Indonesian Text Summarization

Liputan 6 Dataset

NLP-A

18 Oktober 2024

Meet the Team

Kaira Milani Fitria

Team Member

Kevin Sutanto

Team Member

M. Erwinsyah H

Team Member

Muhammad Darussalam

Team Member

Contribution

Research

- Background
- Hipotesis
- Problem Identification
- Project Purposes
- Theoretical Basis

Data Preparation

- Data Conversion
- Data Cleaning
- Data Normalization
- Data Sampling

Exploratory Data Analysis

- Text Length
- Summary Coverage
- Overlap Distribution
- Distribution & Compression Ratio

Modelling BERT

- BERT for Abstractive & Extractive Summarization
- T5 for Abstractive & Extractive Summarization

Benchmarking Model

- ROUGE Metrics for Extractive
- BERTScore Metrics for Abstractive
- Model Documentations

Conclusion

- Conclusion of Project
- Future Improvements
- Recommendations
- Code + Result Documentations

Introduction

News Access

80% pengguna di Indonesia mengandalkan media online sebagai sumber berita utama antara 2021-2023

Source: Reuters Digital News Report 2023

News Media

Hingga Januari 2023, terdapat 1.711 perusahaan media yang telah terverifikasi, di mana 902 di antaranya adalah perusahaan media digital

Source: Data Dewan Pers Indonesia, 2023

Challenges

Banyaknya konten berita menciptakan kebutuhan teknologi AI yang mampu secara otomatis merangkum artikel berita agar lebih mudah dipahami

Automatic Text Summarization

Extractive Summarization

Memilih kalimat-kalimat penting langsung dari teks.

Abstractive Summarization

Menghasilkan rangkuman singkat dengan kalimat baru.

Model Used

T5
Text-to-Text Transfer Transformer

Raffel et al. (2020)

BERT
Bidirectional Encoder Representations from Transformers

Devlin et al. (2019)

Problem Identification

Jumlah Berita yang Banyak

Keterbatasan Waktu

Variasi Kualitas Berita

Kebutuhan untuk Efisiensi

Penurunan Minat Baca

Hipotesis

Model NLP dalam Automatic Text Summarization dapat menjadi solusi dalam permasalahan rangkuman berita pada media online.

Model Automatic Text Summarization yang akan dikembangkan pada permasalahan rangkuman berita ini adalah model BERT dan T5.

Metriks Evaluasi performa model yang dikembangkan adalah ROUGE untuk ekstraktif model, BERTScore untuk abstraktif model.

Dataset yang digunakan pada project ini adalah dataset berita Liputan 6 yang telah di collect dan terserdia di Hugging Face.

Data Preparation

Dataset Information

Large-scale Indonesian corpus for Abstractive and Extractive summarization from Liputan6 news portal.

Data	Train	Dev	Test
Canonical	193,883	10,972	10,972
Xtreme	193,883	4,948	3,862

Train : Training set
Dev : Validation set
Test : Testing set

Dataset example:

```
{
  'clean_article': 'Liputan6.com, Ambon: Partai Bulan Bintang wilayah Maluku bertekad membantu pe
  'clean_summary': 'Konflik Ambon telah berlangsung selama tiga tahun. Partai Bulan Bintang wilay
  'extractive_summary': 'Liputan6.com, Ambon: Partai Bulan Bintang wilayah Maluku bertekad membar
  'id': '26408',
  'url': 'https://www.liputan6.com/news/read/26408/pbb-siap-membantu-penyelesaian-konflik-ambon'
}
```

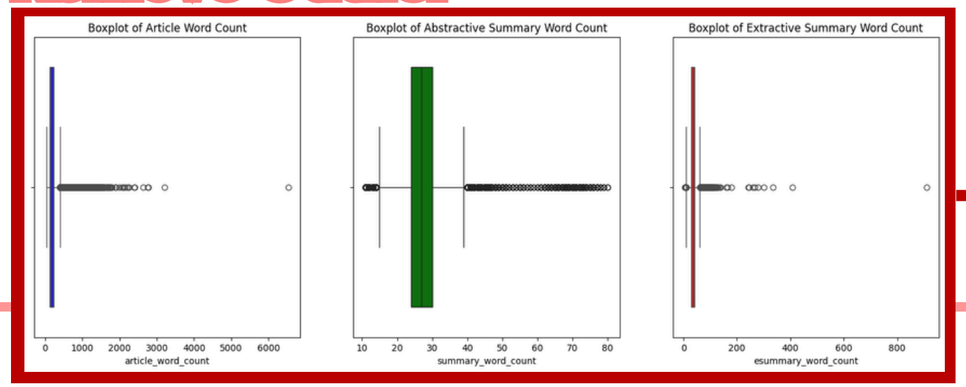
Data Conversion:

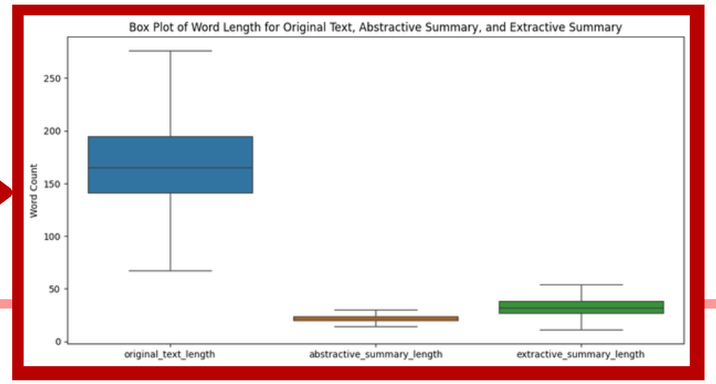
JSON

```
{"id": 296046, "url": "https://www.liputan6.com/news/read/296046/xxxxxxxxxx",
"clean_article": [["text", "text"], ["text", "text"], ["text", "text"], ["text", "text"]],
"clean_summary": [["text", "text", "text", "text"]],
"extractive_summary": [0, 3]}
```

	original_text	abstractive_summary	extractive_summary
0	liputan6 com jakarta sejumlah tokoh masyarakat...	sejumlah tokoh masyarakat riau membantah perny...	liputan6 com jakarta sejumlah tokoh masyarakat...
1	liputan6 com surabaya tokoh madura dan kaliman...	gubernur jawa timur meminta para gubernur se k...	hingga kini ada sekitar 90 ribu pengungsi masu...
2	liputan6 com medan ratusan kepala keluarga nya...	warga binjai timur sumut nyaris bentrok dengan...	liputan6 com medan ratusan kepala keluarga nya...

Remove Outlier





Random Sampling

```
train_df = (29842, 3)
dev_df.shape = (3000, 3)
test_df.shape = (2991, 3)
```

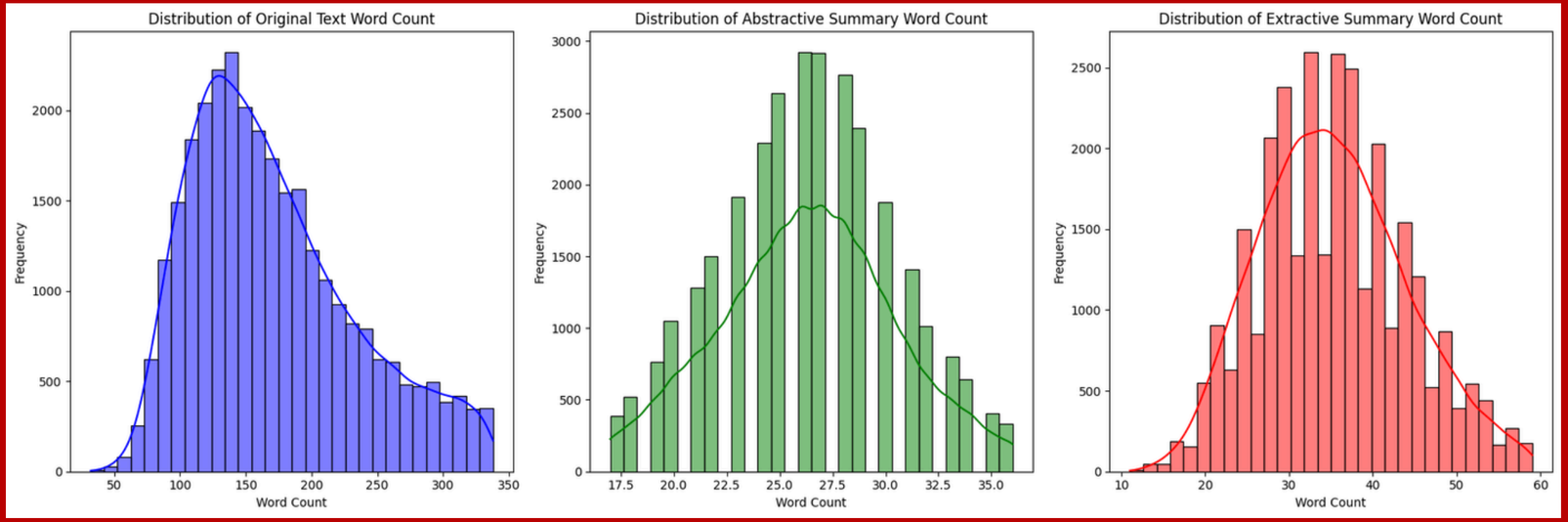
EDA

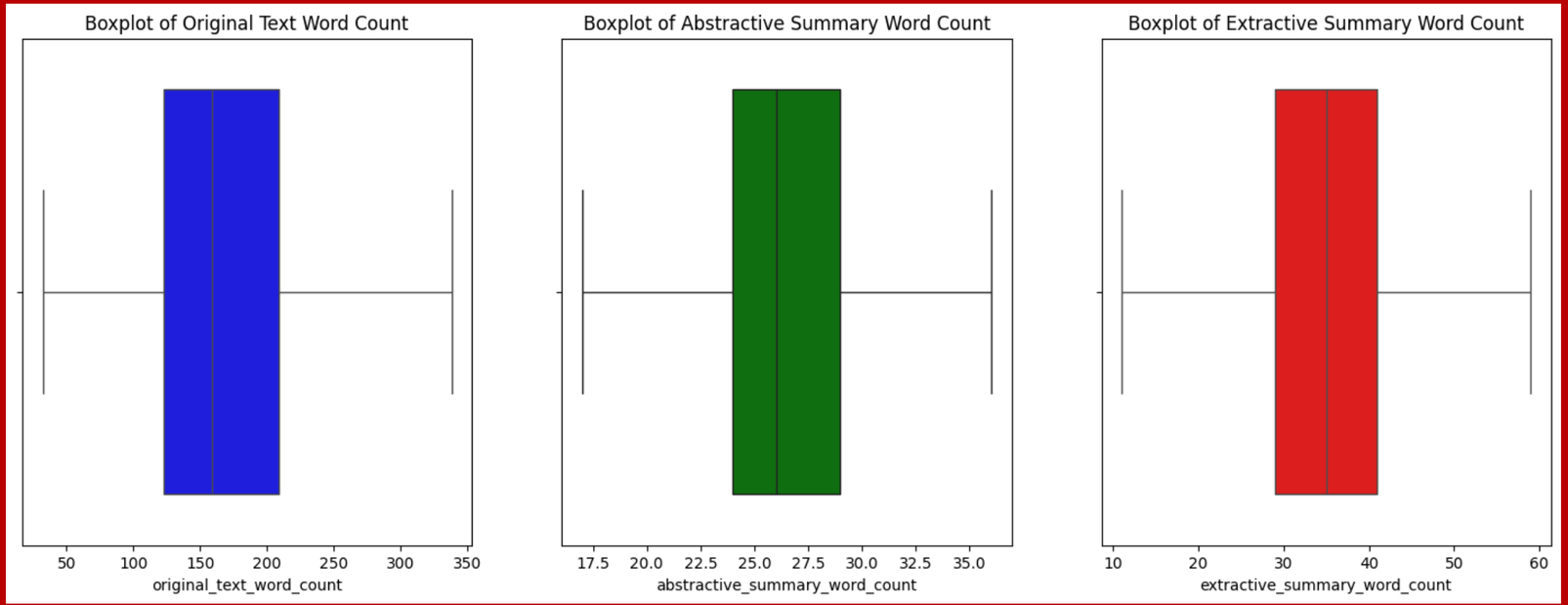
Text Length

```
def word_sentence_count(text):
    words = word_tokenize(text)
    sentences = sent_tokenize(text)
    return len(words), len(sentences)
```

- **Original texts average 150-200 words.**
- **Abstractive summaries are 25-30 words, highly compressed.**
- **Extractive summaries are 30-45 words, preserving more content.**
- **No extreme outliers, indicating clean and consistent data.**

Text Length Analysis





EDA

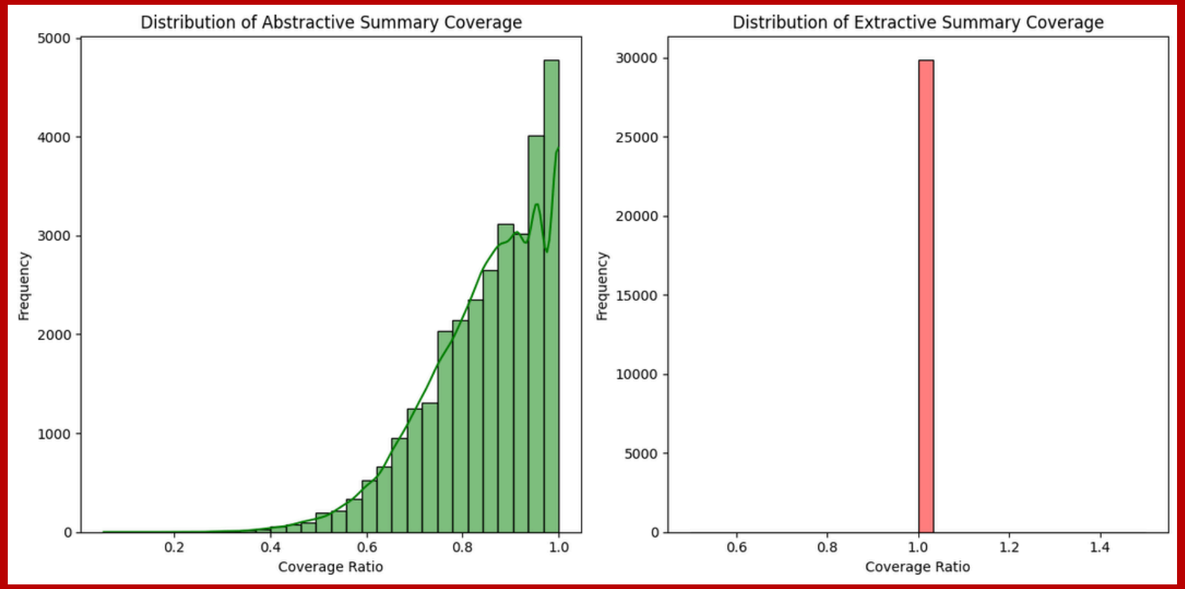
Summary Coverage

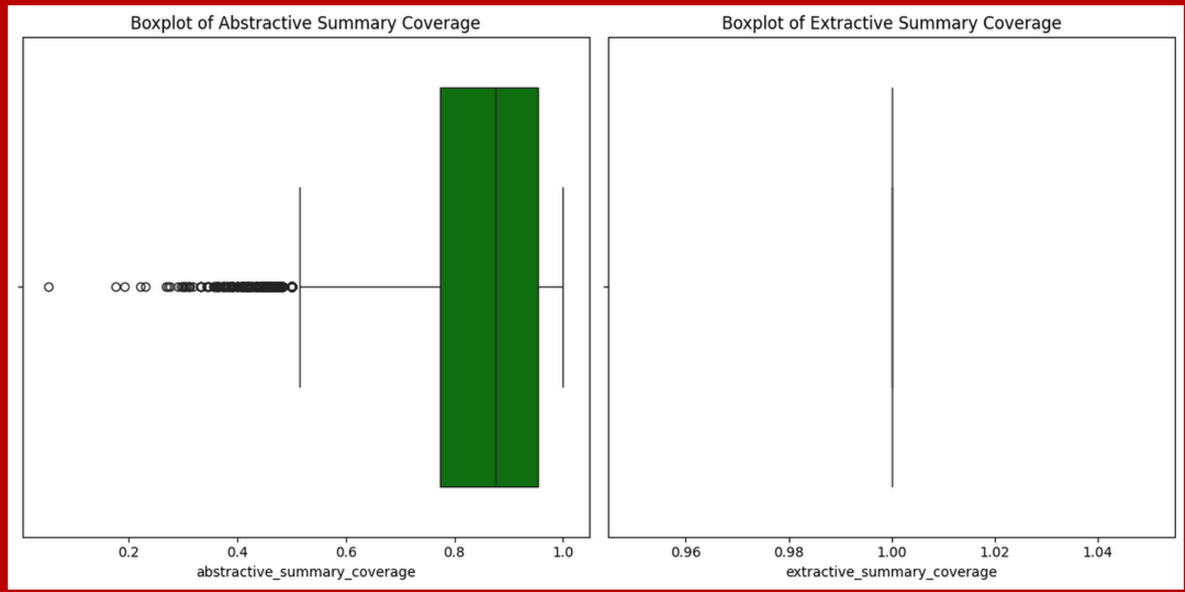
```python
def coverage_ratio(summary, article):
    article_words = set(word_tokenize(article.lower()))
    summary_words = set(word_tokenize(summary.lower()))

    if len(summary_words) == 0:
        return 0
    overlap = summary_words.intersection(article_words)
    return len(overlap) / len(summary_words)
```

- **Abstractive summaries show variable coverage, with some lower-performing cases.**
- **Abstractive coverage is mostly near 1 but varies, with some as low as 0.4 and several outliers below 0.5**
- **Extractive summaries are highly consistent with near-perfect coverage, with no variation.**

Summary Coverage





EDA

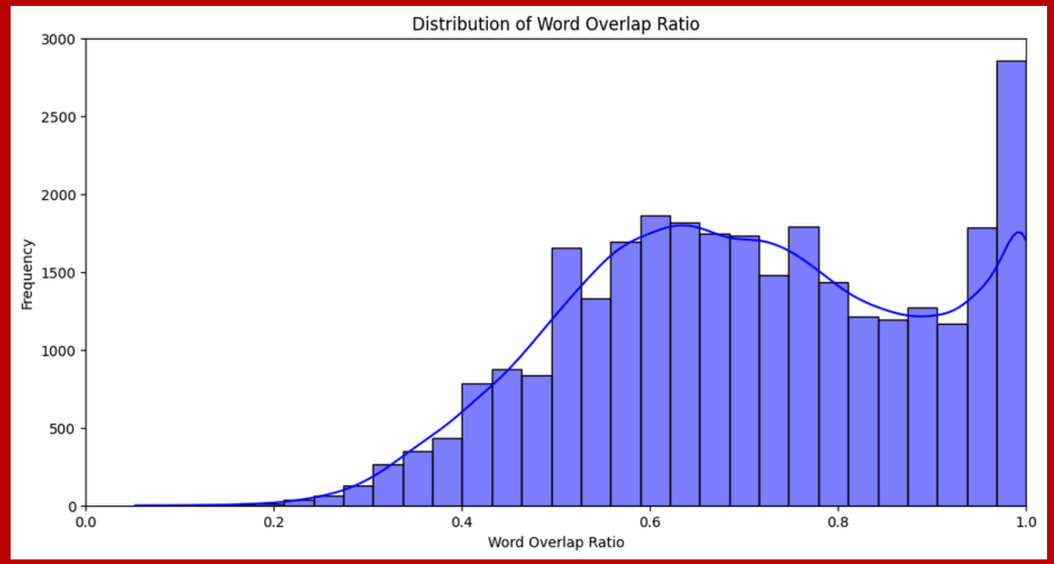
Overlap Distribution

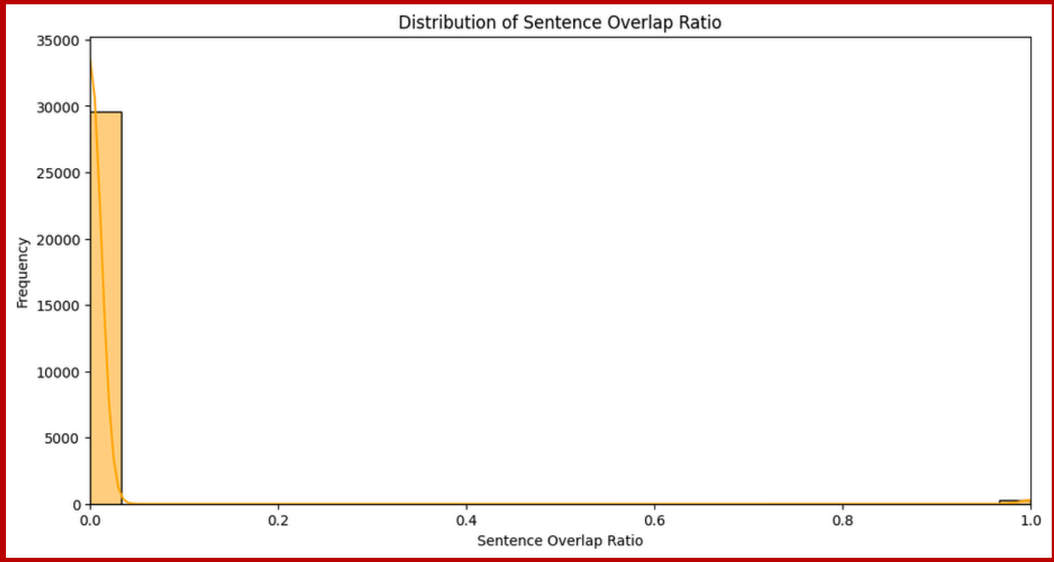
```python
def calculate_word_overlap(abstractive, extractive):
    abstract_words = set(abstractive.split())
    extractive_words = set(extractive.split())
    overlap = abstract_words.intersection(extractive_words)
    return len(overlap), len(abstract_words), len(extractive_words)

def calculate_sentence_overlap(abstractive, extractive):
    abstract_sentences = set(abstractive.split('. '))
    extractive_sentences = set(extractive.split('. '))
    overlap = abstract_sentences.intersection(extractive_sentences)
    return len(overlap), len(abstract_sentences), len(extractive_sentences)
```

- **Most abstractive and extractive summaries have 40-100% word overlap, peak at 60% and some matching perfectly (100%).**
- **Minimal overlap at the sentence level near 0, indicating significant rephrasing in abstractive summaries.**
- **Abstractive summaries share some words with extractive ones but restructure sentences more significantly.**

Overlap Distribution

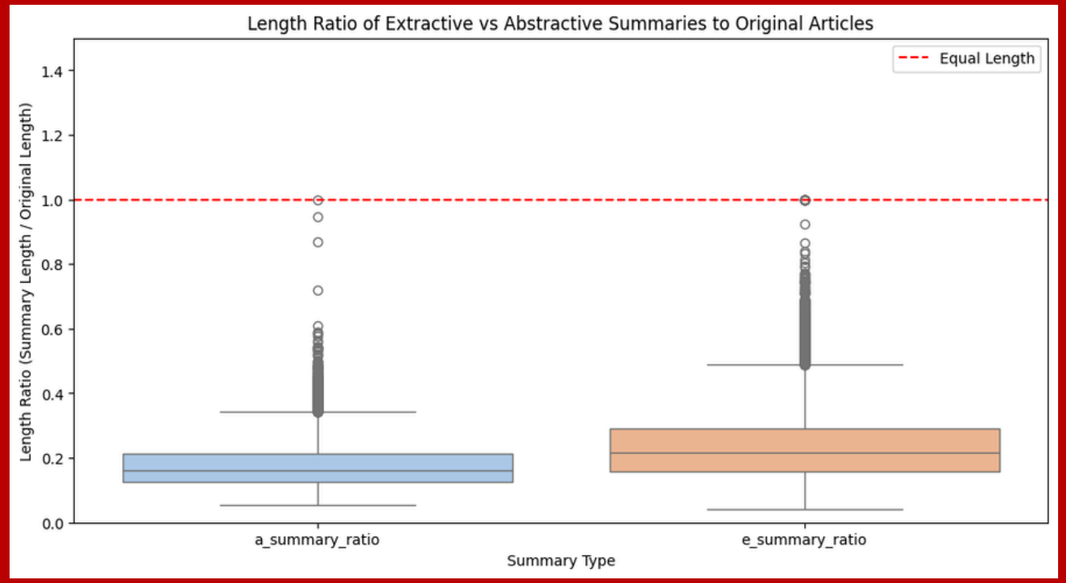


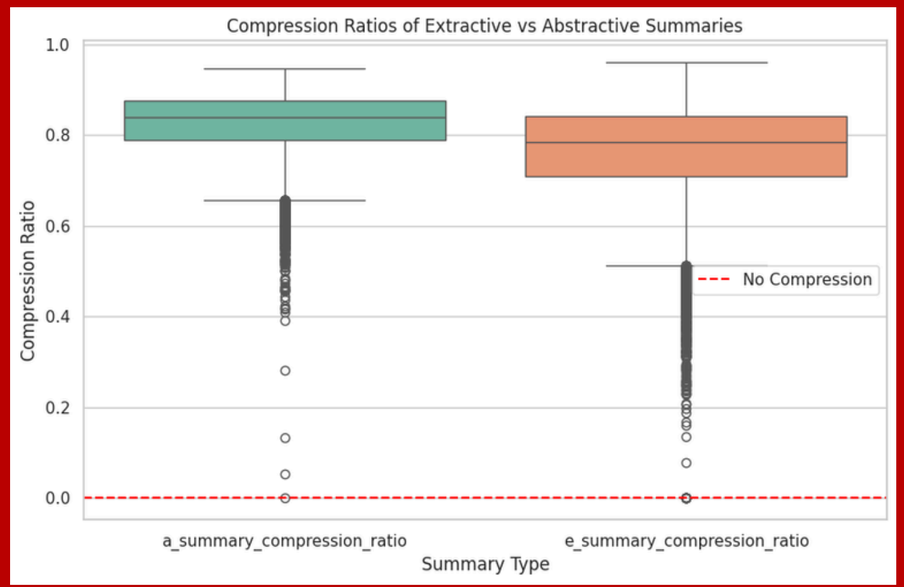


EDA

Summaries Distribution

Extractive summaries retain more content but are less concise, while abstractive summaries provide more compact representations with less variation.

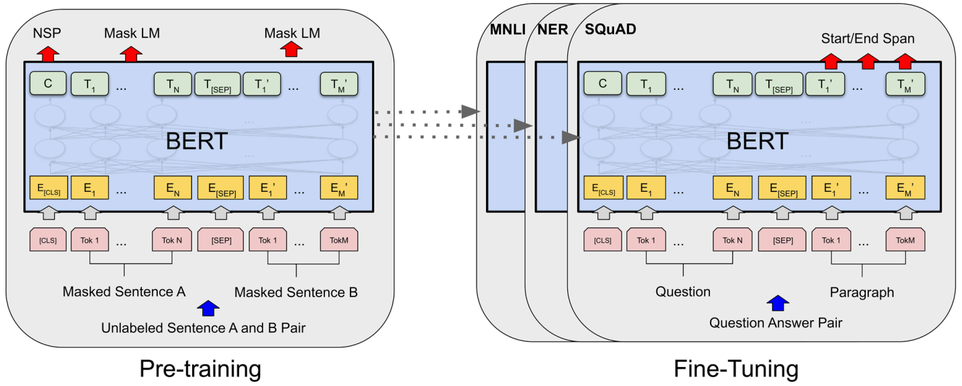




Summary Compression Ratio

Compression ratio measures the reduction in text length between the original and summary. A higher ratio indicates more summarization. Abstractive summaries have a higher compression ratio than extractive ones.

Model Development



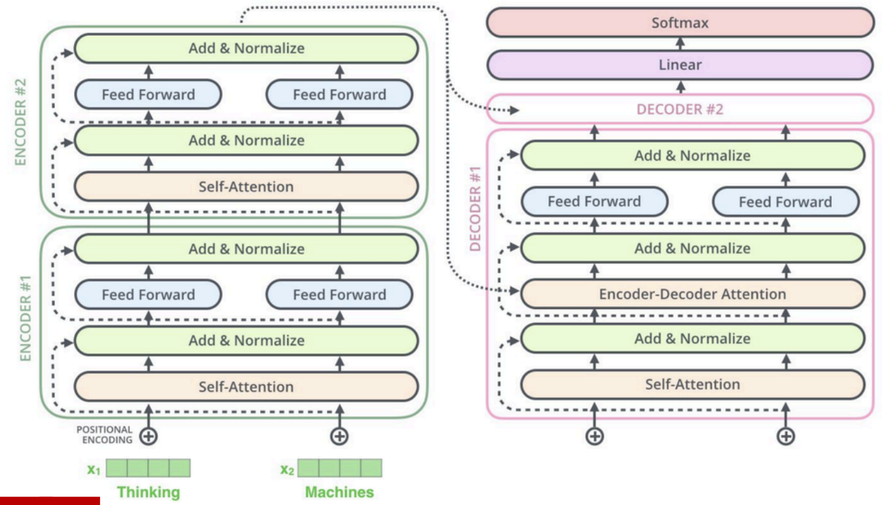
BERT Bidirectional Encoder Representations from Transformers

Architecture:
Uses only the encoder, focused on text understanding

Training Objective:
Masked language model for comprehension tasks.

Output:
Produces embeddings for tasks like classification.

Pre-Training Dataset:
BooksCorpus and Wikipedia



T5 Text-to-Text Transfer Transformer

Architecture:
Uses both encoder and decoder, suitable for a wide range of tasks.

Training Objective:
Sequence-to-sequence, used for both understanding and generation.

Output:
Directly generates text, better for generative tasks.

Pre-Training Dataset:
C4 (Colossal Clean Crawled Corpus) dataset

Flow Training

generate summarization

Data Preparation

- JSON dataset to .csv
- Data Cleaning (Column + Outliers)
- Data Sampling

Data Used

- test.csv (3k data)
- dev.csv (30k data)
- train.csv (30k data)

Modelling

- BERT : Extractive
- T5: Extractive
- BERT : Abstractive
- T5: Abstractive

Published Model

- kairaamilanii/BERT2BERT-ExtractiveSummarizer — Summarization • Updated 3 days ago • 22
- kairaamilanii/T5-ExtractiveSummarizer — Summarization • Updated 3 days ago • 13
- kairaamilanii/BERT2BERT-AbstractiveSummarizer — Updated 3 days ago • 9
- kairaamilanii/T5-AbstractiveSummarizer — Updated 3 days ago • 8

EDA

- Text Length Analysis
- Overlap Distribution
- Summaries Distribution
- Summary Coverage
- Compression Ratios

Result

- Precision, Recall, F1
- ROUGE-1, ROUGE-2, ROUGE-L

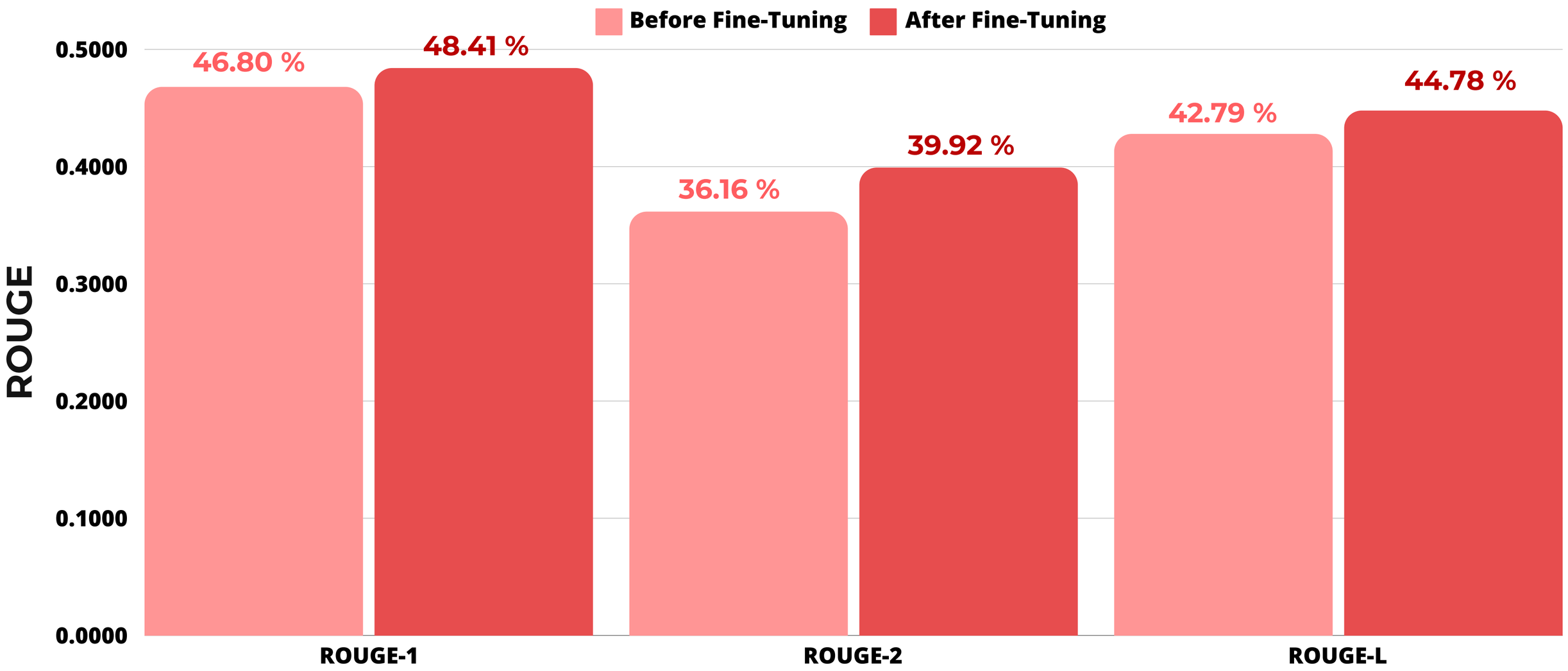
Testing

- BERTScore
- ROUGE

- A_T5.csv
- A_BERT.csv
- E_T5.csv
- E_BERT.csv

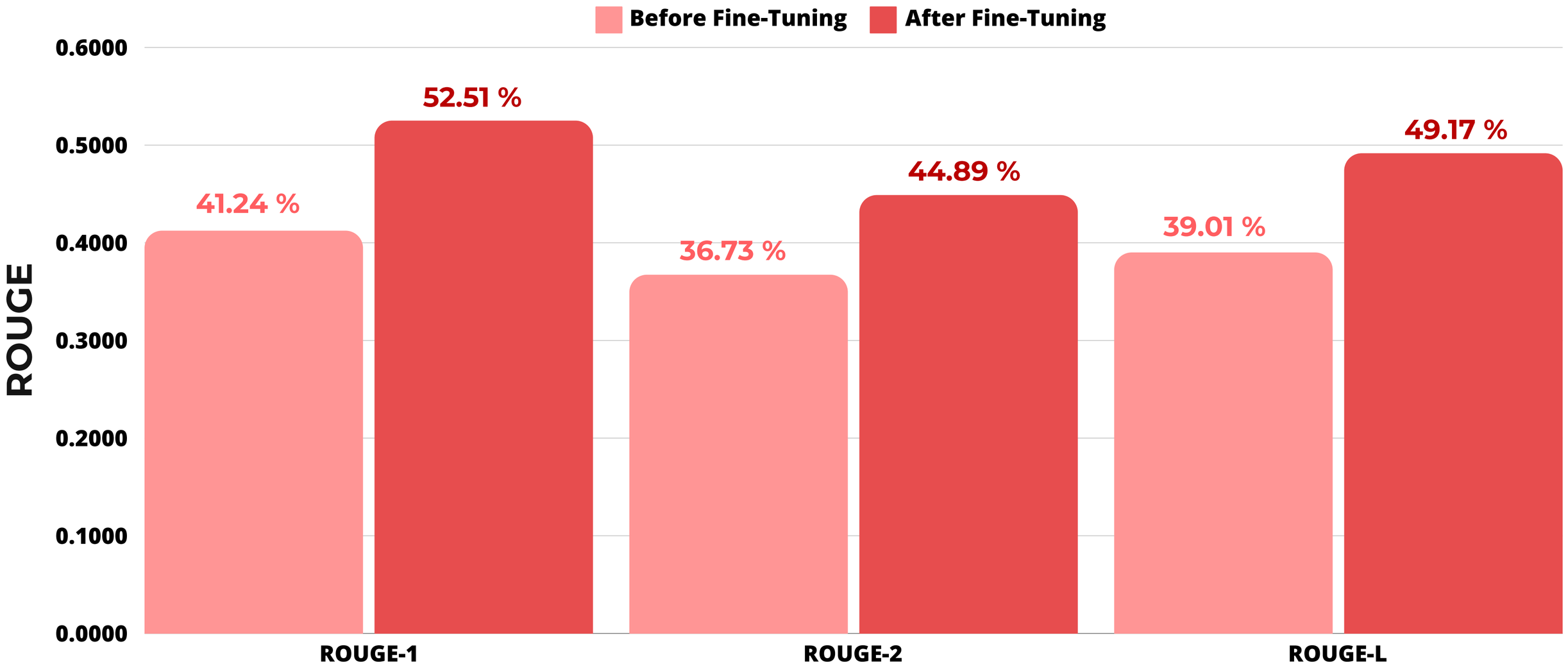
Result 1.

BERT : Extractive Summaries



Result 2.

T5 : Extractive Summaries

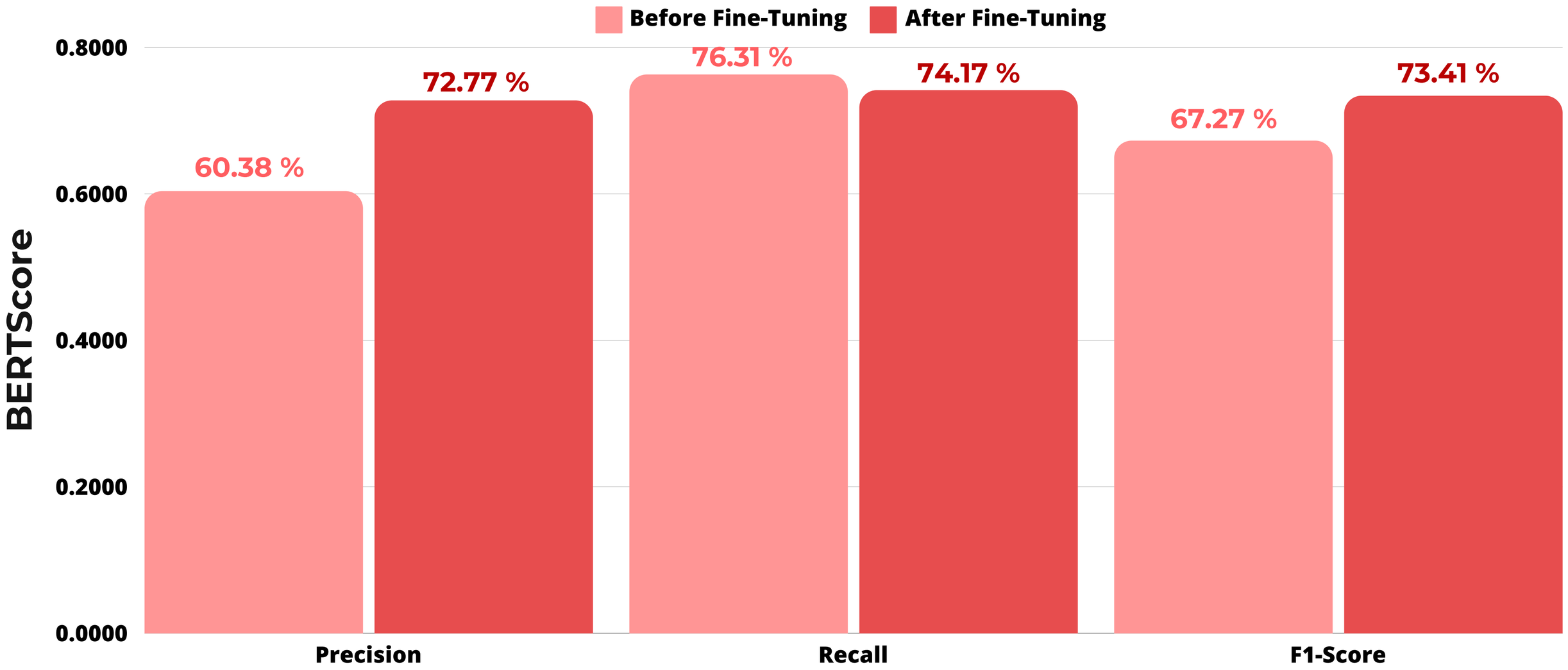


Result 3.

BERT : Abstractive Summaries

Result 4.

T5 : Abstractive Summaries

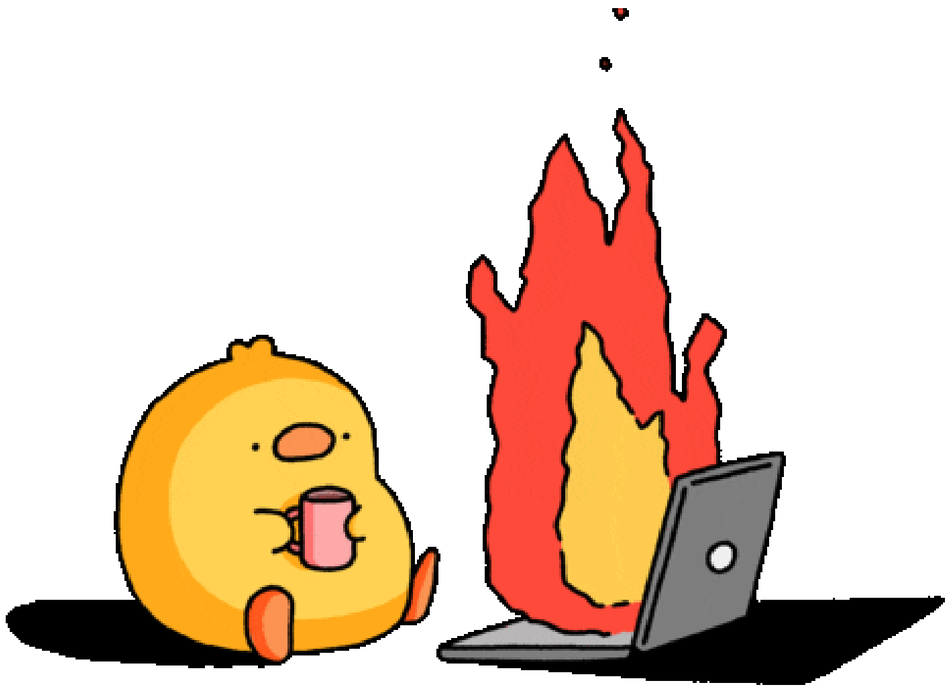


Future Improvement

Fine-tuning on the full training dataset could improve performance by increasing model robustness and generalization.

Explore other pre-trained models specifically designed for text summarization, which may offer better performance for this task.

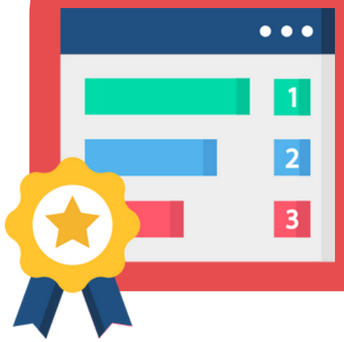
Further hyperparameter tuning and techniques like knowledge distillation could boost the model's overall performance.



Conclusion

Fine-tuning with domain-specific data significantly boosted both T5 and BERT models' performance in summarization tasks

Pre-trained models provided strong baselines, with fine-tuning enhancing results, particularly for domain-specific tasks.

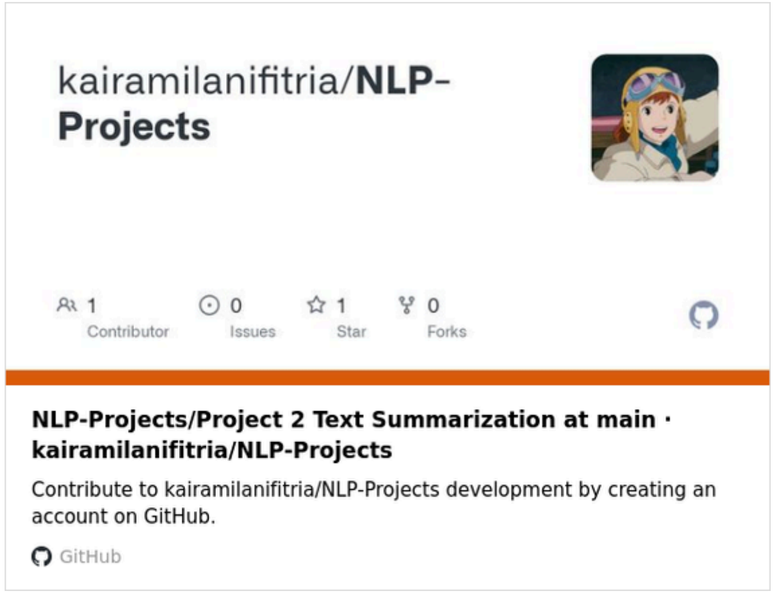


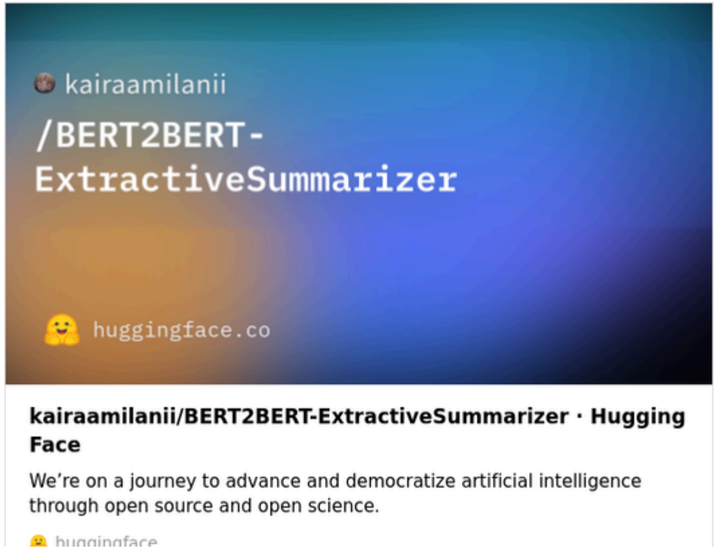
T5 consistently outperformed BERT, likely due to its architecture being better suited for summarization.

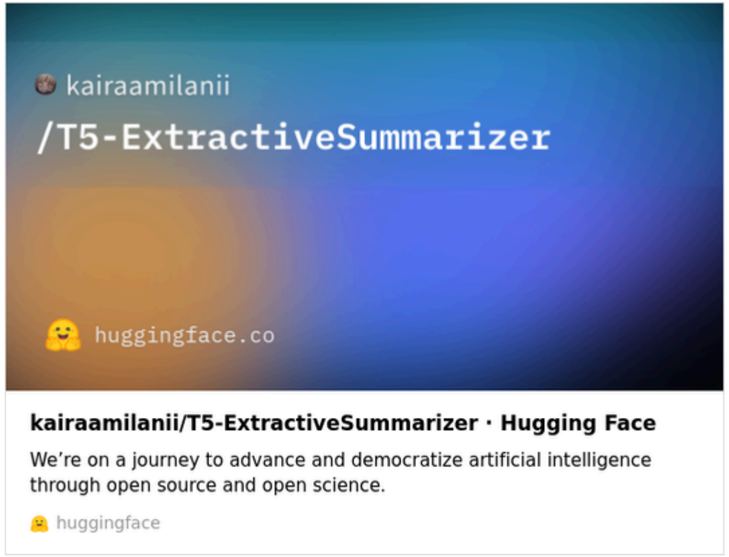
Random sampling 30k data points from the 200k dataset effectively balanced performance with computational limitations.

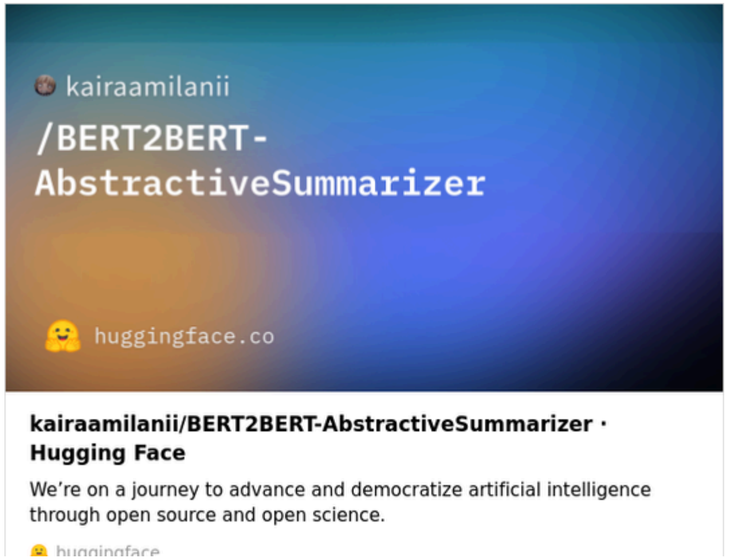
Thank You

Documentations:









Component (Opsional)

Table

Diagram