Perspective-Based Summarization of Healthcare Answers

Group No. 79

1 Introduction

Online healthcare forums contain large volumes of user-generated content where individuals seek medical advice and share experiences. While these platforms provide valuable insights, the responses are often lengthy, repetitive, and contain multiple viewpoints. To enhance readability and information retrieval, this project focuses on developing an NLP-based system that generates concise, perspective-based summaries of healthcare-related answers while preserving essential information and intent.

2 Problem Definition

The primary goal of this task is to automatically generate high-quality summaries of medical Q&A discussions while capturing diverse perspectives, including factual information, user experiences, medical suggestions, causes, and follow-up questions. The challenge lies in ensuring that the summaries maintain clarity, coherence, and relevance to the original context. The effectiveness of the system is evaluated using BLEU and BERT scores, ensuring linguistic fluency and semantic alignment with human-written references.

3 High-Level Plan

- **Dataset Preparation:** Process raw Q&A data, ensuring text consistency and structure.
- **Preprocessing:** Perform text normalization, stopword removal, tokenization, and lemmatization.
- **Summarization Model:** Implement transformer-based abstractive summarization models, including BART, T5, and PEGASUS.
- **Evaluation:** Assess generated summaries using BLEU and BERT scores.
- **Baseline Implementation:** Compare results with extractive summarization methods like TextRank.

4 Approach

Preprocessing: The dataset is cleaned by removing special characters, redundant text, and irrelevant content. Text is tokenized and categorized into five perspectives: Information, Cause, Suggestion, Experience, and Question. This structured format improves input representation for summarization models.

Model Selection: The system leverages PLASMA, a transformer-based model integrating prefix tuning and energy-controlled loss functions to generate perspective-aware summaries. Additionally, BART and T5 serve as baseline models to compare performance.

Insights from the Paper: The research study introduces the PUMA dataset, consisting of 3167 CQA threads annotated with five perspective types. It proposes PLASMA, a novel approach utilizing prompt-based summarization with optimized loss functions, achieving significant improvements over traditional models.

5 Conclusion

By structuring Q&A preprocessing and leveraging state-of-the-art transformer-based models, this project aims to generate high-quality, perspective-based summaries that improve information accessibility. The integration of robust evaluation metrics ensures the system's effectiveness in summarizing diverse healthcare discussions while maintaining readability and coherence.