# SEQUENCE-TO-SEQUENCE NETWORKS FOR MULTI-TEXT DOCUMENT SUMMARIZATION

MINI PROJECT-02


UNDER THE GUIDANCE OF

DR. PAVAN KUMAR C

HOD, CSE

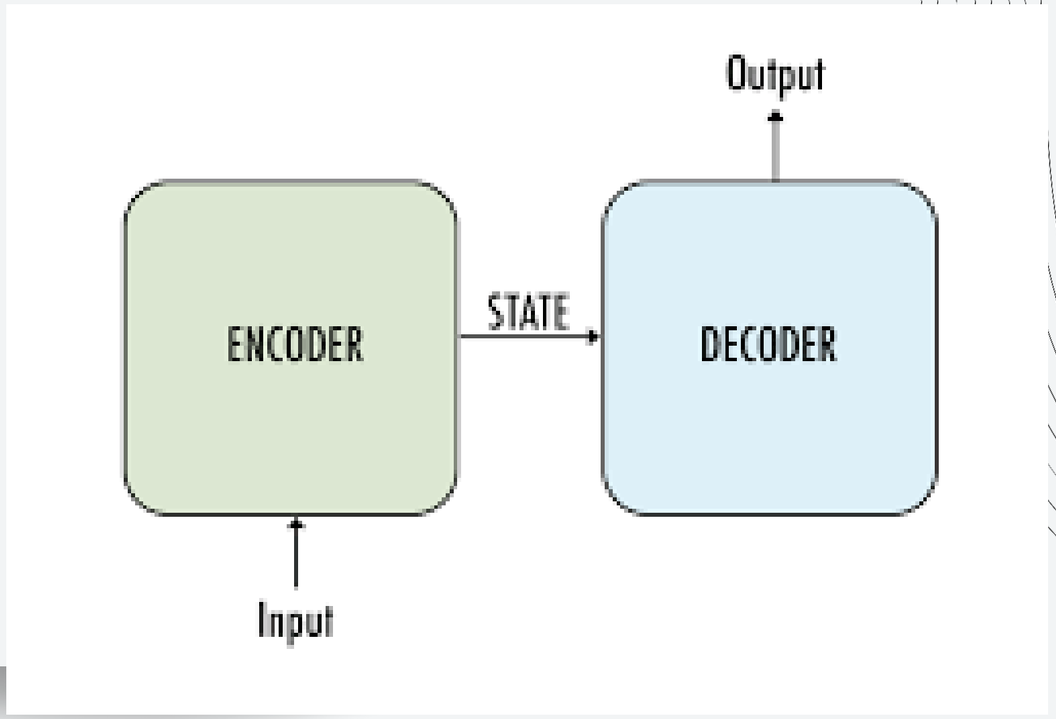
# CONTENT

# INTRODUCTION

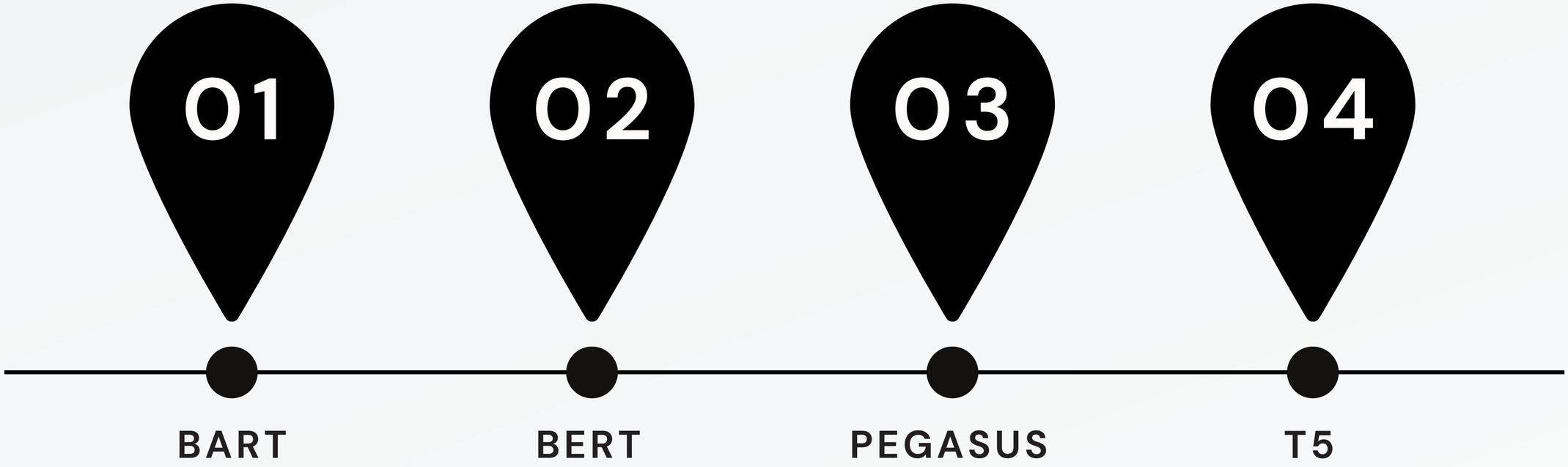
- The medical field is evolving rapidly, producing a massive volume of data. Staying updated is crucial for healthcare professionals, researchers and policymakers.
- Multi-text document summarization in the medical domain addresses this challenge.
    1. concise
    2. coherent
    3. informative
- Recent advancements in Natural Language Processing have greatly enhanced text summarization capabilities.
- Researchers have explored advanced machine learning models such as **BERT**, **BART**, **PEGASUS** and **T5**.
- Multi-text document summarization in the medical domain goes beyond single-document summarization.
- Involves identifying key themes, important findings, and relevant details across multiple documents.
- We aim to shed light on the capabilities of these advanced techniques, paving the way for improved information extraction and knowledge dissemination in the ever-evolving field of medical research and healthcare.

# SEQUENCE-TO-SEQUENCE MODELS

- Sequence-to-sequence networks is a powerful architecture for various NLP (Natural Language Processing) tasks.
- In the context of medical domain text summarization, it provide a flexible framework for capturing the relationships and dependencies among sentences in different documents.
- These networks consist of two main components:
    1. Encoder
    2. Decoder
- By leveraging the sequential nature of textual data, Seq2Seq networks can produce coherent and contextually relevant summaries.



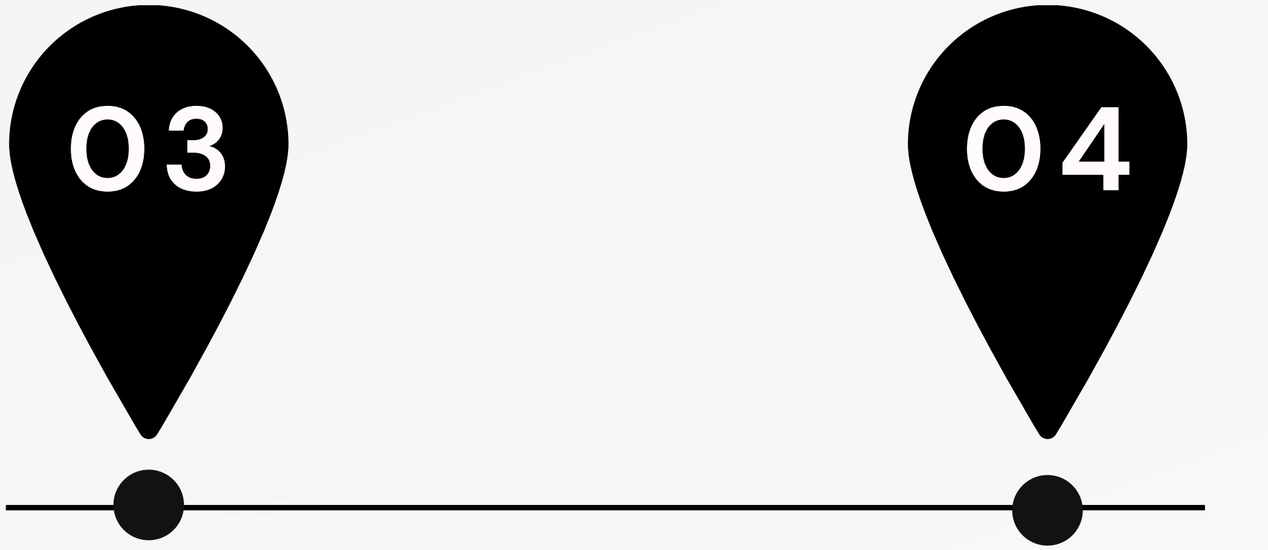
# TRANSFORMER MODELS



# LITERATURE REVIEW

- Researchers adapt Seq2Seq architectures to capture semantic relationships and context between medical documents.
- Attention mechanisms are employed to focus on relevant information.
- Domain-specific embeddings are incorporated to enhance summarization quality.
- Challenges persist, including improving abstractive summaries, addressing data sparsity, and enhancing adaptability to diverse medical sub-domains.
- Hybrid models, combining Seq2Seq networks with reinforcement learning or pre-trained language models, are proposed for further enhancement.
- Literature review demonstrates the evolution of Seq2Seq networks in multi-text document summarization within the medical domain.
- Despite challenges, continuous research and innovations propel the field forward, offering more effective solutions for knowledge dissemination among healthcare professionals and researchers.

# DATASETS



# EVALUATION METRICS

- The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric is a set of metrics used for evaluating the quality of summaries generated by automatic summarization systems.
- ROUGE metrics include measures such as ROUGE-N, ROUGE-L, and ROUGE-W, among others:

1. **ROUGE-N** (N-gram Overlap): ROUGE-N includes various values of n, such as ROUGE-1 (unigrams), ROUGE-2 (bigrams), etc.
2. **ROUGE-L** (Longest Common Subsequence):ROUGE-L measures the longest common subsequence
3. **ROUGE-W** (Weighted Longest Common Subsequence): ROUGE-W is an extension of ROUGE-L that gives more weight to contiguous and in-order common subsequences.

# IMPLEMENTATION

# RESULTS AND ANALYSIS

| MODEL | DATASET | Base model | | | | Fine tuned model | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Rouge-1 | Rouge-2 | Rouge L | Rouge L sum | Rouge 1 | Rouge 2 | Rouge L | Rouge L sum |
| Pegasus | medical_meadow_cord19 | 0.004794 | 0.000055 | 0.004728 | 0.004711 | 0.005054 | 0.000068 | 0.004982 | 0.004988 |
| | medical_cord19 | 0.004429 | 0.000154 | 0.004365 | 0.004375 | 0.005123 | 0.0 | 0.00512 | 0.005134 |
| BART | medical_meadow_cord19 | 0.004602 | 0.000162 | 0.004623 | 0.004578 | 0.006755 | 0.000387 | 0.006798 | 0.006733 |
| | medical_cord19 | 0.003995 | 0.000075 | 0.003963 | 0.003972 | 0.005844 | 0.000154 | 0.005781 | 0.005834 |
| BERT | medical_meadow_cord19 | 0.002795 | 0.000065 | 0.002729 | 0.002713 | 0.003054 | 0.000078 | 0.003982 | 0.003988 |
| | medical_cord19 | 0.002429 | 0.000254 | 0.002365 | 0.002375 | 0.003123 | 0.000389 | 0.00312 | 0.003134 |
| T5 | medical_meadow_cord19 | 0.004381 | 0.000025 | 0.004366 | 0.004321 | 0.005123 | 0.000048 | 0.004388 | 0.004369 |
| | medical_cord19 | 0.004349 | 0.000123 | 0.004338 | 0.004349 | 0.005112 | 0.0 | 0.00511 | 0.005122 |

fig.1 ROUGE scores obtained

# CONCLUSION

- This research explores advanced transformer models, including Pegasus, BART, BERT, and T5, for biomedical text summarization.
- These models empower timely decision-making, fostering accelerated advancements in medical research and healthcare.
- The study's insights and comparative analyses provide valuable tools, transforming how medical knowledge is accessed and utilized.
- Embracing challenges, fostering collaborations, and pushing technological boundaries are essential for enhancing healthcare quality and global health outcomes.

# THANK YOU!