Text Classification Toolkit: User Manual

Machine Learning Project

March 27, 2025

1 Overview

This toolkit provides two text classification scripts:

- BERT-based Deep Learning Classification
- Naive Bayes Optimized Classification

2 BERT Classification Script (bert_classification.py)

2.1 Purpose

A deep learning approach for text classification using BERT (Bidirectional Encoder Representations from Transformers).

2.2 Input Requirements

- CSV file: `datasets/pytorch.csv`
- Required columns:
  - `Title`: Text document title
  - `Body`: Text document body
  - `class` or `related`: Binary label column

2.3 Usage

```
python bert_classification.py
```

3 Naive Bayes Classification Script (br_classification_optimized.py)

3.1 Purpose

A traditional machine learning approach using TF-IDF vectorization and Support Vector Machine classification.

3.2 Input Requirements

- CSV file: `datasets/pytorch.csv`
- Required columns: Same as BERT script

3.3 Usage

```
python br_classification_optimized.py
```

4 Common Preprocessing Steps

Both scripts perform similar text preprocessing:

- Convert text to lowercase
- Remove URLs
- Remove punctuation
- Tokenization
- Remove stopwords

5 Output

Scripts will print evaluation metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- AUC (Area Under ROC Curve)

6 Troubleshooting

- Ensure all dependencies are installed
- Check dataset format and column names
- Verify file paths