**VLG Project Report**

-By Vidhi Gupta

Electrical Engineering Sophomore

vidhi\_g@ee.iitr.ac.in

**Title: Email Summarizer**

# **Project Aim:**

Emails are a ubiquitous communication medium in professional settings, often containing extensive information. Efficiently summarizing email content can significantly enhance productivity. The goal of this project is to implement an email summarization system using the BERT(Bidirectional Encoder Representations from Transformers) and BART (Bidirectional and Auto-Regressive Transformers) models. Also, Compare the results with the transformer’s pipeline.

## 

## **Brief Walkthrough:**

* Loading of dataset
* Pre-processing of the cleaning
* Implementation of the BART Model
* Saving and Real-time implementation
* Implementation of transformer pipeline
* Comparing the results
* Real-time testing of the model

**Tech Used:**

Following are the libraries in Python 3.11.1 that were used in this project:

* PyTorch
* Python Standard Library (Modules like pandas, io, etc.)
* Google BERT
* BART
* Transformers

Following are the Environments and Apps we used in this project:

* Google Colab

## 

## **Working:**

Dataset Handling:

* Dataset used: <https://www.kaggle.com/code/xokent/email-thread-summary-nlp>

(publicly available)

* Specifications: **5000** emails originally; I used only **2000** Emails for computational ease.
* The file format is **JSON,** which was handled using **Pandas**.
* Data Preprocessing: Text cleaning on the 'body' column of a dataset, removing numbers, currency signs, and punctuation.

The approach I used:

* Initially I took inspiration from <https://www.sciencedirect.com/science/article/pii/S2665963823000805/pdfft?md5=e9c43b2d6d7459d6e7daded559504a79&pid=1-s2.0-S2665963823000805-main.pdf>.
* The main idea was to use a clustering algorithm to obtain different clusters representing all the semantic information for that cluster. Then, generating big vectors from these vectors.
* Then, assign sentence ranks to the sentences of emails according to different surface-level features. The features that we can use here are sentence position, Frequency(TF-IDF), proper nouns, and cosine similarity.
* As I researched further, I learned that Transformer models offer better performance, scalability, and parallelization capabilities than RNNs for email summarization. Their ability to capture long-range dependencies, use attention mechanisms, and scale well with model size and computational resources make them the preferred choice.
* Then, I shifted my approach to implementing and testing different Transformers for the best performance.

Model Used:

1. BERT: Bidirectional Encoder Representations from Transformers

* Used Hugging Face Transformers library to load a pre-trained **BERT model and tokenizer** for **sequence classification** to generate summaries for a subset of email data.
* The results and inferences generated are **not idea**l, as they simply select **the top 5 sentences based on their logit scores**. This approach does not necessarily produce meaningful or coherent summaries. The output sentences are just a subsequence of the input email body, and they may **not capture the main idea** or provide a comprehensive summary of the email content.

1. BART: Bidirectional and Auto-Regressive Transformers

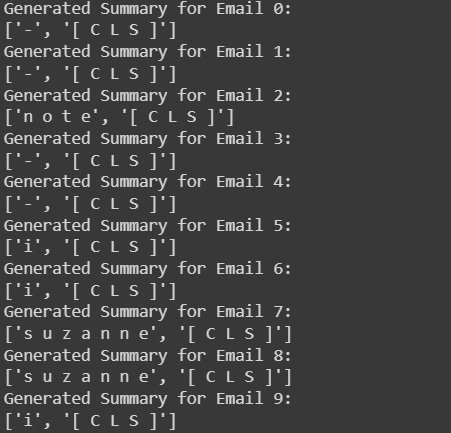
* Used Hugging Face Transformers library to load a **pre-trained BART model** and tokenizer for **conditional generation** to generate summaries for a subset of email data.
* The generated summaries for the emails are **primarily coherent** and **capture the main ideas** of the emails. The BART model, which is used for conditional generation, has provided better summaries compared to the previous example with BERT. However, the quality of the summaries may still vary depending on the content and structure of the input emails.

1. Transformer’s Pipeline:

* used the Hugging Face Transformers library's pipeline feature to perform text summarization. This approach simplifies using a pre-trained model for summarization and allows to quickly generate summaries for your input text.
* The generated summaries generally exhibit good quality by succinctly capturing the main points of the original emails. The content retention is notable, with summaries providing condensed yet relevant information, such as key dates and terms.

## **Results:**

Results using BERT:



Email: Please move the file I sent you from the Testing tab to the West directory under the Pipeline tab. I have updated it in the Testing tab so discard the version i emailed. Thanks.

Results using BART:



Results using Transformer’s Pipeline:

