CS445 Group Projects

A. Project Selection

In this project, you are asked to choose one of the three tasks;

1. Stance Detection
2. Intent Detection
3. Fake News Detection

You will create a system to solve your chosen task using relevant literature. This is a group project where undergraduate students will form four-member groups. You can access the group sheet here. Please write your group number on your chosen task row. There is a quota for how many groups can take each project and it is first come first serve.

B. Finding a dataset

After deciding on the project, you must find a dataset relating to your chosen task. You are allowed to curate your dataset in any way. You can use one from the paperswithcode task page, use a dataset paper, or combine multiple datasets. Check the copyrights of the dataset you will use and comply with it.

The dataset should be split from the start into 3 sets - train/dev/test. If the dataset is small, you can use cross-validation instead of the development set. The split ratios are up to you but normally a split of train/dev/test: 70/15/15 or train/test:80/20 is usually good enough. Use a predetermined seed for splitting - you will have the same split each time you run your code.

You should only test your system once and report the results!!!

C. Relevant Literature

To solve the task you have chosen, you need to conduct a literature search. This is the initial step of the project: to review the methodologies researchers have used to solve the problem.

- You can select one or a few papers to follow and implement them. Discuss why you have chosen these papers.
- You can also create your novel approach to solve the task but you still have to show why you have chosen this particular approach. This requires a small discussion of the existing approaches in the literature.

Suitable Venues

You are only allowed to use papers from the following venues. If you find a really good paper that is not included in the list but you want to implement it, please contact Dilara Keküllüoğlu and state the reasoning behind your choice.

- SCI-Expanded list journals
- ACL
- EMNLP
- NeurIPS
- CoNLL
- NAACL
- EACL
- COLING
- LREC

Use of Existing Codes

You can use any external libraries or open source codes, e.g. from paperswithcode, to implement your approach. If you are using an open-source code, clearly show your contribution. If you use an existing code without disclosing the source or passing it off as your own code, that would count as plagiarism.
We will consider your system's performance in the evaluation, however, your contribution is more important. For example, if you achieve an F1 score of .95 but the system already had an F1 score of .90 before your contribution, you would receive fewer points compared to a system that was written from scratch and received an F1 score of .90.
Your novel contribution to the project is more important than having better F1 scores.

D. Report Results

To relay your results, you will have a presentation and a project report. The presentation should be reflective of your project report. In the project report, you are asked to have the following sections:

1. Introduction

You should give a task description. This is an introduction to the problem, the chosen methodology, and the most important results. The introduction should give a quick overview of your project.

2. Methodology

What dataset are you using? Which papers have you chosen to implement? Explain why you use these specific datasets and papers.
How do you train your system? What is the split ratio between your train/dev/test datasets? Give enough detail that people can replicate your approach.

3. Results

Give the results of your system on the task. Use plots/graphs to show important findings. Give a confusion matrix for the classification. Please give F1-score, macro precision, and macro recall. You will also give precision-recall curves. Compare your results to the Naive Bayes classifier. Compare it to the state-of-the-art as well as the literature you are using in your approach. Give details of the experiment. You will also write any other tried methods, even if you decide not to use them in the end. Show your journey.

4. Discussion

Please discuss these points in this section:

- The selection of your dataset and how that affected the system performance.
- The selection of your particular approach - what are the advantages/disadvantages of your method?
- Comparison of your results to the existing systems - discuss the results that have been reported in the Results section.
- Limitations of the proposed system.

- Possible improvements on the system if you had more time and resources.

5. Conclusion

Give another quick overview of the problem and list the most important results/discussion points to conclude the report.

—--------

The reports should be a maximum of 5 pages (excluding references and appendix). You can add extra results/plots in the appendix if you cannot make them fit 5 pages. Make sure to put the most important results into the main report and leave the extra ones in the appendix.

In addition to the project report, there will be a final individual report submission (only at the final submission - 6th Jan) about your unique contribution to the project and any other remarks you would like to add about the project. (max 2 pages)

Milestone Reports

For the milestone report, please use the following outline.

1. Introduction

You should give an introduction to the task you have chosen. Follow with a summary of what you have done so far.

2. Dataset Selection

What datasets are you using? Give details on why you have selected these particular datasets. How did you split the dataset?

3. Approach Plan

What is the progress on the methodology decision? What are some of the main papers from which you are getting inspiration? Have you finalized your approach or what else will you consider to finalize your approach?

4. Next Steps

What are your next steps? How will you divide the workload between the team members?

Milestone reports should be a maximum of 3 pages (excluding references) detailing following the outline given above.

E. Deliverables

There will be 1 milestone presentation and 1 final presentation. You need to submit reports along with the presentations and for the final submission you are required to submit an individual report. You will also submit your well-commented working system (Jupyter notebook and any files needed to run your code) in the final submission.

Milestone presentations (25th November)

For the milestone, you are asked to submit a project report that will detail the selection of the dataset. At this point, I expect you to have the dataset ready and split. You should also have a draft of your plan for the methodology and who will do what. However, you can change your approach after the presentation if you think it would be better for the project. (Submission for the project report - 25th November)

You are also asked to give a 5 minute presentation detailing your project and approaches. You will schedule this with your assigned TA.

Deliverable: 5 minute presentation + milestone report

Final presentations (6th January)

For the finals, you will submit your project report and a jupyter notebook with your well commented code on the 6th of January. The project report should have all the components detailed in the “Report Results” section. You will also have an

individual report submitted, detailing your own contributions to the team and adding any other remarks you want to make individually.

You are also asked to give a 5 minute presentation detailing your project and approaches. You will schedule this with your assigned TA.

Deliverable: 5 minute presentation + project report + individual report + project code and files (Jupyter notebook and any files needed to run your code)

Please compress your reports and project code into a single file while submitting for the final submission.

F. Grading

- 20% from the milestone presentation and report
- 60% from your final presentation, code, and group report
- 20% from your individual report + your contribution

G. Graduate Students

Graduate students can choose any of the default projects or give their own proposals about a project related to NLP. Please bring your idea to Dilara Keküllüoğlu for approval.

In addition, graduate students can choose to do one of the SemEval 2025 tasks. SemEval is a workshop on semantic evaluation and every year there are several tasks released. The timeline of the dataset release and submission of results aligns nicely with the project timeline of CS445. However, this might be harder to do than other projects. As long as you make a genuine effort to solve the task, it will be enough for this project. You might even get a paper out of this!

Graduate students can make 2-3 member groups. In some situations, you can make the project individually. Please discuss with Dilara Keküllüoğlu if you want to work individually.

The milestone presentations will be made to Dilara Keküllüoğlu directly. The final presentations might be arranged where the whole class will join and listen to the different projects - subject to changes.

H. AI Use

You are allowed to use AI systems to spell-check your English and improve the flow of your report.

You are **NOT** allowed to write your report from scratch using an AI system. Write the report first and then improve on it with AI on your wording and flow.

You are **NOT** allowed to use AI systems to write your code. Write your own code and detail which parts were taken from where you are using another person's code.