

Artificial Intelligence II
Deep Learning for Natural Language Processing
Fall Semester 2023
Homework 4
25% of the course mark
Announced: February 5, 2024
Due: March 10, 2024 23:55.

Description

You will develop a **sentiment classifier** by fine-tuning two models for the **Twitter dataset about the Greek general elections**, that has been provided in the previous homework. The models are:

- The pretrained model **GreekBERT**, available on HuggingFace, documentation.
- The pretrained model **DistilGREEK-BERT**, available on HuggingFace, documentation.

In this homework, you should use the machine learning framework PyTorch (<https://pytorch.org/>).

Before you start the homework, make sure that you have studied the relevant slides of the course (PDF files “Machine Translation”, “Transformers”, and “BERT”), 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

You should plot learning curves that show that your models are not overfitting or underfitting. You should evaluate your models using *precision*, *recall* and *F-measure*.

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 and notebook 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.

Since you have two models for this homework, please produce the output file for GREEK-BERT, but also have available (and commented out) the code and the execution results for DistilGREEK-BERT as well.

- You must 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 datasets from your Kaggle notebook. No need to download/upload them.

Report

For this project, you are asked to create a detailed report. For this reason we provide you with a template in L^AT_EX. 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 L^AT_EX, you can write it to word/docs following the template in L^AT_EX. However we are strongly advice you, to create it in L^AT_EX, as Overleaf now provides you with many shortcuts and abilities making it easier for you.

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** in the competition (and make is available to supervisors) in **Kaggle** and **only**.*
2. Create a report in a **.pdf** format and submit it to e-class. Name your report like: **[full-id].pdf** (e.g. ZZZZZXYYYYY.pdf if you are a bachelor student in this department).

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

Support

Despina-Athanasia Pantazi (dpantazi[at]di.uoa.gr) will be supervising this assignment. Please submit your questions on Piazza under the corresponding directory (**hw4**).