Replication Guide

This document provides step-by-step instructions to replicate the results reported in the study.

System Requirements

Ensure your system meets the following requirements:

- Python 3.8 or later
- GPU (recommended) or CPU with sufficient processing power
- At least 16GB of RAM for handling large datasets
- Internet connection to download dependencies

Install Dependencies

Run the following command to install all required dependencies:

- pip install pandas numpy torch transformers scikit-learn nltk

This will install libraries such as:

- transformers for BERT model
- scikit-learn for machine learning
- pandas and numpy for data processing
- torch for deep learning computations
- nltk for natural language processing

Download the Dataset

The dataset files are located in the datasets/ folder. Ensure that it contains the following CSV files:

- pytorch.csv
- tensorflow.csv
- keras.csv
- incubator-mxnet.csv
- caffe.csv

If any dataset is missing, you may need to obtain them from the original source.

Running the Tool

To run the tool, execute the following command:

- python main.py

The script will:

- Load and preprocess the selected dataset.
- Extract BERT embeddings from the text data.
- Train a logistic regression model with hyperparameter tuning.
- Evaluate the model across multiple experimental runs.
- Save results to a CSV file.

Reproducing the Results

To replicate the exact results:

Set the project name in main.py by modifying the line:

- project = 'caffe' # Change to desired dataset

Run main.py as described above.

The results will be saved in a CSV file at ../caffe_BERT_LogReg_results.csv (for the caffe dataset).

Troubleshooting

ModuleNotFoundError: Ensure all dependencies are installed correctly

CUDA Out of Memory: Reduce batch size or run on CPU (device = 'cpu' in main.py).

FileNotFoundError: Ensure datasets are placed in the datasets/ directory.