK 1: CROSS-DOMAIN CLASSIFICATION

1. SVM:

I applied the TF-IDF vectorization technique to convert the text data into numerical features. I initialized a Support Vector Machine (SVM) classifier with a linear kernel and created a processing pipeline that includes TF-IDF vectorization and the SVM classifier. Then, I performed cross-validation on the PubMed dataset to evaluate the SVM model's performance.

Model	Metric	Test Accuracy
SVM	Cross-Validation	[0.7032, 0.6951, 0.6941]

The accuracy values range from approximately 0.6941 to 0.7032. The model exhibits consistent accuracy across folds, with relatively small variations.

SVM Model trained on “annotated_pubmed.csv’ and tested on “annotated_eureka.csv”:

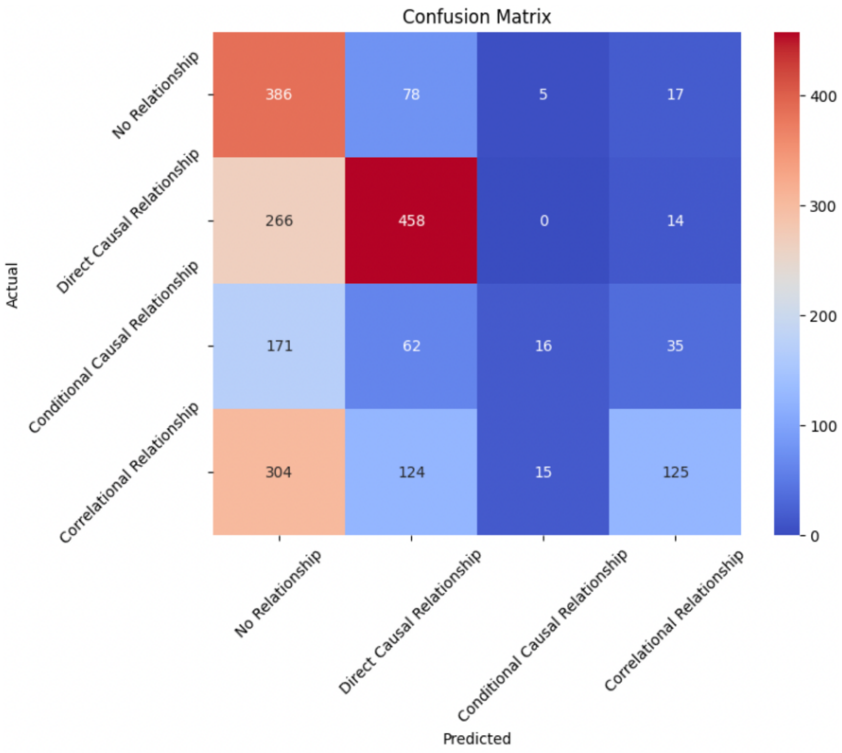
Class	Precision	Recall	F1-Score	Support
No Relationship (Class 0)	0.34	0.79	0.48	486
Direct Causal (Class 1)	0.63	0.62	0.63	738
Conditional Causal (Class 2)	0.44	0.06	0.10	284
Correlational (Class 3)	0.65	0.22	0.33	568
Accuracy			***0.47***	2076
Macro Avg	0.52	0.42	0.38	2076
Weighted Avg	0.55	0.47	0.44	2076

Error Analysis:

During misclassification analysis, I created "actual_labels" and "predicted_labels". A eeries to store actual and predicted classifications. Then, I combined them with sentences to create "misclassified_samples." I added a "Correct Prediction" column with "YES" for matching labels and "NO" for mismatches.

The count revealed 985 correct predictions and 1,091 misclassifications, offering insights into model performance. Also I did a confusion matrix as follows:

Name: ADITI PALA



Model is particularly good at predicting direct causal relationships (171 out of 304) and the absence of causal relationships (35 out of 400).

It has trouble with correlational (78 out of 124) and conditional (125 out of 386) causal relationships, though.

While the model works well overall, it could do a better job of predicting complex relationships.

Top 20 words for each category:

- Top 20 words with No Relationship:
  needed, studies, research, necessary, assessment, treatment, follow, safety, monitoring, performed, require, evaluation, understanding, impact, prevention, adequate, implications, national, focus, suboptimal

- Top 20 words with Direct Causal Relationship:
  needed, studies, research, syndrome, understanding, need, order, clinically, require, warrant, children, identify, remain, nurses, high, consider, medical, assess, findings, proposed

- Top 20 words with Conditional Causal Relationship:
  studies, needed, interventions, needs, research, assessment, need, achieve, warrant, recommended, future, strategies, appropriate, group, exist, required, potential, considered, proposed, postpartum

- Top 20 words with Correlational Relationship:
  associated, association, factor, predict, relationship, marker, predictor, related, older, associations, significantly, women, month, predictive, increased, years, activity, higher, preoperative, study

2. BERT

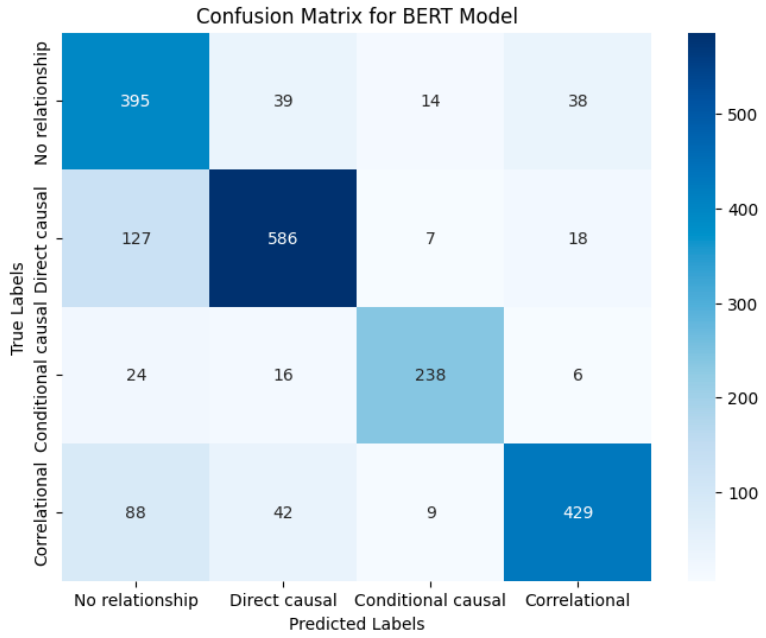
I initialized a BERT model by creating an instance of the BERT Classifier, specifying 2 training epochs. Then, I trained the BERT model using the PubMed dataset, with sentence data as features and labels for classification. I performed cross-validation to assess the model's performance and displayed the cross-validation scores, along with the mean F1 score, for evaluation.

Model	Cross-Validation Scores	Mean F1 Score
BERT	[0.8562, 0.8649, 0.8767]	0.8659

Name: ADITI PALA

BERT Model trained on “annotated_pubmed.csv’ and tested on “annotated_eureka.csv”:

Causal Relationship	Precision	Recall	F1-Score	Support
No Relationship	0.62	0.81	0.71	486
Direct Causal	0.86	0.79	0.82	738
Conditional Causal	0.89	0.84	0.86	284
Correlational Causal	0.87	0.76	0.81	568
Accuracy			**0.79**	2076
Macro Avg	0.81	0.80	0.80	2076
Weighted Avg	0.81	0.79	0.80	2076



Error Analysis:

The BERT model achieves an overall accuracy of 79% on the Eureka dataset with 428 misclassified samples. It excels in predicting direct causal relationships, which are simpler, but performs less effectively on conditional and correlational relationships due to their complexity. The class distribution in the dataset may contribute to this performance variation.

It is worth noting that the Eureka dataset exhibits class imbalance, with a greater number of examples of direct causal relationships compared to conditional and correlational causal relationships. This class imbalance may lead to a bias in the BERT model's predictions.

Comparative Analysis – SVM vs BERT:

Several important findings come from the comparison of SVM and BERT for cross-domain classification. First off, BERT performs better across a range of domains, demonstrating its cross-domain versatility by applying to the Eureka dataset and maintaining a strong accuracy of 79% while SVM shows a notable decline in accuracy. BERT excels at distinguishing between direct and non-direct relationships, highlighting its ability to capture nuanced text patterns.

Additionally, both models' performance is significantly impacted by the class imbalance in the Eureka dataset.

Predictions from BERT may be biased in favor of the majority class, highlighting the necessity of addressing class distribution issues in practical applications. Overall, BERT outperforms traditional SVM in these

crucial areas, making it an invaluable tool for a variety of NLP and text mining tasks. Its versatility, skill in managing complex relationships, and sensitivity to class distribution position it as a potent choice for cross-domain classification.

TASK 2: ZERO-SHOT CLASSIFICATION

Using two labelled datasets, PubMed and Eureka, I predicted causal claim strength using the Huggingface zero-shot-classification pipeline. I used candidate labels such as "No relationship," "Direct causal," "Conditional causal," and "Correlational" to train the zero-shot classifier. I created classification reports and classification report for every category after completing zero-shot classification on the two datasets.

Classification Report for PubMed:

Metric	No Relationship	Direct Causal	Conditional Causal	Correlational	Macro Average	Weighted Average
Precision	0.1637	0.3750	0.0573	0.1564	0.1908	0.2197
Recall	0.0206	0.0150	0.0751	0.8138	0.2292	0.1531
F1 Score	0.0367	0.0289	0.0650	0.2623	0.0982	0.0752
Accuracy	**0.1531**	-	-	-	-	-
Support	1356	998	213	494	3061	3061

Classification Report for Eureka:

Metric	No Relationship	Direct Causal	Conditional Causal	Correlational	Macro Average	Weighted Average
Precision	0.1455	0.3590	0.1414	0.2635	0.2298	0.2506
Recall	0.0165	0.0190	0.0951	0.8310	0.2404	0.2481
F1 Score	0.0296	0.0360	0.1137	0.4002	0.1449	0.1448
Accuracy	**0.2531**	-	-	-	-	-
Support	486	738	284	568	2076	2076

It excels in predicting direct causal relationships but struggles with the more complex conditional and correlational causal relationships. This suggests that the model performs better when dealing with straightforward relationship patterns, highlighting the nuances in classifying diverse textual data.

Does the zero-shot classifier perform equally well in these two domains?

The use of the Huggingface zero-shot classification pipeline enabled us to classify causal claim strength in both the "PubMed" and "Eureka" datasets. The metrics demonstrate that while the model's performance is modest, it shows slight improvement in the "Eureka" dataset. Zero-shot classifier performed poorly in both datasets, with low precision, recall, and F1-scores, especially for Conditional causal and Correlational categories. The model's inability to generalize to these complex causal relationships suggests a need for further fine-tuning and domain-specific training.

TASK 3: CLUSTERING

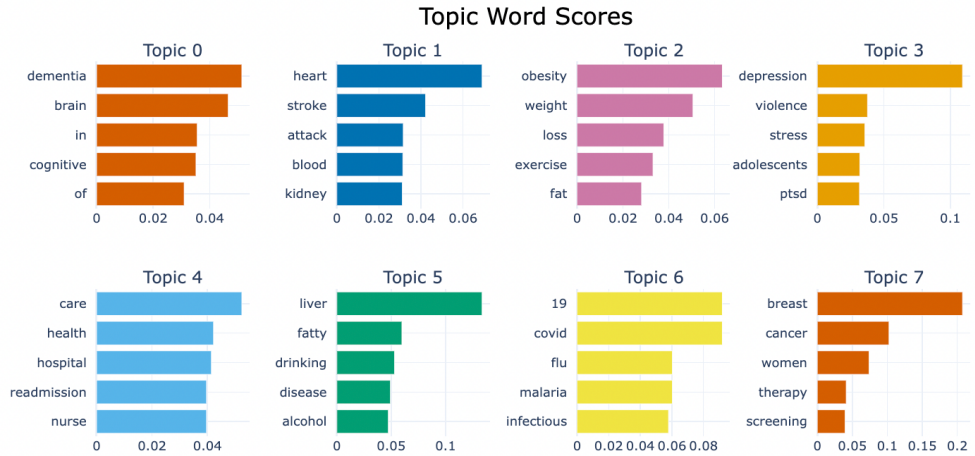
In my analysis, I used SBERT embeddings to represent headlines and applied KMeans clustering to create 10 clusters. I examined cluster sizes, found documents close to each cluster's centroid, and used the Elbow method to determine the optimal cluster count. I noticed variations in performance, with some documents being similar to centroids of other clusters.

```
======cluster # 0 , cluster size: 529
```

Name: ADITI PALA

```
[ this doc is in a different cluster # 4 >> Who are you kidding?
[ this doc is in a different cluster # 4 >> When you always gotta go...
[ this doc is in a different cluster # 4 >> Lollipop or edible?
[ this doc is in a different cluster # 4 >> Also in the May 27 JNCI
Heart failure after first heart attack may increase cancer risk
```

For BERTopic, I employed the BERTopic package for topic modeling. I generated topics, calculated their probabilities, and explored frequently occurring topics and their representative documents. Visualizations were created to depict topic distribution, hierarchy, and for clarity.



Do these two models provide same insights? If not, how do they differ in the health topics that they discovered?

Cluster	SBERT + KMeans	BERTopic
0	Heart health, cardiac risk, and prevention.	Cognitive health in Alzheimer's disease.
1	Child development and risks of premature birth.	Heart health, cholesterol, and kidney failure.
2	Hospital conditions, healthcare accessibility, and...	Obesity, childhood weight gain, and stress.
3	Brain, cognitive health, and dementia prevention.	Teenage depression and ketamine treatment.
4	Healthy lifestyle, obesity, diet, and health effects.	Primary care support, hospital readmissions, and...
5	Different cancer types, cancer research, survival,...	Liver disease, alcohol consumption, and meat...
6	Smoking cigarettes and its relation to other health...	COVID-19 vaccine research and pandemic impact...

By highlighting particular research papers and their significance to the cluster topic, BERTopic seems to shed light on the intricacies of each theme. On the other hand, SBERT+KMeans provides more comprehensive themes that cover different facets of every subject, including risk factors, research, and prevention. These variations in concentration and level of detail draw attention to the adaptability and subtleties of the clustering methods, providing insightful information about the many facets of public health and medical research.