Artificial Intelligence II

Deep Learning for Natural Language Processing

Spring Semester 2024-2025

Homework 1

Weight: 25% of Course Grade

Announced: February 25, 2025

Due: March 27, 2025, before 23:55

Assignment Overview

In this assignment, you are required to develop a **sentiment classifier** using **only Logistic Regression** and **only TF-IDF** in Python on a given English-language **Twitter dataset**. The dataset consists of three columns:

- **ID** - A unique identifier for each text.
- **Text** – Contains the content of the tweet.
- **Label** – Represents sentiment, where **1** indicates a **positive** sentiment and **0** indicates a **negative** sentiment.

The assignment is divided into two main components:

1. **Model Implementation and Kaggle Submission:** You are required to implement your sentiment classifier and submit your predictions to the Kaggle competition.
2. **Report Submission:** In addition to your implementation, you must submit a detailed report explaining your thought process, methodology, and the reasons behind your model's parameter choices.

Ensure that your report is well-structured and clearly justifies your design decisions. Good luck!

Relevant Literature

Before starting this assignment, ensure that you have studied the relevant course materials, including the slides on *"Introductory Concepts of Machine Learning"* and *"Regression."* Additionally, refer to Chapters 4 and 5 of *"Speech and Language Processing"* by Jurafsky and Martin (`http://web.stanford.edu/~jurafsky/slp3/`), or any other relevant resources you find useful.

Guidelines

This task consists of the following main steps:

1. **Exploratory Data Analysis (EDA)** – Perform an initial analysis of the dataset, including descriptive statistics and visualizations.

2. **Text Preprocessing** – Apply necessary preprocessing steps such as removing unsuseful parts of text.

3. **Feature Extraction** – Convert the text into an appropriate format for a classifier. You can use **only** the TF-IDF method.

4. **Model Development and Evaluation** – Implement **only Logistic Regression** classifier and evaluate its performance using metrics. 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.

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, and how to choose the hyperparameters of the algorithm). Additionally, for implementation purposes, students **must use a random seed** to ensure reproducibility.

Presentation of Results: 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.

Additional Considerations: 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 Competition

Submit your code as a **Jupyter Notebook** via the Kaggle competition. Follow these rules:

- Your team name must be your academic identification number (Αριθμός Μητρώου - sdiXXYYYYY or the one for graduate students).

- Your solution must be submitted as a Notebook that outputs a result file named *"submission.csv"*, **NOT AS A FILE UPLOAD**! The resulting 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 share your Notebook on Kaggle with the Teaching Assistant responsible for grading this assignment. **DO NOT SHARE YOUR NOTEBOOK PUBLICLY**!

Report

For this project, and the next ones, you are asked to create a detailed report. For this reason, we provide you with a template in LaTeX. You may use the 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 advise you, to create it in LaTeX, as Overlead now provides you with many shortcuts and abilities making it easier for you.

Grading

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

- EDA and 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 in 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

Yorgos Pantis is supervising this assignment. For any questions, please post them on Piazza.