**README**

**Project Overview**

This project applies natural language processing (NLP) techniques to classify and analyze research papers in the fields of virology and epidemiology. The main goals are to filter out irrelevant papers, classify relevant ones based on the types of deep learning methods used, and extract specific method names from the papers. This project leverages transformer-based models, such as DistilBERT and DeBERTa, to efficiently handle text embeddings and classification tasks.

**Objectives**

This notebook addresses the following questions:

1. **Filter Relevant Papers**: Identify papers that employ deep learning approaches within virology or epidemiology using semantic filtering methods.
2. **Classify by Method Type**: Classify each relevant paper into one of the following method categories:
   * text mining
   * computer vision
   * both
   * other
3. **Extract Method Names**: For each relevant paper, extract the specific deep learning method mentioned.

**Methodology**

**1. Filtering Relevant Papers**

To filter out papers that do not meet the project criteria, the notebook uses **DistilBERT** embeddings for semantic representation. DistilBERT, a lighter and faster variant of BERT, captures contextual information from text and key phrases. By calculating similarity between embeddings, the notebook filters out papers that lack content related to deep learning applications in virology or epidemiology.

**2. Classifying Papers by Method Type**

For classification of the relevant papers, this project utilizes **DeBERTa (small long)** with a **zero-shot classification** approach using natural language inference (NLI). DeBERTa is used to predict which of the four categories—text mining, computer vision, both, or other—best fits each paper’s methods based on its content.

**3. Extracting Deep Learning Method Names**

Once classified, the notebook identifies and extracts the names of specific deep learning methods employed in each paper. This step provides insights into the specific techniques and models utilized in virology and epidemiology research.

**Project Structure**

* **Data Loading and Preprocessing**: Loads and preprocesses the input dataset of research papers.
* **Semantic Filtering with DistilBERT**: Embedding generation and similarity calculations to filter papers.
* **Classification with DeBERTa**: Uses zero-shot NLI-based classification to categorize each paper.
* **Method Extraction**: Extracts and reports specific deep learning methods used in relevant papers.

**Requirements**

This project requires Python 3.7+ and the following libraries:

* transformers (for model embeddings and classification)
* scikit-learn (for handling similarity measurements)
* pandas (for data manipulation)
* numpy (for numerical operations)

**Usage**

1. **Clone or Download** the notebook and data file.
2. **Install Dependencies**: Run the following command to install the required packages:

bash

Copy code

pip install transformers scikit-learn pandas numpy

1. **Run the Notebook**: Open and execute each cell in the notebook to replicate the filtering, classification, and extraction process.

**Results**

The output includes:

* A filtered dataset of papers related to deep learning in virology/epidemiology.
* Classification labels (text mining, computer vision, both, other) for each relevant paper.
* The specific deep learning methods mentioned in each paper, extracted and organized for further analysis.