Infosys Springboard Internship

Text Summarization Project Report

I. Introduction

Project Overview
This report explores the development of both abstractive and extractive text summarization systems using the T5 model and spaCy, respectively, to condense lengthy texts into concise summaries.

Background and Context
Text summarization is crucial in natural language processing (NLP) for simplifying large volumes of text. Abstractive summarization generates new sentences, while extractive summarization selects key sentences from the text.

Objectives and Scope
The project aims to implement and evaluate both abstractive and extractive summarization techniques, demonstrating their effectiveness in generating coherent summaries.

II. Methodology

Approach and Techniques Used

- Abstractive Summarization: Utilizes the T5 model.
- Extractive Summarization: Leverages spaCy to select important sentences.

Tools and Technologies Employed

- T5 model (Transformers library)
- spaCy (NLP library)
- ROUGE metric for evaluation

Data Sources and Collection Methods

- Samsum dataset for training and testing the models.

III. Results

Summary of Key Findings

- Abstractive summarization effectively generates coherent and concise summaries.
- Extractive summarization accurately identifies key sentences.

Data Visualizations and Tables

- Summaries compared with original text using structured tables.
- Performance metrics displayed for both techniques.

	Original Dialogue	Original Summary	Generated Summary
0	Amanda: I baked cookies. Do you want some? \r\...	Amanda baked cookies and will bring Jerry some...	Amanda: I baked cookies. Do you want some? Jer...
1	Olivia: Who are you voting for in this electio...	Olivia and Olivier are voting for liberals in ...	Oliver: I'm not a liberal democrat, but i'm no...
2	Tim: Hi, what's up?\r\nKim: Bad mood tbh, I wa...	Kim may try the pomodoro technique recommended...	Kim: I was going to do lots of stuff but ended...
3	Edward: Rachel, I think I'm in ove with Bella....	Edward thinks he is in love with Bella. Rachel...	Edward: Rachel, I think I'm in ove with Bella....
4	Sam: hey overheard rick say something\r\nSam:...	Sam is confused, because he overheard Rick com...	rick was talking on the phone with someone Sam...

Performance Metrics and Evaluation Results

- ROUGE scores for abstractive summarization.
- Sentence scoring for extractive summarization.

	Metric	Score
0	ROUGE-1 (R1)	0.293556
1	ROUGE-2 (R2)	0.103223
2	ROUGE-L (RL)	0.232902
3	ROUGE-Lsum (RWs)	0.233232

```
#printing the scores of the words/tokens
sentence_scores
```

```
{Text summarization is the creation of a short, accurate, and fluent summary of a longer text document.: 0.3966942148760330
7,
 Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online.:
0.4049586776859504,
 This could help to discover relevant information and to consume relevant information faster.: 0.19008264462809918,
 Consider the internet, which is made up of web pages, news stories, status updates, blogs, and many other things.: 0.074380
1652892562,
```

IV. Discussion

Interpretation of Results

- The T5 model excels in generating human-like summaries.
- spaCy-based extractive summarization is efficient and straightforward.

Implications and Significance

- Both approaches significantly reduce the effort required to understand lengthy texts.

Limitations and Potential Improvements

- Abstractive model requires fine-tuning for specific datasets.
- Extractive method may miss nuanced information.

V. Conclusion

Summary of Main Points

- Successful implementation of both summarization techniques.
- Effective summarization demonstrated through evaluation metrics.

Achievements and Outcomes

- Developed and tested summarization functions.
- Evaluated performance using ROUGE scores and sentence scoring.

Recommendations for Future Work

- Further fine-tuning of the T5 model.
- Exploration of hybrid summarization methods.

VI. References

- Papers and articles on T5 model and spaCy.
- Documentation for the Samsum dataset and ROUGE metric.

VII. Timeline

Key Milestones and Deadlines

- Data preparation: Week 1
- Model implementation: Weeks 2-3
- Testing and evaluation: Week 4
- Report compilation: Week 5

Project Schedule and Progress Tracking

- Detailed schedule with task breakdowns and completion status.

VIII. Team and Contributions

List of Team Members and Their Roles

Project Lead: Oversee project progress & Documentation and report compilation

Team collaborators: Model implementation and testing, Data preprocessing and evaluation and helped the team lead.

IX. Challenges and Lessons Learned

Description of Obstacles Faced

- Data preprocessing complexities.
- Fine-tuning model performance.

Lessons Learned and Best Practices

- Importance of robust data preparation.
- Continuous evaluation and fine-tuning for improved results.\

This report captures the essential elements of the text summarization project, highlighting the methodologies, results, and insights gained throughout the development process.

THANKYOU

PROJECT MENTORS- NARENDRA KUMAR SIR & NISHANT SIR