Summarizing Clinical Text: An NLP Approach to Simplifying Doctor Notes

Aditya Jadhav
Department of Computer Science
Virginia Tech
adityasj@vt.edu

Jay Sarode
Department of Computer Science
Virginia Tech
jaysarode@vt.edu

Rashmi Kulkarni
Department of Computer Science
Virginia Tech
rashmi.kulkarni@vt.edu

I. Project Idea

The project focuses on developing a model that can automatically generate concise summaries from extensive doctor notes, a crucial aspect of medical documentation. Doctor notes typically include detailed observations, patient histories, diagnostic information, treatment plans, and follow-up recommendations. While these notes are essential for delivering comprehensive patient care, the sheer volume and complexity of information can be overwhelming, especially in fast-paced clinical environments. Reviewing such detailed documentation quickly and accurately poses a significant challenge for healthcare providers, who must make timely and well-informed decisions. An automated summarization system aims to address this challenge by condensing doctor notes into shorter, easily digestible summaries that capture the key points of each case. This would allow clinicians to rapidly access vital patient information without sifting through lengthy documentation, helping them maintain a focus on critical details. Ultimately, an effective summarization model supports more efficient patient care, enhances communication among healthcare teams, and can contribute to improved decision-making in treatment and diagnosis. This approach holds great potential for optimizing workflows in healthcare settings, ensuring that essential patient information is accessible when it is needed most.

II. Motivation

Medical documentation requires significant time and attention to interpret, particularly in fast-paced clinical environments. Summarizing doctor notes into concise representations allows healthcare providers to access essential information more efficiently. Automatic summarization offers several benefits: it improves efficiency by providing clinicians with a rapid overview of patient information for quicker decision-making; it enhances communication within healthcare teams by focusing on critical details; and it supports patient safety by highlighting important conditions, treatments, and diagnoses. With these advantages, this project aims to aid healthcare providers in managing patient information effectively.

III. Literature Review

The systematic review by [1] explores the integration of Natural Language Processing (NLP) within Mental Health Interventions (MHI), highlighting its applications, trends, and limitations. The authors analyze 102 studies, identifying rapid adoption of NLP since 2019, with increased sample sizes and use of large language models. Digital health platforms have emerged as significant data sources, providing conversational data for interventions. Key findings indicate that NLP models focusing on text-based features outperform those using audio markers for clinical accuracy. The study identifies common challenges, including biases, limited reproducibility, and a lack of linguistic diversity. To address these issues, the authors propose a framework, NLPxMHI, designed to improve clinical utility, access to data, and fairness in NLP applications for mental health. This comprehensive review outlines a roadmap for future research by bridging gaps between clinical and computational sciences.

Another paper [2] provides a scoping review on digital mental health tools specifically for young people, aged 0-25, analyzing both the benefits and ethical challenges of these technologies. The study discusses how digital mental health can enhance accessibility, therapy facilitation, and patient empowerment. Despite these benefits, ethical issues remain, particularly regarding data privacy, clinical validation, and user-centered design. The authors stress the importance of implementing robust ethical standards and technical guidelines to address these challenges and ensure the safe deployment of digital mental health solutions in clinical settings [2]. The review serves as a foundation for stakeholders involved in developing ethically-aligned mental health technologies targeted at youth.

Another paper talks about how the field of automatic text summarization (ATS) has undergone significant development, with advancements in both methodologies and applications. Summarization techniques are commonly divided into extractive and abstractive approaches. Extractive summarization focuses on selecting key sentences directly from the text, while abstractive methods involve rephrasing content to create more concise representations. ATS typically follows a workflow that includes preprocessing, feature extraction, and applying summarization algorithms, which has evolved from traditional statistical methods to modern deep learning approaches. Recent progress in the field has seen the use of transformer-based models, which have shown promising results for producing coherent summaries. However, several challenges remain, including handling linguistic diversity, addressing domain-specific needs, and managing computational complexity. This comprehensive review by [3] serves as a foundational framework for our project, as it encapsulates key methodologies and advancements that can inform our approach to generating concise and informative summaries for specialized domains.

The paper ”Text Summarization Techniques: A Brief Survey” by [4] provides a concise overview of text summarization methods, focusing on the primary approaches, algorithms, and evaluation techniques. The authors categorize summarization techniques into extractive and abstractive methods, with extractive methods focusing on selecting key sentences directly from the original text and abstractive methods generating summaries by rephrasing and compressing information. The survey discusses classical methods, including statistical and graph-based algorithms, as well as the recent adoption of deep learning models, particularly neural networks and transformer architectures, which have significantly advanced summarization capabilities. El-Kassas et al. also examine the challenges in summarization, such as maintaining coherence, handling domain-specific language, and evaluating generated summaries. This survey serves as a foundational reference for our project, highlighting the strengths and limitations of different summarization approaches and informing the choice of techniques suited to our objectives in generating concise, informative summaries in specialized domains.

IV. Dataset

This study employed two distinct datasets to investigate patient information and medical transcription data. The first, the Open Patients dataset, combines 180,142 patient descriptions from four open-source collections: TREC Clinical Decision Support (CDS) and Clinical Trials (CT) tracks, MedQA-USMLE, and PMC-Patients. Each entry contains an `_id` indicating its origin and a comprehensive patient note. This dataset encompasses a mix of synthetic and authentic patient notes, clinical questions from USMLE exams, and case reports from PubMed Central, making it valuable for evaluating information retrieval and decision support systems. The second dataset consists of medical transcriptions categorized by medical specialty and keywords. It includes attributes such as brief descriptions, sample names, and complete transcriptions. This collection offers a detailed perspective on various medical cases across different specialties, facilitating analysis of transcription content and classification. Combined, these datasets provide extensive insights into patient information and clinical documentation, enabling a comprehensive examination of medical data across various contexts and specialties.

V. Evaluation Metrics

To evaluate the quality of the generated summaries, we will use a combination of automated metrics to measure accuracy and consistency. ROUGE scores will assess overlap between the generated and reference summaries, capturing content accuracy through unigram and bigram matches and sequence coherence with the longest common subsequence. BLEU scores will calculate n-gram precision from unigrams to four-grams, evaluating fluency and readability, with a brevity penalty to ensure concise output. Additionally, we will conduct a structured error analysis to identify and understand common mistakes, such as missing critical information, inclusion of irrelevant details, inaccurate phrasing, or length discrepancies. This process will involve comparing generated summaries to reference summaries to pinpoint frequent errors, tracking their frequency for quantitative insight, and examining specific cases for qualitative understanding. The findings from error analysis will guide iterative improvements, such as refining the training data, adjusting model parameters, and fine-tuning objectives, ensuring the model produces accurate and clinically consistent summaries.

VI. Methodology

A. Summarization

Summarization is the process of condensing a larger body of text into a shorter version while retaining its main ideas and key points. In natural language processing (NLP), there are three main approaches to summarization:

*1) **Extractive Summarization:*** Extractive summarization involves identifying and extracting the most significant sentences or phrases from the source text to create a concise summary. The primary goal is to rank and select the most relevant content based on linguistic, statistical, or semantic importance. This approach preserves the original text's integrity by directly selecting portions of it, making it computationally efficient and suitable for structured data processing. The methodology involves several computational techniques, including keyword extraction, sentence embedding generation, and ranking mechanisms like Maximal Marginal Relevance (MMR) to balance relevance and diversity.

Preprocessing is a critical step in transforming raw text into a structured format for computation. It involves:

- **Removing stopwords:** Filters out common words like “the” and “and,” reducing noise and improving keyword extraction.
- **Tokenization:** Splits text into sentences or words, enabling sentence-level computations.

Dynamic Keyword Extraction identifies critical terms in the document using a combination of TF-IDF (Term Frequency-Inverse Document Frequency) and Named Entity Recognition (NER). TF-IDF assigns scores to words based on their uniqueness in the document relative to the corpus, ensuring that document-specific words are emphasized over generic ones. The TF-IDF (Term Frequency-Inverse Document Frequency) metric is computed as:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \cdot \text{IDF}(t, D)$$

where:

$$\text{TF}(t, d) = \frac{\text{Frequency of } t \text{ in } d}{\text{Total terms in } d}$$

$$\text{IDF}(t, D) = \log\left(\frac{|D|}{1 + |\{d \in D : t \in d\}|}\right)$$

NER further enriches this process by identifying semantic entities, such as diseases or chemicals, ensuring domain-specific relevance. Additionally, a predefined list of medical or domain-specific keywords is incorporated to enhance the robustness of the extraction process. These predefined keywords act as a foundation, ensuring that critical terms which may not be statistically significant in the current text are still considered. This combined approach of TF-IDF, NER, and a predefined keyword list captures statistical, semantic, and domain-specific importance, ensuring the extracted keywords reflect the unique and critical content of the document.

Sentences are represented as high-dimensional embeddings using pre-trained models, capturing semantic relationships between words. These embeddings are normalized to unit vectors for consistency:

$$\vec{s}_{\text{normalized}} = \frac{\vec{s}}{||\vec{s}||}$$

This semantic representation allows for precise sentence comparisons, enabling accurate assessments of relevance and diversity.

Maximal Marginal Relevance (MMR) is a key component of this summarization approach, designed to balance two competing objectives: relevance and diversity. Relevance ensures that selected sentences are closely aligned with the document's main topic, while diversity minimizes redundancy, ensuring that the summary covers distinct and complementary information. The MMR score for a sentence is computed as:

$$\text{MMR} = \lambda \cdot \text{Relevance} - (1 - \lambda) \cdot \text{Diversity}$$

Relevance is calculated as the cosine similarity between the query embedding and the candidate sentence embedding, representing how closely the sentence relates to the core content. Relevance is calculated as:

$$\text{Relevance}(q, s) = \cos(\vec{q}, \vec{s}) = \frac{\vec{q} \cdot \vec{s}}{||\vec{q}|| \, ||\vec{s}||}$$

Diversity, on the other hand, measures how different the candidate sentence is from the already selected sentences, ensuring that redundant or repetitive information is avoided. Diversity is measured as:

$$\text{Diversity}(s, S_{\text{selected}}) = \max_{s' \in S_{\text{selected}}} \cos(\vec{s}, \vec{s'})$$

The trade-off parameter λ controls the balance between these two objectives: a higher λ prioritizes relevance, while a lower λ emphasizes diversity.

Dynamic adjustment of λ based on the number of sentences in the document is critical for achieving optimal summarization results. In shorter documents, fewer sentences inherently reduce the risk of redundancy, so a higher λ ensures that the summary remains tightly focused on the main topic. Conversely, longer documents are richer in content and more prone to redundancy, necessitating a lower λ to promote diversity and better cover the broader range of information. By adapting λ dynamically, the approach ensures a balanced and effective summary irrespective of the input size, allowing for tailored relevance and diversity trade-offs suited to the document's characteristics. This adaptability makes MMR particularly powerful in handling varied datasets.

*2) **Abstractive Summarization:*** Abstractive summarization involves generating a summary that may include new sentences or paraphrased content not explicitly found in the original text. This process requires the model to understand the text, extract core information, and rephrase it fluently and coherently. In this study, we utilized two state-of-the-art models for abstractive summarization: PEGASUS-XSum and T5. Both models leverage transformer architectures but are tailored for distinct summarization styles and datasets, providing a robust comparative framework.

PEGASUS-XSum (Extreme Summarization) is a model explicitly fine-tuned on the XSum dataset, which emphasizes concise, single-sentence summaries. PEGASUS employs a novel pretraining objective called ”Gap Sentence Generation” (GSG), where key sentences from the input are masked, and the model is trained to predict them. This approach makes PEGASUS particularly adept at summarizing long documents into short, highly condensed summaries while preserving critical information.

In contrast, **T5 (Text-to-Text Transfer Transformer)** is a versatile transformer model trained with a unified framework that treats all NLP tasks as text-to-text transformations. Fine-tuned for summarization, T5 excels at generating longer, more detailed summaries with a balance between informativeness and readability. The two models were then compared using standard summarization evaluation metrics BLEU (Bilingual Evaluation Understudy).

*3) **Hybrid Summarization:*** Hybrid summarization combines the strengths of extractive and abstractive methods, aiming to generate concise, coherent, and contextually rich summaries. While extractive summarization selects sentences directly from the original text based on their importance, abstractive summarization rephrases the content to create a more natural and fluent summary.

By merging these approaches, hybrid summarization leverages the precision of extractive methods to identify key content and the flexibility of abstractive methods to generate human-like summaries. This combination ensures that the final output is both accurate and linguistically refined.

Why Hybrid Summarization?

While abstractive summarization alone can produce summaries that are fluent and contextually rich, it comes with challenges such as potential hallucinations—generating information not present in the original text. This is particularly problematic in domain-specific contexts, such as medical or scientific summarization, where accuracy is critical. Extractive summarization, on the other hand, guarantees factual correctness by selecting verbatim content from the source. However, it often lacks fluency and coherence, especially when the selected sentences are concatenated without further processing. Hy-

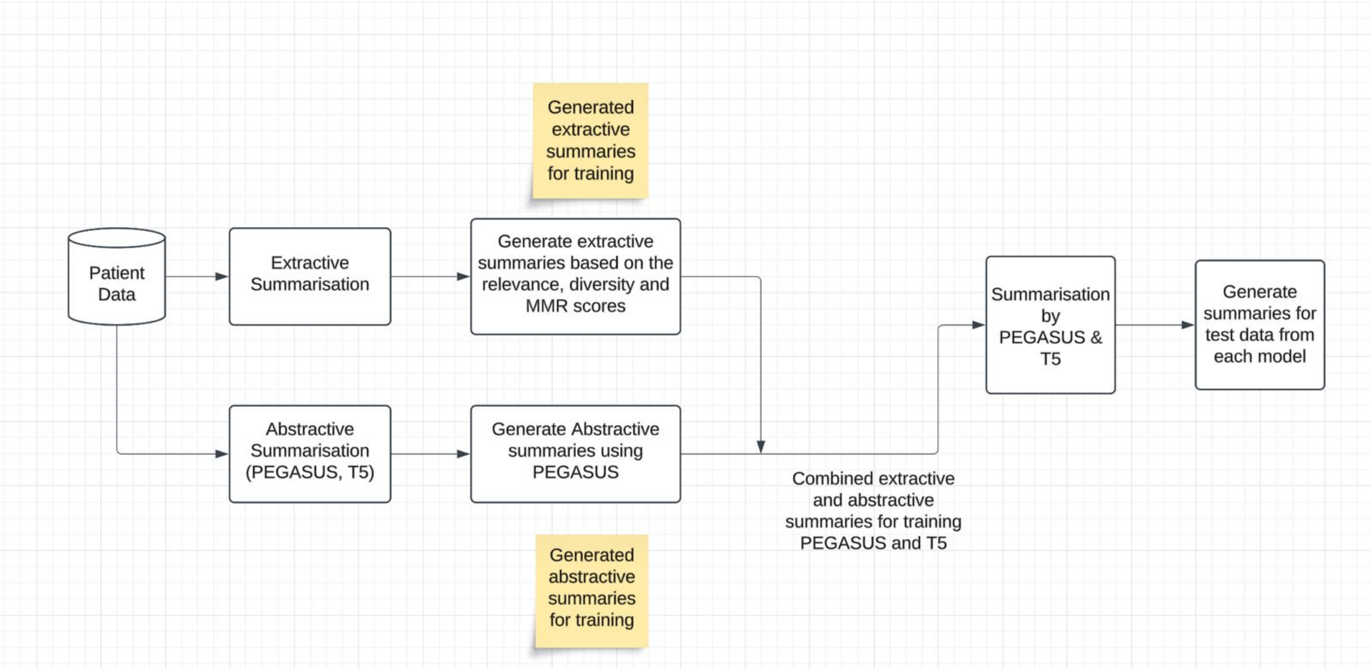

Generated extractive summaries for training
Patient Data
Extractive Summarisation
Generate extractive summaries based on the relevance, diversity and MMR scores
Summarisation by PEGASUS & T5
Generate summaries for test data from each model
Abstractive Summarisation (PEGASUS, T5)
Generate Abstractive summaries using PEGASUS
Combined extractive and abstractive summaries for training PEGASUS and T5
Generated abstractive summaries for training

Fig. 1: Flow diagram for the Hybrid Summarisation approach

brid summarization addresses these shortcomings by ensuring factual accuracy through extractive steps while leveraging the abstractive process to enhance fluency, coherence, and readability.

The decision to employ hybrid summarization was driven by the need to balance accuracy with natural language generation. Extractive summarization, powered by MMR (Maximal Marginal Relevance), was used to identify the most relevant and diverse sentences, ensuring that the key points were captured without redundancy. These sentences, being contextually significant, form an optimal input for abstractive summarization. By feeding these high-quality extracted sentences into an abstractive model like PEGASUS-XSum or T5, the summarization pipeline ensures that the generated output is both precise and linguistically polished.

Hybrid Summarization Workflow: The hybrid summarization process begins with extractive summarization using the MMR method. In this step, sentences are ranked based on their relevance to the document's main topic (measured by cosine similarity) and their diversity relative to already selected sentences. This ensures that the top-n sentences chosen are not only highly relevant but also non-redundant, effectively summarizing the document's core content. These top-n sentences are then passed as input to the abstractive summarization step.

In the abstractive phase, models like PEGASUS-XSum or T5 take the extracted sentences and generate a cohesive summary. PEGASUS-XSum, optimized for creating highly condensed summaries, and T5, capable of generating longer and more detailed summaries, both excel in rephrasing and refining the content. By using the extracted sentences as input, the abstractive model focuses on condensing and rephrasing the most critical information, reducing the risk of including irrelevant or hallucinated content.

Synthetic Summary Generation:

To generate synthetic summaries, we evaluated the scores and quality of different summarization approaches and decided to adopt the hybrid method for this task. This approach combines the factual accuracy of extractive summarization with the fluency and coherence of abstractive summarization, resulting in summaries that are both accurate and linguistically refined. For synthetic summary generation, we used the PEGASUS-XSum model, which is particularly effective at creating concise and high-quality summaries. The model was fine-tuned to generate these summaries, which became the ground truth for the subsequent training phase. This step serves as a critical preprocessing stage, ensuring the availability of high-quality labeled data for the training process. The decision to adopt the hybrid approach for generating synthetic summaries ensures a reliable and contextually rich dataset, forming the basis for further refinement during training.

Training and Testing: For the training and testing phases, we utilized both the PEGASUS-Large model and the T5 model. This dual-model approach was chosen to leverage the complementary strengths of these architectures, ensuring a more robust summarization framework. Using the same model family for preprocessing (PEGASUS-XSum) and training (PEGASUS-Large) ensures consistency in the summarization pipeline, aligning the ground truth synthetic summaries with the training objectives. Similarly, incorporating T5 adds versatility due to its ability to handle diverse input structures and generate detailed, contextually rich summaries.

Why Use the Same Model for Preprocessing and Training?

By using PEGASUS for both preprocessing and training, we ensure that the synthetic summaries generated in the preprocessing phase are inherently compatible with the training phase. This alignment reduces model adaptation time, enhances learning efficiency, and ensures seamless integration between phases. The PEGASUS-Large model, being more robust than PEGASUS-XSum, is capable of handling complex summarization tasks during training and testing, enabling it to generalize effectively to unseen data.

Differences Between Preprocessing and Training The primary distinction between preprocessing and training lies in their objectives. During preprocessing, PEGASUS-XSum is fine-tuned to generate high-quality synthetic summaries that act as the ground truth. This phase focuses on creating reliable, accurate datasets. During training, both PEGASUS-Large and T5 are fine-tuned on these synthetic datasets to generalize summarization capabilities across unseen data. While preprocessing generates labeled data, training refines the model to perform optimally in real-world scenarios, ensuring scalability and accuracy in generating contextually relevant summaries. This dual-model strategy ensures that the summarization pipeline is both versatile and highly effective.

B. Multi-Document Summarization:

Multi-document summarization (MDS) is the process of generating a concise and coherent summary from multiple related documents, capturing the key information while minimizing redundancy and irrelevant content. Unlike single-document summarization, which focuses on one source, MDS handles the challenges of synthesizing information from diverse sources, often with overlapping or contradictory content.

The workflow employs a combination of advanced NLP techniques, including **Zero-Shot Learning (ZSL)**, **embedding generation**, **clustering**, and **Maximal Marginal Relevance (MMR)** ranking, to achieve precise and efficient multi-document summarization. Multi-document summarization is motivated by the need to distill large volumes of information spread across multiple documents into concise, coherent, and non-redundant summaries. This is particularly valuable in domains such as medicine, law, and research, where information is scattered across various sources, often containing overlapping or conflicting content. The goal is to create summaries that provide comprehensive insights, resolve contradictions, and reduce redundancy, making them more accessible to users.

Zero-Shot Learning (ZSL) is a powerful machine learning approach that allows models to make predictions on tasks or classes they have not been explicitly trained on. In the context of summarization and classification, ZSL enables the model to categorize or process unseen documents using pretrained knowledge and contextual understanding. Pretrained transformer models, such as `facebook/bart-large-mnli`, encode both the input text x and the target label y (e.g., “This text discusses cardiology”) into vector representations, $\vec{x}$ and $\vec{y}$, respectively. These are mapped into a shared semantic space where their similarity is measured using cosine similarity:

$$\text{Similarity}(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|}$$

The similarity scores are normalized using the softmax function to compute the probabilities for each label:

$$P(y|x) = \frac{\exp(\text{Similarity}(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\text{Similarity}(x, y'))}$$

The label with the highest probability is selected as the prediction:

$$y^* = \arg\max_{y \in \mathcal{Y}} P(y|x)$$

ZSL is invaluable for handling diverse and unseen tasks, enabling scalability and flexibility in workflows where labeled data is unavailable. Once categorized, the text undergoes **embedding generation**, where sentences are transformed into high-dimensional vectors using pretrained models.

To reduce redundancy and identify thematic overlaps, **clustering** is applied using KMeans. This technique groups embeddings into k clusters by minimizing intra-cluster distances:

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{\vec{s} \in C_i} \|\vec{s} - \vec{\mu}_i\|^2$$

where $\vec{\mu}_i$ is the centroid of the i-th cluster. This step ensures that sentences with similar content are grouped together, enabling diverse topic coverage. To refine the selection of sentences for summarization, **Maximal Marginal Relevance (MMR)** ranks sentences based on relevance and diversity.

Multi-Document Summarization workflow: The multi-document summarization pipeline begins with unlabelled patient data containing textual descriptions, such as doctors' notes. Since these descriptions lack predefined labels, the pipeline starts by assigning labels through zero-shot classification using a pre-trained model like facebook/bart-large-mnli. This model evaluates each description against a set of predefined candidate labels (e.g., medical conditions) and assigns the most relevant label if the confidence score surpasses a specified threshold. Descriptions meeting this criterion are categorized as ”high-confidence” data. For low-confidence data, *embeddings* are generated using a model like BiomedBERT. These embeddings are high-dimensional and undergo Principal Component Analysis (PCA) to reduce dimensionality while retaining the most important features. PCA transforms the original d-dimensional embeddings $\vec{X}$ into a lower p-dimensional space using the following linear transformation:

$$\vec{Z} = \vec{X}\vec{W}$$

where, $\vec{Z}$ is the transformed data in the reduced p-dimensional space, $\vec{W}$ is the matrix of the top p eigenvectors of the covariance matrix of $\vec{X}$, $\vec{X}$ is the original data matrix with n samples and d features.

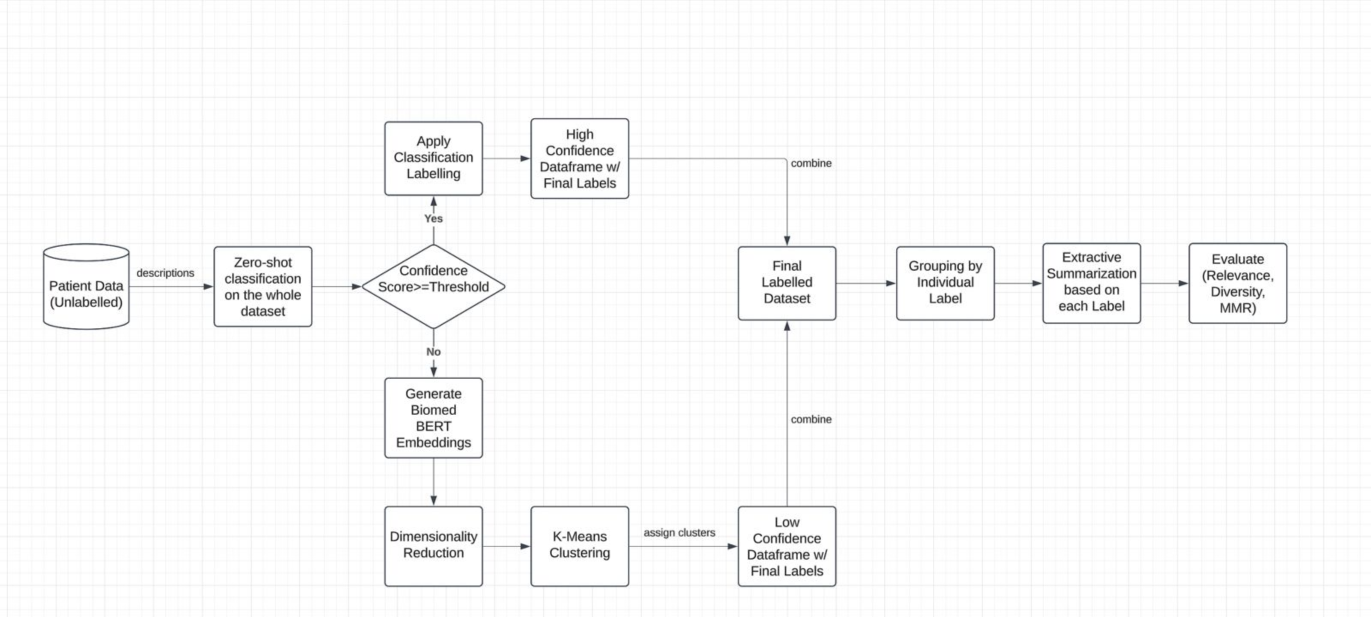

Patient Data (Unlabelled)
descriptions
Zero-shot classification on the whole dataset
Confidence Score>=Threshold
Yes
No
Apply Classification Labelling
High Confidence Dataframe w/ Final Labels
combine
Generate Biomed BERT Embeddings
Dimensionality Reduction
K-Means Clustering
assign clusters
Low Confidence Dataframe w/ Final Labels
combine
Final Labelled Dataset
Grouping by Individual Label
Extractive Summarization based on each Label
Evaluate (Relevance, Diversity, MMR)

Fig. 2: Flow diagram for Multi-Document Summarization

After dimensionality reduction, the *K-Means algorithm* is applied to group the low-confidence data into k clusters. K-Means minimizes the within-cluster variance by solving the following optimization problem:

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{\vec{s} \in C_i} \|\vec{s} - \vec{\mu}_i\|^2$$

where, C_i represents the i-th cluster, $\vec{s}$ is a data point in cluster C_i, $\vec{\mu}_i$ is the centroid (mean) of cluster C_i, k is the number of clusters.

The clusters are mapped to predefined labels, forming the ”low-confidence” dataset. The high- and low-confidence datasets are then combined into a unified ”final labelled dataset,” where all descriptions have been assigned a label through either classification or clustering. Once the data is labelled, descriptions are grouped by their final labels to form consolidated collections. This transition from multi-document to grouped-document format facilitates deeper analysis of the data. Within each group, all associated descriptions are concatenated into a single text. Extractive summarization is then applied to the grouped texts using Maximal Marginal Relevance (MMR). This technique selects the most relevant and diverse sentences to create concise summaries. Relevance is determined based on the similarity of each sentence to the overall document embedding, while diversity is ensured by minimizing redundancy. The balance between relevance and diversity is controlled through a tunable parameter, ensuring high-quality summaries that capture the essence of the grouped data while maintaining variability in the selected content. The final step involves evaluating the quality of the generated summaries using metrics like relevance, diversity, and MMR scores. These scores allow for an objective assessment of how well the summaries encapsulate the grouped texts.

VII. Result

A. *Results from Summarization:*

The chart illustrates the variation in Relevance, Diversity, and Maximal Marginal Relevance (MMR) scores across different data points, where the x-axis denotes the range in thousands and the y-axis represents the scores between 0 and 1. The Relevance score (blue line) remains consistently high, fluctuating around 0.8, indicating that the system reliably identifies sentences closely aligned with the main topic or query.

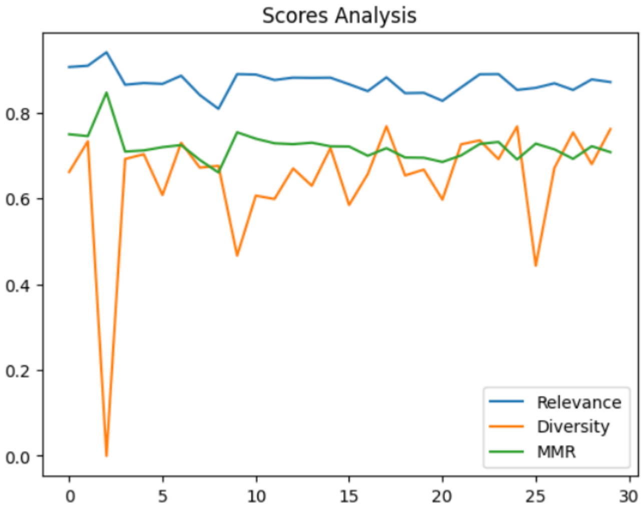
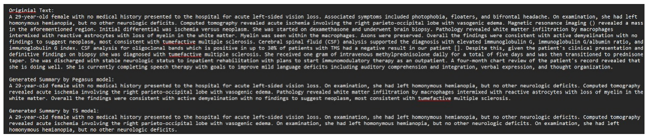


Fig. 3: Score Analysis in Extractive Summarization

In contrast, the Diversity score (orange line) exhibits significant variability, with sharp drops in some data ranges, reflecting challenges in selecting distinct sentences when input data is highly redundant. Peaks in diversity scores show the system's ability to reduce redundancy when the input data offers more variability.

The MMR score (green line), which balances relevance and diversity, stays relatively stable between the two, demonstrating that the system maintains an effective trade-off to optimize both factors.



```
Originial Text:
A 29-year-old female with no medical history presented to the hospital for acute left-sided vision loss. Associated symptoms included photophobia, floaters, and bifrontal headache. On examination, she had left
homonymous hemianopia, but no other neurologic deficits. Computed tomography revealed acute ischemia involving the right parieto-occipital lobe with vasogenic edema. Magnetic resonance imaging () revealed a mass
in the aforementioned region. Initial differential was ischemia versus neoplasm. She was started on dexamethasone and underwent brain biopsy. Pathology revealed white matter infiltration by macrophages
intermixed with reactive astrocytes with loss of myelin in the white matter. Myelin was seen within the macrophages. Axons were preserved. Overall the findings were consistent with active demyelination with no
findings to suggest neoplasm, most consistent with tumefactive multiple sclerosis. Cerebral spinal fluid (CSF) analysis supported the diagnosis with elevated immunoglobulin G, immunoglobulin G/albumin ratio, and
immunoglobulin G index. CSF analysis for oligoclonal bands which is positive in up to 30% of patients with TMS had a negative result in our patient []. Despite this, given the patient's clinical presentation and
definitive findings on biopsy she was diagnosed with tumefactive multiple sclerosis. She received one gram of intravenous methylprednisolone daily for a total of five days and was then transitioned to prednisone
taper. She was discharged with stable neurologic status to inpatient rehabilitation with plans to start immunomodulatory therapy as an outpatient. A four-month chart review of the patient's record revealed that
she is doing well. She is currently completing speech therapy with goals to improve mild language deficits including auditory comprehension and integration, verbal expression, and thought organization.

Generated Summary by Pegasus model:
A 29-year-old female with no medical history presented to the hospital for acute left-sided vision loss. On examination, she had left homonymous hemianopia, but no other neurologic deficits. Computed tomography
revealed acute ischemia involving the right parieto-occipital lobe with vasogenic edema. Pathology revealed white matter infiltration by macrophages intermixed with reactive astrocytes with loss of myelin in the
white matter. Overall the findings were consistent with active demyelination with no findings to suggest neoplasm, most consistent with tumefactive multiple sclerosis.

Generated Summary by T5 model:
A 29-year-old female with no medical history presented to the hospital for acute left-sided vision loss. On examination, she had left homonymous hemianopia, but no other neurologic deficits. Computed tomography
revealed acute ischemia involving the right parieto-occipital lobe with vasogenic edema. On examination, she had left homonymous hemianopia, but no other neurologic deficits. On examination, she had left
homonymous hemianopia, but no other neurologic deficits.
```

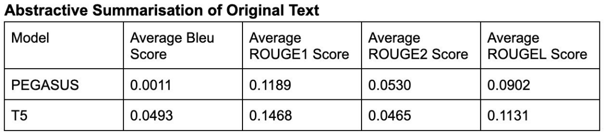
Fig. 4: Model Performance on both models

The performance of PEGASUS and T5 in generating summaries for the given medical case demonstrates distinct strengths and weaknesses. PEGASUS excels in preserving critical details, such as findings related to demyelination and tumefactive multiple sclerosis, ensuring that the summary retains the essential medical context. However, this comes at the cost of verbosity and occasional redundancy, making the output less concise and harder to read. On the other hand, T5 produces a more compact and readable summary, which is suitable for general-purpose applications. However, it omits important clinical details, such as demyelination and multiple sclerosis, which significantly reduces its utility in domain-specific contexts like medical summarization. While T5's brevity enhances readability, it compromises the factual accuracy and completeness required in medical scenarios. PEGASUS, despite its verbosity, aligns better with the needs of the medical domain by ensuring no critical information is lost, making it a more reliable choice for tasks where factual integrity and comprehensive coverage are paramount.

B. ***Comparative Analysis:***

The performance of abstractive models, PEGASUS and T5, on the original text demonstrates limitations in handling domain-specific data. PEGASUS achieves low scores across all metrics, with an average BLEU score of 0.0011 and ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.1189, 0.0530, and 0.0902, respectively. T5 performs slightly better, achieving an average BLEU score of 0.0493 and ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.1468, 0.0465, and 0.1131, respectively. Despite these improvements, the abstractive models struggle to align with reference summaries due to their tendency to generate hallucinated or contextually irrelevant content, particularly when dealing with complex medical text.

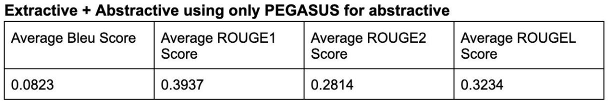
In contrast, the hybrid approach significantly outperforms the abstractive models in all metrics, demonstrating the effectiveness of combining extractive and abstractive techniques. The hybrid method, which uses extractive summarization to preprocess input for PEGASUS, achieves an average BLEU score of 0.0823 and ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.3937, 0.2814, and 0.3234, respectively. These improvements highlight the hybrid approach's ability to retain factual accuracy and generate summaries that are both contextually relevant and linguistically coherent.



Abstractive Summarisation of Original Text

Model	Average Bleu Score	Average ROUGE1 Score	Average ROUGE2 Score	Average ROUGEL Score
PEGASUS	0.0011	0.1189	0.0530	0.0902
T5	0.0493	0.1468	0.0465	0.1131

Fig. 5: Abstractive Summarisation of Original Text

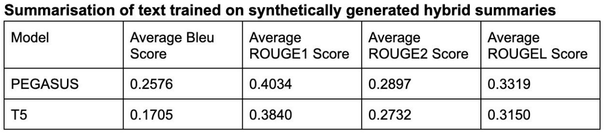


Extractive + Abstractive using only PEGASUS for abstractive

Average Bleu Score	Average ROUGE1 Score	Average ROUGE2 Score	Average ROUGEL Score
0.0823	0.3937	0.2814	0.3234

Fig. 6: Hybrid Summarisation of Original Text

The superior performance of the hybrid approach is attributed to the strengths of both methods. The extractive step ensures that only the most relevant and diverse sentences, directly sourced from the input text using Maximal Marginal Relevance (MMR), are passed to the abstractive model. This reduces the risk of hallucination and guarantees factual correctness. The abstractive step further enhances these sentences by rephrasing them for improved fluency and coherence. This combination addresses the limitations of purely abstractive summarization, which often sacrifices factual accuracy for readability, and extractive summarization, which lacks flexibility and fluency. By leveraging the strengths of both approaches, the hybrid method provides a robust solution for summarization, particularly in domain-specific contexts like medical text, where precision and linguistic quality are critical.



Summarisation of text trained on synthetically generated hybrid summaries

Model	Average Bleu Score	Average ROUGE1 Score	Average ROUGE2 Score	Average ROUGEL Score
PEGASUS	0.2576	0.4034	0.2897	0.3319
T5	0.1705	0.3840	0.2732	0.3150

Fig. 7: Predicted Summarization results

The results above demonstrate the effectiveness of training and testing PEGASUS and T5 models on synthetically generated hybrid summaries, which act as high-quality ground truth data. PEGASUS shows significant improvements across all metrics compared to the earlier hybrid summarization results, achieving a BLEU score of 0.2576 (up from 0.0823), a ROUGE-1 score of 0.4034 (up from 0.3937), a ROUGE-2 score of 0.2897 (up from 0.2814), and a ROUGE-L score of 0.3319 (up from 0.3234). T5 also exhibits notable gains, with a BLEU score of 0.1705 (up from 0.0823), a ROUGE-1 score of 0.3840 (slightly below PEGASUS), a ROUGE-2 score of 0.2732, and a ROUGE-L score of 0.3150. These results highlight that training on synthetic summaries enables the models to produce outputs that are more aligned with the ground truth, improving both lexical and contextual accuracy. PEGASUS consistently outperforms T5 across all metrics, particularly in BLEU and ROUGE-L, suggesting its superiority in generating concise and factually accurate summaries. The improvements in ROUGE-2 and ROUGE-L scores, in particular, demonstrate the models' enhanced ability to capture both local contextual relationships and global sequential structures. This validates the utility of the hybrid summarization pipeline not only as a method for generating high-quality summaries but also as a preprocessing step to create reliable training data. Overall, the approach demonstrates how synthetically generated summaries can significantly enhance model performance, ensuring better alignment, coherence, and relevance in the generated summaries.

C. Results for Multi-Document Summarization:

The plot below shows the **relevance scores** across different final labels. The relevance score indicates how closely the selected sentences in the extractive summary relate to the overall document.

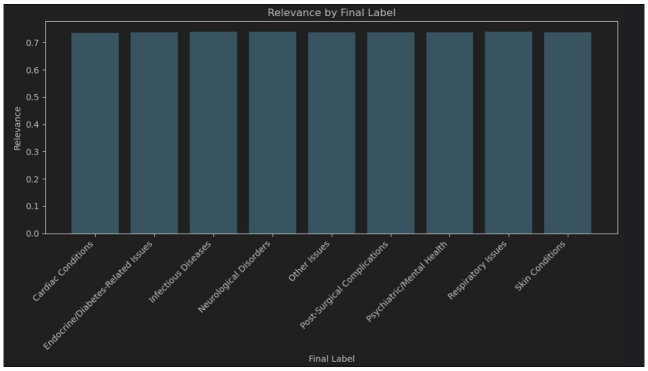


Fig. 8: Relevance scores for labels

The bars are nearly uniform, suggesting that the selected summaries are consistently relevant across all labels. This uniformity reflects that the summarization process effectively captures the core context regardless of the medical condition category.

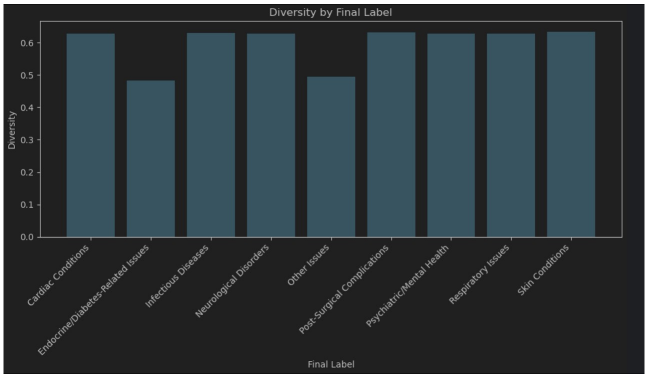


Fig. 9: Diversity scores for labels

The plot above visualizes the **diversity scores** across labels, measuring how distinct the selected sentences are within each summary. Categories such as ”Infectious Diseases” and ”Post-Surgical Complications” have lower diversity compared to others like ”Cardiac Conditions” or ”Neurological Disorders.” This indicates that summaries for the former categories may include more repetitive information, potentially due to the nature of the data or the limited variation in the original documents.

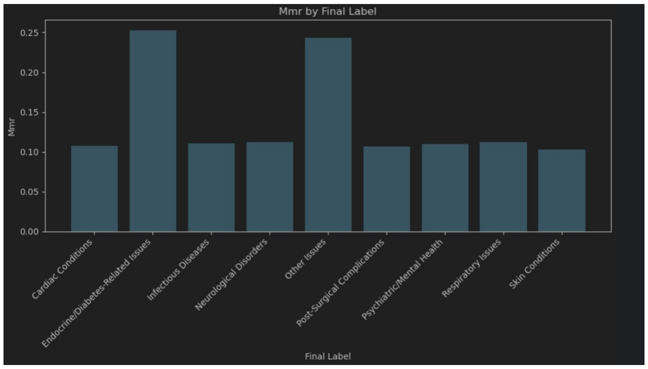


Fig. 10: MMR scores for labels

The plot above displays the Mean Maximal Relevance (MMR) scores. Higher scores indicate a better balance between relevance and diversity. Categories like ”Other Issues” and ”Post-Surgical Complications” have notably higher MMR scores, suggesting that the summarization model performs well in balancing relevance and diversity for these labels. Lower scores, as seen in ”Infectious Diseases,” might suggest a challenge in maintaining this balance due to repetitive or homogeneous text.

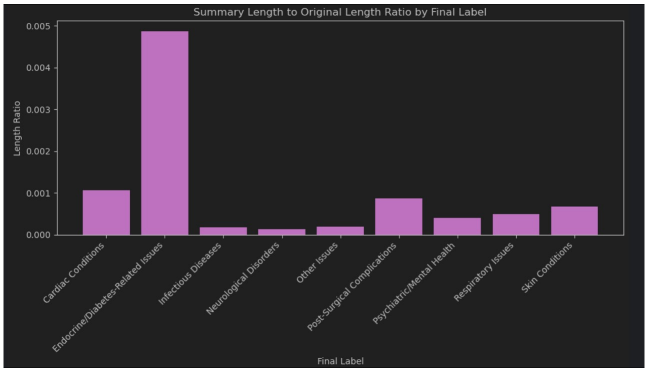


Fig. 11: Summary scores for labels

This plot compares the ratio of the summary length to the original text length for each final label. Categories like ”Endocrine/Diabetes-Related Issues” and ”Infectious Diseases” have higher ratios, implying that their summaries are relatively longer compared to their original documents. This could be due to detailed documentation for these conditions. Conversely, categories like ”Neurological Disorders” and ”Cardiac Conditions” have lower ratios, indicating more concise summaries.

VIII. Conclusion

The results obtained from both Summarization and Multi-Document Summarization emphasize the effectiveness of leveraging hybrid approaches that combine the strengths of extractive and abstractive techniques. These methods not only enable the generation of concise and coherent summaries but also ensure factual accuracy and contextual relevance, addressing some of the critical limitations observed in standalone summarization approaches.

The hybrid methodology, which employs extractive summarization to identify the most relevant and diverse sentences followed by abstractive summarization to refine fluency and coherence, consistently outperformed purely abstractive models. This dual-layered strategy mitigates common challenges like hallucination—where abstractive models introduce content not present in the original text—and redundancy, by ensuring that the core ideas are well-represented and linguistically polished. Such an approach proved particularly beneficial in domain-specific contexts, such as medical data, where factual integrity is paramount, as incorrect or incomplete information could lead to adverse outcomes. The substantial improvements in BLEU and ROUGE metrics further validate the hybrid approach's capability to balance precision and readability, demonstrating its suitability for handling complex, information-dense datasets.

In Multi-Document Summarization, the combination of advanced NLP techniques—such as Zero-Shot Learning for initial classification, clustering to group semantically related content, and Maximal Marginal Relevance (MMR) for ensuring relevance and diversity—created a robust pipeline for synthesizing information from multiple sources. This pipeline successfully addressed the challenges posed by diverse and overlapping content in multi-document contexts, producing summaries that were contextually rich, informative, and non-redundant. The flexibility and scalability of this framework make it highly adaptable to various real-world applications, especially in domains like medicine and research, where high-quality, accurate summarization is crucial for decision-making and knowledge dissemination. These findings underscore the potential of hybrid methodologies and multi-document workflows to transform how complex textual data is summarized, making the information more accessible and actionable.

IX. Future Work

While the current hybrid summarization framework demonstrates strong performance and effectiveness, there remain several avenues for improvement and exploration to enhance its robustness, adaptability, and applicability across diverse domains. One key area for development is the inclusion of domain-specific fine-tuning for models such as PEGASUS and T5. While these models perform well on general datasets, fine-tuning them on specialized datasets, such as medical, legal, or scientific corpora, could significantly enhance their ability to capture nuanced domain-specific language and improve the accuracy and relevance of generated summaries. This customization would ensure that the summaries are better aligned with the terminology, context, and conventions of specific fields, thereby increasing their reliability and practical utility.

The framework could also benefit from a deeper integration of evaluation metrics that go beyond standard BLEU and ROUGE scores. Metrics that assess factual accuracy, contextual relevance, and user satisfaction could provide a more comprehensive understanding of model performance. Additionally, incorporating user feedback into the training loop through human-in-the-loop approaches could iteratively improve the quality of generated summaries. By addressing these areas, the hybrid summarization framework has the potential to become a more robust, scalable, and versatile tool, adaptable to a wide range of complex real-world scenarios. This progression would solidify its role as a transformative technology in summarizing and synthesizing large-scale, complex textual datasets.

In addition, the incorporation of more sophisticated clustering techniques in the Multi-Document Summarization pipeline offers a promising direction. Current methods, such as K-Means, provide efficient grouping of semantically similar content but may struggle with highly complex or overlapping datasets. Advanced clustering techniques, such as hierarchical clustering, density-based clustering (e.g., DBSCAN), or graph-based approaches, could enhance the precision of grouping by capturing intricate relationships between data points. This would be particularly beneficial in cases where multiple documents contain heterogeneous or contradictory information, allowing for a more nuanced understanding and summarization of the data.

X.

References

[1] J. M. Z. M. Malgaroli, T. D. Hull and T. Althoff, "Natural language processing for mental health interventions: A systematic review and research framework," *Translational Psychiatry*, vol. 13, no. 309, 2023. [Online]. Available: https://doi.org/10.1038/s41398-023-02592-2

[2] C. L. B. Wies and M. Ienca, "Digital mental health for young people: A scoping review of ethical promises and challenges," *Frontiers in Digital Health*, vol. 3, p. 697072, 2021. [Online]. Available: https://doi.org/10.3389/fdgth.2021.697072

[3] K. N. S. C. D. M. H. M. F. Mridha, A. Akter Lima and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156 043–156 063, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3129786

[4] A. R. W. S. El-Kassas, C. Salama and H. F. Mohamed, "Text summarization techniques: A brief survey," *Future Computing and Informatics Journal*, vol. 6, no. 1, pp. 3–17, 2021. [Online]. Available: https://doi.org/10.1016/j.fcij.2020.12.001