Effect of Pre-Trained Models in Text Summarization for proceedings

Student ID: 19059835

Name: Aishwarya John Pole Madhu

Supervisor Name: Christoph Salge

Table of Contents

List of tables

No table of figures entries found.

List of figures

1 Background research

1.1 Introduction

Summarizing is shortening a longer document while keeping its main ideas and structure intact. This technique can be useful in a variety of contexts, including studying, research, and communication. Summarizing a text can help break down complex ideas and make them easier to comprehend. It may help people save time since they don't have to read the complete text to acquire the important elements (Fabbri, et al, 2019). The main advantage of text summarization is that it enables readers to quickly get the main points of anything without reading the whole thing. This is especially useful when dealing with lengthy texts such as research papers, articles, and books. Summarization can make it easier to understand a text's main points and to identify important details. In addition, the time spent reading may be cut in half with a good summary (Liu, et al, 2019).

Text summarization can also help to improve communication. By summarizing a text, a reader can make it easier for others to understand the main ideas. Summarizing can also help to identify the key points of a text and to make it easier to compare and contrast different sources (Kryściński, et al, 2019). Finally, text summarization can help to improve understanding and knowledge by making it easier to identify important facts and details. Summarizing a text can help to identify the main points and to make it easier to think critically about the source material. This can help to improve comprehension and to better retain the information that is presented in a text (Sharma, et al, 2020).

1.2 Need for text summarization in court proceedings

Text summarization is becoming increasingly important in court proceedings. It is a process of automatically reducing a document to its most important points. To do this in just a small percentage of the time it takes to read and absorb the complete document, a succinct and precise summary of the content is required. Text summarization is especially useful in courts because it can help reduce the amount of time spent on tedious document review (Kanapala, et al, 2019). Lawyers and judges need to review vast amounts of material in a short amount of time, and text summarization can help them quickly identify the most important points in a document. This can help them to focus on the most relevant facts and arguments, and quickly identify any inconsistencies or inaccuracies in the document.

Text summarization is also beneficial in court proceedings because it can help the parties involved in a case better understand the opposing side's arguments. By summarizing the opposing party's arguments into concise points, both sides can gain a better understanding of the other's position. This can lead to more efficient and effective negotiations between the parties, and ultimately result in a better outcome (Keneshloo, et al, 2019). Text summarization can also be used to help identify key evidence in a case. By summarizing a document, lawyers can quickly identify any relevant evidence that could be used to support their argument. This can be especially helpful in cases involving large volumes of evidence. Overall, text summarization is an invaluable tool for court proceedings. It can help reduce the time spent on document review, provide a better understanding of the opposing party's arguments, and identify key evidence in a case. As text summarization technology continues to improve, it is likely to become even more important in court proceedings in the years to come (Kouris, et al, 2022).

1.3 Pretrained models

Pre-trained models for text summarization are models that had been trained on a huge body of data before being applied to a particular summarising job. Pre-trained models are particularly useful for text summarization because to summarise well, one must be fluent in the language and culture of the original material (Zhang, et al, 2019). Pre-trained models are able to capture the underlying semantics of the text and are more accurate than models that are trained from scratch. There are normally two parts to every pre-trained model for summarizing text: an encoder and a decoder. The encoder works to understand the structural and semantic features of the text, such as word order and relationships between words, while the decoder works to generate a summary of the text (Miller, et al, 2019).

1.4 ML and DL in text summarization

The purpose of the article by Alami et al. (2019) is improve automated text classification (ATS) by combining unconstrained deep neural network methods with a word embedding approach. Word2Vec format is shown to be superior to the more popular BOW model, and therefore the authors begin by developing a text summary approach dependent on word embeddings. Second, the authors provide supplementary models that include data from several sources by using a mixture of word2vec and unsupervised feature learning. In this work, the authors show that Word2Vec-trained unsupervised NN models outperform BOW-trained

models. Finally, authors suggest three different ensemble methods. The first ensemble uses a majority vote approach to combining BOW and word2vec. Information from the BOW method and unsupervised NN are combined in the second ensemble. Thirdly, authors have an ensemble that takes the data authors have gotten from Word2Vec and unsupervised neural networks and combines them. Authors demonstrate that ensemble approaches enhance ATS quality, and in particular, that an ensemble based on the word2vec approach produces superior results. Lastly, authors assess the models' effectiveness using a variety of experiments.

The HeterSumGraph neural network, developed by Wang et al. (2020), is a heterogeneous graph-based approach to extractive summarization. Sentences aren't the only thing that make up this network; there are also semantic nodes with varying fineness. These extra nodes serve as the go-between between sentences, enriching the cross-sentence interactions, and function as an intermediate among sentences. In addition, the logical expansion of the graph structure from a setting with a single document to one with several documents is possible thanks to the introduction of document nodes. Authors believe that authors are the first people to perform a comprehensive qualitative investigation into the benefits of utilising various types of nodes within graph-based NN for the purpose of extractive document summarization. This is based on the fact that authors were the ones who introduced these nodes for the first time.

Goularte et al (2019) approach to using a minimal set of fuzzy rules, we can summarise texts has the potential to be used in the creation of expert systems that can autonomously grade written work in the future. The suggested technique of summarizing has been taught and evaluated in trials utilising a dataset of texts written in Brazilian Portuguese that was submitted by students as a response to assignments that were assigned to them in a VLE. The suggested technique was evaluated with a number of different approaches, such as a naïve baseline, Score, Model, and Sentence, with the assistance of ROUGE measurements. A summarization system's effectiveness may be measured with the help of the Rouge metric. A system's ability to properly extract important information from a text is evaluated using precision and recall. Metric Rouge calculates an overall score based on a number of parameters, including the summary's length, the degree to which it overlaps with the reference summary, and other considerations. The findings indicate that the proposed technique yields a more accurate f-measure (with a confidence interval of 95%), in comparison to the methods described before.

In Song, et al, (2019) study, the authors provide an LSTM-CNN based ATSDL that may generate novel sentences by delving into smaller units than sentences, such as semantic phrases. This allows for the construction of new sentences. In contrast to other methods that are based

on abstraction, ATSDL is made up of two primary phases: the process begins with the identification of target phrases in the input sentences, and the latter of these stages provides text summaries via the use of deep learning. The ATSDL framework beats the frameworks within the context of meaning and syntax, and it obtains successful performance on manual linguistic quality assessment, as shown by experimental findings on the CNN and DailyMail datasets. This project used ROUGE score to evaluate the algorithms.

Author	**Objective**	**Parameter used**	**Result Achieved**
Alami et al.	Improving text summarization using unsupervised neural networks by including word embeddings and ensemble learning	Word Embedding and Ensemble Learning	Improved precision and recall by 8.72% and 27.08%, respectively.
Zhang et al.	Text summary using MLP and pre-training	Pretraining-based natural language generation	Improved ROUGE scores of up to 4.57%.
Goularte et al.	An approach to text summarising using fuzzy rules for use in machine evaluation	Fuzzy Rules	High accuracy of 98.50%.
Song et al.	Deep learning-based LSTM-CNN abstract text summarization	LSTM-CNN based DL	ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.44, 0.21, and 0.41, respectively.
Ramesh et al.	Abstractive Text Summarization Using T5 Architecture	T5 Architecture	ROUGE-1 and ROUGE-2 scores of 0.72 and 0.51, respectively.

1.5 Pre-trained models for text summarization

In Zhang, et al, (2019) research, authors offer a new encoder-decoder architecture that is based on pretraining and features a two-stage process that produces an output sequence based on an input sequence. In order to transform the input sequence into representations of context, the authors of this model employ the encoding method BERT. There are two stages of the design that focus on the decoder. To begin, an initial output sequence draught is generated using a

Transformer-based decoder. Step two involves writers hiding individual words in the draught sequence before sending it to BERT. The authors then use the BERT-provided draught representation in conjunction with an input Transformer-based decoder to make predictions about the improved word at each masked point. The methodology is, as far as authors are aware, the first way that integrates the BERT into the process of text creation jobs. In order to take the initial step in this direction, authors will assess the suggested approach by applying it to the text summarising problem. The results of our experiments indicate that our model reaches new levels of excellence using information from both the CNN/Daily Mail and New York Times datasets. The algorithms were graded using the ROUGE score in this assignment.

1.6 Summarization using T5 technique

Today, especially on the internet, there is a plethora of text due to the proliferation of information and communication technologies. The text must be summarised so that it is easier to read and understand without losing the essential meaning or context. In order to quickly and easily find the most relevant and important information in a large body of text, automatic text summarization is a useful tool. In this paper, authors propose a T5 and NLP-based text summarization model that can grasp the big picture, extract the most vital information, and produce coherent summaries. The algorithms were graded using the ROUGE score in this assignment (Ramesh, et al, 2022).

Ghadimi et al. (2022) provide HMSumm, a method for abstractively summarising the contents of several documents. The proposed method combines extractive summarising with abstractive summarization. The abstractive summary is created by first creating an extractive summary from the source materials. As a first step, writers deal with duplicate information, a common challenge in summarising many documents. The DPP is used to avoid repetition. At this point, we may also choose a maximum length for the input sequence that will be used in the abstractive summarization procedure. This may go one of two ways. The major motivation is to save time throughout the computation process. Second, an abstractive summarizer must retain the most important parts of the texts it is given to summarise. The authors assess the extracted summary phrases using a deep submodular network (DSN), and they determine redundancy using BERT-based similarities. By feeding the collected extractive summary into pre-trained models, two abstract summaries are created (BART and T5). The number of words in an abstract summary helps authors choose the best one. The performance of HMSumm is compared to that of several other methods, including evaluations from both humans and

ROUGE. The algorithms are evaluated using the DUC 2002, DUC 2004, Multi-News, and CNN/DailyMail datasets. Experiments show that HMSumm outperforms competing algorithms.

1.7 GPT2 for summarization

Kieuvongngam, et al, (2020) states that in light of the current COVID-19 pandemic, it is crucial for the medical community to keep up with the ever-increasing volume of published research on coronaviruses. Because of this disconnect between researchers and the ever-increasing volume of publications, the COVID-19 Open Research Dataset Challenge has released a dataset of scholarly articles and is calling for the use of machine learning techniques to help bridge the existing knowledge gap. Authors address this difficulty by conducting text summarization using two pre-trained NLP models, BERT and OpenAI GPT-2. ROUGE scores and visual examination are used in our analysis. The algorithm, using keywords pulled from the source articles, gives abstracted and detailed data. To aid the medical community, the effort will provide concise summaries of works for which an abstract is not yet accessible.

Rinse, et al, (2019) research paper takes an extractive and an abstractive approach to automating text summarizing. The former strategy makes use of submodular components and the BERT model of linguistic representation, whereas the GPT-2 model is used in the latter. Authors use two distinct kinds of datasets in this work: the CNN/DailyMail dataset, which is a standard for news article datasets, and the Podcast dataset, which is made up of transcripts of podcast episodes. On the CNN/DailyMail dataset, the GPT-2 achieves results that are on par with methods. A qualitative study, in the shape of a human evaluation, is also conducted by the authors, and they evaluate the trained model to show that it learns plausible abstractions, in addition to the quantitative evaluation.

Reading the news in little doses, understanding it "from long to short," and quickly, properly, and thoroughly getting important information are all aided by employing an automatic text summary, as stated by Yang et al (2021). The GPT2 model is used to power the authors' automatic summarising technique and modify the loss calculation section of the GPT2LMHeadModel included in the transformers package. The authors then conduct tests on both the original news information as well as the split-word data, collecting data for both 5 and 10 epochs, respectively, and evaluate the results with the ROUGE evaluation value. The measurements for both Rouge-1 and Rouge-2 have significantly increased when compared to previous studies. This model is an improvement on the BERT benchmark model's

representation of the summary text, which, in the context of automated extraction of news headlines, is insufficient and inaccessible.

1.8 Research gap

Despite the recent advancements in the discipline of summarizing text, there is still a lack of research focused on the effects that pre-trained models have on the summarization of real time court proceedings. However, few studies have explored the use of pre-trained models such as BERT, GPT-2, and T5 in text summarization in various fields, there is a lack of implementation of the techniques in court proceedings. These models have already proven useful for a variety of NLP applications, but their effectiveness for text summarization has yet to be adequately assessed. Thus in thins research, the gap of using pre-trained models on the summarization of real time court proceedings is carried out in detail. Another aspect of identifying if the size of the article changes the accuracy obtained by the models is also studied in detailed which was not explored in previous researches. The results will be evaluated using evaluation metrics such as accuracy, precision, recall, F1 score and Gouge score.

1.9 Significance

The findings of this study will provide an understanding of the effects of pre-trained models on the summarization of proceedings. This understanding can be used to develop more effective summarization methods that can be used to quickly and accurately summarize proceedings. Furthermore, the results of this study can be used to develop new techniques and strategies for summarizing proceedings that can be used in a variety of applications. Additionally, the findings of this study can be used to develop more accurate algorithms for summarizing proceedings that can be used in a variety of domains.

1.10 Problem statement

The purpose of text summarising is to mechanically produce a brief abstract of a given text material. Without reading the whole thing, this may assist readers get the gist of what the material is about (Kumar, et al, 2021). The use of pre-trained algorithms for text summarising has grown in popularity in recent years due to the fact that they may increase the precision and efficiency of summarization tools. Pre-trained models are models that are developed using large datasets and are trained to produce a certain output (Zhang, et al, 2019). The current limitations of text summarization in proceedings are a major challenge to effectively

understanding the main points of a text document. However, pre-trained models are often limited to large documents and struggle to accurately summarize new proceedings with both small and large documents (Fabbri, et al, 2019). Therefore, it is important to study how well pre-trained models perform when applied to Old Bailey proceedings, as well as to assess their performance when applied to proceedings of different sizes. The purpose of this research is to examine the usefulness of pre-trained models in this context in text summarization for new proceedings, and evaluating how well different pre-trained algorithms can summarise the content of Old Bailey hearings. Additionally, the study aims to analyze the effects of the size of the article on the accuracy of the model, as well as to determine the effectiveness of pre-trained models for summarizing new proceedings. The results will be evaluated using evaluation metrics such as accuracy, precision, recall, F1 score and Gouge score.

1.11 Aim

The purpose of this research is to examine how well-known pre-trained models (such BERT, BART, GPT2, and T5) perform when applied to Old Bailey proceedings and to study the performance of the models when applied to smaller and bigger proceedings.

1.12 Research question

- Do the size of the article affect the accuracy of the pre-trained models for text summarization?
- Are pre-trained models more effective for summarizing both long and short proceedings?

1.13 Research hypothesis

Hypothesis 1: Pre-trained models are more effective for summarizing long proceedings than short proceedings than long proceedings.

1.14 Objectives

- Evaluate how well various pre-trained models do at summarizing texts of Old Bailey proceedings and other proceedings.
- To analyze the effects of the size of the article on the accuracy of the model.

- To determine how well-suited pre-trained models are for summarizing new proceedings.

2 Summary of progress till date

My research aim is to investigate how well-trained models that have already been developed, such BERT, BART, GPT2 and T5 algorithms when applied to Old Bailey proceedings and to study the performance of the models when applied to smaller and bigger proceedings. I have to submit the proposal of the research, problem statement and objectives also. I have completed 15% of my research and continue with pre-trained models of BERT, BART, GPT2 and T5.

2.1 Data Description

The research data is collected from the Old Bailey website. The Old Bailey Proceedings xlsx is a dataset containing archival records of criminal trial proceedings at the Old Bailey Courthouse in London, England between 1674 and 1913.

https://www.oldbaileyonline.org/static/Data.jsp

The data includes information on the court proceedings such as the names of the people involved, the charges, the verdicts, and the sentences. It also contains supplementary information such as the occupations of the defendants, the dates of the trials, and the location of the courthouse. This valuable dataset gives a unique insight into the criminal justice system of the period, and is useful for researchers studying the social, cultural, and legal history of the time.

The data contains 100 rows and 4 columns in the format ‘xlsx’ to read the function read_xlsx().

2.2 Number of Unique Elements

There are 9 unique elements in the attribute of ‘Category’. They are 'Breaking Peace', 'Theft', 'Killing', 'Deception', 'Royal Offences', 'Violent Theft', 'Sexual Offences', 'Miscellaneous', 'Damage to Property'.

2.3 Value Counts in the Output Attribute

To find the value counts in the output attribute ‘Category’ described in the following.

```
Theft                  36
Royal Offences         18
Breaking Peace         15
Deception              15
Killing                 7
Violent Theft           4
Sexual Offences         2
Damage to Property      2
Miscellaneous           1
Name: Category, dtype: int64
```

Figure 1 Output value counts

The ‘Theft’ category contains maximum amount of data and ‘Miscellaneous’ contains minimum amount of data.

2.4 Information or Null

To find the information about the data or checking the null values in the data frame.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Date                100 non-null    object
 1   link                100 non-null    object
 2   Category            100 non-null    object
 3   Proceedings Content  100 non-null    object
dtypes: object(4)
memory usage: 3.2+ KB
```

Figure 2 Information or Null

There are no null values in the pre-trained models in text summarization data.

2.5 EDA of Pre-trained models

Visualizing Count Plot for the attribute ‘Category’ is in the following.

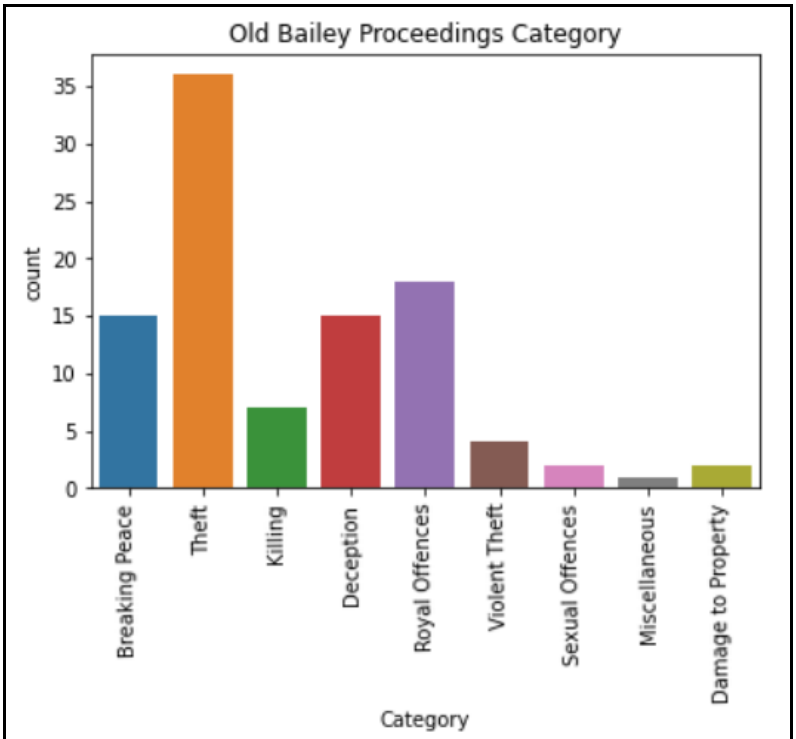


Figure 3 Output Attribute 'Category' of proceedings

From the above count plot category variable 'Theft' has maximum amount of data. The variables 'Breaking Peace', 'Deception' and 'Royal Offences' contain more than 15 numbers of data.

2.6 Dropping Unwanted Attributes

To drop or delete the unwanted columns which are not giving any predictions. In this text summarization for proceedings, the attributes named 'Date' and 'link' are deleted.

3 Considering issues

3.1 Social issues

The study's goal is to promote the ethical and appropriate use of the collected data. This study checks to see that information isn't misused for things like prejudice or discrimination. In addition, the study takes precautions to protect the privacy of its participants by never disclosing their personal information to other parties. The study is conducted in accordance with the standards of data privacy, data security, and data integrity to guarantee these aims are satisfied. This comprises access control rules, data minimization practises, data retention regulations, and the use of proper security mechanisms and data encryption methods. In addition, the study checks if sufficient measures are in place to prevent data from being accessed, used, disclosed, or destroyed without authorization.

3.2 Ethical issues

As correct and timely information is provided to all participants, the project is transparent in its activities. The initiative promotes a safe working environment by providing the tools to prevent accidents. Since it does not include any potentially dangerous activities, the project will not have a detrimental impact on the environment. The project properly secures sensitive data by ensuring the privacy of all project workers.

3.3 Legal issues

All applicable privacy laws and standards have been followed during the course of this project. To the best of my knowledge, I have followed all applicable guidelines for collecting data and conducting analysis. It is unlikely that any antitrust issues will arise. I have ensured that the project has no antitrust issues by taking the required measures.

3.4 Professional issues

The investigation is very honest and precise. The greatest standards are applied to the study process, and every effort is made to detect and rectify any mistakes that may have been made. All sources are properly cited, and plagiarism is not accepted. All study materials, including data and software, are properly referenced and given credit where credit is due. All attempts

are taken to assure the greatest quality of research, and the results are presented in an honest and truthful manner.

4 Project plan

4.1 Short description of your project

The use of pre-trained models in text summarization has many advantages. Firstly, pre-trained models can provide better accuracy and performance compared to traditional methods. Secondly, pre-trained models are more easily deployable, since they require less training data and can be applied to a variety of tasks. BERT is trained on large datasets, making it more accurate and reliable than other models. BERT can be used to understand the context of words and phrases, allowing it to better understand language. BERT is trained to use pre-trained parameters, making it easier to fine-tune and adapt to new tasks. BART is a seq2seq model, meaning it can generate text from a source text. It is pre-trained on large datasets, making it more accurate and reliable than other models. It is able to use both encoder and decoder for text summarization tasks, allowing for more accurate results. GPT2 is a transformer-based model, meaning it can process natural language more effectively than other models. GPT2 is able to capture long-term dependencies, allowing for more accurate results. T5 is a multi-task learning model, meaning it can be used for multiple tasks. T5 may acquire new skills and apply them to previously completed tasks, making it easier to fine-tune and adapt to different tasks. This study's goal is to find out how well pre-trained models perform when applied to Old Bailey proceedings and to study the performance of the models when applied to smaller and bigger proceedings. The evaluation of the model is done by calculating the loss and accuracy. By comparing the accuracy of all this model, the most efficient pre-trained models can be identified when applied to legal data. The old court dataset offers a unique insight into the history of the court system and the evolution of certain legal practices and principles over time. It also provides an invaluable resource for researchers and historians to explore how the court system has changed and adapted over time. Thus summarizaion in the old court processdings are yet to explore in detail and thus choosing the dataset.

Percentage of the accuracy is calculated as (correctly summarized /Total) * 100

Percentage of the loss is calculated as (incorrectly summarized /Total) * 100

4.2 Plan to conduct the research

The plan for this study will involve a quantitative approach to assess pre-trained models' usefulness in practise in text summarization for new proceedings. Specifically, the research methodology will involve the following steps:

Data Collection: The dataset for this study will be collected manually from Old Bailey proceedings. The proceedings will be of varying sizes, ranging from short proceedings (less than 500 words) to long proceedings (over 2000 words).

Pre-trained Model Selection: Pre-trained models such as BERT, BART, GPT2 and T5 algorithms will be selected for this study.

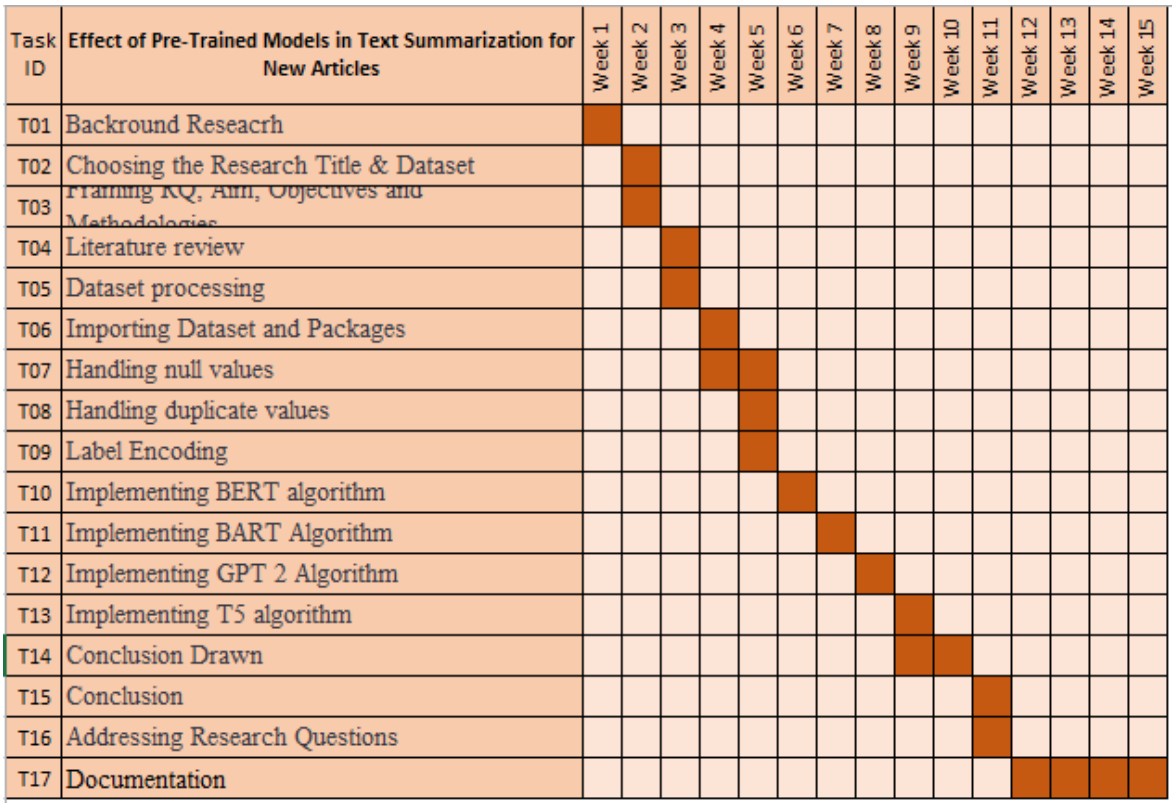
Model Training: The gathered data will be used to further hone the pre-trained models.

Evaluation and Analysis: The accuracy, loss and performance of the pre-trained models when applied to Old Bailey proceedings and other proceedings of varying sizes will be evaluated and analyzed.

Results: The results of the study will be reported and discussed.

This study will provide valuable insights into the efficiency of text summarization using pre-trained models of new proceedings.

4.3 Gantt chart



Task ID	Effect of Pre-Trained Models in Text Summarization for New Articles	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15
T01	Backround Reseacrh															
T02	Choosing the Research Title & Dataset															
T03	Framing RQ, Aim, Objectives and Methodologies															
T04	Literature review															
T05	Dataset processing															
T06	Importing Dataset and Packages															
T07	Handling null values															
T08	Handling duplicate values															
T09	Label Encoding															
T10	Implementing BERT algorithm															
T11	Implementing BART Algorithm															
T12	Implementing GPT 2 Algorithm															
T13	Implementing T5 algorithm															
T14	Conclusion Drawn															
T15	Conclusion															
T16	Addressing Research Questions															
T17	Documentation															

4.4 Tools used

Hardware Tools

1. Desktop or laptop computer

2. External hard drive

Software Tools

1. Python programming language

2. Jupyter Notebook

3. Anaconda Data Science Platform

5 Appendix

Appendix- 1 Code

1- Old Bailey Proceedings Files collection

```
import pandas as baile_ProcedPs
##import Old Bailey Proceedings Files
baile_Proced = baile_ProcedPs.read_excel('Old Bailey Proceedings.xlsx')
baile_Proced.shape

baile_Proced[:5]# 5 top data

baile_Proced[-5:]# 5 tail data
```

*** the data contains date of the bails, link, category and proceedings contents.

```
baile_Proced['Category'].nunique()## number of case category

baile_Proced['Category'].unique()
```

** cases are under 9 different categories.

```
baile_Proced['Category'].value_counts()

baile_Proced['link'][:10] #some link

baile_Proced['Proceedings Content'][3]# 3rd case proceedings.

baile_Proced.info()
```

** infor: All the attributes are object type data.
** And non_ull.

```
## EDA

import seaborn as baile_ProcedBSr
## Plot to check the category
baile_ProcedBSr.countplot(x ='Category', data = baile_Proced).set(title='Old Bailey
Proceedings Category')

import matplotlib.pyplot as baile_ProcedYP
baile_ProcedYP.xticks(rotation=91)
```

** 'Theft' kind of cases are more.

```
##Drop date and link attributes...
del baile_Proced['Date']
del baile_Proced['link']
```

baile_Proced.to_csv('bailey_Proceedings.csv', index=False)

Appendix- 2 Code Screenshot

```
import pandas as baile_ProcedPs
##import Old Bailey Proceedings Files
baile_Proced = baile_ProcedPs.read_excel('Old Bailey Proceedings.xlsx')
baile_Proced.shape
```

```
(100, 4)
```

```
[2] baile_Proced[:5]# 5 top data
```

	Date	link	Category	Proceedings Content
0	15th January 1900	https://www.oldbaileyonline.org/browse.jsp?id=...	Breaking Peace	RICHARD BOWERS . I live at 1, Chadwick Street,...
1	15th January 1900	https://www.oldbaileyonline.org/browse.jsp?id=...	Theft	WILLIAM JOHN WHYTE . On December 13th, between...
2	15th January 1900	https://www.oldbaileyonline.org/browse.jsp?id=...	Theft	JAMES AMOS (City Detective). On January 1st, a...
3	15th January 1900	https://www.oldbaileyonline.org/browse.jsp?id=...	Killing	about 4.30 p.m. on August 30th I was driving m...
4	12th February 1900	https://www.oldbaileyonline.org/browse.jsp?id=...	Deception	GEORGE ENGLISH BOYLE . I am a messenger at the...

```
baile_Proced[-5:]# 5 tail data
```

	Date	link	Category	Proceedings Content
95	11th October 1910	https://www.oldbaileyonline.org/browse.jsp?id=...	Royal Offences	Sergeant HENRY GARANRD, G Division. About 11 a...
96	11th October 1910	https://www.oldbaileyonline.org/browse.jsp?id=...	Breaking Peace	BENJAMIN WRAGG , 18, Risinghill Street, Clerke...
97	15th November 1910	https://www.oldbaileyonline.org/browse.jsp?id=...	Theft	Inspector WILLIAM EVANS (Birmingham Police). P...
98	6th December 1910	https://www.oldbaileyonline.org/browse.jsp?id=...	Theft	Mr. Huntly Jenkins prosecuted. Prisoner was tr...
99	10th January 1911	https://www.oldbaileyonline.org/browse.jsp?id=...	Breaking Peace	BABS MILLAR . I live at 81, Tabard Street, Bor...

*** the data contains date of the bails, link, category and proceedings contents.** * from year 1900 to 1910, totally 100 cases are available.

```
[4] baile_Proced['Category'].nunique()## number of case category
```

```
9
```

```
baile_Proced['Category'].unique()
```

```
array(['Breaking Peace', 'Theft', 'Killing', 'Deception',
       'Royal Offences', 'Violent Theft', 'Sexual Offences',
       'Miscellaneous', 'Damage to Property'], dtype=object)
```

** cases are under 9 different categories.

```
[6] baile_Proced['Category'].value_counts()
```

```
Theft                36
Royal Offences       18
Breaking Peace       15
Deception            15
Killing               7
Violent Theft         4
Sexual Offences       2
Damage to Property    2
Miscellaneous         1
Name: Category, dtype: int64
```

```
baile_Proced['link'][:10] #some link
```

```
0    https://www.oldbaileyonline.org/browse.jsp?id=...
1    https://www.oldbaileyonline.org/browse.jsp?id=...
2    https://www.oldbaileyonline.org/browse.jsp?id=...
3    https://www.oldbaileyonline.org/browse.jsp?id=...
4    https://www.oldbaileyonline.org/browse.jsp?id=...
5    https://www.oldbaileyonline.org/browse.jsp?id=...
6    https://www.oldbaileyonline.org/browse.jsp?id=...
7    https://www.oldbaileyonline.org/browse.jsp?id=...
8    https://www.oldbaileyonline.org/browse.jsp?id=...
9    https://www.oldbaileyonline.org/browse.jsp?id=...
Name: link, dtype: object
```

```
baile_Proced['Proceedings Content'][3]# 3rd case proceedings.
```

```
'about 4.30 p.m. on August 30th I was driving my van along the Caledonian Road in the direction of Albion Street—outside a public-house there I saw four men
conversation—I saw an old gentleman walking along the road—the four men followed him down—my van got close up to the old gentleman, and I saw Barrett (See S
aper, Vol. CXXX., page 923) separate from the others—he ran behind my van, then ran across the road in front of it, and came in front of the old gentleman—G
e a signal with his left arm for Barrett to do something to the old gentleman—Barrett made a snatch at the old gentleman\'s chain, and hit him under the chi
s other hand; then the other three rushed upon him and hustled him, and threw him down, with his legs in the gutter and his head on the pavement—I shouted o
you brutes!" and jumped off my van—one of the men said to the deceased, "Mind where you are going to"—then they all four ran round a public-house on the oth
picked the old gentleman up—he had a wound on the back of his head—I tied my handkerchief round his head, and called for assistance—part of his watch-chain
s pocket, and part not—I picked Thompson out at the Police-court, but at the Police-station I failed to identify him—I was examined here last October, when
men, Barrett, Grande, and Jones, were tried, and convicted—I say now that the prisoner is the fourth man.\n\nBy the COURT. Barrett was the man who hit the o
man under the chin—as he fell back he fell on the other three, and they knocked him into the gutter.\n\nCross-examined by the Prisoner. I did not hear you s
rande.\n\nGEORGE FREDERICK BARNES . I am a railway porter—on the afternoon of Wednesday, October 30th, I was walking down Caledonian Street in the direction
n Street—I saw the old gentleman and four men round him hustling and struggling with him—one of them struck his arm as if to make a snatch, and he fell on t
f his head in the gutter—three of the men ran away, one stopped a second or two as if to pick him up, then he ran after Jones or Bliss, as he is called—I id
three of the men, and the fourth is the prisoner.\n\nGEORGE JARED . I was next to the Albion public-house in Albion Street, Caledonian Road, on August 30th,
fternoon—I saw three men running away; I followed them as fast as I could, but they got away—I heard Barrett say, "It is no good running now "—the third man
running was the prisoner—some time after December 10th I was taken to the Police-station, and amongst them I saw the third man who was running away; that wa
n.\n\nBy the COURT. I saw the old gentleman on the ground with his head in the gutter.\n\nJAMES HARVEY . I am a house agent—I knew Mr. Benbow—on August 30th
3.30 p.m., I was called to see him in Albion Street—he was about 64 years old—he was bleeding; I got a cab and took him to the hospital—his chain was missin
tch was there—he had a £2 piece, which was gone.\n\nTHOMAS WALTER COFFIN, F.R.C.S . I practise at 31, Maiden Terrace, Haverstock Hill—I was the regular medi
dant of Mr. Benbow—he enjoyed pretty good health, and was a fairly strong man—I saw him\n\na short time before August 30th; he was then in fairly good healt
alled on the evening of August 30th to his house, and I found him suffering from a very severe injury go his brain—I did all I could for him; he lingered on
early morning of October 1st, when he died—I made a post-mortem examination and found that the cause of death was a compound comminuted fracture of the skul
tensive fractures to the base of the skull—those would quite naturally result from the deceased having been pushed down on to a kerbstone and coming in cont
the stone—I have no doubt that it was from those injuries that he died.\n\nTHOMAS PERCY LEGG . I was the house surgeon at the Royal Free Hospital on August
n I attended to and dressed Mr. Benbow\'s wounds—he had an injury to the base of the skull, from which he was bleeding freely—he desired to go away and have
```

```
baile_Proced.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 4 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   Date                 100 non-null    object
 1   link                 100 non-null    object
 2   Category             100 non-null    object
 3   Proceedings Content  100 non-null    object
dtypes: object(4)
memory usage: 3.2+ KB
```

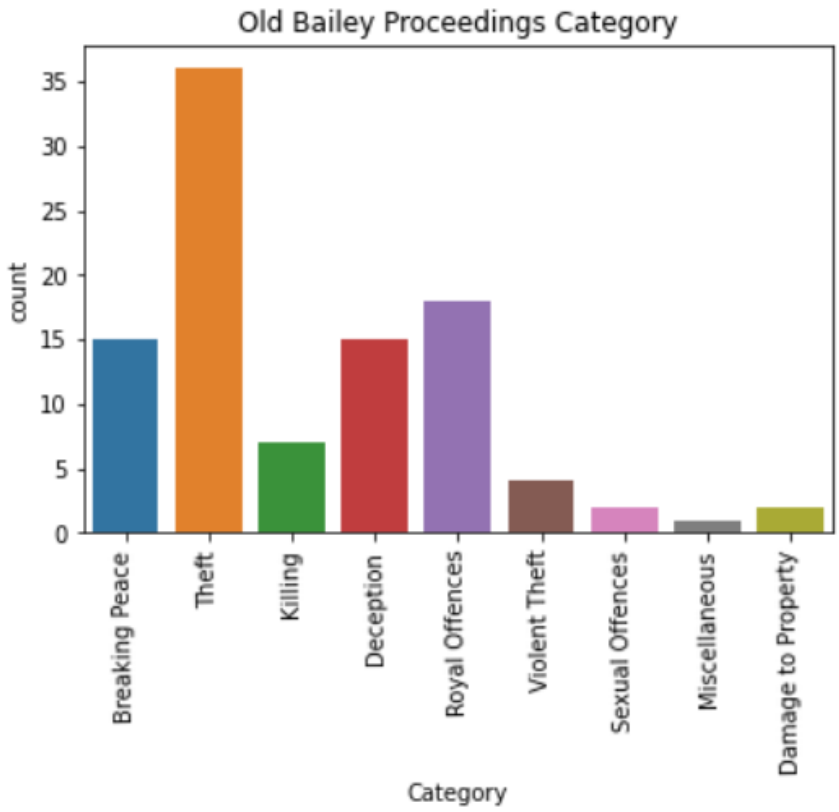
** infor: All the attributes are object type data. ** And non_ull.

EDA

```
[10] import seaborn as baile_ProcedBSr
     ## Plot to check the category
     baile_ProcedBSr.countplot(x ='Category', data = baile_Proced).set(title='Old Bailey Proceedings Category')

     import matplotlib.pyplot as baile_ProcedYP
     baile_ProcedYP.xticks(rotation=91)
```

```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8]),
 [Text(0, 0, 'Breaking Peace'),
  Text(1, 0, 'Theft'),
  Text(2, 0, 'Killing'),
  Text(3, 0, 'Deception'),
  Text(4, 0, 'Royal Offences'),
  Text(5, 0, 'Violent Theft'),
  Text(6, 0, 'Sexual Offences'),
  Text(7, 0, 'Miscellaneous'),
  Text(8, 0, 'Damage to Property')])
```



** 'Theft' kind of cases are more.

```
[11] ##Drop date and link attributes...
     del baile_Proced['Date']
     del baile_Proced['link']
```

0s

```
[12] baile_Proced.to_csv('bailey_Proceedings.csv', index=False)
```

0s

6 References

Alami, N., Meknassi, M. and En-nahnahi, N., 2019. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. Expert systems with applications, 123, pp.195-211.

Fabbri, A.R., Li, I., She, T., Li, S. and Radev, D.R., 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. arXiv preprint arXiv:1906.01749.

Ghadimi, A. and Beigy, H., 2022. Hybrid multi-document summarization using pre-trained language models. Expert Systems with Applications, 192, p.116292.

Goularte, F.B., Nassar, S.M., Fileto, R. and Saggion, H., 2019. A text summarization method based on fuzzy rules and applicable to automated assessment. Expert Systems with Applications, 115, pp.264-275.

Kanapala, A., Pal, S. and Pamula, R., 2019. Text summarization from legal documents: a survey. Artificial Intelligence Review, 51, pp.371-402.

Keneshloo, Y., Ramakrishnan, N. and Reddy, C.K., 2019, May. Deep transfer reinforcement learning for text summarization. In Proceedings of the 2019 SIAM International Conference on Data Mining (pp. 675-683). Society for Industrial and Applied Mathematics.

Kieuvongngam, V., Tan, B. and Niu, Y., 2020. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. arXiv preprint arXiv:2006.01997.

Kouris, P., Alexandridis, G. and Stafylopatis, A., 2022. Abstractive text summarization based on deep learning and semantic content generalization.

Kryściński, W., Keskar, N.S., McCann, B., Xiong, C. and Socher, R., 2019. Neural text summarization: A critical evaluation. arXiv preprint arXiv:1908.08960.

Kumar, Y., Kaur, K. and Kaur, S., 2021. Study of automatic text summarization approaches in different languages. Artificial Intelligence Review, 54(8), pp.5897-5929.

Li, J., Sun, A., Han, J. and Li, C., 2020. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), pp.50-70.

Ramesh, G.S., Manyam, V., Mandula, V., Myana, P., Macha, S. and Reddy, S., 2022. Abstractive Text Summarization Using T5 Architecture. In Proceedings of Second

International Conference on Advances in Computer Engineering and Communication Systems (pp. 535-543). Springer, Singapore.

Rinse, V. and Siitova, A., 2019. Text summarization using transfer learnin: Extractive and abstractive summarization using bert and gpt-2 on news and podcast data.

Sharma, B., Katyal, N., Kumar, V. and Lathwal, A., 2020. Automatic Text Summarization Using Fuzzy Extraction. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2019, Volume 1 (pp. 395-405). Springer Singapore.

Song, S., Huang, H. and Ruan, T., 2019. Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications, 78, pp.857-875.

Wang, D., Liu, P., Zheng, Y., Qiu, X. and Huang, X., 2020. Heterogeneous graph neural networks for extractive document summarization. arXiv preprint arXiv:2004.12393.

Yang, Z., Dong, Y., Deng, J., Sha, B. and Xu, T., 2021, October. Research on Automatic News Text Summarization Technology Based on GPT2 Model. In 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (pp. 418-423).

Zhang, H., Xu, J. and Wang, J., 2019. Pretraining-based natural language generation for text summarization. arXiv preprint arXiv:1902.09243.

Zhang, X., Wei, F. and Zhou, M., 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. arXiv preprint arXiv:1905.06566.