CSCE 5290: Natural Language Processing

Project Proposal

Title: Enhancing Scientific literature exploration using unified categorization and multimodal summarization

1. Motivation

Being a researcher, it's really a challenging task to extract information from vast amounts of literature, find most relevant scientific papers and prepare a concise summary to understand and refer to. The motivation for this project is to address the aforementioned challenge by efficiently navigating through enormous amounts of text data and provide actionable insights leveraging advanced NLP techniques, enhancing exploration and understanding for researchers.

2. Significance

Efficiently categorizing research articles just by using abstracts and generating effective multimodal summaries from the lengthy scientific documents, provides relevant and comprehensive insights to researchers. This approach not only saves time and provides convenience, but also empowers researchers to efficiently utilize their time and delve deeper in most relevant research resulting in acquired knowledge and new contributions, which makes it significant.

3. Objectives

The first goal for the project is to implement and train a model for abstract based classification, where the model accurately predicts the category of scientific papers.

The next goal is to train a model (BERT/BART/T5) to perform PDF based multimodal summarization while putting identified categories into context.

The step would be to integrate both the models and produce category based concise summaries for scientific articles.

To determine success, the plan is to follow both quantitative and qualitative metrics. On the quantitative front, for classification, accuracy, precision and recall are few metrics which will be used. For summarization, relevance using ROUGE and BLEU scores will be referred to. Evaluation using a qualitative metric is complex and requires domain experts to subjectively evaluate the relevance of generated summaries within specific categories. Few summary samples can be evaluated by colleagues for determining the quality.

4. Features

Technical characteristics involve

- Utilizing pre-trained efficient models in NLP, customized through transfer learning.
- For classification which uses scientific paper abstract, plan is to use Bidirectional Encoder Representations from Transformers (BERT) model.
- For summarization using pdfs of scientific papers, the plan is to compare Bidirectional and Auto-Regressive Transformers (BART) and Text-to-Text Transfer Transformer (T5) and use the one which produces a high relevance score.
- Efficient preprocessing of arXiv Dataset with 1.7 million articles using different NLP libraries is also a must to produce better results.

From deliverables aspect

- Pre-trained classification and summarization models must be finetuned for specific requirements of abstract based classification and multimodal pdf summarization, producing high accuracy and relevance scores.
- Summarized output must be based on the given context of the predicted category from the classification model and SME feedback must be considered for a few sample summaries.

Milestones are same as 3 goals mentioned in objectives, which are:

- Milestone-1: finetune model-1
- Milestone-2: finetune and finalize model-2
- Milestone-3: integration of both models

Working with such a large corpus with multimodal summarization while choosing a model which performs better, adding qualitative feedback are few of **distinctive** elements. These points will definitely help to produce better category based multimodal summary which helps provide better understanding to researchers with least efforts possible, for example large corpus helps classify better while combating overfitting, and multimodal summarization considers not only text but the complete information from the article.

5. Dataset.

The dataset of scientific research papers: **arXiv Dataset**

Size: 1.7 million articles

Type: dataset has scholarly articles, from the different fields such as physics, computer science, math, statistics, electrical engineering, quantitative biology, and economics. Data repository has both metadata and full-text PDFs.

Sources: freely available via Google Cloud Storage buckets, maintained and operated by Cornell University

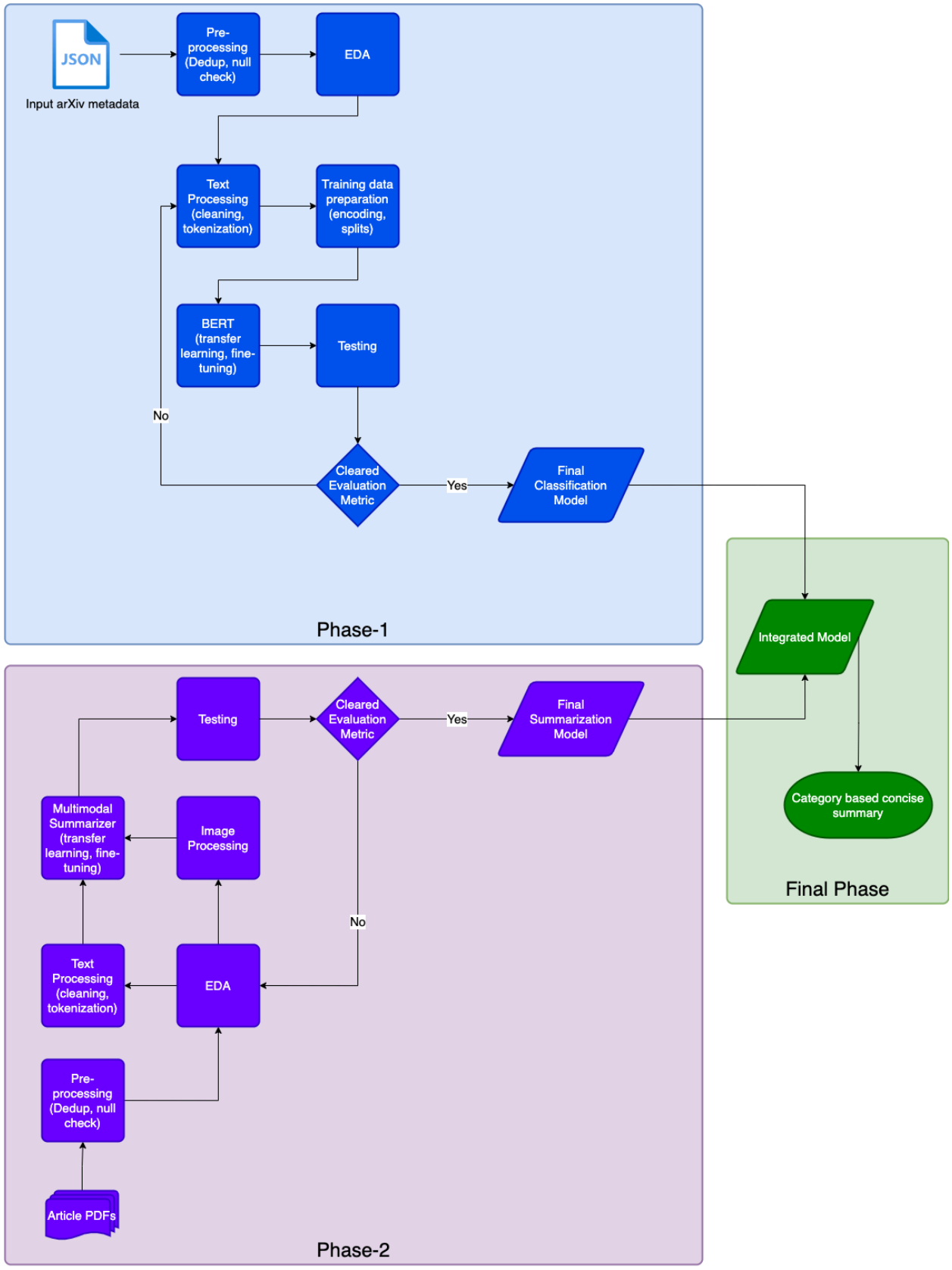
Dataset details: Metadata has below columns: (details provided where required)

- Id, submitter, authors, title, abstract, categories, versions
- comments: Additional info, such as number of pages and figures,
- journal-ref: Information about the journal the paper was published in
- In addition to metadata, each paper pdf can directly be accessed.

Generic outline for **pre-processing** requirement: (actual list can have more steps as required during implementation phase)

- Text cleaning and tokenization
- Encoding for category label
- Training on small subsets
- Train - test splits
- Handling PDFs for summarization

6. Visualization



Project Flow Diagram

Above project flow diagram follows the overall comments covered in different sections. Inputs are covered in dataset sections and 3 phases represent the 3 milestones mentioned. Rest all points are already covered in pre-processing and feature sections. Additionally, these are the initial thought and design, approach and models might improve with learning and experimentation.

GitHub Link:

SharmaAshwini/NLP Project