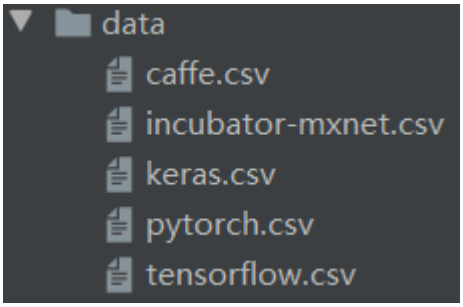
Purpose:

This is a tool used for bug reports classification.

How to use:

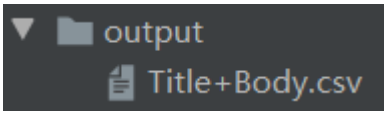
1. The data directory

The data directory contains five datasets. You can also add your own datasets to this directory and use the tool to classify.



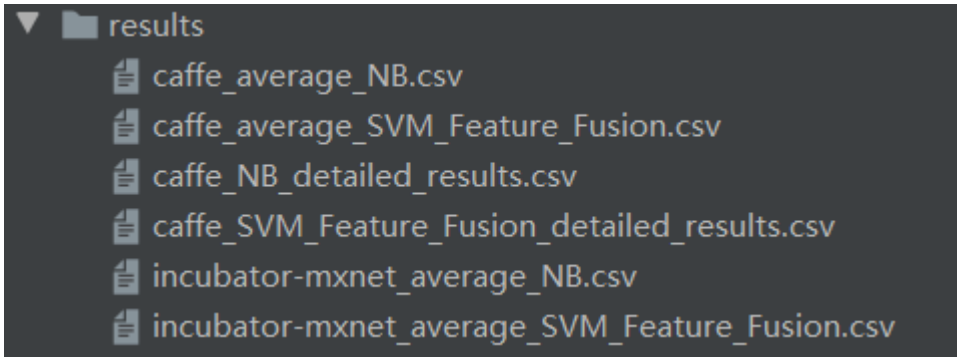
2. The output directory

The output directory contains new file formed by extracting key information columns from the dataset. If the file name (Title+Body.csv) exists, running the code again will overwrite the contents of the original file.



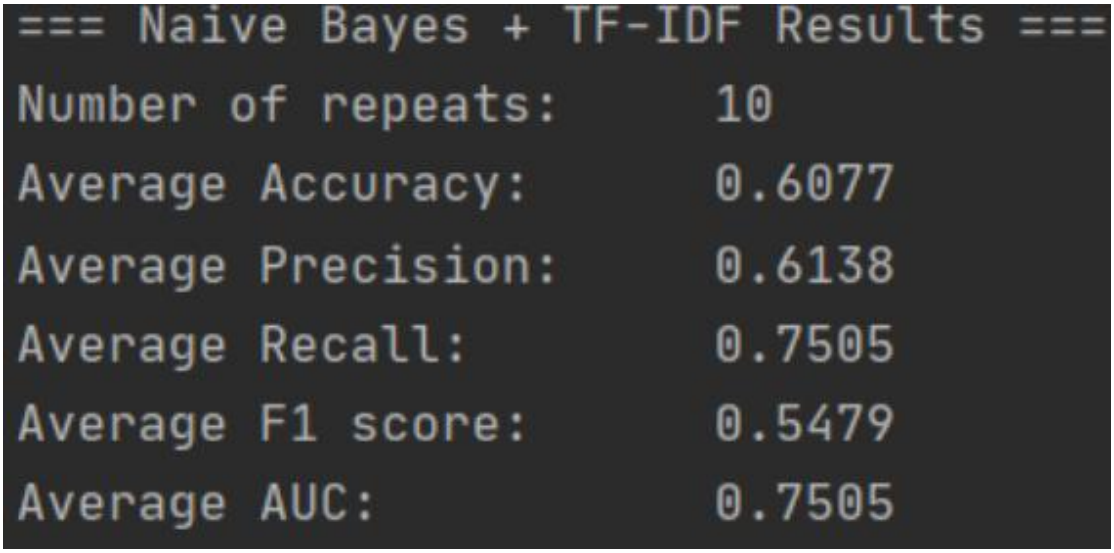
3. The results directory

①The results directory contains many files about the results of the experimental runs. The existing files are the data I show in the report.
②The first word of the file name is the name of the dataset project.
③The file with the word NB in the name is the result of the baseline method.
④The file with the word SVM_Feature_Fusion in the name is the result of the method explained in report.
⑤The file with the word average in the name is the average of the model performance of the ten runs. For this type of file, if the name of this file exists, the new data will be appended to the original data when the code is rerun.
⑥The file with the word detailed in the name is the result data of each of the ten runs. For this type of file, if the name of this file exists, the new data will overwrite the original data when the code is rerun.



4. The baseline.py file

The baseline.py file is a file of the baseline method. Once this file is run, the dataset is classified using the baseline method. You can modify the value of the project variable in line 81 to the name of the target dataset. When the code is run, the machine will classify this target dataset. Each time this file is run, three files will be generated, namely the Title+Body.csv file in the output directory, and the other two files are the average performance results of ten experiments in the results directory and the performance results of each experiment in the results directory. This picture below is an example of the console output results.



5. The coursework.py file

The coursework.py file is a file that uses proposed method in the report. Once this file is run, the dataset is classified using the method proposed in the report. The value of the project variable in line 93 can be modified to the name of the target data set. When the code is run, the machine will classify the target dataset. Each time this file is run, three files will be generated, namely the Title+Body.csv file in the output directory, and the other two files are the average performance results of ten experiments in the results directory and the performance results of each experiment in the results directory. This picture below is an example of the console output results

```
Number of repeats:      10
Average Accuracy:       0.8586
Average Precision:      0.6568
Average Recall:         0.5924
Average F1 score:       0.5916
Average AUC:            0.7532
```

6. The statistical tests.py file

The statistical tests.py file performs statistical tests on the results of ten experiments of the two methods. You can modify the value of the project variable in line 5 to the name of the target dataset, so that each time you run the code, it is for the target dataset. To run this file, you need to ensure that the detailed result files of the two methods on the target dataset exist. After running, it will output whether our method has significantly improved the performance of each indicator compared to the baseline method. This picture below is an example of the console output results.

```
Metric: Accuracy
Test used: Paired t-test
Test statistic: 7.216309165331452
p-value: 4.994388813086891e-05
Result: Significant difference!
------
```