Requirements

Python Version

- Python 3.8 or higher

Required Libraries and Versions

- numpy>=1.20.0
- pandas>=1.3.0
- scikit-learn>=1.0.0
- scipy>=1.7.0
- joblib>=1.1.0
- matplotlib>=3.4.0
- seaborn>=0.11.0
- tqdm>=4.62.3
- statsmodels>=0.13.0

Text Processing Libraries

- nltk>=3.6.0
- gensim>=4.0.0
- spacy>=3.2.0
- regex>=2021.8.3
- en-core-web-sm (spaCy model)

Class Imbalance Handling

- imbalanced-learn>=0.8.0

Installation Instructions

1. Clone the repository using HTTPS:

```
git clone https://git.cs.bham.ac.uk/ceg212/Caleb_ISE_Coursework.git
```

Or using SSH:

```
git clone git@git.cs.bham.ac.uk:ceg212/Caleb_ISE_Coursework.git
```

2. You have two options for installing dependencies:

Option A: Using a virtual environment (recommended for isolation)

```
python -m venv venv
source venv/bin/activate  # On Windows: venv\Scripts\activate
pip install -r requirements.txt
```

Option B: Installing directly in your environment or IDE You can also install the dependencies directly in your Python environment or through your IDE's package manager using the requirements.txt file:

```
pip install -r requirements.txt
```

3. Download NLTK resources (required for text preprocessing):

```
python download_nltk_resources.py
```

4. Download spaCy model (required for the Hybrid model):

```
python -m spacy download en_core_web_sm
```

System Requirements

- At least 8GB of RAM is recommended for processing larger datasets and training models
- Approximately 1GB of disk space for the datasets, models, and outputs
- GPU is not required but can speed up model training

Project Structure

```
Caleb_ISE_Coursework/
├── lab1/                        # Main project directory
│   ├── __pycache__/             # Python cache files
│   ├── comprehensive_results/  # Detailed output from all model runs
│   │   └── plots/               # Framework-specific visualisations (F1 scores,
precision/recall)
│   ├── datasets/                # Dataset files
│   │   ├── caffe.csv
│   │   ├── incubator-mxnet.csv
│   │   ├── keras.csv
│   │   ├── pytorch.csv
│   │   └── tensorflow.csv
│   ├── plots/                   # Generated visualisation plots
│   │   ├── F1-Score_comparison.png
│   │   ├── model_comparison_metrics.png
│   │   ├── Precision_comparison.png
│   │   ├── Recall_comparison.png
│   │   ├── Training_Time_comparison.png
│   │   └── various framework-specific visualisations
│   ├── results/                 # Results from evaluations
│   │   ├── all_frameworks_multiple_runs.csv
│   │   ├── evaluation_summary.md
│   │   └── framework-specific result files
│   ├── test_results/            # Results from individual tests
│   ├── baseline_model.py        # Naive Bayes + TF-IDF implementation
│   ├── br_classification.py     # Bug report classification implementation
│   ├── bug_report_classifier.py # Main classifier implementation
│   ├── create_sample_dataset.py # Utility to create test datasets
│   ├── evaluation.py            # Evaluation metrics and visualisation
│   ├── evaluation_framework.py # Framework for comprehensive evaluation
│   ├── feature_extraction.py    # Feature extraction methods
│   ├── hybrid_model.py          # Ensemble model implementation
│   ├── intermediate_model.py    # SVM + Word2Vec implementation
│   ├── preprocessing.py         # Text preprocessing utilities
│   ├── test_all_models.py       # Script to test all models
│   ├── test_framework.py        # Script to test framework components
│   ├── test_hybrid_model.py     # Script to test hybrid model vs baseline
│   ├── test_intermediate_model.py # Script to test intermediate model vs baseline
│   ├── test_model_preprocessing.py # Preprocessing for model tests
│
├── download_nltk_resources.py  # Utility to download NLTK data
├── requirements.txt             # List of dependencies to download
├── requirements.pdf             # This file - installation instructions
├── manual.pdf                   # User manual for the tool
└── replication.pdf              # Instructions for replicating experiments
```