Project Structure

```
.
└── AML_project/
    └── project/
        ├── utils/
        │   ├── utils.py
        │   ├── spelling.py
        │   └── ...
        ├── tokenizers/
        │   ├── fitted_multilabel_tokenizer_lstm
        │   └── ...
        ├── predictions/
        │   ├── lstm_multi.csv
        │   └── ...
        ├── models/
        │   ├── lstm_multi.h5
        │   └── ...
        ├── data/
        │   ├── glove
        │   ├── fasttext
        │   ├── train.csv
        │   ├── test.csv
        │   ├── test_labels.csv
        │   └── ...
        ├── preprocessing.ipynb
        ├── ...
        └── 2_phases_classification.ipynb
```

- **project**: contains all the notebooks and files necessary to their execution
- **utils:** contains all python scripts with auxiliary functions
- **tokenizers:** contains tokenizers used in each model
- **predictions:** contains saved predictions on test set
- **models:** contains all the trained models
- **data:**
  - contains original train, test and target labels and also the processed ones
  - **glove:** contains glove pretrained embedding vectors
  - **fasttext:** contains fasttext pretrained embedding vectors
- ***.ipynb***:*

    - **preprocessing:** used for preprocessing of text and performing data augmentation
    - **data_exploration:** used for analyzing the dataset
    - **compare_results:** used to compare results coming from different models
    - ***_binary:** those notebooks contain the development and training of models for the binary classification task (toxic / not toxic)
    - ***_multilabel:** those notebooks contain the development and training of models for the multilabel classification task (toxicity types)
    - **2_phases_classification:** contains the combination of binary and multilabel models

Datasets

The main dataset used is the one provided by the kaggle challenge available at https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data.

- **train.csv** (training dataset):
    - **id**: comment identifier
    - **comment_text**: raw english text of a comment
    - **toxic**: binary target label
    - **severe_toxic**: binary target label
    - **obscene**: binary target label
    - **threat**: binary target label
    - **insult**: binary target label
    - **identity_hate**: binary target label
- **test.csv** (test dataset):
    - **id**: comment identifier
    - **comment_text**: raw english text of a comment
- **test_labels.csv** (ground truth for test set examples):
    - **id**: comment identifier
    - **toxic**: binary target label
    - **severe_toxic**: binary target label
    - **obscene**: binary target label

    - **threat**: binary target label
    - **insult**: binary target label
    - **identity_hate**: binary target label

Back translations of the original datasets are available at https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/48038.

How to run the code

Inside each of the notebooks there is an initial “*Configuration”* block which contains paths to the base directories.

To run the code we recommend to follow these steps:

1. Add the shared folder to your drive as a shortcut
    a. Right click on the folder name
    b. Select **Add shortcut to Drive**
2. Change the BASE_DIR variable in the “*Configuration”* block to reflect the path where you choose to add the shortcut
    a. **NB**: if you choose to add the shortcut to the root of your Drive this second step shouldn’t be necessary