Course Info: **CSCE 5290 – Natural Language Processing (Fall’23)**

Group: *Bharath Narayana Juthuka Srinivasa (11603312)*
Sanjeev Reddy Siripinane (11656848)
Anirudha Kulakarni Karanam (11656382)

GitHub Link: *https://github.com/Anirudh-kulakarni-a11/CSCE-5290-Project.git*

Dataset: **PMC Open Access Subset**

Proposal: **Scientific Paper Categorization and Key Point Extraction**

PROJECT PROPOSAL

Contents

Motivation:

Many scientific journals/white papers and articles are written and published every single day. Although this is a good thing, the amount of data being created by these is massive. And due to this reason, it can be hard for people to read or index and understand all of them. And that is why we need tools that can summarize these papers/journals/articles and give people what they need which is the summarized key takeaways from the content. This way, readers can quickly understand what the paper is referring to or is about. Since the world of scientific literature is expanding at a massive rate and there is clear indication of growth and advancements in science and technology, it puts us in a tricky position to deal with and causes problems for researchers, students, and professionals.

Filtering through this heap of data to find relevant research papers can be time-consuming and energy-draining. So, there is a good need for tools or processes that can categorize scientific articles and summarize them in a concise manner, allowing users to quickly understand the context and key points of a paper, article, or journal.

Significance:

Efficiency: Automated categorization and key point extraction will drastically reduce the time researchers spend on literature reviews.

Accessibility: By categorizing papers, researchers can directly access papers that are most relevant to their domain of interest.

Informed Decision Making: Summaries and key points will enable researchers to make quick decisions on the relevance of a paper without delving deep into it.

Open-Source Contribution: Our work aims to provide a comprehensive mechanism/process for the vast PMC dataset, which could be a significant contribution to the open-source and research community.

Objectives:

Data Processing: Parse and preprocess the PMC dataset to make it suitable for NLP tasks.

Model Development: Create and train models for both categorizing scientific papers and extracting their key points or summaries.

Evaluation: Establish metrics and evaluate the performance of the models to ensure accuracy, reliability.

Features:

Dynamic Input System: A platform where users can upload scientific papers in text format.
Category Prediction: Once a paper's content is uploaded, the system will predict its category.
Key Point Extraction: Along with categorization, the tool or model will extract and display key points or findings from the papers.
Advanced Search: Users can search for papers based on categories, keywords, authors, or even specific phrases from key points.

Deliverables:

- A ready-to-use web based or computer-based tool.
- A report on how we did the project and our results.

Uniqueness:

Our project is special because it uses new NLP methods to understand big science journals or papers or articles. We focus on giving quick key-points or summaries from long papers. This is not common in many tools today.

Feedback Mechanism:

In the computer-based or web-based tool, we will add a way for users to give feedback. This feedback will tell us if the tool is good or if we need to make it better.

Milestones:

- Weeks 1-3: Prepare the data and understand it.
- Weeks 4-7: Build the NLP models.
- Weeks 8-10: Make the computer tool and test it.

Dataset Description:

Source: The dataset we have chosen is from the National Center for Biotechnology Information (NCBI) and is known as the PubMed Central Open Access Subset (PMC).

Link: PubMed Central Open Access Subset

Nature of Data:

Format: The dataset contains structured information about each scientific paper and is stored in the format of XML files.

Content: Each XML file holds information about the paper's title, abstract, content, authors, journal name.

Relevance to NLP:

The dataset is rich in textual information, making it ideal for Natural Language Processing (NLP) tasks. The diverse range of topics covered in the papers provides a broad scope for NLP techniques, from basic text processing to advanced model training.

Size and Volume:

The PMC dataset is vast, containing a significant number of scientific articles from various biomedical and life sciences domains. For the initial stages of our project, we may consider working with a subset to streamline our processes before scaling to the entire collection.

Use in the Project:

Categorization: We'll use the content of the papers to train our NLP models to group them into various categories.

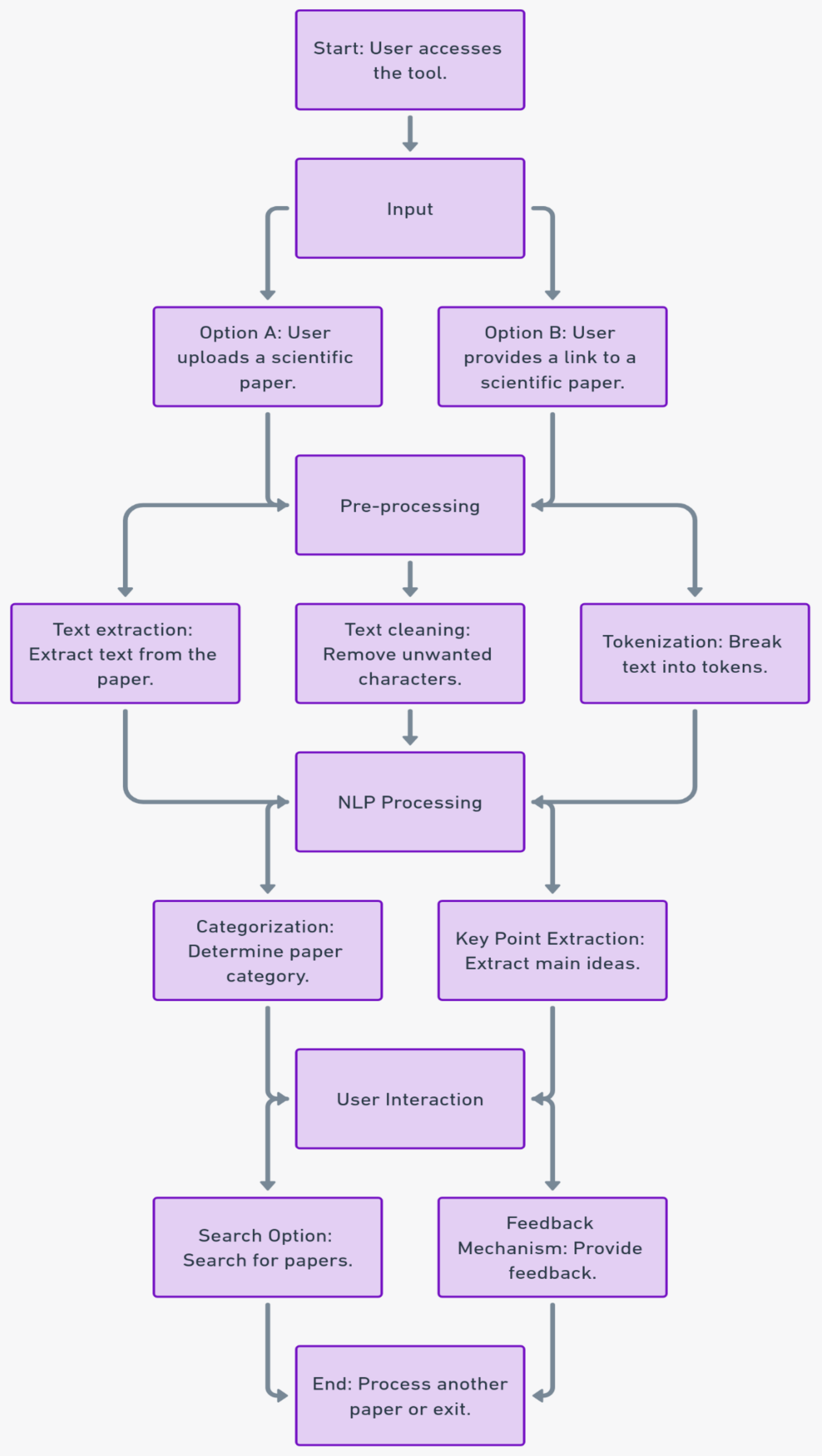
Key Point Extraction: The abstracts and main content of the papers will be essential for the NLP models to extract the main ideas or highlights.

Challenges:

Data Cleaning: Given the structured nature of XML, parsing the data to extract relevant information will be our first challenge. We need to ensure that we filter out any unnecessary details.

Diversity of Topics: The wide range of topics in the dataset can pose challenges in defining clear categories or in understanding the context for key point extraction.

User and Process Workflow – Visualization:


Start: User accesses the tool.
Input
Option A: User uploads a scientific paper.
Option B: User provides a link to a scientific paper.
Pre-processing
Text extraction: Extract text from the paper.
Text cleaning: Remove unwanted characters.
Tokenization: Break text into tokens.
NLP Processing
Categorization: Determine paper category.
Key Point Extraction: Extract main ideas.
User Interaction
Search Option: Search for papers.
Feedback Mechanism: Provide feedback.
End: Process another paper or exit.

References:

- NCBI. (2021). *PubMed Central Open Access Subset. It provides a comprehensive collection of biomedical and life sciences journal literature. This dataset will serve as the foundation of our project. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/*
- *Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. This toolkit will be instrumental in the preprocessing and initial stages of the project. It offers tools and Python libraries for processing and analyzing human language data.*
- *Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. BERT might play a crucial role in our model development, especially for key point extraction.*
- *Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. This paper provides insights into using pretrained models for extractive summarization, a technique we plan to explore for key point extraction.*