NLP Project 2

Multi-label classification

In the files tweets_train.csv and tweets_test.csv you will find tweets from celebrities, each line containing a tweet and the associated ID for this person.

In the files labels_train.csv and labels_test.csv you will find the gender, the generation and the occupation of these persons. The three target variables can have the following values:

target	values
`gender`	*male, female*
`generation`	*Silent, Boomers, Generation X, Millennials, Generation Z*
`occupation`	*politics, science, creator, sports, performer*

Develop three different classifiers for:

- the identification of female and male persons,
- identifying the generation of the persons,
- the identification of the occupation and
- a fourth classifier prediction (`gender`, `generation`, `occupation`) for each ID.

Calculate the confusion matrix, accuracy, precision, recall and F1-score for each classifier and additionally the weighted means.

Optimize the three first classifiers for greatest possible **accuracy**.

Optimize the fourth classifier such that all three target variables are correctly predicted for as many tweets as possible.

Hints:

- For each person you have a set of tweets. The idea is to predict the target variables with the help of a full set. If you train your classifier on a single tweet you would need for example some majority voting at the end to end up with a single prediction.
- It could be useful to implement some measures against imbalance in the target variables