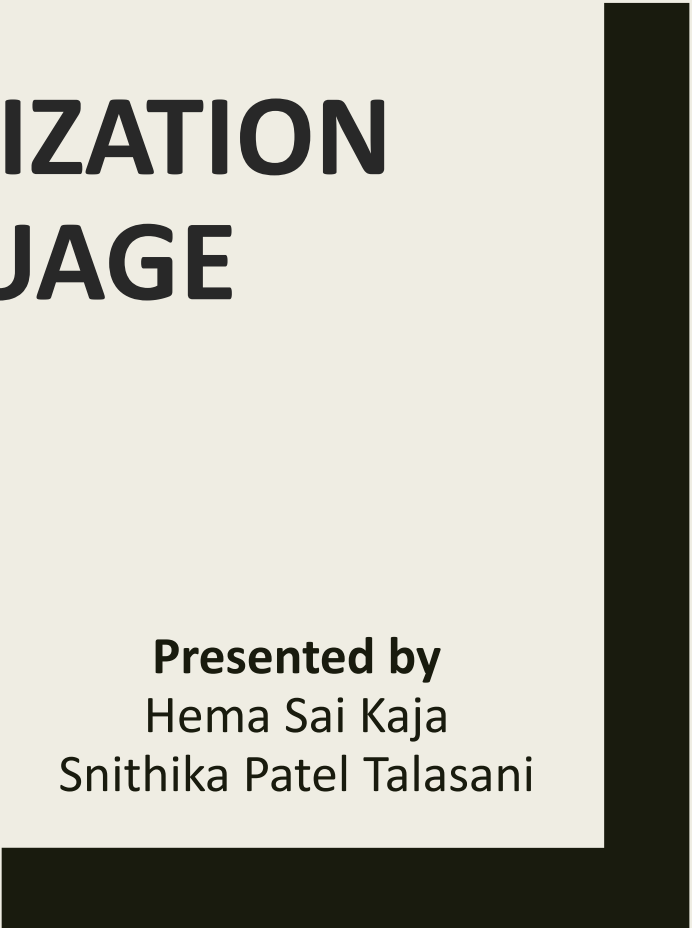


# **EFFICIENT TEXT SUMMARIZATION WITH NATURAL LANGUAGE PROCESSING**

**Presented by**  
Hema Sai Kaja  
Snithika Patel Talasani



# Contents

Introduction

Task Objective

Datasets

Evaluation Metrics

Models and Methods

Comparison between Models

Conclusion/Results

# Introduction

- Text summarization is the creation of a short, accurate, and fluent summary of a longer text.
- This method is greatly needed to address the ever-growing amount of text data available online.
- The use of text summarization makes it easier for the users to collect the important data from huge information.
- The whole idea of text summarization is to collect the necessary summary from a large amount of data.

# Task Objective

- Develop an advanced text summarization system using NLP techniques.
- Improve the efficiency and accuracy of text summarization algorithms.
- Explore novel approaches for generating concise and informative summaries from large text documents.

# Datasets

- BBC News Summary([BBC News Summary \(kaggle.com\)](https://www.kaggle.com/datasets/bbc-news-summary)): The dataset comprises political news articles from BBC spanning 2004 to 2005, housed in the News Articles folder. Each article is accompanied by summaries in the Summary folder.
- These summaries offer extractive text summarization, where sentences from the articles themselves are utilized as summaries. This method, favored by researchers in automatic text summarization, involves scoring sentences and selecting those with the highest scores as summaries.
- While this approach is simpler and less computationally intensive, it may sacrifice smoothness and coherence in the summaries, occasionally resulting in a lack of readability due to disconnected adjacent sentences.

# Evaluation Metrics

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** This metric measures the overlap between the generated summary and the reference summary in terms of n-gram overlap and sequence similarity. It provides insights into how well the generated summary captures the essential information from the source document.
- BLEU (Bilingual Evaluation Understudy):** This metric evaluates the precision of the generated summary by comparing it with reference summaries based on n-gram matches. It gives an indication of how well the generated summary matches the reference summaries.

# Models and Methods

Various NLP models and techniques are employed, including preprocessing of texts, training a Word2Vec model, and implementing an LSTM neural network for generating summaries.

## **Word2Vec Model**

The code trains a Word2Vec model to generate word embeddings from the preprocessed training texts. These embeddings serve as input to the LSTM model and are useful for representing words as numerical vectors.

## **LSTM-Based Summarization Model**

A simple LSTM-based model (LSTM Model) is defined with:

1. An embedding layer to convert word indices into embeddings.
2. A Long Short-Term Memory (LSTM) layer for sequential data processing.
3. A linear output layer to produce the model's predictions.

# Models and Methods

## **GPT2LMHeadMode**

- GPT-2 is a powerful language model by OpenAI designed for generating text. It uses the transformer architecture and a causal language modeling approach, predicting the next word based on the previous ones. This allows GPT-2 to generate coherent, contextually relevant text, making it useful for various natural language processing tasks, including text summarization.
- Its performance in summarization depends on factors like data quality, task formulation, and fine-tuning strategies. However, GPT-2 is known for producing natural-sounding summaries and can be adapted to a variety of contexts.



# Comparison between Models

```
# Evaluation
rouge_scores = calculate_rouge(validation_summaries, generated_summaries)
bleu_score = calculate_bleu(validation_summaries, generated_summaries)

from tabulate import tabulate
# Print results
table_data = [
    ["ROUGE-1", rouge_scores['rouge1'].fmeasure],
    ["ROUGE-2", rouge_scores['rouge2'].fmeasure],
    ["ROUGE-L", rouge_scores['rougeL'].fmeasure]
]
print(tabulate(table_data, headers=["Metric", "Score"]))

Metric      Score
-----
ROUGE-1     0.338822
ROUGE-2     0.023963
ROUGE-L     0.0849856

print("BLEU score:", bleu_score)

BLEU score: 0.0038161839957274062
```

Both metrics are important, but for text summarization, higher ROUGE scores are generally more indicative of a successful summary, while higher BLEU scores are more commonly used in machine translation.

# Conclusion/Results

- Long-Term Dependency:** LSTMs may struggle with capturing long-term dependencies in text, leading to loss of important context and impacting summary quality, particularly for longer texts.
- Limited Context Understanding:** Due to the fixed-size hidden state, LSTMs may have difficulty understanding complex contextual relationships, resulting in less coherent summaries.
- Data Efficiency:** LSTMs require large amounts of annotated data for training, making them challenging to use in domains with limited resources or specialized content.
- Inference Speed:** LSTMs are computationally intensive, with slower inference times due to the sequential processing of tokens, which is problematic for real-time applications.
- Evaluation Metrics:** Traditional metrics like ROUGE may not effectively capture summary quality in terms of coherence and fluency, leading to challenges in evaluating LSTM-based summaries.

These limitations suggest that while LSTMs have potential, they may not always be the optimal choice for text summarization, especially in cases requiring high coherence, efficiency, or large-scale datasets.

# References

1. D. Yadav, J. Desai, and A. K. Yadav, "Automatic Text Summarization Methods: A Comprehensive Review," 2015, [Online]. Available: 20mcs105@nith.ac.in, dsy99@rediffmail.com, ayadav@nith.ac.i.
2. V. K. Gogulamudi "Text Summarizing Using NLP," in *Recent Trends in Intensive Computing*, Dec. 2021, doi: 10.3233/APC210179.
3. I. Z. Khan, A. A. Sheikh, and U. Sinha, "Graph Neural Network and NER-Based Text Summarization," in *Recent Trends in Intensive Computing*, 2019.