Artificial Intelligence II

Deep Learning for Natural Language Processing

Spring Semester 2025-2026

Homework 1

Weight: 25% of Course Grade

Announced: February 24, 2026

Due: March 20, 2026, before 23:55

Assignment Overview

In this assignment, you will develop and evaluate automatic classifiers for the task of **response clarity classification** in political question–answer (QA) pairs.

You will work with the dataset released for **SemEval 2026 Task 6 (CLARITY)**. The dataset consists of political QA pairs extracted from interviews in which journalists ask questions and public figures provide answers. In many cases, answers are not fully responsive: the speaker may be vague, provide only partial/general information, or avoid the question.

The goal is to predict the clarity of the answer with respect to the question (according to the labels provided in the dataset).

You must implement **two text representations** and compare their performance:

1. **TF–IDF + Logistic Regression:** Represent QA text using **TF–IDF** features and train a **Logistic Regression** classifier.
2. **Word2Vec-based representation + Logistic Regression:** Represent QA text using **Word2Vec** embeddings (e.g., by averaging token vectors) and train a **Logistic Regression** classifier. Compare the results against the TF–IDF baseline.

The assignment is divided into two components:

1. **Model implementation and Kaggle submission:** Implement your system and submit predictions to the Kaggle competition.
2. **Report submission:** Submit a report describing your preprocessing, representation choices, modeling decisions, and experimental results, including a clear comparison between the TF–IDF and Word2Vec-based approaches. In addition to reporting your evaluation scores, please include an **in-depth analysis** of your findings (e.g., a comparative discussion of strengths/weaknesses, plausible reasons for performance gaps, and targeted error analysis highlighting which linguistic or task-specific phenomena each approach handles well or poorly). Finally, discuss your **design choices** (e.g., preprocessing decisions, hyperparameters, and representation settings), explain *why* you made them, and summarize what impact they had on the final performance.

Ensure that your report is written in an academic style and supported by quantitative evidence.

Data

This assignment uses a dataset of political **interviews with U.S. presidents**. The dataset is available on `Hugging Face` and provides two splits: **train** and **test**. You are expected to **train** your models on the **train** split and **evaluate** them on the **test** split.

Each record contains metadata (e.g., `title`, `date`, `president`, `url`, `question_order`) and the interview text. Importantly, the dataset distinguishes between:

- `interview_question`: the original journalist turn, which may contain *multiple* questions;
- `interview_answer`: the corresponding answer.

Because a single journalist turn may include more than one question, the dataset also provides a segmented field `question`, which contains **exactly one** question. Therefore, the core supervised instances in this assignment are **(question, interview_answer)** pairs, where `question` is a single-question segment and `interview_answer` is the answer given in the interview.

The prediction target is `clarity_label`, a three-class label: *Clear Reply*, *Ambivalent*, and *Clear Non-Reply*. Each (question, interview_answer) pair is associated with a **single, unique** ground-truth `clarity_label`. Detailed field definitions and dataset documentation are provided in the Hugging Face dataset card, as well as in the accompanying paper.

Relevant Literature

Before starting this assignment, ensure that you have studied the relevant course materials, including the course slides (PDF files) *"Introductory Concepts of Machine Learning"*, *"Regression"*, *"Word Vectors"*, and *"word2vec"*. Additionally, refer to Chapters 4–5 and Chapter 7 of *"Speech and Language Processing"* by Jurafsky and Martin (`https://web.stanford.edu/~jurafsky/slp3/`), or any other relevant literature you may find useful.

Guidelines

This task consists of the following main steps:

1. **Exploratory Data Analysis (EDA):** Perform an initial analysis of the dataset (e.g., label distribution, text length statistics, and informative visualizations).
2. **Text preprocessing:** Apply appropriate preprocessing (e.g., lowercasing, punctuation normalization, and optional stopword removal). Clearly justify each choice.
3. **Input construction:** Decide how you will combine the question and the answer into a single input (e.g., concatenation with a separator token). State your design explicitly.
4. **Feature extraction:** Implement **two representations**:
   (a) **TF–IDF** features using `TfidfVectorizer`.
   (b) **Word2Vec-based** representations (e.g., average Word2Vec vectors).
5. **Model development and evaluation:** Train and evaluate the required models for each representation. For **TF–IDF**, use **Logistic Regression**. For **Word2Vec-based** features, use **Logistic Regression** and compare your results against the TF–IDF baseline. On Kaggle, the evaluation metric is **F1**. In your report, include **accuracy**, **precision**, **recall**, and **F1-score**.

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 give your sdiXXYYYYY as a name to your Notebook and share it on Kaggle with the Teaching Assistant (TA) responsible for grading this assignment. **DO NOT SHARE YOUR NOTEBOOK PUBLICLY**!

Report

For this project (and subsequent ones), you are required to submit a detailed report. A LaTeX template is provided via the course platform (e-class).

If you experience difficulties writing in LaTeX, you may draft your text in another editor first; however, you are strongly encouraged to produce the final report in LaTeX using the provided template.

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%**]

- 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 it available to supervisors) in **Kaggle** 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

Sergios-Anestis Kefalidis is supervising this assignment. For any questions, please post them on Piazza.

About SemEval

SemEval (International Workshop on Semantic Evaluation) is a long-running and widely recognized shared-task competition in Natural Language Processing (NLP), typically **co-located with a major NLP conference**. Each year, SemEval hosts multiple tasks that are announced in advance and attract international participation from research groups and industry teams.

A typical SemEval cycle lasts approximately **six months**. It commonly includes a development period, an evaluation phase that concludes around late January, and a short period (approximately one month) during which participating teams prepare a system description paper. Accepted papers are published in the official SemEval proceedings.

Participating in SemEval-style tasks is valuable experience, as it provides a realistic experimental setting, standardized evaluation, and an opportunity to compare approaches and publish the results and analysis in the field of NLP.

The goal of this year's assignment is also to familiarize you with the complete SemEval workflow: introduction to a new task, data handling, exploratory data analysis, model development and testing, evaluation, improving methodologies to increase performance, and reporting. By completing a full end-to-end run in the style of a shared task, you will be better prepared to participate in future SemEval competitions, should you wish to do so.