NLP 053 - CSE UOI 2025 - Kaggle Challenge

Konstantinos Skianis

May 2025

Due date: 15 June 2025

Only one member of the team is required to upload the solution.

Exam: 23 and 24 June 2025

We will announce a time program, and you will be requested to come within a time slot in 23 and 24 June. Please respect it.

In order to be examined, you first need to have at least one valid submission on the competition Kaggle webpage. Regarding the exam day and process:

- You don't have to get prepared for a formal presentation. I want to understand the steps you followed. Each team will have around **15-20 minutes** to present and be examined.
- Along with your presentation of the solution, I may ask theoretical questions to see if you know and understand what is in the report.
- **If you fail to answer a number of these questions, it either means that you did not participate, did not understand what the course was about, and thus you won't pass the exam.**
- If you fail, you can get reexamined for the same challenge in September.

Deliverables

Each team should deliver by **15 June 2025** a zipped folder (with your names and IDs as filenames) including:

- A report as a .pdf file (see details below). Should be 5 pages in length at least, cover page excluded. Please ensure that both your real name(s), the name of your Kaggle team, and student IDs appear on the cover page.
- A presentation that can be derived from the report.
- A folder named "code" containing all the scripts needed to reproduce your submission. Although Python is preferred, you are free to use any other language like R, Matlab or Java.

Grading

Your grade will be decided by the Kaggle score and the steps you followed, as described in your report (feel free to have these steps as sections in your report):

- Preprocessing (10 points). What preprocessing techniques did you apply to text or/and graph?
- Feature engineering (30 points). Regardless of the performance achieved, intended to reward the research efforts, creativity and rigor put into feature engineering. Best submissions will capture both textual and graph information. You are expected to:
  - Explain the motivation and intuition behind each feature. How did you come up with the feature (e.g., research paper)? What is it intended to capture?
  - Rigorously report your experiments about the impact of various combinations of features on predictive performance, and, depending on the classifier, how you tackled the task of feature selection.
- Models, tuning, and comparison (30 points). Best submissions will:
  - Compare multiple classifiers (e.g., SVM, Random Forest, Boosting, logistic regression...).
  - For each classifier, explain the procedure that was followed to tackle parameter tuning and prevent overfitting.
  - Investigate state-of-the-art models.
- Kaggle score (20 points). We will follow a logarithmic function. There will be small differences in higher scores.
- Report and presentation readability, as well as code completeness and organization (10 points).

Notes

Overall, the best submissions will:

1. Clearly deliver the solution, providing detailed explanations of each step.
2. Provide clear, well-organized, and commented code.
3. Understand which features and models are best.
4. Refer to research papers.

Finally, note that the testing set has been randomly partitioned into public and private. Scores on the leaderboard are based on the public set, but final scores (based on which grading will be performed) will be computed on the private set. This removes any incentive for overfitting the testing set.