Name: Shubh Mody

Health Claim Analysis

In this assignment, I undertook three essential tasks in health claim analysis using natural language processing techniques. Firstly, I conducted cross-domain classification, comparing the performance of Support Vector Machines (SVM) and BERT, a transformer-based model, in categorizing health claim strength in two distinct domains - PubMed research papers and health news headlines. Secondly, I employed the Hugging face zero-shot classification pipeline to predict causal claim strength in both domains, evaluating the classifier's consistency and effectiveness. Finally, I delved into clustering analysis, utilizing SBERT+kMeans and BERTopic to reveal common health topics in a collection of health news headlines, highlighting the differing insights provided by these clustering models.

**Task - 1: Cross Domain Classification:**

I used the Unigram TFIDF vectorizer with a minimum document frequency of 5 to apply to the SVM model

**Model comparison between SVM and BERT for in-domain classification:**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | F-1 Macro | 3-fold Cross-Validation Score |
| SVM | 74.1% | 0.63 | 0.68 |
| BERT | 87.28% | 0.83 | 0.86 |

**Top 20 features for all 4 Categories obtained from SVM model:**

**Label 0 – No relationship:** policy, need, psychosocial, developed, prevent, trials, necessary, obtained, require, achieve, providers, appropriate, focus, assess, performed, implications, required, research, studies, needed.

For Label 0 (No Relationship), the SVM selected words that appear to be contextually unrelated or have no strong causal or correlational association, but it also included terms like "needed" and "research," which might have a weak contextual relationship.

**Label 1 – Direct Causal:** indicator, controls, predicted, vary, strong, blood, difference, rygb, correlation, link, lower, increased, received, predictors, association, correlated, predictor, related, predict, associated.

Label 1 (Direct Causal) features indicate that the SVM successfully identified terms suggesting direct causality, such as "indicator" and "controls," but it mistakenly included words like "correlation" and "related," which are more associated with correlational relationships.

**Label 2 – Conditional Causal:** help, influence, cad, certain, context, useful, affect, cause, contribute, increase, protective, role, responsible, appeared, decrease, appear, result, reduce, play, improve.

The SVM captured words associated with conditional causality, like "help" and "contribute," but it also included "research," which can be more aligned with non-causal contexts.

**Label 3 – Correlational:** fasting, lowering, reducing, successful, effects, biological, terms, effect, does, effectiveness, reduced, benefits, reduces, oral, beneficial, improves, effective, did, improved, resulted.

The SVM's chosen features seem to be related to correlations, with words like "effects" and "biological" representing this category well; however, it mistakenly included words like "improve" and "improves," which could indicate a causal relationship.

**Model comparison between SVM and BERT for cross-domain classification:**

|  |  |  |
| --- | --- | --- |
| Trained On PubMed Research Data and Tested on Health News Headlines | | |
| Model | Accuracy | F-1 Macro |
| SVM | 49.2% | 0.42 |
| BERT | 81.3% | 0.82 |

**Error Analysis:**

**SVM**

Actual Class 3 (Correlation), Predicted as Class 1 (Direct Causal): 118

Actual Class 1 (Direct Causal): Predicted as Class 3 (Correlation): 32

**BERT**

Actual Class 3 (Correlation), Predicted as Class 1 (Direct Causal): 24

Actual Class 1 (Direct Causal): Predicted as Class 3 (Correlation): 28

|  |  |  |
| --- | --- | --- |
| Trained On Health News Headlines Data and Tested on PubMed Research Papers | | |
| Model | Accuracy | F-1 Macro |
| SVM | 55.4% | 0.48 |
| BERT | 79.2% | 0.77 |

**Error Analysis:**

**SVM**

Actual Class 3 (Correlation), Predicted as Class 1 (Direct Causal): 111

Actual Class 1 (Direct Causal): Predicted as Class 3 (Correlation): 149

**BERT**

Actual Class 3 (Correlation), Predicted as Class 1 (Direct Causal): 63

Actual Class 1 (Direct Causal): Predicted as Class 3 (Correlation): 164

**SVM errors:**

SVM tends to make more classification errors, particularly in confusing Class 1 (Direct Causal) and Class 3 (Correlation) categories in both domains.

In the PubMed Research to Health News Headlines setting, SVM misclassifies Class 3 as Class 1 more often.

In the Health News Headlines to PubMed Research Papers setting, SVM misclassifies Class 1 as Class 3 more frequently.

**BERT errors:**

While BERT outperforms SVM overall, it still exhibits some confusion between Class 1 and Class 3 categories.

In both domains, BERT also has some instances where Class 3 is mistakenly classified as Class 1 and vice versa, but the number of errors is generally lower than those made by SVM.

**Comparison Conclusion:**

🡪 BERT is the better-performing model in both cross-domain settings due to its higher accuracy and F-1 Macro scores, as well as the lower number of classification errors.

🡪 Both SVM and BERT make errors, but BERT makes fewer errors in both scenarios. The nature of errors made by both models is somewhat consistent, with both models often confusing Class 1 (Direct Causal) and Class 3 (Correlation) categories.

🡪 The consistent and notably better performance of BERT suggests that BERT can generalise better and it can handle the complexities and variations in text data more effectively than SVM.

**Task - 2: Zero Shot Classification:**

**Classification Report:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Category Class** | **Precision** | **Recall** | **F-1 Score** | **Macro F-1** | **Accuracy** |
| PubMed Research Papers | no relation | 0.19 | 0.00 | 0.01 | 0.10 | 14% |
| direct causal | 0.43 | 0.01 | 0.02 |
| conditional causal | 0.08 | 0.20 | 0.11 |
| correlational | 0.16 | 0.78 | 0.26 |
| Health News Headlines | no relation | 0.20 | 0.01 | 0.01 | 0.15 | 25% |
| direct causal | 0.40 | 0.01 | 0.02 |
| conditional causal | 0.15 | 0.19 | 0.17 |
| correlational | 0.26 | 0.79 | 0.40 |

**Performance Review:**

In the "PubMed Research Papers" dataset, the classifier struggles with low precision and recall, especially in the "no relation" category, resulting in an overall low F1 score and accuracy.

In contrast, in the "Health News Headlines" dataset, the classifier exhibits slightly better performance with higher precision, recall, F1 scores, and accuracy, particularly in the "correlational" category.

However, the classifier's performance in both domains is far from equal, as it faces challenges in both datasets. Therefore, it can be concluded that the zero-shot classifier does not perform equally well in these two domains as SVM and BERT from the previous task.

**Task - 3: Clustering:**

As part of the preprocessing, I removed punctuations from the "title" column. I then opted for the TFIDF vectorizer with a minimum document frequency requirement of 2 and the removal of stop words, resulting in a vocabulary size of 5182. I made this choice due to the substantial size of the corpus, and I believe that TFIDF would be a suitable vectorization method for achieving optimal results.

**SBERT + K-means:**

I decided to go with 10 clusters and these would be suitable topic names for the clusters formed:

Cluster 0 (Size: 891): Marijuana and Drug Use Trends in the United States

Cluster 1 (Size: 945): Cardiovascular Health and Risk Factors

Cluster 2 (Size: 1340): Cancer Research and Predictive Biomarkers

Cluster 3 (Size: 851): Maternal and Child Health Issues

Cluster 4 (Size: 895): Cognitive Health and Dementia Prevention

Cluster 5 (Size: 1049): Obesity, Weight Management, and Diet

Cluster 6 (Size: 1257): Innovations in Joint and Bone Health

Cluster 7 (Size: 942): Global Health, Infectious Diseases, and Pandemics

Cluster 8 (Size: 695): Depression, Mental Health, and Treatment Approaches

Cluster 9 (Size: 1135): Emergency Care, Healthcare Disparities, and Patient Outcomes

The SBERT + K-Means model has performed well in forming appropriate clusters. The 10-cluster configuration proves to be suitable, with each cluster capturing distinct and coherent themes. The balanced distribution of cluster sizes ensures manageable analysis without dominance by a single cluster. These clusters align effectively with the content, indicating successful grouping of related documents.

**BERTopic Modelling:**

A group of colored bars with black text

Description automatically generated

A graph of a number of people with different colored bars

Description automatically generated with medium confidence

Topic 0: Cognition and memory related Diseases

Topic 1: Diabetes and Blood Sugar

Topic 2: Pregnancy and Childbirth

Topic 3: Gastrointestinal Health

Topic 4: Alcohol Consumption

Topic 5: Prostate Health

Topic 6: HIV and AIDS

Topic 7: Opioid Use and Abuse

Topic 8: Vision Health

Topic 9: Pandemic and Infections

The BERTopic model also did an effective job of clustering the data into distinct health-related topics. Each cluster encapsulates specific domains such as Cognitive diseases, diabetes, pregnancy, gastrointestinal issues, alcohol consumption patterns, prostate health, HIV/AIDS, and opioid usage. The keywords within each cluster are highly relevant to the overall theme of the respective cluster, indicating that the topics have been well-defined and are likely to be meaningful in a real-world context.

**Comparative Insights and Conclusion:**

The SBERT + K-Means and BERTopic models offer both similar and distinct insights. They share the commonality of effectively clustering health-related content. However, they differ in granularity, with SBERT + K-Means providing a broad overview of health topics, including drug trends, cardiovascular health, cancer research, and more. In contrast, BERTopic uncovers specific health domains such as cognitive diseases, diabetes, and vision health. While both models offer valuable insights, SBERT + K-Means provides a broader understanding, while BERTopic delves into more specific health conditions and areas, making it a choice depending on the analytical needs.