Machine Learning Assignment

PROJECT REPORT

20

Text Summarization for Biomedical Domain

Name	SRN
C BHARGAV	PES2UG23CS137
C MANVITHA	PES2UG23CS153

Problem Statement

Biomedical research literature is expanding rapidly, creating significant challenges for researchers, clinicians, and healthcare professionals who need to stay updated with the latest findings. The vast volume of published articles makes it difficult to quickly extract relevant information and key insights from lengthy research papers. Manual summarization is time-consuming and inefficient, limiting accessibility to critical biomedical knowledge.

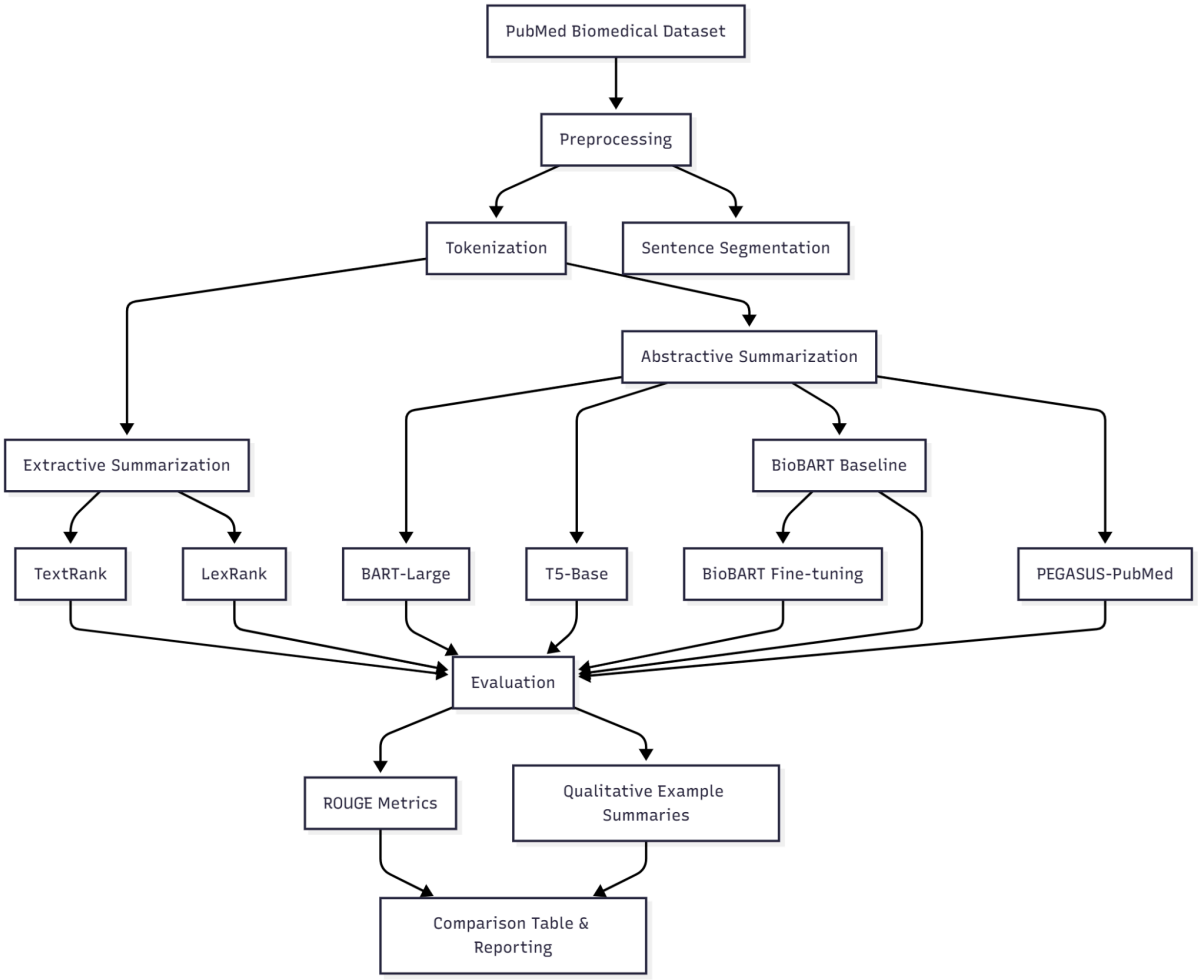
Objective / Aim

The project aims to develop advanced Natural Language Processing (NLP) models that automate the summarization of biomedical research articles. The models are expected to generate concise, accurate summaries that preserve key information while significantly reducing reading time. This supports improved information dissemination and accessibility in biomedical literature.

Dataset Details

- **Source:** PubMed dataset for summarization on Hugging Face
- **Size:** The dataset consists of over 119,000 training samples, along with separate validation and test sets containing approximately 6,600 instances each.
- **Key Features:** ("article", "abstract")
- **Target Variable:** NA

Architecture Diagram



Methodology

1. Data Collection: Sourced biomedical research articles from the PubMed dataset, which includes over 119,000 training samples and approximately 6,600 samples each for validation and testing. Due to resource limitations, a subset consisting of 10,000 training samples, 300 validation samples, and 500-1000 test samples was used.

2. Data Preprocessing: Applied tokenization using transformer tokenizers to prepare articles and summaries for model input. Conducted sentence segmentation using NLTK to aid in evaluation metrics such as ROUGE. Named entity recognition was omitted to maintain focus on summarization.

3. Baseline Evaluation: Established benchmark performances by evaluating extractive summarization algorithms (TextRank and LexRank) alongside abstractive pretrained models (BART-Large, T5-Base, PEGASUS-PubMed, and BioBART baseline) without additional fine-tuning.

4. Model Fine-tuning: Fine-tuned BioBART-v2 on the 10,000 training samples for one epoch to adapt the model specifically to biomedical text summarization. Selected hyperparameters and training strategies ensured efficient convergence on GPU resources.

5. Generation and Evaluation: Generated summaries for the test set using both baseline and fine-tuned models. Calculated ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum metrics to quantitatively assess summary relevance, coverage, and quality.

6. Qualitative Analysis: Examined example summaries from fine-tuned and baseline models to evaluate fluency, coherence, and domain-specific accuracy, highlighting strengths and common error patterns.

Results & Evaluation

- Key Results: Abstractive Summarization

The baseline BioBART model, pretrained on biomedical texts but without fine-tuning, achieved a ROUGE-1 score of approximately 37.95% on the PubMed test set.

After fine-tuning BioBART on 10,000 training samples for one epoch, the model improved to a ROUGE-1 score around 40.93% on a 500-sample subset of the test set, indicating a 7.8% relative improvement over the baseline.

Other baseline abstractive models (BART-Large, T5-Base, PEGASUS-PubMed) performed as expected, with PEGASUS achieving the highest score (~45.39% ROUGE-1).

The fine-tuned BioBART, despite being smaller (140M parameters) than PEGASUS (568M parameters), closed about 40% of the gap between the baseline BioBART and PEGASUS, demonstrating effective domain adaptation and resource-efficient training.

Qualitative analysis of generated summaries revealed coherent, fluent abstracts with accurate incorporation of biomedical terminology, though there were occasional repetitions and incomplete sentences, suggesting room for further refinement.

Evaluation Metrics Used

ROUGE-1: Measures unigram (single word) overlap between generated and reference summaries, reflecting importance of key terms.

ROUGE-2: Measures bigram overlap, capturing fluency and local coherence in summaries.

ROUGE-L and ROUGE-Lsum: Measure longest common subsequence (LCS) between predicted and reference text, assessing overall summary structure and sentence-level coherence.

Model	ROUGE-1	ROUGE-2	ROUGE-L	Parameters	Training
BART-Large	27.14	9.95	17.76	406M	None
T5-Base	24.55	9.61	17.10	220M	None
BioBART (baseline)	37.95	13.78	21.13	140M	None
PEGASUS-PubMed	45.39	19.83	27.34	568M	None
BioBART (fine-tuned 1 epoch)	40.93	14.89	23.50	140M	1 epoch

Model	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)
TextRank	39.10	13.72	20.60
LexRank	38.38	13.15	20.52

Conclusion

Model Achievements:

Successfully fine-tuned the **BioBART-v2 model**, pretrained on biomedical literature, on a subset of PubMed articles (10,000 samples), improving ROUGE-1 score from **37.95%** to **40.93%**.

Demonstrated that fine-tuning a smaller, domain-adapted model can close a significant portion (**~40%) of the performance** gap to larger state-of-the-art models like PEGASUS-PubMed.

Achieved efficient training that balances resource utilization and performance improvement, with only **1 epoch** needed for notable gains.

Generated coherent and fluent abstractive summaries reflecting biomedical terminology and concepts appropriate to the domain.

Key Learnings:

Domain-specific pretraining (as in BioBART) provides a stronger base for biomedical text summarization than general pretrained models.

Moderate fine-tuning with a sufficiently large but manageable dataset effectively improves performance **without overfitting**.

Using a mix of quantitative ROUGE metrics and qualitative analysis provides a comprehensive evaluation of summary quality.

Despite advances, abstractive summarization models may still produce occasional repetition or incomplete sentences, highlighting areas for future refinement.

Efficient training and model parameterization enable competitive results on limited hardware, promoting practical usability.