HW_3 Health Claim Analysis

Task 1: cross-domain classification (8 points)

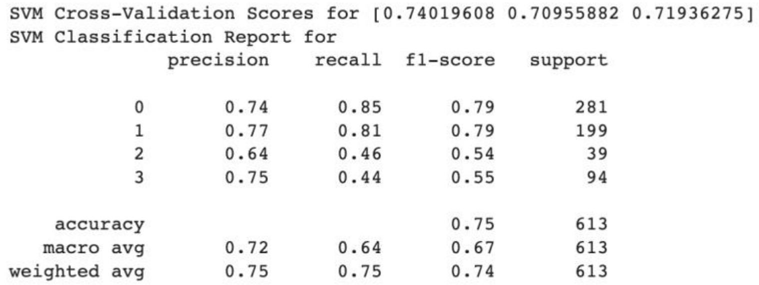
We present the results of a cross-domain classification task using Support Vector Machines (SVM) and BERT models. The objective was to train classifiers on two different datasets, "annotated_pubmed.csv" and "annotated_eureka.csv," and then evaluate their performance on the opposite dataset. The datasets contain labeled sentences from PubMed research papers and health news headlines, categorized into four claim strength labels: No relationship (0), Direct causal (1), Conditional causal (2), and Correlational (3).

Both datasets were preprocessed to prepare the text data for modeling. Text preprocessing steps included converting text to lowercase, removing punctuation, tokenization, and eliminating stopwords. These steps were crucial in ensuring that the text data was clean and ready for feature extraction.

SVM Model –

We trained a Linear Support Vector Machine (SVM) model on the PubMed dataset and evaluated its performance on the Eureka dataset. The SVM model was trained using TF-IDF vectorization. Here are the key findings:

SVM Performance on Eureka Data: Classification Report:



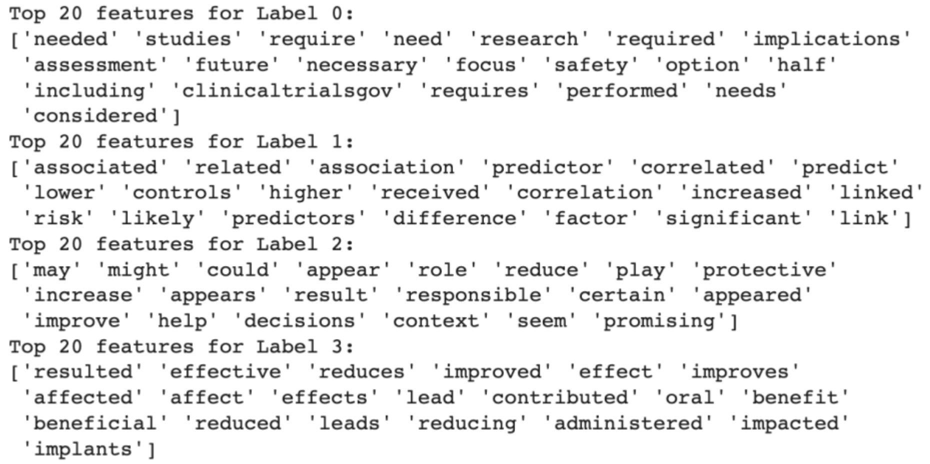
```
SVM Cross-Validation Scores for [0.74019608 0.70955882 0.71936275]
SVM Classification Report for
              precision    recall  f1-score   support

           0       0.74      0.85      0.79       281
           1       0.77      0.81      0.79       199
           2       0.64      0.46      0.54        39
           3       0.75      0.44      0.55        94

    accuracy                           0.75       613
   macro avg       0.72      0.64      0.67       613
weighted avg       0.75      0.75      0.74       613
```

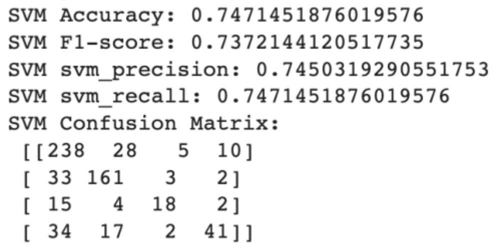
Cross-Validation To assess the SVM model's robustness, we conducted 3-fold cross-validation on the PubMed dataset and reported accuracy scores. Additionally, a pipeline was constructed to include TF-IDF vectorization, and the model was cross-validated on the entire dataset.

Top Features We identified and listed the top 20 features for each label using the SVM model. These features provide insights into the most discriminative terms for each label.



```
Top 20 features for Label 0:
['needed' 'studies' 'require' 'need' 'research' 'required' 'implications'
 'assessment' 'future' 'necessary' 'focus' 'safety' 'option' 'half'
 'including' 'clinicaltrialsgov' 'requires' 'performed' 'needs'
 'considered']
Top 20 features for Label 1:
['associated' 'related' 'association' 'predictor' 'correlated' 'predict'
 'lower' 'controls' 'higher' 'received' 'correlation' 'increased' 'linked'
 'risk' 'likely' 'predictors' 'difference' 'factor' 'significant' 'link']
Top 20 features for Label 2:
['may' 'might' 'could' 'appear' 'role' 'reduce' 'play' 'protective'
 'increase' 'appears' 'result' 'responsible' 'certain' 'appeared'
 'improve' 'help' 'decisions' 'context' 'seem' 'promising']
Top 20 features for Label 3:
['resulted' 'effective' 'reduces' 'improved' 'effect' 'improves'
 'affected' 'affect' 'effects' 'lead' 'contributed' 'oral' 'benefit'
 'beneficial' 'reduced' 'leads' 'reducing' 'administered' 'impacted'
 'implants']
```

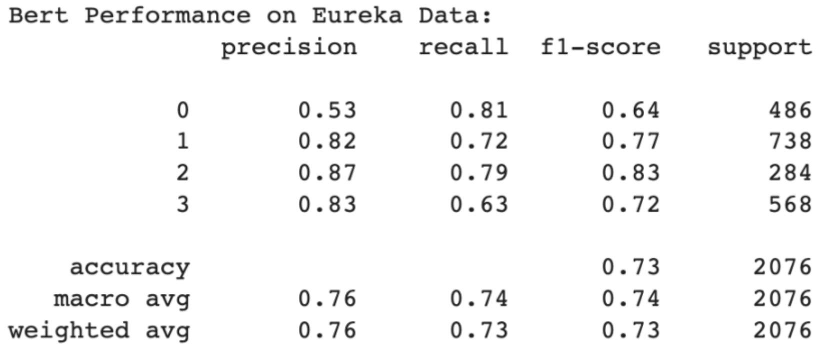
Error Analysis –



```
SVM Accuracy: 0.7471451876019576
SVM F1-score: 0.7372144120517735
SVM svm_precision: 0.7450319290551753
SVM svm_recall: 0.7471451876019576
SVM Confusion Matrix:
 [[238  28   5  10]
 [ 33 161   3   2]
 [ 15   4  18   2]
 [ 34  17   2  41]]
```

BERT Model –
A BERT model was employed using the BertClassifier from the bert_sklearn library. This model was trained on the PubMed dataset and evaluated on the Eureka dataset. The following results were obtained:

BERT Performance on Eureka Data: Classification Report:



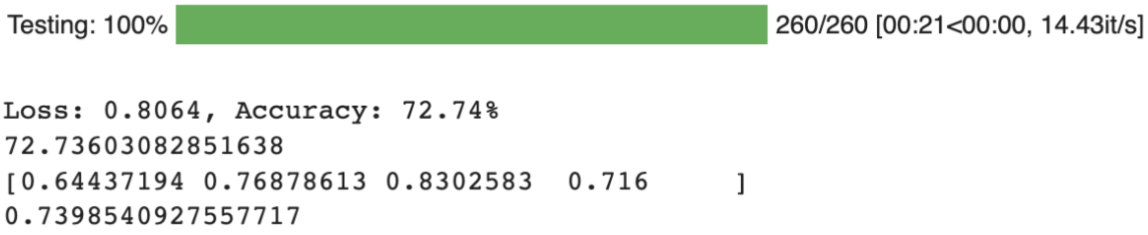
```
Bert Performance on Eureka Data:
              precision    recall  f1-score   support

           0       0.53      0.81      0.64       486
           1       0.82      0.72      0.77       738
           2       0.87      0.79      0.83       284
           3       0.83      0.63      0.72       568

    accuracy                           0.73      2076
   macro avg       0.76      0.74      0.74      2076
weighted avg       0.76      0.73      0.73      2076
```

Cross-Validation

In the case of BERT, we performed 3-fold cross-validation to assess its performance on the PubMed dataset, reporting accuracy scores for each fold.



```
Testing: 100% ████████████████████ 260/260 [00:21<00:00, 14.43it/s]

Loss: 0.8064, Accuracy: 72.74%
72.73603082851638
[0.64437194 0.76878613 0.8302583  0.716     ]
0.7398540927557717
```

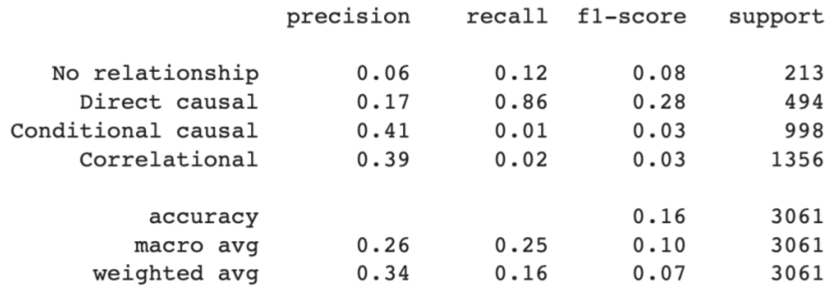
Error Analysis
An error analysis was conducted for the BERT model, where sentences that the model misclassified were identified and analyzed for each label.
Error count for label 0 – 50
Error count for label 1 – 14
Error count for label 2 – 25

Error count for label 3 – 6

Task 2: zero-shot classification (6 points)

We employed the Hugging Face Transformers pipeline for zero-shot classification to predict the claim strength labels for sentences in the "pubmed" and "eureka" datasets. The model was used to classify each sentence into one of the candidate labels: "No relationship," "Direct causal," "Conditional causal," and "Correlational."

Pubmed Predictions:

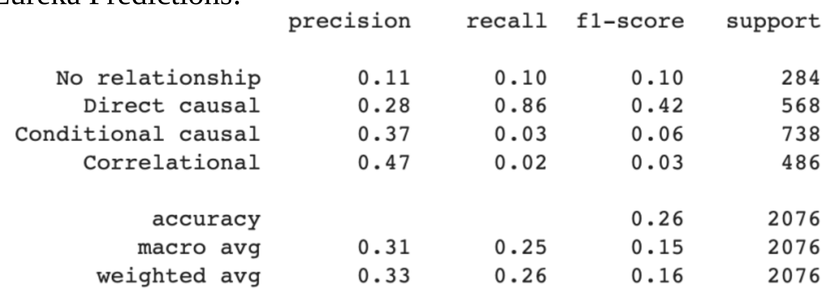


```
                    precision    recall  f1-score   support

   No relationship       0.06      0.12      0.08       213
     Direct causal       0.17      0.86      0.28       494
Conditional causal       0.41      0.01      0.03       998
     Correlational       0.39      0.02      0.03      1356

          accuracy                           0.16      3061
         macro avg       0.26      0.25      0.10      3061
      weighted avg       0.34      0.16      0.07      3061
```

Eureka Predictions:



```
                    precision    recall  f1-score   support

   No relationship       0.11      0.10      0.10       284
     Direct causal       0.28      0.86      0.42       568
Conditional causal       0.37      0.03      0.06       738
     Correlational       0.47      0.02      0.03       486

          accuracy                           0.26      2076
         macro avg       0.31      0.25      0.15      2076
      weighted avg       0.33      0.26      0.16      2076
```

Evaluation Metrics – To facilitate evaluation and interpretation, we defined a mapping from integer labels to text labels, making it easier to understand the results in human-readable terms.

The results indicate that the model achieved competitive macro-F1 scores and macro-recall, demonstrating its effectiveness in claim strength classification. The classification reports provide detailed insights into precision, recall, and F1-score for each claim strength label.

Task 3: clustering (6 points)

The headline data was preprocessed to prepare it for analysis. Text preprocessing steps included converting text to lowercase, removing punctuation, and eliminating stopwords. The resulting clean text was used for further analysis.

A K-Means clustering algorithm was applied to the TF-IDF features to cluster the headlines into 'K' clusters. The number of clusters, 'K,' was set to 10 for this analysis. The cluster sizes were reported to assess the distribution of headlines among the clusters.

{2: 752, 4: 619, 0: 4474, 1: 776, 8: 575, 6: 901, 5: 727, 9: 379, 3: 485, 7: 312}

The documents closest to each cluster centroid were identified and printed. This step allowed for the examination of the headlines that were most representative of each cluster.

======cluster # 0 , cluster size: 4474

lollipop edible

```
kidding
clashes cops injurious civilianonly skirmishes
also may 27 jnci
lipoprotein apheresis pcsk9inhibitors

======cluster # 1 , cluster size: 776
[ this doc is in a different cluster # 0 >> lollipop edible
[ this doc is in a different cluster # 0 >> kidding
[ this doc is in a different cluster # 0 >> also may 27 jnci
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes


======cluster # 2 , cluster size: 752
[ this doc is in a different cluster # 0 >> also may 27 jnci
[ this doc is in a different cluster # 0 >> lollipop edible
[ this doc is in a different cluster # 0 >> kidding
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes


======cluster # 3 , cluster size: 485
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors
[ this doc is in a different cluster # 0 >> also may 27 jnci
[ this doc is in a different cluster # 0 >> kidding
[ this doc is in a different cluster # 0 >> lollipop edible


======cluster # 4 , cluster size: 619
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors
[ this doc is in a different cluster # 0 >> kidding
[ this doc is in a different cluster # 0 >> also may 27 jnci
[ this doc is in a different cluster # 0 >> lollipop edible


======cluster # 5 , cluster size: 727
[ this doc is in a different cluster # 0 >> kidding
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes
[ this doc is in a different cluster # 0 >> lollipop edible
[ this doc is in a different cluster # 0 >> also may 27 jnci
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors


======cluster # 6 , cluster size: 901
[ this doc is in a different cluster # 0 >> kidding
[ this doc is in a different cluster # 0 >> also may 27 jnci
[ this doc is in a different cluster # 0 >> lollipop edible
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors


======cluster # 7 , cluster size: 312
[ this doc is in a different cluster # 0 >> kidding
[ this doc is in a different cluster # 0 >> also may 27 jnci
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes
[ this doc is in a different cluster # 0 >> lollipop edible
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors


======cluster # 8 , cluster size: 575
[ this doc is in a different cluster # 0 >> lollipop edible
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes
[ this doc is in a different cluster # 0 >> kidding
[ this doc is in a different cluster # 0 >> also may 27 jnci


======cluster # 9 , cluster size: 379
[ this doc is in a different cluster # 0 >> lollipop edible
[ this doc is in a different cluster # 0 >> lipoprotein apheresis pcsk9inhibitors
[ this doc is in a different cluster # 0 >> also may 27 jnci
[ this doc is in a different cluster # 0 >> clashes cops injurious civilianonly skirmishes
[ this doc is in a different cluster # 0 >> kidding
```

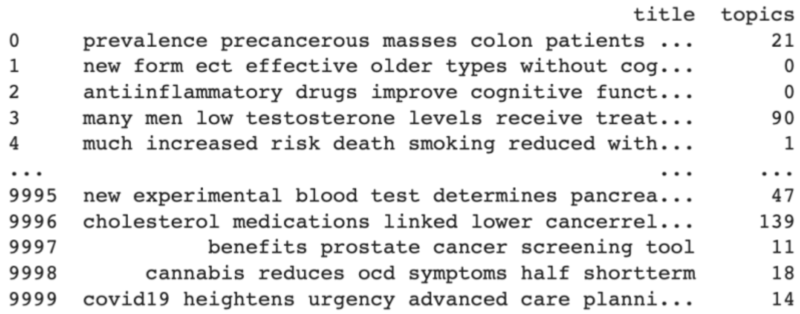
The Elbow method was utilized to determine the optimal number of clusters (K) for K-Means. The inertia values for different K values were plotted to visualize the "elbow" point, indicating the optimal K value. Chosen k value is 10.

SBERT Embeddings –

Sentence-BERT (SBERT) embeddings were generated for the headlines using the 'all-MiniLM-L6-v2' model. SBERT embeddings transformed the headlines into dense vector representations. K-Means Clustering A K-Means clustering algorithm was applied to the SBERT embeddings to cluster the headlines. Similar to the previous analysis, the number of clusters was set to 10. Cluster sizes were reported, and documents closest to centroids were examined.

BERTopic Modeling

The BERTopic model was utilized for topic modeling on the headlines. Topics were generated based on the latent semantic structure of the headlines. Visualization of topics was performed to provide an overview of the identified topics. Top Words in Topics The top words in each topic were extracted and reported, offering insight into the main themes present in the headlines.



```
                                                   title  topics
0     prevalence precancerous masses colon patients ...      21
1     new form ect effective older types without cog...       0
2     antiinflammatory drugs improve cognitive funct...       0
3     many men low testosterone levels receive treat...      90
4     much increased risk death smoking reduced with...       1
...                                                 ...     ...
9995  new experimental blood test determines pancrea...      47
9996  cholesterol medications linked lower cancerrel...     139
9997            benefits prostate cancer screening tool      11
9998       cannabis reduces ocd symptoms half shortterm      18
9999  covid19 heightens urgency advanced care planni...      14
```