

NLP Project Transcript Summarization

CS 410/510: Natural Language Processing
(NLP), Fall 2023



Meet our team



Our Group Chat Summary

The conversation captures a team's collaborative efforts in an NLP project. Upasana Choudhari leads the group, focusing on text summarization. The members, including Shreya Choure, Saurav, and Pranav PSU, share resources and discuss methodologies, with models like BERT and GPT2 being considered. They explore datasets for their project, evaluate using metrics like ROUGE and BLEU scores, and coordinate through meetings. The group effectively divides tasks, aiming to finalize their presentation and report, showcasing teamwork and problem-solving in a technical project context.

Shreya

Pranav



Saurav

Upasana

Agenda

- Project Overview
- Goals and Objectives
- Dataset Overview
- Types of Summarization: Extractive & Abstractive
- Model Performance Metrics
- Visualizations
- Demonstration
- References and Acknowledgments





Project overview

Our NLP Project develops conversation summarization models using TF-IDF, TextRank, and Large Language Models like T5, BART and evaluates its performance with quantitative analyses. Further It is deployed to “streamlit” for User Interaction.



**T5 / BART
Transformers**





Goals and Objectives




- To develop a **summarization model** to distill crucial information from conversations, delivering concise and coherent summaries for enhanced understanding.
- Perform **quantitative evaluation** and comparative analysis across various models to assess their effectiveness.
- **Visualize** diverse aspects of the data through a range of graphical representations.
- Deploying models on **Streamlit** for seamless and user-friendly access.

Industry Applications

- **Automated Call Categorization:** classify call center calls into predefined categories
 - **AI-Driven Issue Extraction:** provide a quick and accurate summary for efficient issue resolution.
 - **Real-time Analytics for Service Enhancement:** identify high-frequency call categories and areas for improvement, contributing to enhanced services.
- 
- 

Dataset Overview

- **DialogSum** is a conversations summarization dataset, consisting more than 13K samples split into train, test and validation with manually labeled summaries and topics.
- **Train Data:** 12.7k conversations
- **Test Data:** 1.5 k conversations
- **Validation Data:** 500

id string · lengths 	dialogue string · lengths 	summary string · lengths 	topic string · lengths 
7 ————— 11	190 ————— 5.18k	31 ————— 1.04k	2 ————— 44
train_0	#Person1#: Hi, Mr. Smith. I'm Doctor Hawkins. Why are you her...	Mr. Smith's getting a check-up, and Doctor Hawkins advises him...	get a check-up
train_1	#Person1#: Hello Mrs. Parker, how have you been? #Person2#:...	Mrs Parker takes Ricky for his vaccines. Dr. Peters checks the...	vaccines
train_2	#Person1#: Excuse me, did you see a set of keys? #Person2#:...	#Person1#'s looking for a set of keys and asks for #Person2#'s...	find keys

src: <https://huggingface.co/datasets/knkarthick/dialogsum/viewer>

Evaluation Methodology

ROUGE Scores: Recall-Oriented Understudy for Gisting Evaluation

- It's a set of metrics for the automatic evaluation of machine-generated text.
- Compares the overlap between the generated and reference summary.
- **ROUGE-N** (n-gram overlap):
 - ROUGE-1: Unigrams (individual words) overlap.
 - ROUGE-2: Bigrams overlap.
 - ROUGE-3: Trigrams overlap, and so on.
- **ROUGE-L** (Longest Common Subsequence):

Measures the overlap in the longest common subsequence between the generated summary and the reference summary.

Text Summarization Approaches

Types of Text Summarization

```
graph TD; A[Types of Text Summarization] --> B[Extractive Summary]; A --> C[Abstractive Summary]; B --> D[TF IDF]; B --> E[Text Rank]; C --> F[BART]; C --> G[T5]; C --> H[GPT 4]
```

Measures the overlap of exact word sequences between the generated summary and the reference text.

Extractive Summary

TF IDF

Text Rank

Measures the longest common subsequence, allowing for flexibility in word order.

Abstractive Summary

BART

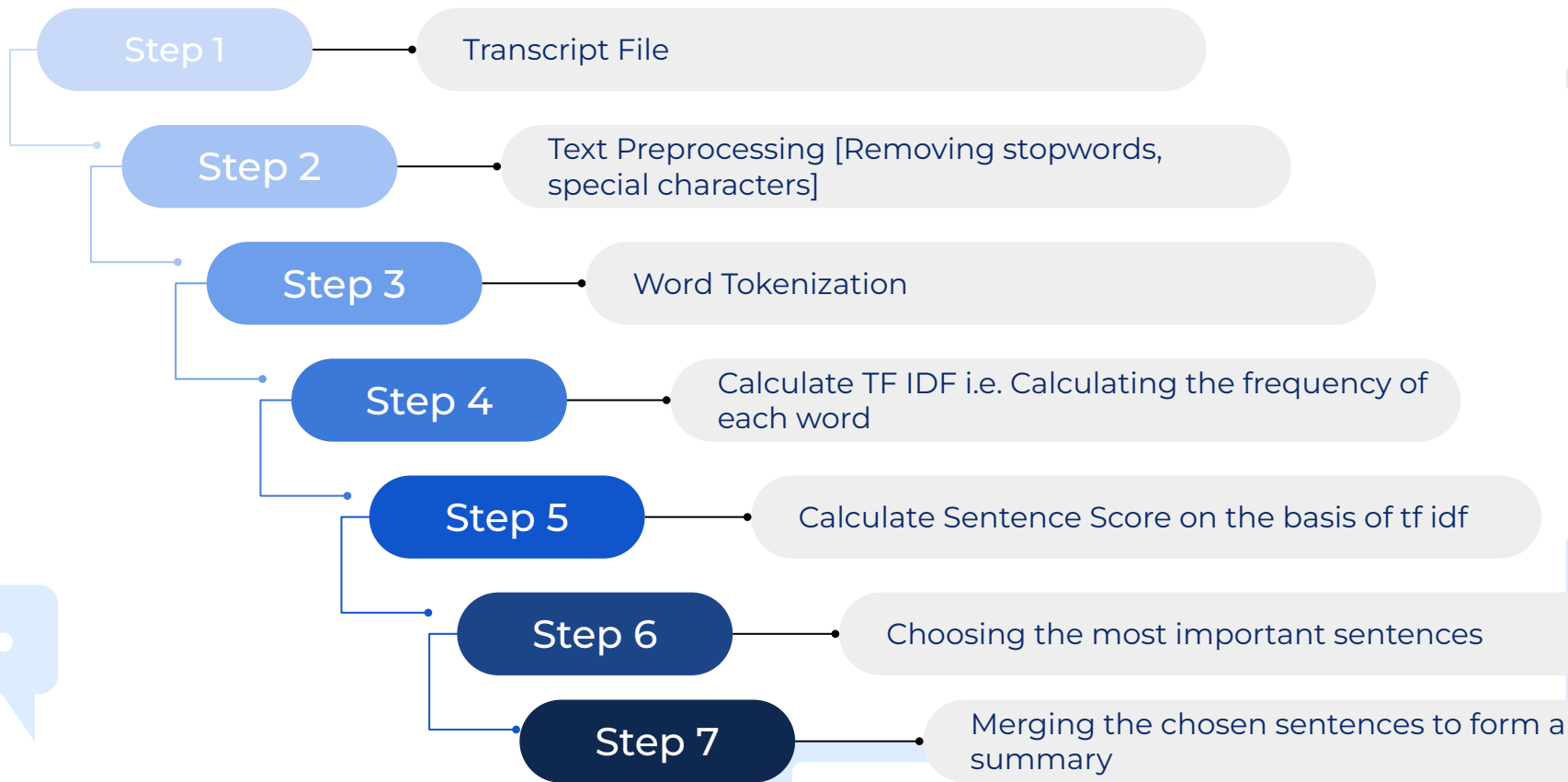
T5

GPT 4



Extractive Summary

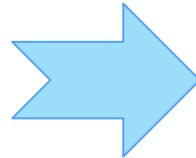
TF-IDF Term Frequency - Inverse Document Frequency



TFIDF

Conversation:

#Person1#: Hello, how are you doing today?
#Person2#: I ' Ve been having trouble breathing lately.
#Person1#: Have you had any type of cold lately?
#Person2#: No, I haven't had a cold. I just have a heavy feeling in my chest when I try to breathe.
#Person1#: Do you have any allergies that you know of?
#Person2#: No, I don't have any allergies that I know of.
#Person1#: Does this happen all the time or mostly when you are active?
#Person2#: It happens a lot when I work out.
#Person1#: I am going to send you to a pulmonary specialist who can run tests on you for asthma.
#Person2#: Thank you for your help, doctor.



Summary:

#Person2#: I ' Ve been having trouble breathing lately. **#Person2#:** It happens a lot when I work out.

Performance

TFIDF

ROUGE-1:

Precision: 0.23

Recall: 0.28

F1-Score: 0.24

ROUGE-2:

Precision: 0.06

Recall: 0.08

F1-Score: 0.07

ROUGE-L:

Precision: 0.18

Recall: 0.22

F1-Score: 0.19

- These scores suggest that the extracted summary has some relevance to the reference summary but is not highly accurate.
- The low ROUGE-2 scores indicate that the summary might be missing the exact phrasing or specific important information from the original text.
- The moderate to low scores indicate that there's significant room for improvement, particularly in capturing the details of the original text.



Text Rank

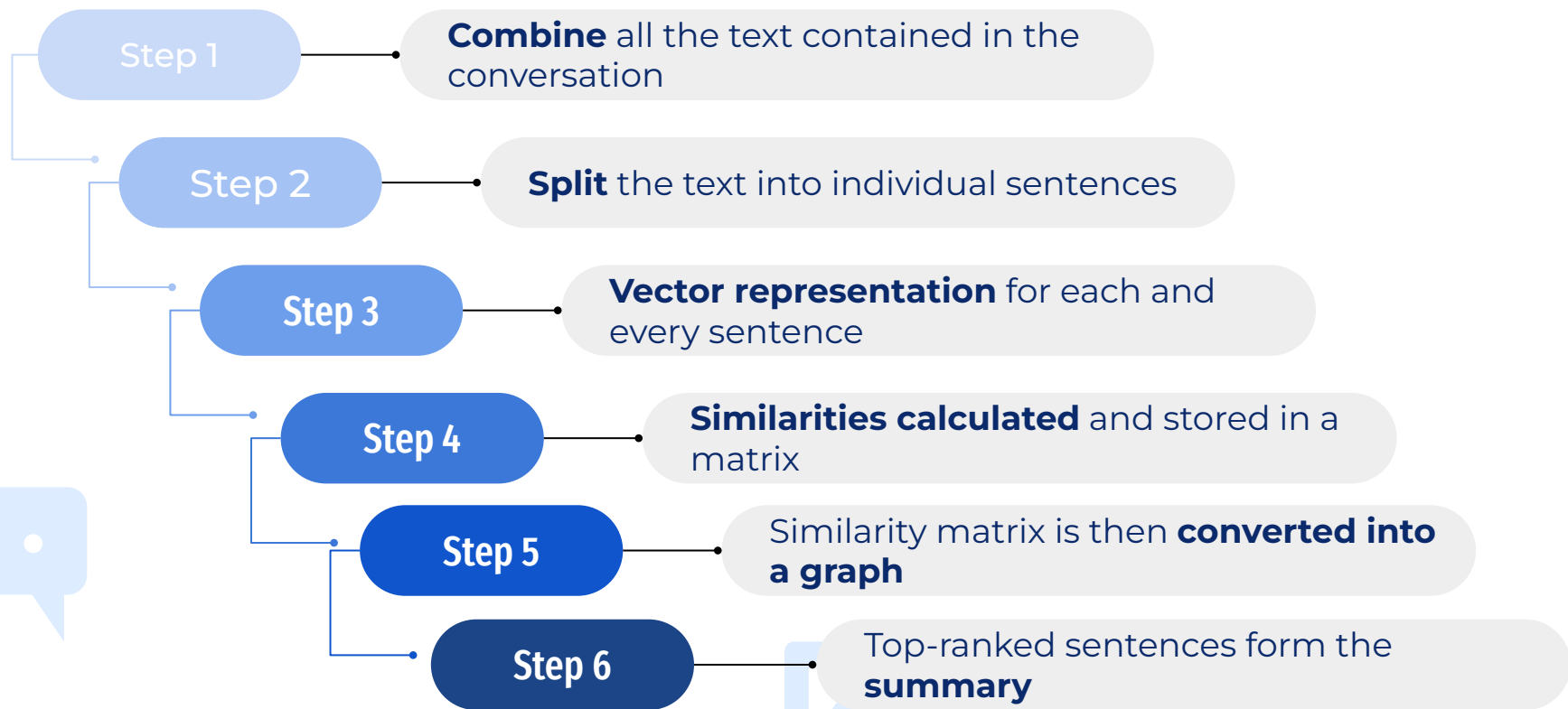
It's a part of Extractive Summarization.

Inspired by PageRank algorithm.

Graph-based ranking algorithms



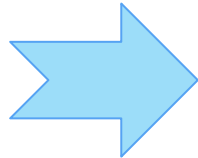
Text Rank: Process Flow



Text Rank

Conversation:

#Person1#: Hello, how are you doing today?
#Person2#: I ' Ve been having trouble breathing lately.
#Person1#: Have you had any type of cold lately?
#Person2#: No, I haven't had a cold. I just have a heavy feeling in my chest when I try to breathe.
#Person1#: Do you have any allergies that you know of?
#Person2#: No, I don't have any allergies that I know of.
#Person1#: Does this happen all the time or mostly when you are active?
#Person2#: It happens a lot when I work out.
#Person1#: I am going to send you to a pulmonary specialist who can run tests on you for asthma.
#Person2#: Thank you for your help, doctor.



Summary:

Have you had any type of cold lately? No, I don't have any allergies that I know of. Does this happen all the time or mostly when you are active?




Performance





Text Rank

ROUGE-1:
Precision: 0.13
Recall: 0.26
F1-Score: 0.17

ROUGE-2:
Precision: 0.03
Recall: 0.07
F1-Score: 0.05



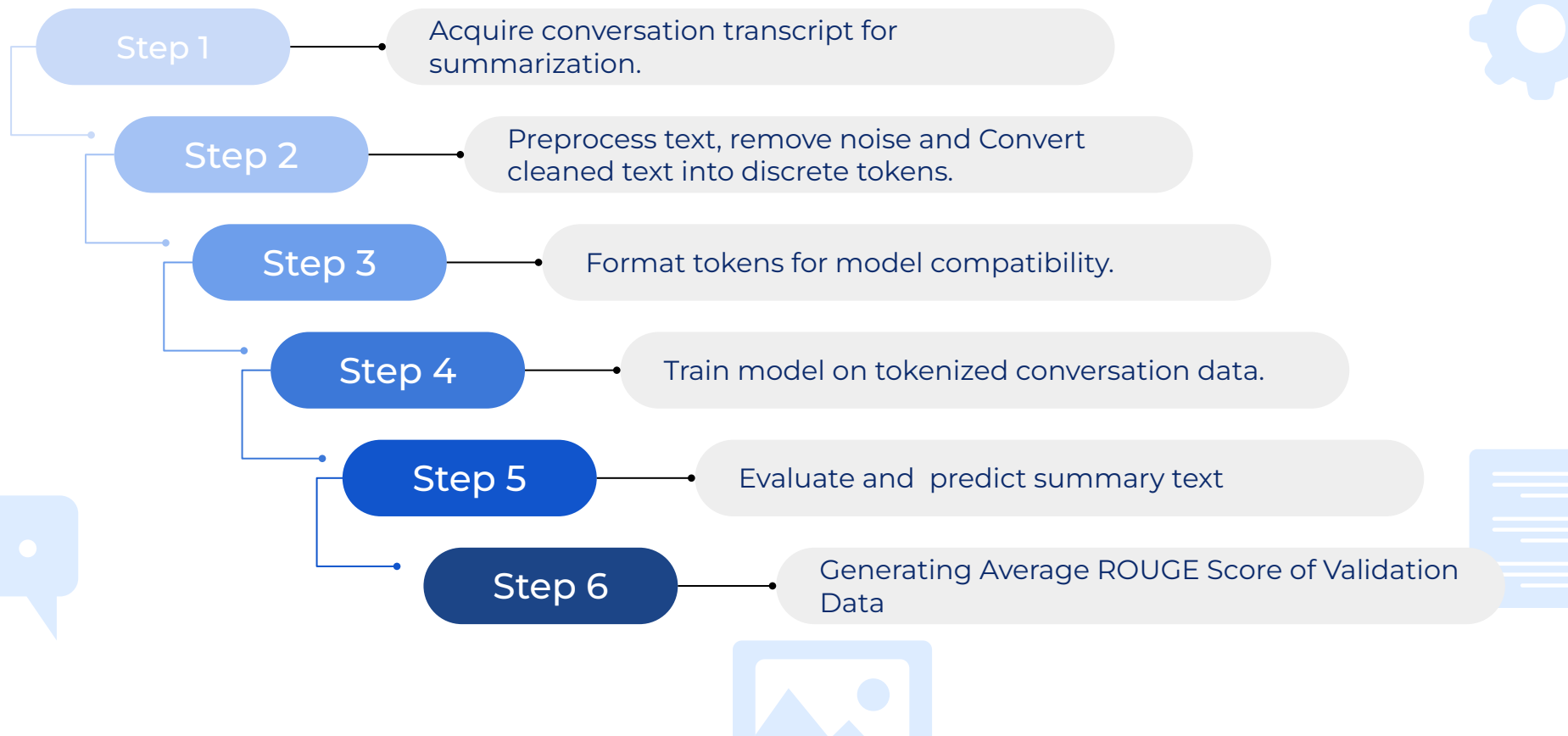
ROUGE-L:
Precision: 0.12
Recall: 0.24
F1-Score: 0.16

1. The **ROUGE-1 scores**
 - Summary captures some key concepts but misses significant details.
 2. The **ROUGE-2**
 - Scores are notably low
 - indicates the summary struggles reproduce detailed elements
 3. The **ROUGE-L**
 - Scores suggests a limited ability in reflecting the overall structure.
 - Summary is not effective in maintaining the flow.
- 
- 



Abstractive Summary

Flow of Abstractive Approach T5 & BART



T5 RAW & T5 Fine Tuned PERFORMANCE

T5 RAW

ROUGE-1:

Precision: 0.15
Recall: 0.28
F1-Score: 0.19

ROUGE-2:

Precision: 0.06
Recall: 0.12
F1-Score: 0.07

ROUGE-L:

Precision: 0.15
Recall: 0.28
F1-Score: 0.19

T5 Fine Tuned

ROUGE-1:

Precision: 0.36
Recall: 0.47
F1-Score: 0.39

ROUGE-2:

Precision: 0.12
Recall: 0.19
F1-Score: 0.14

ROUGE-L:

Precision: 0.33
Recall: 0.44
F1-Score: 0.37

- Fine-tuning has notably improved T5's performance, **doubling or more than doubling precision, recall, and F1-scores across all ROUGE metrics.**
- **The fine-tuned T5 model shows a substantial increase in recall for ROUGE-1 and ROUGE-L,** indicating a better coverage of the reference summary content.
- Despite improvements, **ROUGE-2 scores remain relatively low even after fine-tuning,** suggesting that capturing bigram relationships is still a challenge for the T5 model.

BART RAW & BART Fine Tuned PERFORMANCE

BART RAW

ROUGE-1:

Precision: 0.11
Recall: 0.44
F1-Score: 0.18

ROUGE-2:

Precision: 0.03
Recall: 0.16
F1-Score: 0.05

ROUGE-L:

Precision: 0.10
Recall: 0.42
F1-Score: 0.17

BART Fine Tuned

ROUGE-1:

Precision: 0.34
Recall: 0.51
F1-Score: 0.40

ROUGE-2:

Precision: 0.13
Recall: 0.20
F1-Score: 0.15

ROUGE-L:

Precision: 0.31
Recall: 0.46
F1-Score: 0.36

- Fine-tuning the BART model results in **considerable improvements in precision, recall, and F1-score** for summarization, indicating enhanced model accuracy.
- The consistently **higher recall** compared to precision for both raw and fine-tuned models suggests that the summaries are more exhaustive but potentially include some irrelevant information.
- The **lower scores for ROUGE-2** compared to ROUGE-1 and ROUGE-L in both models highlight the difficulty of accurately matching two-word phrases, although fine-tuning shows a marked improvement in this area.

GPT - 4

ChatGPT-4 is the most advanced multimodal AI from OpenAI that outperforms T5 and BART with superior text understanding, image processing, and contextually aware responses.

ChatGPT - 4

ROUGE-1:

Precision: 0.28

Recall: 0.30

F1-Score: 0.28

ROUGE-2:

Precision: 0.08

Recall: 0.09

F1-Score: 0.08

ROUGE-L:

Precision: 0.24

Recall: 0.27

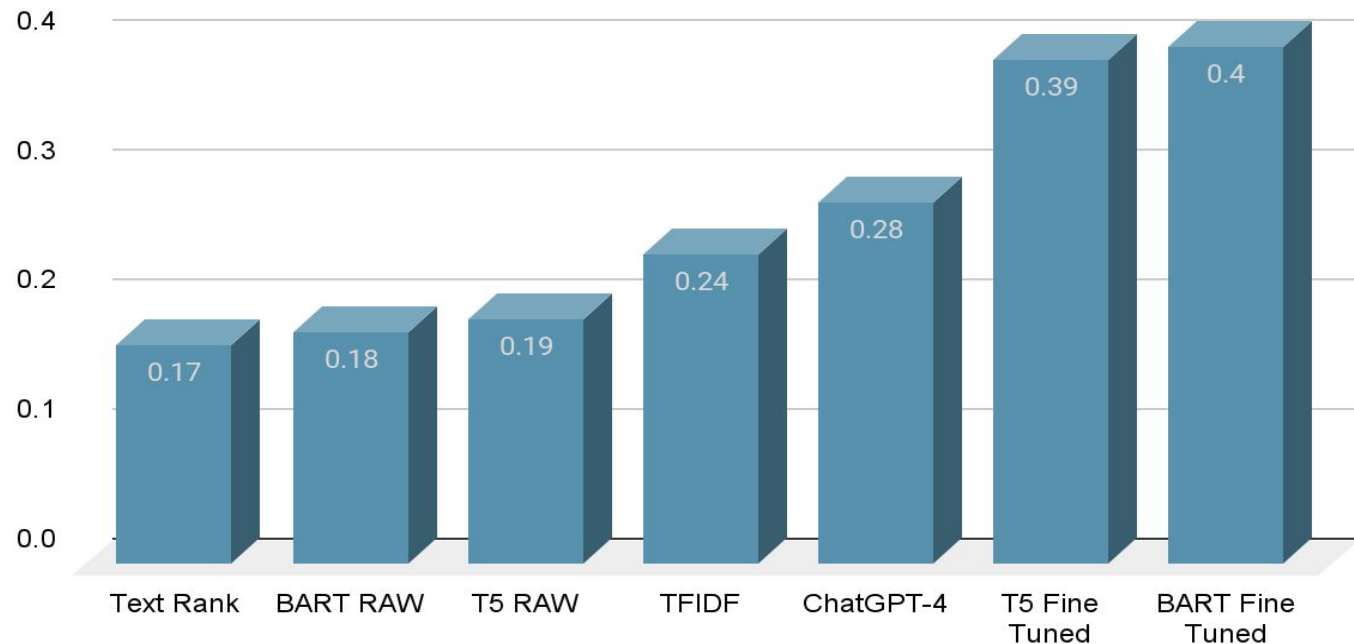
F1-Score: 0.25

- **ChatGPT-4 outperforms BART RAW and T5 RAW** in ROUGE-1 and ROUGE-L precision, indicating more accurate summary content matching with reference summaries.
- **Despite its advancements, ChatGPT-4's recall scores in ROUGE-1 and ROUGE-L are closer to T5 RAW**, suggesting a moderate enhancement in capturing the entirety of reference content compared to BART RAW.

Final Prompt: Summarize the given conversation based on topics discussed, tone and mood of it in 130 to 150 characters.

ROUGE Score Visuals

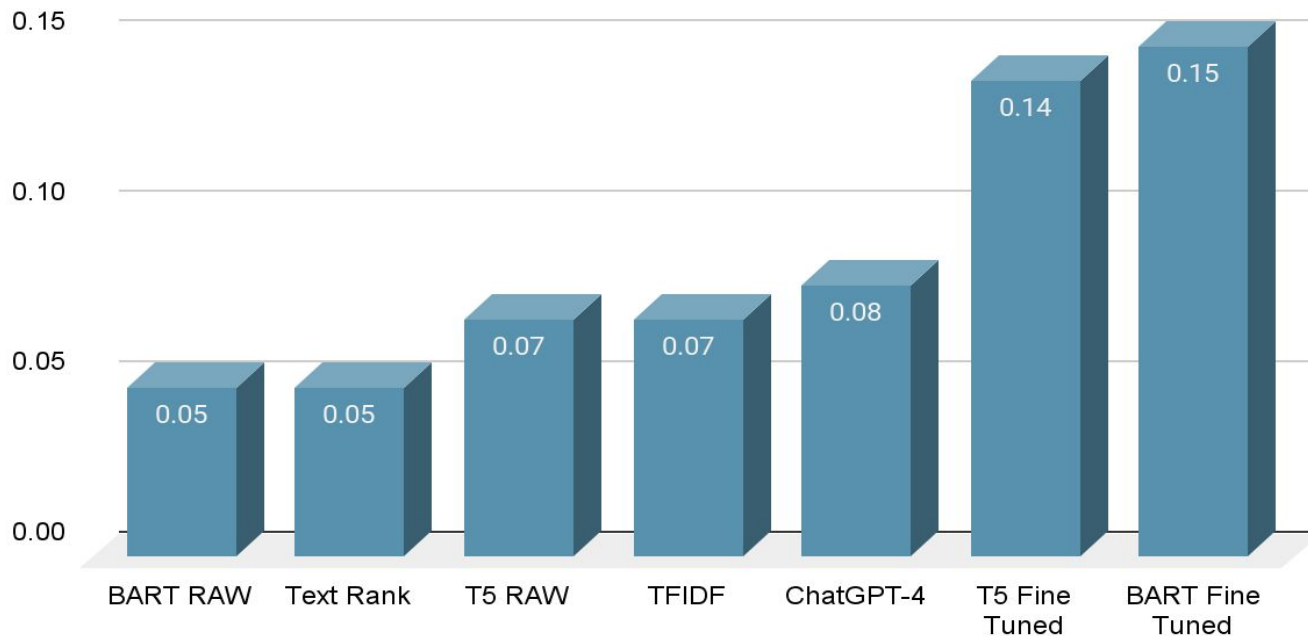
ROUGE 1 F1 Score



Since good summaries & headings can be written differently, Rouge scores around 50 are considered excellent results.

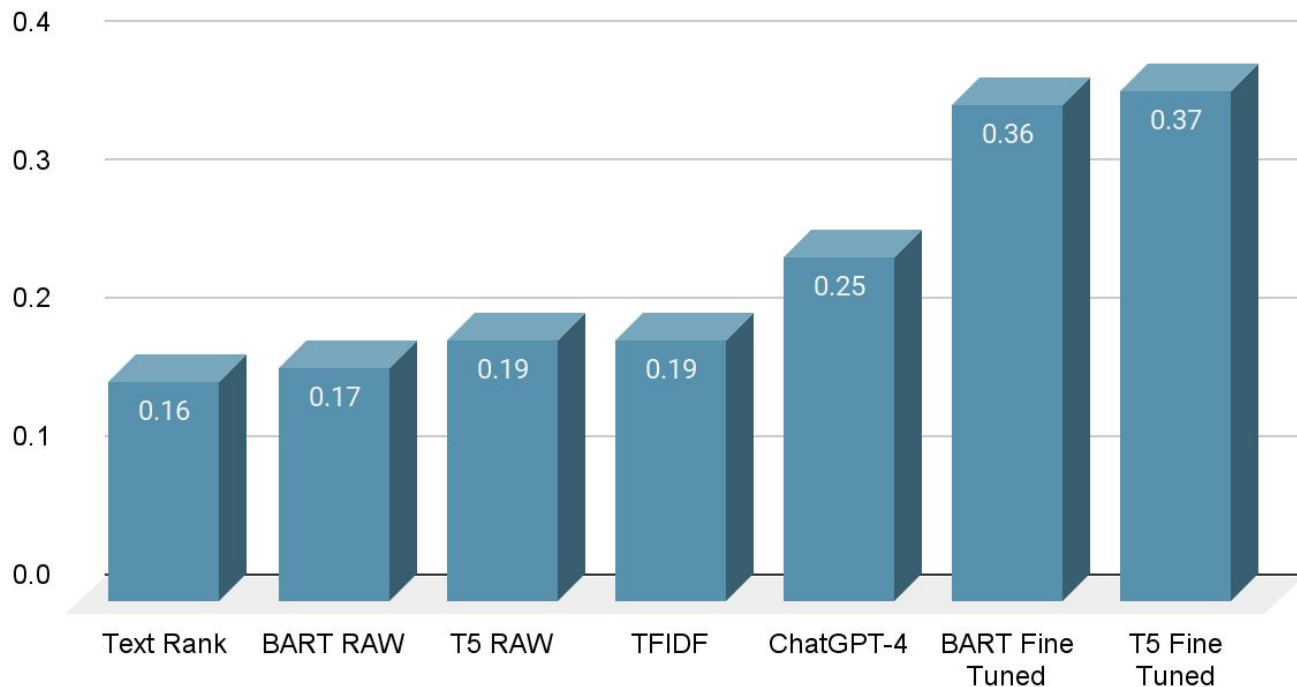
ROUGE Score Visuals

ROUGE R2 F1 Score



ROUGE Score Visuals

ROUGE L F1 Score





Other Performance Metrics



1. BLEU (Bilingual Evaluation Understudy)

Purpose: BLEU is primarily used for evaluating the quality of text which has been machine-translated from one language to another.

How It Works: It compares the machine-generated text to one or more reference texts. BLEU looks at the overlap of n-grams (phrases of n words) between the generated and reference texts.

2. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

Purpose: METEOR is another metric for evaluating machine translation, designed to address some of the shortcomings of BLEU.

How It Works: It compares the generated text with reference texts, focusing on unigrams. METEOR considers not only the exact word matches but also stemmed versions and synonyms, and it incorporates a measure of word order into the evaluation.



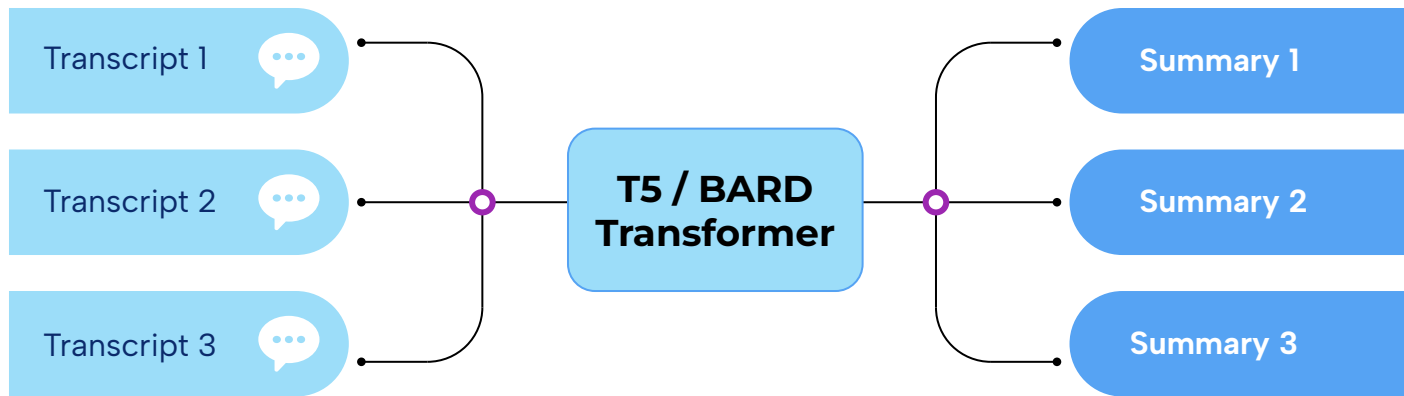
3. Manual Testing: Validating with the Domain experts



Demonstration

Hosted on : **Streamlit**

Link : <https://conversation-summarizer.streamlit.app/>





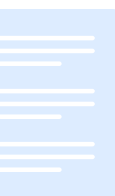

Acknowledgments

**Special Thanks to Professor Ameeta
and TA Bahareh.
Also, All the team members.**





Challenges

1. Infrastructure Limitations while development and deployments (GPU and RAM)
 2. Fine Tuning the T5 and BART with high epoch and choosing perfect batch sizes
 3. Training the models on various dataset to make it generalised
- 
- 



Ethical Considerations



- Removal of PII from the dataset maintaining compliance with other data protection laws like GDPR.
- Confidentiality in Health center or similar transcripts
- Consent and authorized use of training dataset as it can be sensitive and may contain personal information.



[illegible]