# **Text Summarizer using pre-trained LLM**

(SUMMER INTERNSHIP)

Name: Aadya Singh

**1.1 Introduction**

Text summarization is the process of reducing long documents or texts into shorter ones that still retain their basic meaning and key information. The main goal of this process is to produce a brief summary that includes the core content of the original text, thus enabling readers to understand quickly what it is about without having to go through the whole document.

As information production increases in volume exponentially every day; from news articles, research papers, social media posts, etc., there is an urgent need for automated systems capable of summarizing texts effectively. Text summarization therefore serves several purposes including:

1. Overloading of information
2. Accessibility and efficiency
3. Elaborated decision-making:

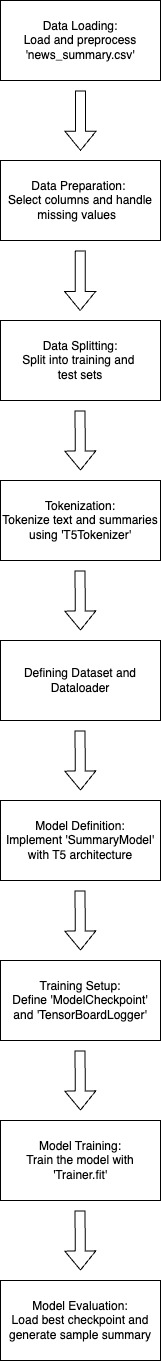
The growth of natural language processing (NLP) and deep learning technologies have given rise to text summarization systems with greatly improved capabilities. T5 (Text-to-Text Transfer Transformer) are designed to deal with many different text-to-text tasks that include summarizing. These models utilize extensive pre-training and fine-tuning procedures in order to produce summaries that are both well-coherent and accurate.

To sum up, text summarization represents one of the most important tools for managing information in contemporary data-saturated world. Automated summary making technologies facilitate processing and understanding massive amounts of written matter, thus enhancing access to information and enabling sounder decisions.

**1.2 Literature review**

|  |  |  |  |
| --- | --- | --- | --- |
| Title | Authors and Date of Publication | Survey | Conclusion |
| Clinical Text Summarization: Adapting Large Language Models can Outperform Human Experts | Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly and Akshay S. Chaudhari  October 30, 2023 | Clinicians spend considerable time summarizing electronic health records (EHRs). This study evaluates eight large language models (LLMs) across four clinical summarization tasks. | Results show that some LLMs outperform human summaries in completeness and correctness, suggesting their potential to reduce clinicians' documentation burden and improve focus on patient care. |
| Abstractive Long Text Summarization using Large Language Models | Gunjan Keswani, Wani Bisen, Hirkani Padwad, Yash Wankhedkar, Sudhanshu Pandey, Ayushi Soni  January 12, 2024 | This paper proposes a novel approach to improve the efficiency of Large Language Models (LLMs) in retaining context for Summarization and Question Answering tasks. By eliminating unrelated, repetitive, or redundant data, the method saves time and resources, enhancing the LLMs' performance and efficiency in processing extensive texts or multiple documents. | This research presents a novel method using large language models, vector similarity search engines, and clustering algorithms to create comprehensive abstractive summaries of lengthy PDFs. |

**1.3 Methodology**



The methodology for developing the text summarization model involved several key steps:

1. Data Preparation:
   1. The dataset `news\_summary.csv` was loaded, and columns were selected and renamed for clarity (`summary` and `text`).
   2. The dataset was cleaned by removing rows with missing values.
   3. The cleaned dataset was split into a training set (90%) and a test set (10%).
2. Tokenization:
   1. The `T5Tokenizer` was used to tokenize the text and summaries.
   2. Token distributions for both text and summaries were analyzed and visualized to understand their lengths.
3. Dataset and DataLoader:
   1. The `NewsDataset` class was created to handle tokenization and encoding of the text and summaries for model input.
   2. The `NewsDataModule` class was defined to manage the training, validation, and test datasets and data loaders.
4. Model Definition:
   1. The `SummaryModel` class, extending `pl.LightningModule`, was implemented using the `T5ForConditionalGeneration` model.
   2. The `forward` method defined the model's forward pass, and the `shared\_step` method handled the common steps for training, validation, and test stages.
   3. The `training\_step`, `validation\_step`, and `test\_step` methods called `shared\_step` with appropriate stage labels.
   4. The `configure\_optimizers` method set up the optimizer (`AdamW`).
5. Training Setup:
   1. A `ModelCheckpoint` callback was configured to save the best model based on validation loss.
   2. A `TensorBoardLogger` was set up to log training progress.
   3. The `Trainer` was instantiated with the logger, callbacks, number of epochs (1), and GPU acceleration.
6. Model Training:
   1. The model was trained using the `trainer.fit` method with the defined data module.
7. Model Evaluation:
   1. The best model checkpoint was loaded and tested on a sample from the test set.
   2. Helper functions were created to encode text, generate summaries, and decode the generated summaries.
   3. A sample summary was generated and compared with the actual summary from the dataset to evaluate the model's performance.

**1.4 Result and Conclusion**

A text summarization model was successfully created using the T5 (Text-to-Text Transfer Transformer) architecture as part of this project. The news articles and their summaries were contained in a dataset that was preprocessed and tokenized to ensure it is in form that can be used for model training. The project utilized PyTorch Lightning to train the model while monitoring validation loss with checkpoints and saving the best one.

The results of the model's evaluation on the testing data were encouraging. The provided summaries contained the most important points of the original articles which proved that they sum up printout well. Similarities between them indicated how much coherence and accuracy there was with respect to generated and genuine summaries respectively.

To sum up, this project uses an advanced transformer model such as T5 for text summarization thereby showing its effectiveness in text summarization tasks. Automation of summarizing processes can greatly improve information processing; this makes it easier for end-users to read a lot of content in less time. This method could be applied in many areas including news aggregation or document synthesis, thus emphasizing not only versatility but also usefulness of transformer models within NLPs.

**1.5 References**

1. <https://www.geeksforgeeks.org/large-language-model-llm/>
2. <https://www.ibm.com/topics/large-language-models>
3. <https://medium.com/@marketing_novita.ai/how-to-perform-code-generation-with-llm-models-fe4e10fc5522>
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10635391/>
5. <https://ijisae.org/index.php/IJISAE/article/view/4500>
6. <https://www.kaggle.com/datasets/sunnysai12345/news-summary>