NLP Project
Group - 20
Topic:- Effective Healthcare answer summarization

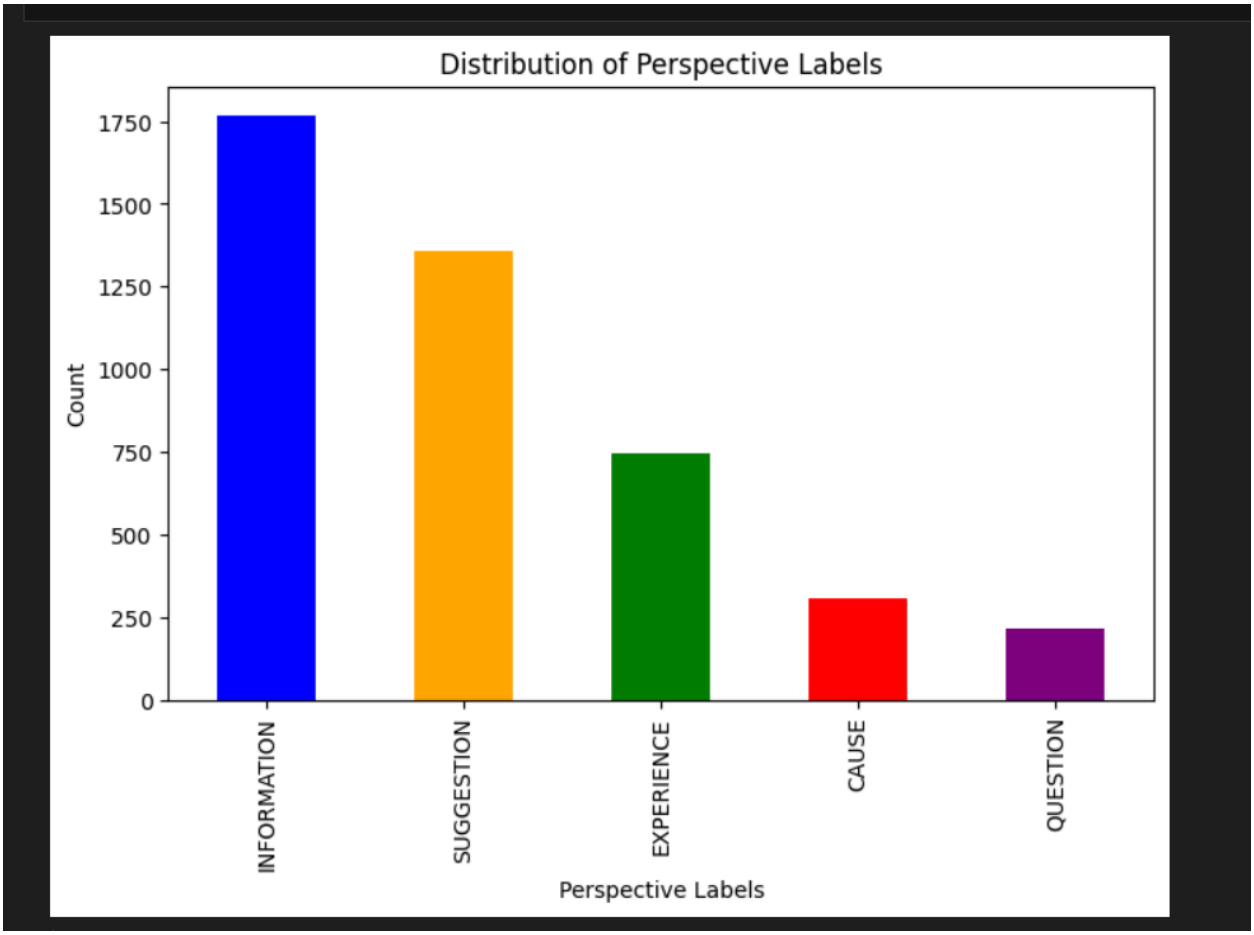
A small description of our NLP project, highlighting the key objectives and methodologies.

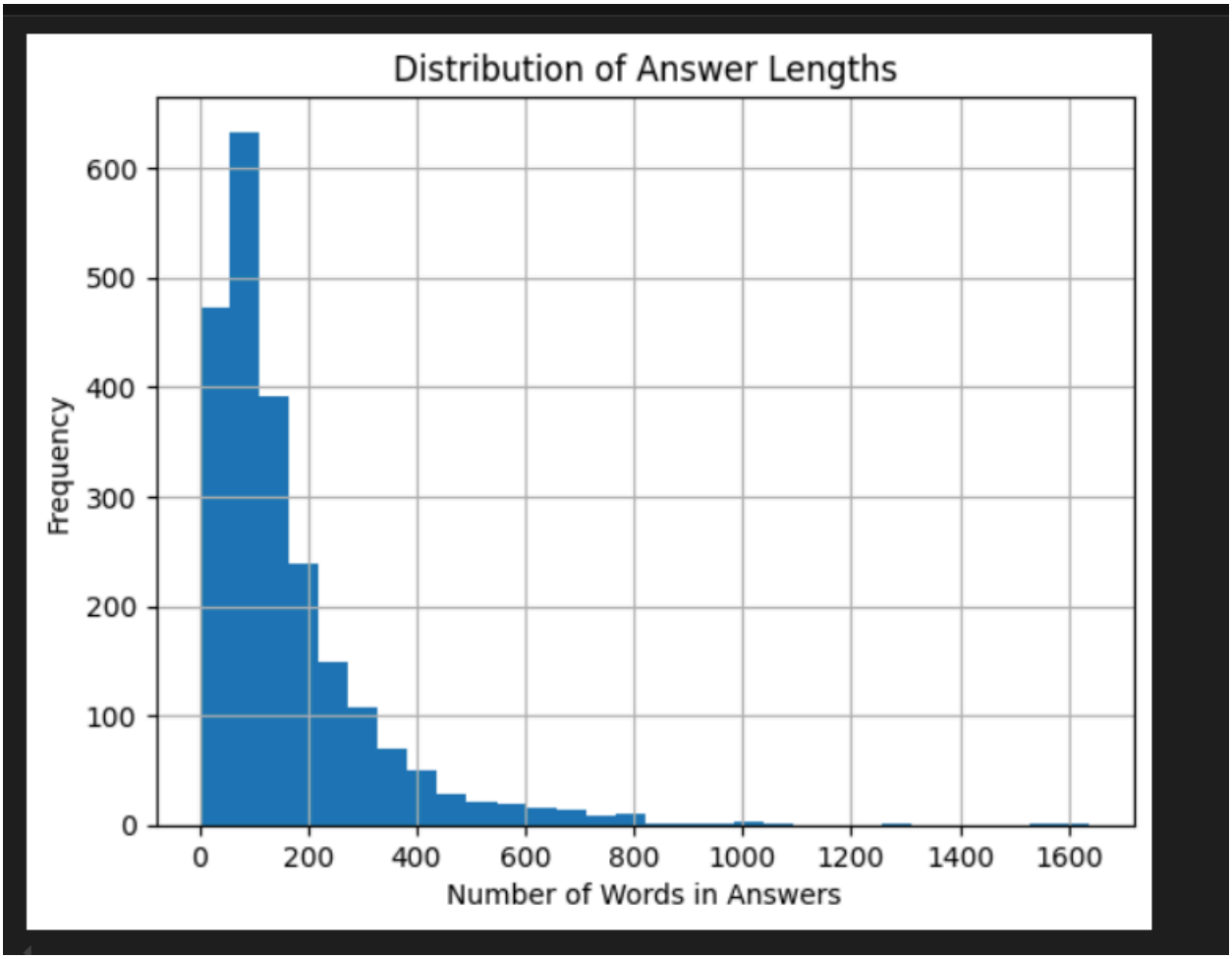
Our project focuses on generating perspective-based summaries of answers by extracting diverse perspectives such as **causes, suggestions, and personal experiences** from multiple user responses to generate a single best summary. For this, we have trained two models:

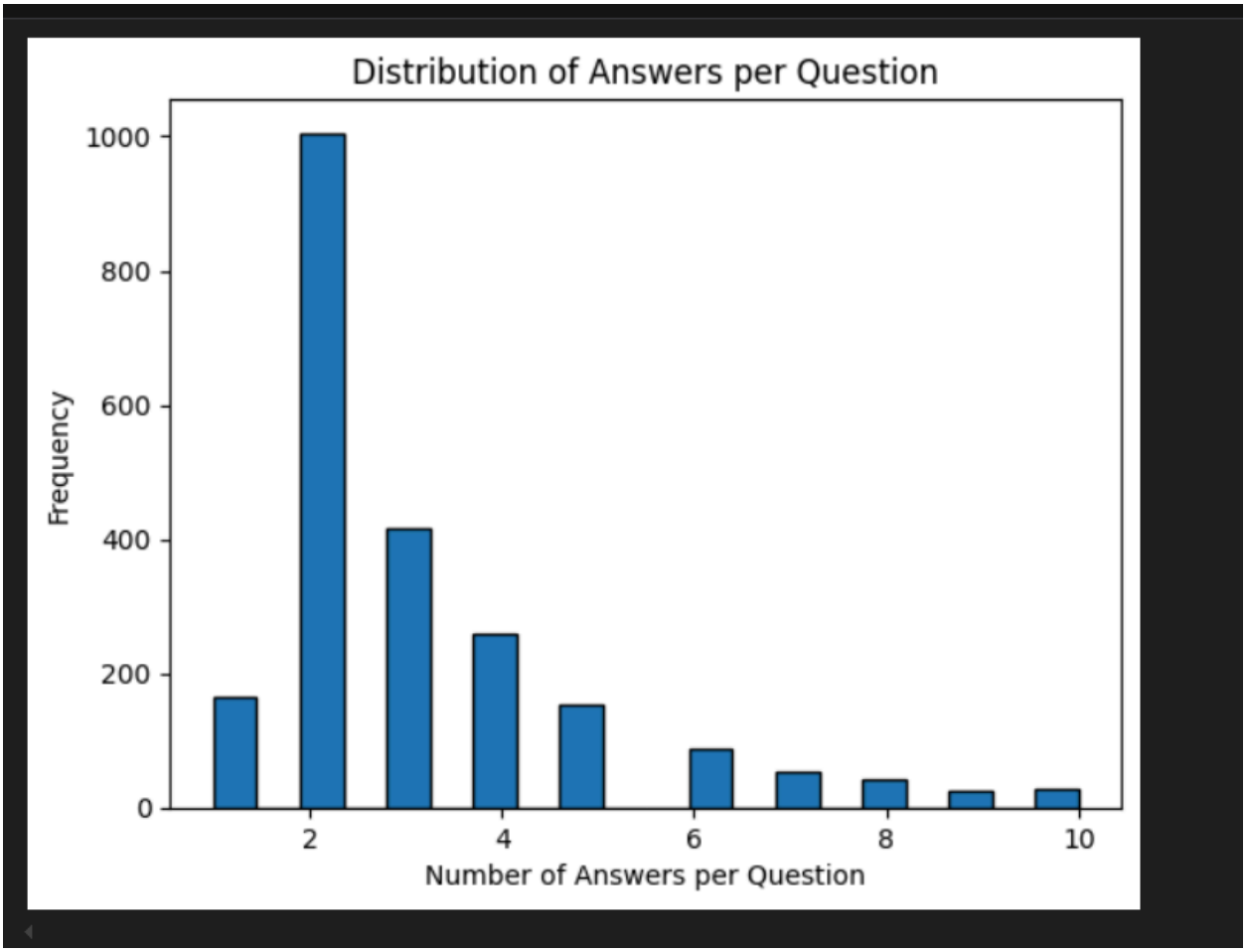
1. **Flan-T5 Model**: We used the pretrained **Flan-T5** model and further fine-tuned it. Summaries generated by both the pretrained and fine-tuned versions are displayed in the submitted notebook.
2. **Plasma Model**: This model is based on the approach used in a research paper. This model is used for perspective based learning. We have attached the results obtained from this model as well.

Workflow

1. **Exploratory Data Analysis (EDA):**
   Performed initial exploration of the dataset (e.g., checking data distribution, missing values, text length).


Distribution of Perspective Labels
Count
1750
1500
1250
1000
750
500
250
0
INFORMATION
SUGGESTION
EXPERIENCE
CAUSE
QUESTION
Perspective Labels


Distribution of Answer Lengths
Frequency
0
100
200
300
400
500
600
0
200
400
600
800
1000
1200
1400
1600
Number of Words in Answers



2. **Data Preprocessing:**
   - Handled missing values in the dataset.
   - Created input-output pairs for training into the required format(question,context,answers,etc)
   - Initialized and applied a tokenizer.
   - Converted tokenized inputs and outputs into tensors.

3. Model Selection and Fine-Tuning:
   - Used the **Flan-T5** model.
   - Implemented zero-shot baseline to see out-of-the-box performance
   - Fine-tuned the model on the prepared dataset to improve performance

4. **Evaluation and Comparison:**
   - Computed BLEU score for both baseline and fine-tuned models.
   - Compared zero-shot vs. fine-tuned results side by side.
   - Displayed sample predictions for better qualitative assessment.

```
========================================
Example 1
----------------------------------------
Input Text:
Summarize from the INFORMATION perspective:  Question: what is orgasm? Answers: An orgasm, also known as a sexual climax

Dictionaries still give the subsidiary meaning, "a similar point of intensity of emotional excitement," but as of 2005 t


General
Both males and females can experience orgasm, but the exact response varies across gender. Generally speaking, orgasm is

Many concepts may play a role in building orgasm, including thinking about making a baby, dominance and submission, etc.

True love might be a component of a good orgasm, but is itself different and in the long term more fundamental than the
----------------------------------------
Reference Summary:
For information purposes, an orgasm is a pleasurable response to sexual stimulation, involving physical, emotional, and
----------------------------------------
Fine-Tuned Model's Prediction:
For information purposes, an orgasm, also known as a sexual climax, is a pleasurable physical, psychological, or emotion
========================================
```

For PLASMA, we have trained the PLASMA model given in this paper

Results for the plasma model:

Epoch1

```
Validation completed. Results saved to './generated/generated_result.csv'
Train Loss: 1.3649 | Validation Loss: 0.8723
Best model saved at ./saved_model
Epoch 2/5
Training: 100%|                                                                 | 559/559 [13:21<00:00,  1.43s/it]
Validation processing...
```

Epoch 2

```
Validation completed. Results saved to './generated/generated_result.csv'
Train Loss: 0.8581 | Validation Loss: 0.8372
Best model saved at ./saved_model
Epoch 3/5
```

Epoch 3

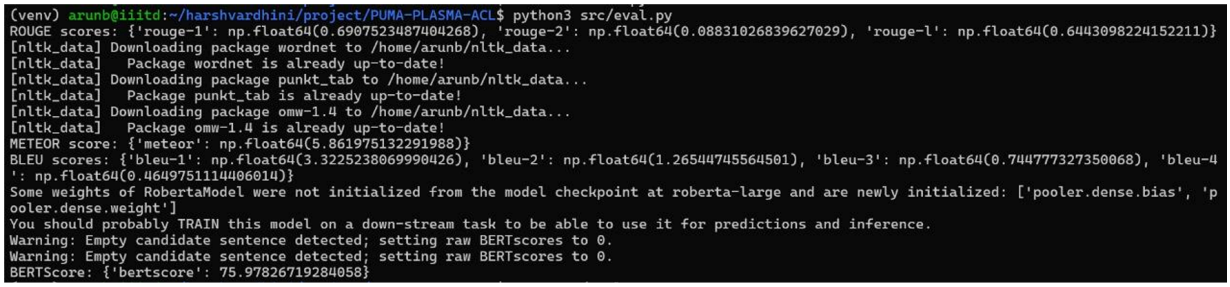
```
Validation completed. Results saved to './generated/generated_result.csv'
Train Loss: 0.8291 | Validation Loss: 0.8213
Best model saved at ./saved_model
Epoch 4/5
Training: 100%|                                                                 | 559/559 [13:24<00:00,  1.44s/it]
Validation processing...
```

Epoch 4

```
Validation completed. Results saved to './generated/generated_result.csv'
Train Loss: 0.8129 | Validation Loss: 0.8140
Best model saved at ./saved_model
Epoch 5/5
```

Epoch 5

Evaluation metrices/ Results



Next Steps and Improvements

So far, we have implemented a basic approach by referencing the research paper, training two models, and evaluating their performance using simple preprocessing techniques. Moving forward, our goal is to enhance the performance of the model by:

1. **Researching Advanced Models**: Exploring state-of-the-art models specifically designed for medical question answering to improve summarization quality.
2. **Stacking Models**: Implementing a combination of models to leverage different strengths and generate better results.
3. **Fine-tuning with Diverse Pretrained Models**: Experimenting with multiple pretrained models to compare their effectiveness in our task.
4. **Optimizing Preprocessing**: Improving data preprocessing techniques to better prepare the input for model training.
5. **Hyperparameter Tuning**: Experimenting with different hyperparameter settings to optimize model performance.
6. **Incorporating External Datasets**: If applicable, expanding our dataset to improve generalization and robustness.

Execution Instructions

Running the Project

This project was executed using **Google Colab** for training the **Flan-T5** model and a **GPU server from our college** for training the **Plasma Model**. To run the project:

1. **Open the Jupyter Notebook (`NLP_Project_20.ipynb`)** in Google Colab.

2. **Install required libraries** by running the necessary setup cells at the beginning of the notebook.
3. **Follow the execution order** and run each cell sequentially to replicate our results.
4. For the **Plasma Model**, refer to the precomputed results attached in the notebook, as it was trained on the college GPU server.