CS F429 Natural Language Processing

Project description and rubrics for assessments

Project Group Size: Minimum 3, Maximum 5

Max Marks: 50

Problem definition for the project: You are free to pick up any problem in the area of Natural Language Processing

Mandatory components for the project are:

1) **Dataset Selection and Justification**: Students must identify and justify the dataset(s) they use. Specify the domain (e.g., news, biomedical, legal, conversational), the size of the dataset, its source, and why it is suitable for the chosen task. Discuss dataset limitations (bias, coverage, imbalance) and potential preprocessing steps.
2) **Baselines (Classical vs. Latest)**: Students must compare at least one classical method with a modern approach. Classical methods include statistical and pre-transformer based models. Modern methods include transformer-based models, instruction-tuned LLMs, or fine-tuned pre-trained embeddings. The comparison should show how NLP has evolved in solving the task.
3) **Model details**: For each model, explain its architecture (layers, embedding types, attention mechanism, etc.), training/fine-tuning strategy (frozen vs. updated embeddings, hyperparameters), and rationale for choice. If pre-trained models are used, explain which checkpoints, how they were adapted, and what modifications were made. Include implementation details (libraries, frameworks, compute environment).
4) **Evaluation Metrics**: Students must choose appropriate metrics for their problem. Metrics should align with the task (e.g., F1/Accuracy for classification, BLEU/METEOR/ROUGE for generation etc). Justify why these metrics are relevant, and reflect on their limitations (e.g., BLEU doesn't capture semantic fidelity).
5) **Error Analysis**: Students must analyze system errors to identify weaknesses. Provide both **quantitative** (confusion matrices, error rates per class, breakdown by entity length/frequency) and **qualitative** (examples of failure cases, ambiguous cases, annotation challenges) analysis. Highlight recurring patterns and discuss how errors could be mitigated.
6) **Novel Contribution:** Students must demonstrate a unique insight or improvement beyond baseline replication. Clearly articulate what is new. This could involve addressing an assumption (e.g., domain independence), filling a gap (e.g., handling long-tail entities), or introducing a new modeling tweak (e.g., hybrid rules + neural models).
7) **Report**: Students must submit a structured report (6–10 pages) similar to a research paper. **Reports written in Latex will get bonus points**. Report template to follow:

a) Introduction to the NLP task
b) Literature survey on why the problem is important
c) Dataset
d) Method
e) Experiments
    i) Set up
    ii) Metrics
    iii) Baselines
    iv) Evaluation, Results, and Discussion
f) Conclusion

Project assessment will be done in 3 parts.

1. **Part 1 (5 marks, Deadline: same as Assignment 1, September 8 04:00 AM IST)** Students will upload a 1-2 page report describing the problem statement, dataset, and plan for project completion. The rubrics used for the report assessment are:
    a. **Clarity (1)**: Is the problem clearly articulated and well-defined?
    b. **Motivation (1)**: Is the problem significant? Does it address a real-world problem?
    c. **Dataset (2)**: Are the datasets clearly identified?
    d. **Plan (1)**: Is there a well-defined roadmap for completing the project?
2. **Part 2 (15 marks, Deadline: Same as Assignment 2, around mid October, 2025)** Students should implement at least one baseline model for the NLP task and submit a report containing Sections a) to d). They will be asked to demonstrate the partially completed system. Rubrics are:
    a. **Literature survey (1)**: Are relevant sources appropriately cited?
    b. **Dataset and Preprocessing (1)**: Is there a description of the dataset (e.g., domain, structure, size, source)? Are any initial observations or statistics provided? Are the preprocessing methods appropriate for the (e.g., tokenization, stemming, stopword removal)? Are they clearly described and implemented? Are examples or outputs provided to demonstrate their effect on the data?
    c. **Metrics (1)**: Have the metrics been identified?
    d. **Model Baseline and Results (2)**: Are model details and hyperparameters provided? Are preliminary results on the dataset provided? Is error analysis done?
    e. **Demonstration and Viva (10)**: Training of the model to be shown during the demo. Results should be shown through a GUI or CLI.
3. **Part 3 (30 marks, Deadline: November last week, 11:55 PM IST)** Students should complete the remaining sections of the report and demonstrate the full fledged system. Rubrics are:
    a. **Model's Novelty (3):** Has some new technique(s) been proposed? Have the hyperparameters been stated? Is the architecture explained?
    b. **Improvement (3)**: Has the new technique led to improvement over the previous technique on the dataset?
    c. **Results (2):** Are latest results included? Is error analysis done?
    d. **Challenges (2)**: Are the challenges faced during different stages of the project identified and described? Are strategies or solutions for

overcoming the challenges explained?

e. **Demonstration and Viva (20)**: Training of the model to be shown during the demo. Results should be shown through a GUI or CLI.

Sample Problem statements

1. **Building Text-to-SQL systems**: More details can be found here. Students are free to choose any other benchmark.
2. **Evaluating LLM-Powered AI Tutors**: This is a live challenge floated as part of IndoML 2025. More details can be found here.
3. **Attribute-Value Prediction From E-Commerce Product Descriptions**: This challenge was floated last year as part of IndoML 2024. More details can be found here.
4. **Any other problem of your choice…….**