CSCE 5290: Natural Language Processing

Project Proposal

Title: Summarization and Classification on WikiText Dataset

GROUP-3

1. Motivation

The growth of text-based data is rapid, especially that of large and structured datasets such as the WikiText. Summarization and classification are turning out to be very big challenges. In fact, the principal challenge remains in the development of automatic systems capable of compressing enormous amounts of text into compact summaries while retaining crucial information for classification applications. The project at hand is designed to study and understand how different summarization models impact downstream tasks-specially text classification.

As a next step, we here investigate how much of the essential information for correct categorization has remained behind by first summarizing the text and then using those summaries for classification. The scope of the current project is to investigate the performance and accuracy of different summarization models and verify their influence with respect to text classification performance.

2. Significance

It is necessary to understand in the view of the fact that it represents two of the most important topics in NLP: summarization and classification. Although these tasks are done separately at most times, how summarization influences performance in classification is an unexplored and unique domain. This research has potential applications regarding news summarization, analysis of legal documents, sentiment analysis, among others, where a quick yet quality decision has to be made based on large volumes of text.

This would reveal how summarization models retain most of the important information necessary for good classification and thus might result in a better performance in many realistic settings.

3. Objectives

• The WikiText dataset can be used to perform a comparison between two or three summarization models including BART, T5, and Pegasus.

• Inference on the quality of summarization models based on ROUGE and BLEU metrics.

• Classifiers like BERT and XGBoost will be used by feeding the summarized output as an input for the classifier. Performances of the models can then be compared.

• The accuracy, precision, and recall of the models can be used as comparison metrics in order to show the effect of summarization on classification.

Success Criteria:

The ROUGE and BLEU scores will be higher for successful summarization models.

The performance in terms of the original text input vs summarized text input will be measured based on its accuracy, precision, and recall. The minimum drop in performance will signify that summarization for classification purposes has been successful.

4. Features

Summarization: This summarization work can be taken over by BART, T5, or Pegasus, which have inducted their knowledge from the WikiText dataset.

Classification: Use the summarized text as input to classification tasks using models such as BERT and XGBoost.

Comparison: Complete comparative detail in the performance of all models on summarization and classification.

Evaluation: Evaluation metrics are used for better performance of the model and to improve the model performance for summarization, and accuracy, precision, recall for classification.

Milestones:

- Week 1-2: Dataset gathering, preprocessing and exploration.
- Week 3: executing and testing multiple summarization models.
- Week 4: Evaluating summarization output using metrics.
- Week 5: Executing classification models and feeding in summarized text.
- Week 6: Comparing end results and evaluating results for any improvements of model performance.

5. Dataset

Link: https://www.kaggle.com/datasets/rohitgr/wikitext

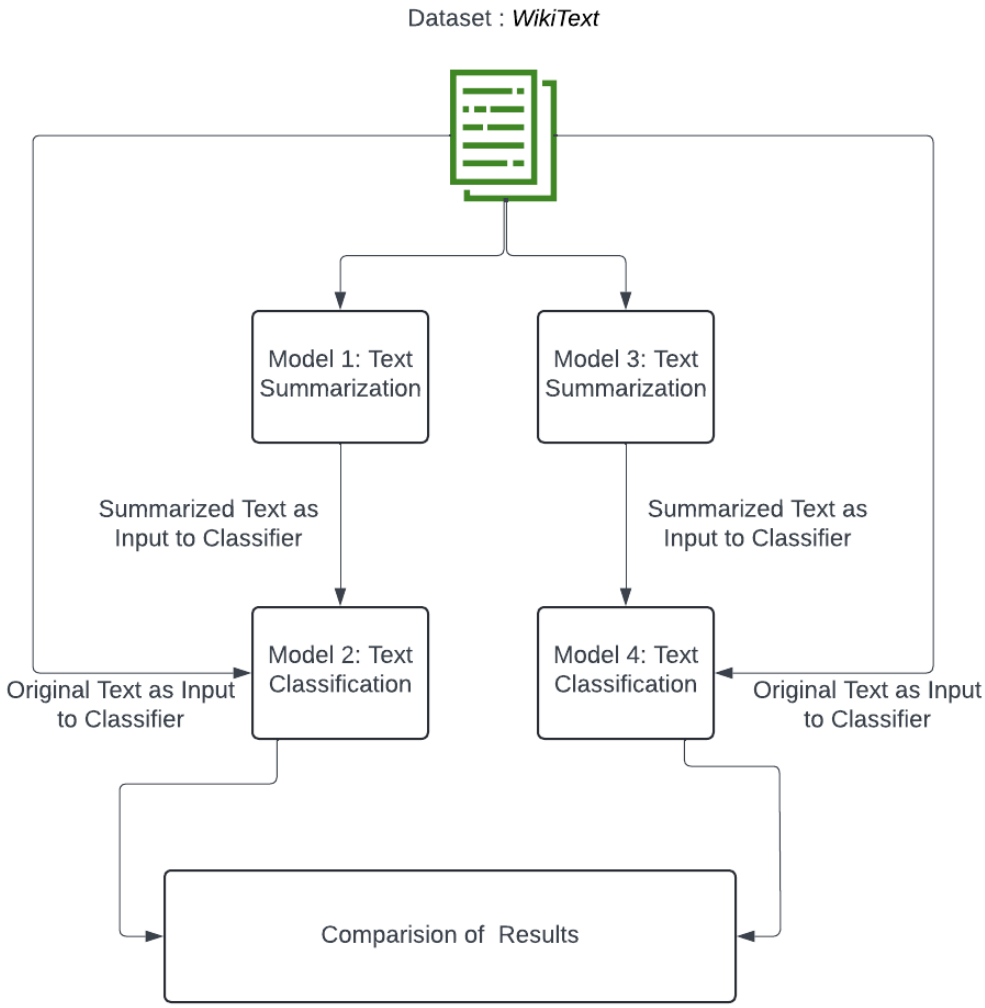
The WikiText dataset is the used in our project, which has over 100 million tokens taken out from Wikipedia articles. The data is best for language modeling and summarization tasks because of its size and complexity. It may require techniques like tokenization, and preprocessing steps like cleaning for any formatting inconstancy. And also, some components may have to be decreased in size for better compatible with the models for precise training and testing.

The WikiText dataset after doing the summarization task then will perform classification ,categorize may be friction or non friction ,writing styles formal or informal and topics like historical or scientific and the dataset is too huge and we may use subsets of the dataset or like specific ones like articles related to particular time period or author etc.

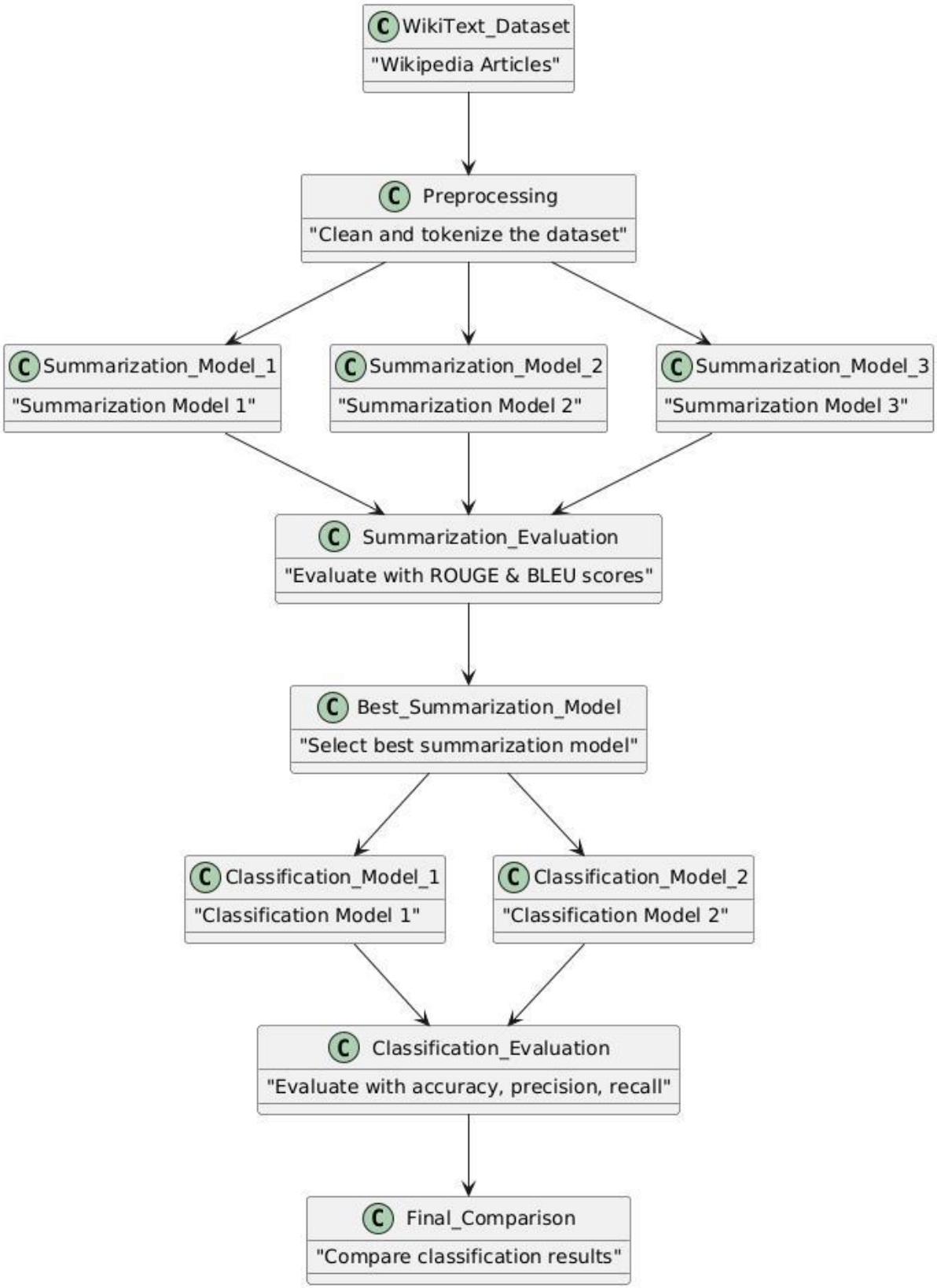
- **Data Types**: The data is well-formatted textual data extracted from Wikipedia articles, providing a mix of long-form content and structured information ,headings.
- **Preprocessing**: Preprocessing may involves tokenization, take of unnecessary symbols or formatting and removing or reducing bigger sections to fit into summarization model input constraints. And also stop words removal and splitting the dataset into train and test and validation .

Feature	Description
Dataset Name	WikiText- 103
Type	Text based (Historical and Scientific context)
Source	Wikipedia
Size	100 million tokens
Number of documents	28,000
Language	English
Format	Plain Text

6. Visualization



A) Proposed Plan of The Model



b) Detailed View of the Model

GitHub Link: https://github.com/Ajaysimha29/Nlp_project_3