Assignment 3 – Natural Language Processing 2023 – 2024 (JM2050)

IMPORTANT DISCLAIMER

The dataset you will be using in this assignment is highly sensitive and proprietary. You are not allowed to use the dataset or any derivatives thereof in any means or applications other than the assignment for the 2023 JM2050 course. You are not allowed to copy, replicate or otherwise (re)distribute the dataset other than solely for fulfilling the assignment set forth in this document. Doing so will result in legal actions.

For this assignment, you will be using a dataset that contains almost 3k records of human-written texts as answers to life-meaning and -purpose questions as well as the outcomes of the same person from an IPIP-NEO personality test which aims to capture a person's OCEAN/BIG5 personality traits

The columns and their meanings are as follows:

- TEXT – The original answers two 20 questions, concatenated into one large human-written text.
- TEXT_NL – The Dutch (automatically generated) translation of the TEXT column.
- cEXT – Whether or not this person scores positive on extraversion (y) or not (introversion (n).
- cNEU – Whether or not this person scores positive on neuroticism (y) or not (n).
- cAGR – Whether or not this person scores positive on agreeableness (y) or not (n).
- cCON – Whether or not this person scores positive on conscientiousness (y) or not (n).
- cAGR – Whether or not this person scores positive on openness (y) or not (n).

Main Goal

The main goal of the assignment will be to create a classifier that predicts each of the OCEAN personality traits from input text statistically significantly better than the baseline.

Note: there is a pre-trained OCEAN-prediction model for English on Huggingface. Only for the **bonus** assignments are you allowed to use that model and compare its performance against other approaches. You are **not allowed** to use it in the regular assignments.

1 Binary Classification – Comparison

You will have to work out a binary classifier for each of the five personality traits separately. Make sure you explore using traditional machine learning based approaches and algorithms and compare them against using deep learning based approaches. For example: use TF-IDF, Part-of-speech tags or Bag-of-Words fed into an SVM model and compare it to using BERT's tokenizer with a BERT-based classifier. Use only the English texts.

Deliverables:

- A Python notebook that is error-free
  - Colab links are allowed
  - ipynb files are allowed
- Clear sections containing
  - Reading and handling of the dataset
  - Preprocessing and cleaning of the documents
  - Actual modelling

        - For ML approach(es)
        - For DL approach(es)
    - Evaluation of the modelling
        - For ML approach(es)
        - For DL approach(es)
    - Comparison of ML vs. DL
- The notebook should contain clear traces of different parameter attempts and in markdown, explanations of the results and why further steps were (not) taken
- In markdown, any challenges you encountered and how you solved them or why you didn't

2 Probabilistic and Multi-label Classification

Based on your findings and results in the first part of the assignment, continue further into 2 directions:

1. Predict and classify the five personality traits in one go using one model.
2. Return some probability or continuously-scaled label for the personality traits, so no longer 1/0 (y(es)/n(o)) but 0.25 or -0.3.

Investigate the (im)possibilities of using machine learning methods for this and explain how they compare to deep learning based methods. Use only the English texts.

The deliverables should be the same as part 1, you can reuse or continue from your previous notebook(s).

3 Multi-lingual Modeling

The dataset also contains Dutch texts. See if and how you can incorporate this into your modeling. Investigate and clearly answer questions like:

- Do you train a cross-lingual model for both Dutch and English at the same time? If so, why? How did you approach this?
- Will your method scale to other language like German or French that are not included? Explain how and why.
- What challenges do you need to overcome that are present when classifying multiple languages vs just English?
- Anything else you encounter.

The same deliverables hold.

4 Ethical Considerations

A dataset like this and a model to predict personality scores can lead to unethical applications, think about Cambridge Analytica: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html. Write a small essay about the nature of this data and potential applications. Explain clearly what good it can do but also what it can lead to in the wrong hands.

The deliverable should be a small document (PDF, Markdown, Word) of at most 1 page A4.

Bonus Assignments

You will not get points deducted in case you fail to make it to these bonus assignments but you will be rewarded if you do and manage to come to solid and sound conclusions.

Bonus 1

Come up with a mechanism to automatically (unsupervised) extend your classifier to any given language, for example German or French. You may use crosslingual models and translation models but if you do, you clearly have to investigate and evaluate the performance and soundness of each of the steps used. Also reason what it would mean to go beyond Latin/Roman languages such as Chinese or Arabic.

Bonus 2

While hard to find and rare, there are some works and datasets on personality prediction from text available in literature, for example the LIWC corpus and the stream-of-consciousness essay dataset. Find out if there are datasets and resources that you can use other than the provided CSV file. Explain how you can incorporate those datasets and resources into your own classifier(s).

You may also include the pre-trained model from Huggingface which is trained on the stream-of-consciousness dataset. If you do, carefully evaluate and explain its performance.