Effect of Pre-Trained Models in Text Summarization for proceedings

7COM1039-0109-2022

Advanced Computer Science Masters Project

Aishwarya John Pole Madhu

Student ID - 19059835

Problem statement

Text summarization is a process of automatically generating a concise summary of a given text document. This can help readers quickly understand the main points of the text without having to read the entire document (Kumar, et al, 2021). The use of pre-trained models in text summarization has become increasingly popular in recent years as it can help improve the accuracy and performance of summarization systems. 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). As a result, there is a need to investigate the effectiveness of pre-trained models when applied to Old Bailey proceedings, as well as to assess their performance when applied to proceedings of different sizes. This study aims to address this issue by investigating the effectiveness of pre-trained models in text summarization for new proceedings, and to compare the performance of pre-trained models on text summarization of Old Bailey proceedings. 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.

Aim

The aim of this study is to investigate the effectiveness of pre-trained models such as 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.

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?

Research hypothesis

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

Objectives

- To compare the performance of pre-trained models on text summarization 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 the effectiveness of pre-trained models for summarizing new proceedings.

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 (Bidirectional Encoder Representations from Transformers) 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 (Bidirectional Encoder Representations from Transformers) 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 (Generative Pre-trained Transformer 2) 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 (Text-To-Text Transfer Transformer) is a multi-task learning model, meaning it can be used for multiple tasks. T5 is able to use transfer learning to adapt to new tasks, making it easier to fine-tune and adapt to different tasks. Thus the study is to investigate the effectiveness of pre-trained models when applied to

Old Baileyproceedings 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.

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

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

Plan to conduct the research

The plan for this study will involve a quantitative approach to assess the effectiveness of pre-trained models 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 pre-trained models will be trained on the collected dataset.

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 effectiveness of pre-trained models for text summarization of new proceedings.

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															

Dataset description

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

The Old Bailey Online XML data is a huge dataset. Under the terms of a Creative Commons Non-Commercial license, all of the project's data is available for re-use (CC-BY-NC). The Old Bailey Application Programming Interface (OBAPI) gives users the ability to interact directly with the text of individual trials as well as complete sessions that have been published as a part of the Proceedings. Only the trials themselves are accessible via this method; front matter, ads, and the Ordinary's Accounts are not included. It is not feasible to download more than ten full trials at a time using a single query. The complete data is available in XML format only.

The dataset includes some documentation on the data structure and variables, as well as 2,163 editions of the Proceedings and 475 Ordinary's Accounts that have been marked up using TEI-XML. Additionally, the dataset includes all of the editions. Each file in the Proceedings section represents one session of the court that took place between 1674 and 1913, and each file in the Ordinary's Account section represents a single pamphlet (1676-1772).

References

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.

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

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.