Text Summarization Using Transformer Models

In an age of information overload, text summarization emerges as a critical tool. This project explores the power of Transformer-based models, specifically T5 and PEGASUS, to condense vast amounts of text into concise, meaningful summaries.

Name : AHMED HASSAN

Date 12 November 2025

Data Science Buildables

The Challenge: Information Overload

The digital era brings an unprecedented volume of information—news, research, reports—making it impossible for individuals to consume everything. This constant influx leads to:

- Time constraints for comprehensive reading
- Difficulty in extracting core insights rapidly
- Increased cognitive load from excessive content

Our motivation was to combat this by developing an automated summarization system.

Project Goals & Exploration

Our primary objective was to build a system that could:

Understand Key Meaning

Accurately grasp the core message of any given article.

Generate Readable Summaries

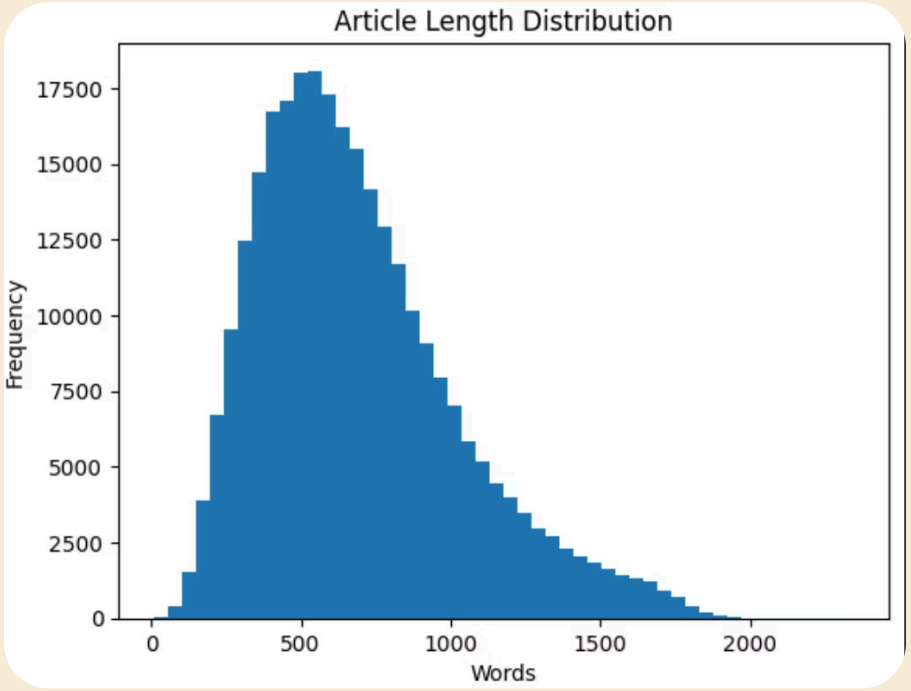
Produce concise, fluent, and coherent summaries automatically.

Save Time & Effort

Enable users to quickly absorb long-form content, significantly reducing reading time.

This project also served as an opportunity to delve deep into Natural Language Processing (NLP) and the intricacies of Transformer architectures, which are foundational to modern AI language systems.

Dataset: CNN/DailyMail — The Benchmark



We utilized the widely recognized **CNN/DailyMail dataset**, a standard benchmark for text summarization. It comprises thousands of news articles from CNN and Daily Mail, each meticulously paired with a human-written summary, known as "highlights."

- **Article:** The complete news story text.
- **Highlights:** The corresponding human-crafted summary.

To optimize training, smaller subsets were selected: **5,000 samples for training** and **500 for validation**. Prior to model training, crucial **data cleaning and Exploratory Data Analysis** (**EDA**) were performed, including checks for null values, duplicates, and article length analysis.

T5-small vs PEGASUS

T5 (**Text-to-Text Transfer Transformer**) is a groundbreaking model that unifies all NLP tasks into a single text-to-text format. For summarization, this means:

> *"summarize: [article text]"*

The model then learns to generate the summary directly as its output. We fine-tuned **T5-small** due to its efficiency and suitability for resource-constrained environments like Google Colab GPUs. Our training settings were:

- **Epochs:** 2
- **Batch size:** 8
- **Learning rate:** 5e-5
- **Optimizer:** AdamW

PEGASUS, developed by Google, is specifically engineered for **abstractive text summarization**. Its strength lies in understanding sentence-level relationships to produce summaries that closely mimic human writing, often generating novel phrases not present in the original text.

During our evaluation, PEGASUS consistently outperformed T5 in terms of **ROUGE scores**, indicating higher quality summaries. However, its significant model size (~2.1 GB) presented deployment challenges, particularly for GitHub hosting and easy integration. This trade-off between performance and practicality led us to a strategic decision for deployment.

Evaluation: ROUGE Metrics

To objectively assess model performance, we employed **ROUGE** (**Recall-Oriented Understudy for Gisting Evaluation**) metrics, which measure the overlap between generated summaries and human-written reference summaries.

ROUGE-1

Measures the overlap of unigrams (single words).

ROUGE-2

Measures the overlap of bigrams (pairs of words).

ROUGE-L

Measures the longest common subsequence, capturing sentence-level structure.

While PEGASUS yielded superior ROUGE scores, **T5-small struck an optimal balance** between summarization quality, inference speed, and model footprint, making its generated summaries clear, concise, and effective in capturing main ideas.

Model	Step	Training Loss	Validation Loss	ROUGE-1	ROUGE-2	ROUGE-L
Pegasus	500	5.640900	6.339778	0.353699	0.156405	0.262721
T5-small	500	1.097200	0.847001	0.239550	0.088880	0.192200

Deployment: Bringing T5-small to Life

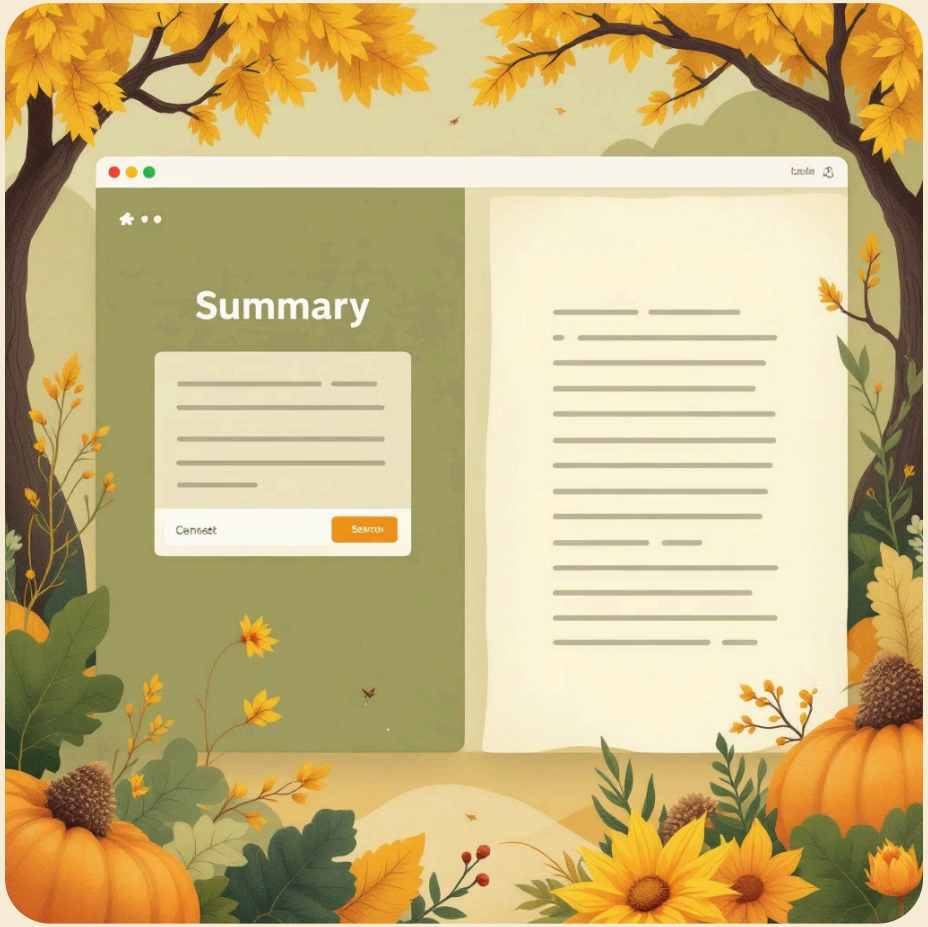
Given the practical considerations of model size and performance balance, the **T5-small model was chosen for deployment**. We leveraged **Flask**, a lightweight Python web framework, to create a user-friendly interface.

- Users can **input any paragraph**.
- The model **instantly generates a concise summary**.

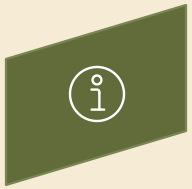
This deployment makes the power of Transformer-based summarization accessible, demonstrating its real-world utility.

GitHub Repository

Live Demo

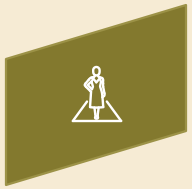


Key Learnings & Future Outlook



Data Preprocessing

Mastered techniques for cleaning and preparing text data for NLP tasks.



Transformer Architectures

Gained deep insights into the mechanics of T5 and PEGASUS models.

Model Evaluation

Understood the critical role and application of metrics like ROUGE.

Deployment Strategy

Learned fine-tuning and deployment of large NLP models, including crucial trade-offs.

Size vs. Performance

Recognized the balance between model size and performance in practical applications.

Conclusion: The Power of Automated Summarization

This project successfully demonstrated the transformative power of AI in automating text summarization using state-of-the-art Transformer models. While PEGASUS showcased superior accuracy, T5-small proved to be the more practical choice for deployment due to its efficiency and smaller footprint.

Future Work:

- Expansion to **multi-language summarization**.
- Development of **topic-based summaries** for specialized content.
- Exploration of **more advanced hardware** for larger models.

Text summarization is an evolving field within NLP, promising even more sophisticated and impactful applications in managing and understanding vast information landscapes.