AML - Challenge 3, Text-based sentiment analysis

Challenge Description

Sentiment analysis is a rapidly growing field, as large databases become increasingly available, and companies have an interest in **automatically extracting sentiment from internet users**.

With the increasing amount of content and debate on social media platforms such as Twitter, there is an interest in **automatically extracting meaning and sentiment from users' posts**, to be able to **evaluate** the **aggregate opinion of a large number of users**. Capturing sentiment in language is important in these times where decisions and reactions are created and updated in seconds.

For this challenge, students are provided with tweets from Figure Eight's Data for Everyone platform. Your task will be to **classify tweets with positive, neutral, or negative labels**, reflecting the sentiment of the user writing the tweet.

For example:

- *"I love this ice-cream, it is so good ! :-)"* -> positive.
- *"Last week I visited San Francisco"* -> neutral.
- *"Disappointed by the last Avengers movie..."* -> negative.

You are provided with an example notebook which covers the basic elements of NLP models for sentiment analysis, along with a baseline. The evaluation criterion is based on your model's ability to predict the correct labels for the testing data.

At the end of the challenge, if you wish to do so, there is a bonus task which consists of detecting the relevant words inside the tweet, which are most responsible for the attributed sentiment. For example, in the phrase *"I really like this song"*, the relevant words are: *"really like"*.
This is not evaluated with the kaggle platform however; the testing labels are provided in the train.csv file (under selected_text), for you to compute the performance of your own model.

Evaluation Metric

The evaluation metric for this competition is Macro F1-Score. The F1 score, commonly used in information retrieval, measures accuracy using the statistics precision p and recall r.

Precision is the **ratio of true positives (tp) to all predicted positives (tp + fp)**.

$$Precision = \frac{T_P}{T_P + F_P}$$

PRECISION used to know HOW WELL/ACCURATE a POSITIVE IS DETERMINED.

Recall is the **ratio of true positives to all actual positives (tp + fn)**.

$$Recall/Sensitivity = \frac{T_P}{T_P + F_N}$$

RECALL or SENSITIVITY used to know HOW MUCH POSITIVES ARE DETERMINED.

The Macro F1-Score takes into account class imbalances, in the case of a multi-class classification problem.

Bonus:

For the bonus task, your can estimate the overlap between your prediction, and the selected words ground truth using the Jaccard coefficient, which measures the ratio of intersection to union of predicted and label sets.

Submission Format

Submission files should contain two columns: textID and sentiment. Due to constraints from the Kaggle platform, your submitted results should be in the class label format: {1, 0, -1} instead of the strings provided in the training data {'positive', 'neutral', 'negative'}.

Your submission.csv file should look like:

```
textID, sentiment

353dd866e2,1

c2e5d07506,0

0cb05b5520,-1

c34de3065b,0

bcae98309c,-1

...
```

You can download an example submission file (sample_submission.csv) on the Data page.