# Artificial Intelligence II
# Deep Learning for Natural Language Processing
# Spring Semester 2024-2025
# Homework 2
# 25% of the course mark
# Announced: March 21, 2025
# Due: April 28, 2025 before 23:59

## Description

As in homework 1, in this homework, you have to develop a **sentiment classifier** using **deep neural networks** for the English-language **Twitter dataset**, that has been provided in the previous homework.

In this homework, you should use the machine learning framework **PyTorch**, and the inputs to your model must be **Word2Vec word embeddings**.

Before you start the homework, make sure that you have studied the relevant slides of the course (PDF files "Feed Forward neural networks", "Perceptrons", "Backpropagation", "Training DNNs", "Word Vectors" and "word2vec") and the relevant chapter 7 of the "Speech and Language Processing" book of Jurafsky and Martin (`http://web.stanford.edu/~jurafsky/slp3/`) or any other relevant literature you may find useful.

It is your responsibility to choose all the details of developing a good model (e.g., whether to do cross-validation, whether to do regularization, which gradient-based training algorithm to use, how to choose the hyperparameters of the algorithm, how to make sure that your model does not underfit or overfit etc.).

## Evaluation

On Kaggle, the evaluation metric must be **accuracy**, while in your report, you should include **accuracy**, **precision**, **recall**, and **F1-score** to assess model performance comprehensively.

Ensure that your results are supported with clear, high-quality, and well-labeled plots that effectively illustrate your findings. For example, verify that your model does not suffer from underfitting or overfitting. Additionally, for implementation purposes, you **must use a random seed** to ensure reproducibility. You are encouraged to mention in the appendices of your report any other approaches you explored that did not improve the model's performance.

## Kaggle

You will submit your code (in the form of a Jupyter Notebook) through a Kaggle competition. Make sure to do the following:

- Your team name must be your academic identification number (Αριθμός Μητρώου).

- Your solution must be submitted as a Notebook that outputs a result file named "submission.csv", **NOT AS A FILE UPLOAD**! The result file must follow the format specified in the provided "sample_submission.csv" file and must contain the predictions that your model makes over the test set.
- You must give your sdiXXYYYYY as a name to your Notebook and share your Notebook on Kaggle with the Teaching Assistant responsible for grading this assignment. **DON'T SHARE YOUR NOTEBOOK PUBLICLY**!

## Data

You can view the data here. You should read your dataset from your kaggle notebook. No need to download/upload it.

## Report

For this project, you are asked to create a detailed report. For this reason we provide you with a template in LaTeX. You may use Overleaf online editor. Find the template **here**. Open OverLeaf, create an account if you don't have one already, and then upload the zip file by selecting: New project; Upload project; Select a .zip file; (it uses a pdfLaTeX compiler).

If you are having any issues in writing with LaTeX, you can write it to word/docs following the template in LaTeX. However we strongly advice you, to create it in LaTeX, as Overlead now provides you with many shortcuts and abilities making it easier for you.

The report must be written in English for students in the Master's program in Data Science and Information Technologies. All other students may choose their preferred language.

## Grading

**Implementation**: Code, kaggle submission [**Total 70%**]

- Data processing: [**10%**]
- Model creation: [**20%**]
- Experiments: [**30%**]
- Fine-tuning & Optimization: [**10%**]

**Report**: Analysis and Presentation [**Total 30%**]

- Experiments: [**10%**]
- Analysis: [**15%**]
- Plots: [**5%**]

## Submission guides

We expect you to:

1. Submit your **Jupyter Notebook** (and make is available to supervisors) in **Kaggle** and **only**.*
2. Submit your report in a **.pdf** format from e-class. Name your report like: **[full-id].pdf** (e.g. ZZZZZZXXYYYYY.pdf if you are a bachelor student in this department).

**We won't accept code submissions from e-class/e-mails, etc.*

## Support

Myrto Tsokanaridou will be supervising this assignment. Please post your questions on Piazza under the corresponding directory (hw2).