Replication

Make sure you have Python 3.6+ installed and the libraries from requirements

```
pip install pandas numpy scikit-learn nltk
```

Place your dataset in:

```
/datasets/{dataset_name}.csv
```

How to run:

1. Open the script - `bra_class.py`
2. Set the `project = dataset_name e.g. pytorch`
3. Run the script:

```
python bra_class.py
```

Each classifier (Naive Bayes, Decision Tree, Random Forest) will be run REPEAT (default 10) times across:

- Original TF-IDF
- Improved TF-IDF
- Enhanced TF-IDF

Results will be automatically saved to:

```
/results/mean/
```

```
/results/raw/{project}/
```

Where mean contains the mean results over REPEAT runs and raw contains every run measurement. Results will also be printed to the console.

Additional Experimentation:

- Can change parameters inside
- `nb_params,`
- `dt_params,`
- `rf_params,`
- `functiobs original_tfidf,`
- `improved_tfidf,`
- `enchanced_improved_tfidf`

Changing said parameters will have marginally different results depending on the changes.

For a larger dataset, increasing the max features would potentially improve most of the results (decreasing the time). The opposite happens for a smaller dataset.

To try and compare my results with yours:

- Use keras.csv first, then use tensorflow.csv, one is a much larger dataset

Additional Notes:

- Preprocessing includes HTML/emoji removal, stopword filtering, and text cleaning.
- GridSearchCV is used for hyperparameter tuning for all models except TF-IDF, where configurations are manually varied.
- Make sure you have write permissions to the `results/` folder.