A logo with blue text

Description automatically generatedA close up of a logo

Description automatically generated

**Module**

**2024S-T3 AML 2304-Natural Language Processing 01**

**(DSMM Group 1 & Group 3)**

**Project Title**

**Text Summarization**

**Intake**

**Term Three**

**Submitted By**

**Bipin Pandey (C0907008)**

**Nakul Budhathoki (C0901205)**

**Pujan Shrestha(C0901167)**

**Bronia John (C0912834)**

**Diksha Gusain(C0908141)**

**Harish Kundal(C0906990)**

**Parminder Kaur(C0908143)**

**Simarjeet Kaur (C0894309)**

**Priya Mehta (C0906992)**

Date: 15 Aug 2024

Contents

[Introduction 4](#_Toc174659439)

[Problem Statement 4](#_Toc174659440)

[Proposed Solution 4](#_Toc174659441)

[Objectives 4](#_Toc174659442)

[Secondary Objectives: 4](#_Toc174659443)

[Literature Review 4](#_Toc174659444)

[Text Summarization Techniques: 4](#_Toc174659445)

[Transformer Models in Summarization: 4](#_Toc174659446)

[Evaluation Metrics: 4](#_Toc174659447)

[Methodology 5](#_Toc174659448)

[Dataset Description 5](#_Toc174659449)

[Preprocessing 5](#_Toc174659450)

[Text Cleaning 5](#_Toc174659451)

[Tokenization 5](#_Toc174659452)

[Lemmatization 5](#_Toc174659453)

[N-gram Extraction 5](#_Toc174659454)

[Model Selection: 6](#_Toc174659455)

[Model Implementation: 6](#_Toc174659456)

[Evaluation Metrics: 6](#_Toc174659457)

[Results and Analysis 7](#_Toc174659458)

[ROUGE Scores 7](#_Toc174659459)

[T5 Model: 7](#_Toc174659460)

[PEGASUS Model: 7](#_Toc174659461)

[Combined Model (T5 + PEGASUS): 7](#_Toc174659462)

[Interpretation of ROUGE Scores: 7](#_Toc174659463)

[Summary Length Distribution: 7](#_Toc174659464)

[ROUGE Score Comparison: 7](#_Toc174659465)

[Summary Length Distribution: 8](#_Toc174659466)

[Discussion 9](#_Toc174659467)

[Strengths of the Combined Model 9](#_Toc174659468)

[Challenges 9](#_Toc174659469)

[Interpretation of Results 9](#_Toc174659470)

[Deployment 10](#_Toc174659471)

[Deployment Procedure 10](#_Toc174659472)

[Conclusion 10](#_Toc174659473)

[Summary of Findings 10](#_Toc174659474)

[Implications 10](#_Toc174659475)

[Future Work 10](#_Toc174659476)

[Model Improvements: 10](#_Toc174659477)

[Deployment Enhancements: 10](#_Toc174659478)

[Team Contributions 11](#_Toc174659479)

[MECE Table of Contributions 11](#_Toc174659480)

[Reference: 12](#_Toc174659481)

# Introduction

Summarization is crucial in transforming big volumes of text into summaries that are shorter, but cohesive. This is more especially important in areas such as journalism, research and the legal profession where is often required to understand large documents within a least amount of time.

## Problem Statement

The most traditional techniques of summarizing fail to satisfy the criterion of consciousness as well as informativeness at the same time. This work aims in overcoming these challenges through the use of two latest transformer models; T5 and PEGASUS.

## Proposed Solution

Since both T5 and PEGASUS models performed fairly well in our analysis, the hybrid model provides highly informative summaries while also reducing the summarization information. T5 model excels in multiple NLP distinctions and PEGASUS is build specifically for abstractive summarization.

# Objectives

Primary Objective: Our goal is to create a strong summarization system which will combine T5 and PEGASUS, to have such summaries that would be as precise as possible.

## Secondary Objectives:

* To quantify how well the model performs, we ‘ll use what are called ROUGE scores (you’ll learn more about those next).
* We will try various degrees to determine what provides us with the finest summaries.
* In a bid to enhance the comprehension of the results, we will generate displays.
* Last but not the least, the model will be deployed in order to be used by other people.

# Literature Review

## Text Summarization Techniques:

Text summarization is of two types, namely the extractive and the abstractive. On its turn, extractive summarization takes sentences directly from the text, which may not sound so smooth. On the contrary, Abstractive Summarization attempts at rewriting the text which while it can give summaries that sound more natural, it is much more difficult to get right.

## Transformer Models in Summarization:

In spoken content, automatic summarization has suddenly become far more viable with transformers. As with T5 model, I found it quite useful because it can be fine tuned for several tasks including summarization. PEGASUS is slightly more particular; it is trained for text summarization, which does make it very good at it.

## Evaluation Metrics:

What we do to assess how effective is a summary we use something called ROUGE scores. They are used to assess the degree of the matching between the elements that constitute the generated summary and elements in a reference summary. This is basically what ROUGE-1 considers and focuses solely on single words, ROUGE-2 looks for pairs, and ROUGE-L looks for sequences of length L.

# Methodology

## Dataset Description

The data set applied in this work consist of news and the summaries of those news. It was further split into training, validation and test set to carry out the training and assessment of the model. For each article, a reference summary was provided, which was used as the gold standard when comparing results of the models.

## Preprocessing

Preprocessing is an essential process to follow when dealing with the data to make it fit for model building. The following preprocessing steps were applied to ensure the quality and consistency of the input data:The following preprocessing steps were applied to ensure the quality and consistency of the input data:

### Text Cleaning

* Lowercasing: To standardize the vocabulary, all the text data was converted to lower case and also this helped in reducing the vocabulary size.
* Removing Punctuation: Symbols and punctuations were also omitted from the algorithm as they hardly do any meaningful job during the summarization process.
* Removing Stopwords: Examples include using ‘and’, ‘the’, and ‘is’ of which the NLTK stopwords list was used to remove them. They are not crucial at all and do not supply tangible added value to the summary; their exclusion allows to concentrate on the critical words.

### Tokenization

Word Tokenization: The lower level of two primary processes involved in text analysis Text was split into individual words (tokens). This step is rather important because models need input to be tokenized to be able to generate summaries.

### Lemmatization

Word Normalization: Lemmatization was used in an effort to stem words to their base forms (such as “running” to “run”). This process tends to narrow down the vocabulary and thereby emphasise on the fundamental meaning of the word.

### N-gram Extraction

Bigram and Trigram Extraction: It is to note that only bigrams or bi-grams (two successive words) and trigrams or tri-grams (three successive words) were used in the analysis. This is quite helpful when it comes to identifying composite words, that is those formed from more than one word.

## Model Selection:

* T5 Model: T5 is just the like the Swiss Army knife for language tasks. It can translate languages, answer questions or summarize text, and I found it quite good at the last one.
* PEGASUS Model: In turn, PEGASUS is like a specialist – this model has been designed for the summarization, so it is knowing how to create the summaries, which will be both brief and informative.

## Model Implementation:

* For training both models, we had a set of working parameters which we tuned for higher accuracy. We then took the output of T5 and integrated it with PEGASUS so as to have a summary which integrates the best of the two models.

## Evaluation Metrics:

ROUGE Scores: To assess how our summaries appeared like, we had the following: We evaluated how well our summaries came out by using the ROUGE scores. These scores permitted determining how much of the pertinent information from the initial text was retained in the summary.

# Results and Analysis

## ROUGE Scores

here’s how the models’ scores looked after fine-tuning and running through the models:

### T5 Model:

* ROUGE-1: 0. 3387 (This is how well it transcribed individual words from the text.)
* ROUGE-2: 0. 1311 (This shows how well it captured pairs of words or phrases.)
* ROUGE-L: 0. 2419 (This quantifies its capacity to maintain the exact sequential pattern of the fundamental text.)

### PEGASUS Model:

* ROUGE-1: 0. 3485
* ROUGE-2: 0. 1231
* ROUGE-L: 0. 2121

### Combined Model (T5 + PEGASUS):

* ROUGE-1: 0. 3387
* ROUGE-2: 0. 1311
* ROUGE-L: 0. 2419

## Interpretation of ROUGE Scores:

* ROUGE-1: T5 is very effective to obtain the key terms, and PEGASUS; but the employed fusion algorithms does not enhance this aspect.
* ROUGE-2: The result for phrases is also higher for the combined model indicating that the information is preserved and not repeated.
* ROUGE-L: The combined model is also quite good here, which means it retains the flow and organisation of the original text fairly fairly well.

## Summary Length Distribution:

Visualization: Next we made histograms in order to display how long the summaries are. Again, the combined model is much more in the middle in terms of output size and is quite good in this sense.

## ROUGE Score Comparison:

A graph of different colored bars

Description automatically generated

The bar chart aptly supports the fact that it is possible enjoin the two models T5 and PEGASUS to enhance the summarization since the ROUGE-L score measures the basic summary cohesion and structure. T5 model itself seems to do fairly good when tested separately, particularly for ROUGE-1 scores , I also see that PEGASUS is quite good to fair as well. Combined model benefits from the strengths of both models with the idea that the overall ROUGE scores is/are higher.

## Summary Length Distribution:

A bar chart with different colored bars

Description automatically generated

This histogram is the best way to compare the T5 and PEGASUS models in the context of the length of the summaries they produce. In some cases one might select a model depending on the number and type of summary which is to be generated. If one prefers more abridged versions, then it appears that T5 might be more helpful. But if more detailed summaries are required then PEGASUS may turn out to be beneficial. Further, this visualization makes one realize that there is need to know the output characteristics of given models in case they are to be cascaded together or selected for certain tasks.

# Discussion

## Strengths of the Combined Model

Extending the T5 model and the PEGASUS model gives us the best of both worlds. Another positive of T5 is that it can maintain the summary’s short length; at the same time, PEGASUS guarantees a detailed summary. The summaries they generate are therefore balanced.

## Challenges

Thus, it was quite challenging to opt for T5 or PEGASUS at the right proportion. A great deal of trial and error was needed to ensure that the summaries were relatively short, yet varied from one another in terms of length, at the same time. It also takes a lot of computational resources to run both models, this is something one has got to bear in mind especially if one is restricted in terms of computational power.

## Interpretation of Results

Some of the strengths which the results indicate are that the combined model efficiently identifies key phrases and is capable of preserving the natural flow of text. Even though it does nothing much for precise words recognition it does a comparatively better job in generating summarized and informative reports.

# Deployment

## Deployment Procedure

We put the model online using Flask and Gradio, which allowed it to be run using only a web browser. This is especially useful when you can just feed it text and get a summary, and optionally ROUGE scores if you provide a reference summary.

Gradio Interface: It is easy to use and all one has to do is typing the text then the interface provides you with the summary. If you wish, you can also compare the generated summary with a reference summary in using the ROUGE scores.

Performance Considerations: Making them run faster though involved a technique which as much as has to do a summary job doesn’t do all the calculations. We also adjusted settings of a model to achieve a good compromise in speed up of the process and quality of the results.

# Conclusion

## Summary of Findings

T5 and PEGASUS are suitable to be used together in summarization since it generates easy to understand and relevant summaries. According to the generated result of ROUGE, it promote that the most important message are taken and, at the same time, the summary is coherently formed.

## Implications

I think this model could be very helpful in the professions, which involve working with the mass text, for example, in the journalist’s profession, in the legal profession, or in academic writing. It is also very convenient to use as it has a friendly web interface for use by anyone.

# Future Work

## Model Improvements:

We could also in the future attempt to combine even more models or as an addition, utilize quite a number of techniques in order to make the given summaries even better.

## Deployment Enhancements:

We could also improve the times of the deployment and also its scalability if for instance we wish to deal with a number of summarizations at one time.

# Team Contributions

## MECE Table of Contributions

|  |  |  |
| --- | --- | --- |
| **Team Member** | **Task Category** | **Specific Contributions** |
| **Pujan Shrestha,**  **Nakul Budhathoki** | **Data Preparation** | - Collected and organized the dataset  - Conducted initial data exploration and cleaning |
| **Pujan Shrestha,**  **Diksha Gushain,**  **Harish Kumdal** | **Preprocessing** | - Implemented text cleaning and tokenization  - Applied stopword removal and lemmatization |
| **Bipin Pandey,**  **Nakul Budhathoki,**  **Bronia john** | **Model Selection & Training** | - Selected the T5 and PEGASUS models  - Fine-tuned models on the preprocessed data |
| **Priya,**  **Bronia John,**  **simarJeet Kaur** | **Evaluation & Analysis** | - Calculated ROUGE scores  - Performed model comparison and analysis |
| **Bipin Pandey,**  **Parmindar Kaur** | **Visualization** | - Generated visualizations for data distribution  - Visualized ROUGE score comparisons |
| **Bipin Pandey** | **Deployment** | - Developed and implemented the Gradio interface  - Managed the Flask deployment process |
| **Parmindar kaur,**  **priya** | **Documentation** | - Compiled the project report  - Ensured consistency and clarity in the documentation |

# Reference:

**Hugging Face Documentation.** *Transformers Documentation.* Available at: https://huggingface.co/docs/transformers

* Comprehensive documentation on the Transformers library, covering model usage, training, and deployment

**Gradio Documentation.** *Gradio: Build Machine Learning Web Apps with Python.*

* Documentation for the Gradio library, used for deploying machine learning models as interactive web applications.

**Flask Documentation.** *Flask: Web Development, One Drop at a Time.*

* Official documentation for Flask, the micro web framework used to deploy the model.

**Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020).** *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* *Journal of Machine Learning Research*.

* This paper details the T5 model, its architecture, and its applications in various NLP tasks, including text summarization.

**Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020).** *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization.* *Proceedings of the 37th International Conference on Machine Learning*.

* This paper explains the PEGASUS model, designed specifically for abstractive summarization, and discusses its performance across different datasets.