**NAME:**

**MUKARRAM MUSHTAQ**

**ROLL NO:**

**FA22-BCS-094**

# **INTRODUCTION**

The need for efficient text summarization has grown quickly, especially in industries like digital news and media. With the continuous surge of information, the ability to create clear and concise summaries of lengthy documents has become essential for both readers and businesses. The field of text summarization has advanced considerably since [1] early work, where he introduced one of the first methods for automatic summarization based on word frequency in his paper "The Automatic Creation of Literature Abstracts." This pioneering model laid the foundation for future developments in the area.

Over time, advancements in summarization techniques have resulted in more sophisticated models. In 1989, [2], a major contributor to the development of text retrieval systems, significantly advanced the field with his paper "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer." Salton’s work introduced vector space models, employing statistical methods to enhance the effectiveness of text processing.

By the 2000s, more advanced and automated systems had emerged. [3], in his paper "Automated Text Summarization and the SUMMARIST System," contributed to the creation of SUMMARIST, a system combining statistical and symbolic approaches for automated summarization. Around the same time, [4] introduced the ROUGE evaluation metric in his paper "ROUGE: A Package for Automatic Evaluation of Summaries," which became a widely accepted standard for assessing the quality of generated summaries. ROUGE is now extensively used to evaluate the performance of modern models, including the T5 model fine-tuned in this study.

The field continued to evolve with [5] paper "Centroid-based Summarization of Multiple Documents," which introduced extractive methods for summarizing multiple documents simultaneously. These key studies, along with progress in deep learning, have opened the door for modern models like T5, which can be fine-tuned for greater accuracy and relevance.

This study builds upon these historical advancements by fine-tuning the T5-small model on a reduced version of the CNN/Daily Mail dataset. Its goal is to overcome the limitations posed by smaller datasets, which often lead to less effective summarization. The improvements seen in ROUGE scores through this method show notable progress in the model’s ability to produce accurate, contextually relevant summaries with minimal computational resources. These results emphasize the potential of such models for applications like real-time news aggregation, personalized content delivery, and more.

This study seeks to address the challenges associated with traditional text summarization models, particularly when working with smaller datasets. By fine-tuning the T5-small model using a condensed version of the CNN/Daily Mail dataset, the study aims to enhance the model's performance, as measured by ROUGE scores—a standard metric for evaluating summarization quality. The improvements in ROUGE-1, ROUGE-2, and ROUGE-L scores demonstrate the model’s improved ability to capture key information while keeping computational costs low. These findings hold significant value for the digital technology sector, where real-time summarization, news aggregation, and personalized content delivery are essential to meeting user needs.

1. Luhn, H.P., *The automatic creation of literature abstracts.* IBM Journal of research and development, 1958. **2**(2): p. 159-165.

2. Gerard, S., *Automatic Text Processing: The Transformation.* Analysis, and Retrieval of Information by Computer, 1989.

3. Hovy, E. and C.-Y. Lin. *Automated text summarization and the SUMMARIST system*. in *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*. 1998.

4. Chin-Yew, L. *Rouge: A package for automatic evaluation of summaries*. in *Proceedings of the Workshop on Text Summarization Branches Out, 2004*. 2004.

5. Radev, D.R., et al., *Centroid-based summarization of multiple documents.* Information Processing & Management, 2004. **40**(6): p. 919-938.