# Replication Guide: Bug Report Classification Tool

## System Requirements

To ensure that all libraries and dependencies are downloaded, install Python 3.x and enter this line of code into the interpreter terminal:

- pip install gensim==4.3.3 nltk==3.9.1 numpy==1.26.4 pandas==2.2.3 scipy==1.13.1 scikit-learn==1.6.1 gdown==5.2.0

## Dataset

To replicate the reported results, ensure that the project variable in line 78 contains the value 'combined_dataset'.

## Reproducibility: Controlled vs Random Conditions

To replicate the results of the paired t-test, set the random_state parameter in line 216 to the value repeated_time. This makes sure that the train/test splits are fixed and consistent.

It isn't possible to exactly replicate the results of the unpaired t-test as the train/test splits are randomly generated and not fixed. To achieve this, omit the random_state parameter in line 216.

## Running the Tool

Execute the script by entering this line of code into the interpreter terminal:

- python proposed_solution.py

## Expected Outputs

The script will perform the following:

- Preprocess the data
- Train a Word2Vec Model using the preprocessed data
- The first time the program is run, the Google News Word2Vec model embeddings will be downloaded using the gdown package.
- Perform a grid search to optimise hyperparameters and find the best set of word embeddings for the logistic regression model
- Train and evaluate the model over 20 iterations
- Display performance metrics in the terminal
- Save results to a CSV file called "combined_dataset_LR+W2V.csv"