CS412 Machine Learning
2025 Spring Term Project

Due Date: May 25, 2025

1 Goal

The course project will allow you to work on a real machine learning problem using any of the methods you have learned in the course or outside. You will work in groups of 3-5 people. The finalized groups could be found here. The finalized groups file can also be found on SuCourse under the "Project" tab, with the name *Finalized Project Groups*.

2 Problem Description

In this project, you will work on a **binary sentiment analysis** task. Given a movie review, the goal is to predict whether the review is `positive` or `negative`.

3 Dataset

The dataset that you will work on in this project is the **IMDB movie reviews dataset**. The dataset contains movie reviews annotated with binary sentiment labels: `positive` or `negative`. The dataset is provided as a CSV file with two columns:

- **review:** The review.
- **sentiment:** Either `positive` or `negative`.

An example review:

> *"I have now seen quite a few films by Pedro Almodóvar, but this would have to be the most disappointing so far..."*
> **Sentiment:** negative

The dataset contains 50,000 reviews about movies (positive 25,000, negative 25,000) The dataset can be found in this link.

4 Your Task

In this project, you are expected to complete the following steps:

- Present basic statistical information about the dataset. (It is good practice to explore and analyze the dataset you will work on.)
- Preprocess the data appropriately for your models. (This may include tokenization, text cleaning, or feature extraction.)
- Implement three different appropriate machine learning or deep learning methods for sentiment classification. (Examples include Naive Bayes, SVM, Logistic Regression, RNNs, fine-tuning BERT, or using static word embeddings (e.g., Word2Vec, GloVe) with any classifier.)
- Evaluate and compare the performance of your models using appropriate metrics such as accuracy, precision, recall, F1-score, or AUC. Official evaluation will be based on the macro-F1 score.
- Provide a detailed analysis and discussion of your experimental results and observations. State advantages, disadvantages and results of the approaches you adopted.
- Prepare a Google Colab notebook and a PDF report, clearly explaining all steps taken, along with your results and discussion throughout the project.

You are free to design and implement models that can solve this problem using techniques learned in class or external methods.

---

5 Project Report and Code Submission Instructions

You are required to submit both a Jupyter Notebook containing your code and a PDF report documenting your project. The report should be written in a **technical report format** and organized as follows:

- **Cover Page:** Include a cover page that clearly states your group number, the full names, and student IDs of all group members.
- **Introduction:** A brief summary of the problem, the methods you applied, and the main results obtained.
- **Problem Description:** A formal definition of the problem. Clearly state whether it is treated as a classification, regression, or other machine learning task.
- **Methods:** A detailed description of the methods you implemented. This section should include:
  - Statistical analysis of the dataset (e.g., class distributions, average review lengths)
  - Preprocessing steps (e.g., tokenization, text cleaning)
  - Explanation of the methods applied (e.g., description of classifiers, neural network architectures, transfer learning setups)

  - Model architectures and hyperparameter choices

- **Results and Discussion:** Presentation and analysis of your experimental results. Include:
  - Evaluation metrics (accuracy, precision, recall, F1-score, AUC, etc.)
  - Precision-recall curves, confusion matrices (if applicable)
  - Comparative discussion across different methods and feature representations

- **Conclusion:** A brief summary of the overall work. Highlight the methods that performed best, discuss why they might have worked better, and reflect on the key lessons learned during the project.

- **Appendix:** Clearly describe the contributions of each team member. You may also include supplementary material such as additional plots, tables, or detailed hyperparameter configurations.

Additional Submission Instructions

- **Jupyter Notebook:** Ensure that all code cells and their outputs are included. (Your notebook will not be re-run during grading.)

- **Notebook Link:** At the beginning of your PDF report, provide a shareable link to your Colab notebook (accessible to anyone with the link).

- **Submission Files:**
  - Jupyter Notebook: `CS412-Project-GroupNumber.ipynb`
  - PDF Report: `CS412-Project-GroupNumber.pdf`

- **Late Submissions:** **No late submissions will be accepted.**

6 Grading

The grading will be based on the following components with given rough grade percentages:

- **40% - Methodology and Experimentation:** Appropriateness of the chosen methods, whether a sufficient number of different methods were tried (at least three), correctness and necessity of preprocessing, proper use of data splits (train/validation/test), and soundness of experimental setups...

- **30% - Performance:** The official evaluation will be based on macro F1 score over the test set. The best system will obtain the max pts (30 pts), while others will have 10-28pts depending on their relative performance.

- **30% - Report Quality:** Clarity, organization, and thoroughness of the report. Proper explanation of methods, inclusion of relevant tables, figures, and well-structured presentation of findings.