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

Instructor: **Dr. Zeenat Tariq (*Zeenat.Tariq@unt.edu*)**

Dataset: **PMC Open Access Subset**

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

PROJECT PROPOSAL

Contents

Motivation 2
Significance 2
Objectives 2
Features 3
Dataset Description 3
References 4

Motivation:

Many science papers are written every day. This is good but can be hard for people to read all of them. We need tools that can group these papers and tell us the key points. This way, readers can quickly understand what the paper is referring to or is about. Since the realm of scientific literature is expanding at an unprecedented rate and the growth is indicative of the rapid advancements in science and technology, it also presents a daunting challenge for researchers, students, and professionals.

Sifting through vast amounts of data to find relevant research papers can be time-consuming and overwhelming. There is a pressing need for automated tools or processes in place that can categorize scientific articles and distill their essence in a concise manner, allowing users to swiftly gauge the relevance and significance of a paper or article.

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.

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

Interface Development: Design a user-friendly interface to make the tool accessible to researchers and professionals.

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.

Dataset Description:

Source:

The dataset we will be using for this project comes 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 is provided in XML format, which contains structured information about each scientific paper.

Content: Each XML file typically holds data about a paper's title, abstract, content, authors, journal name, and other metadata.

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 different 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.

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.*