**Project Data Management Plan**

Project Overview

Project Title: **Traditional Vs. Transformer-Based Models in NLP: A Comparative Study on Extractive Text Summarization**

Project Summary:

In this era of big data, summarizing information efficiently has become more crucial than ever. With the sheer volume of data generated daily across the web, news outlets, social media, and various industries, it's increasingly difficult to extract meaningful insights without some form of automated processing. One of the core tasks in natural language processing (NLP) is text summarisation, which is of two types: extractive and abstractive summarization.

Extractive text summarization entails selecting the most informative sentences from a given text to produce a concise summary whereas in abstractive summarization new sentences are generated to give the most critical information. While more recent approaches to summarisation make use of deep learning architectures, traditional approaches frequently rely on statistical techniques and feature engineering.

Word2Vec is a popular technique for learning distributed representations of words. It has been successfully applied to various NLP tasks, including text summarization. BERT, a transformer-based model pre-trained on a massive amount of text data, has demonstrated state-of-the-art performance on a wide range of NLP tasks. This project aims to conduct a comprehensive comparison of Word2Vec and BERT for extractive text summarization, considering factors such as model architecture, training data, and evaluation metrics. By understanding the strengths and weaknesses of these models, we can contribute to the development of more effective and efficient text summarization techniques.

Research Question:

How does the performance of traditional word embedding models like Word2Vec compare to state-of-the-art transformer-based models like BERT for extractive text summarization?

Project Objectives:

* **Extractive Text Summarization using Traditional Models**:

Implement and fine tune extractive summarization model using traditional word embedding technique (Word2Vec) on the CNN/Daily Mail dataset.

* **Extractive Text Summarization using Transformer-Based Models**:

Implement and fine tune extractive summarization using transformer-based model, such as BERT, on the same dataset to facilitate direct comparison.

* **Evaluate and Compare Model Performance:**

Compare the accuracy, relevance, and coherence of summaries generated by Word2Vec and BERT using evaluation metrics such as ROUGE, BLEU etc.

* **Analyse Trade-offs in Model Efficiency and Complexity:**

Analyse the computational trade-offs between traditional and transformer-based models in terms of training time and computational complexity.

References:

* Rajesh, B., Chaitanya, K.N., Govardhan, P.T., Mahesh, K.K. & Sudarshan, B., 2024. Text Summarization using NLP. *International Research Journal of Engineering and Technology (IRJET)*, 11(03), pp.702. Available at: <https://www.irjet.net/archives/V11/i3/IRJET-V11I392.pdf>
* Abdel-Salam, S. & Rafea, A., 2022. Performance Study on Extractive Text Summarization Using BERT Models. *Information*, 13(2), p.67. Available at: <https://doi.org/10.3390/info13020067>
* Karjule, V., Dange, J., Thange, S. & Kokate, N., 2023. A Survey on Text Summarization Techniques. *Journal of Natural Language Processing*. Available at: <https://www.researchgate.net/publication/375120640_A_Survey_on_Text_Summarization_Techniques>

Project Timeline

Data Management Plan

Dataset overview:

The CNN/DailyMail Dataset is an extensive collection of over 300,000 unique English-language news articles sourced from CNN and the Daily Mail. The dataset was originally created between April 2007 and April 2015, with articles collected through automated web scraping techniques. The dataset was curated by researchers from Google DeepMind and modified for summarization by teams from IBM Watson, Université de Montréal, and Stanford University. Researchers aimed to support machine reading comprehension and question answering tasks, but subsequent versions have shifted focus to text summarization tasks, allowing for both extractive and abstractive summarization applications.

Data Collection:

The data was originally collected by Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom of Google DeepMind, with Tomáš Kočiský and Phil Blunsom also affiliated with the University of Oxford.For the purpose of this project , the dataset with version 1.0.0 was collected from HuggingFace.

Link to the dataset: <https://huggingface.co/datasets/abisee/cnn_dailymail>

Metadata:

The data is in JSON (JavaScript Object Notation) format. The data set is split into test, train and validation data:

* **Training Set**: 95,705 instances (size: 244MB)
* **Test Set**: 11,490 instances (size: 28.6MB)
* **Validation Set**: 13,368 instances (size: 33MB)

|  |  |
| --- | --- |
| Data Fields | Data Structure |
| Id | A unique identifier (SHA1 hash of the article URL) |
| Article | The full text of the news article |
| Highlights | The summary of the article, usually a single sentence or a few key points |

Average token count of the articles and highlights:

| **Feature** | **Mean Token Count** |
| --- | --- |
| Article | 781 |
| Highlights | 56 |

Document Control:

* Files within the dataset will follow a standardized naming format to ensure clarity and consistency.
* Version Control: Version control will be maintained using Git, enabling tracking of changes, collaboration, and easy access to previous iterations of the dataset.
* GitHub Repository: [Link to the Repository](https://github.com/sandrabinu3/Text-Summarization-NLP-Project). The GitHub repository is linked to Google Colab, allowing for seamless access to the dataset directly from Colab notebooks.

Readme file:

The ReadMe file will include a comprehensive overview of the project, information about the dataset, usage guidelines and instructions of the coding environment, dependencies and model details.

Security & storage:

Daily backup of data files to ensure data integrity. Secure OneDrive storage and public access limited in GitHub repository.

Ethical Requirements:

The dataset includes anonymization measures that align with GDPR requirements, ensuring ethical use in research. Proper licensing and adherence to institutional policies further enhance its suitability for academic studies. The dataset is released under the **Apache-2.0 License**. permitting academic and commercial use.