University of Windsor
Faculty of Science
School of Computer Science

COMP-8730
Natural Language Processing & Understanding
Winter 2024

#	Title	Due Date	Grade Release Date
5	**Evaluation**	March 17, AoE	March 25

This course is research-oriented and project-driven in which a research project should be defined and completed in the field of NLP within one semester. The objectives of the research project are to provide graduate students with:

- An experience with research procedure, in general, and research in NLP, in particular.
- Hands-on experience with NLP.
- Advancing state of the art in NLP while passing a grad course.
- An opportunity to present a research outcome at an international computer science conference
- An opportunity to meet with scholars in the NLP community

The research project proposes a solution(s) to a problem by implementing an algorithm like a software project. However, there are differences in some respects. For instance, while a software project may implement an existing algorithm, a research project should propose and implement a *new* algorithm that improves or addresses a particular aspect of a problem that the current algorithms overlook. Roughly, a research project has the following milestones (phases):

1) Proposal
2) Literature Review
3) Proposed Method (Formal + Code)
4) **Experiment (Evaluation)**
5) Presentation (Paper + Talk)

In this course, a manual is prepared to guide the students through each milestone. The current manual is for the fourth milestone: Experiment (aka Evaluation), which further has the following steps:

Evaluation Strategy

The evaluation strategy is the roadmap you choose to invite the reader/reviewer judgment about how the results capture the truth about your proposed method, the competitors, and the claims you put forward. Let's understand it by examples:

Example 1. One of the immediate examples is mathematical proof for a proposition. For instance, you claim that a number is (not) prime. Your evaluation strategy (the roadmap you show your claim is true) may include 'direct proof,' 'induction,' 'contraposition,' etc[1].

Example 2. Another example would be the efficiency (speed) of a method. You claim that your proposed method is faster than the baselines. Then, your evaluation strategy may include either theoretical or empirical steps. For theoretical, you must mathematically prove that your method's time complexity is of less order in terms of Big O[2]. In an empirical strategy, you may show your method's elapsed time and the competitors based on *wallclock*, *given the same running platform for all methods*.

Example 3. In some research areas, the researchers agree on an evaluation strategy (ies) in those areas. For instance, in the broad scope of Artificial Intelligence, the closer the output of a method to human answer, the better. In other words, the human answers are taken as the *Gold Standard* and calculating the similarity/distance of the AI-model's output to the Gold Standard is the evaluation strategy. In a supervised classification task, the evaluation strategy includes calculating the confusion matrix of a method on a held-out test set unseen to the

[1] https://en.wikipedia.org/wiki/Mathematical_proof#Methods
[2] Refer to courses such as Design and Analysis of Algorithms, Theory of Computation, and Theory of Computer Science

method. In our course, we have covered the evaluation strategies of language models, binary classifiers, multi-class classifiers, and vector representation methods.

Example 4. In the presence of Gold Standard, the evaluation strategy is also called *intrinsic* evaluation. In contrast, *extrinsic* evaluation strategies deal with showcasing the claims in underlying applications. For instance, to show that removing a specific list of words will improve the sentiment analysis, you can select a sentiment analysis method and run it with and without the words and evaluate the sentiment analysis method's output using an accepted evaluation strategy for sentiment analysis. If removing the words yields better performance, your claim is accepted yet *only in this sentiment analysis method*.

~~**Example 5. RQ-based (tbd in future)**~~

The evaluation strategy is the most crucial part of a research project based on which we accept a research project's claims. An incorrect step will make all our findings unreliable. *Empirical* evaluation strategy includes (1) datasets, (2) metrics, (3) baselines, and (4) result presentation.

(1) Datasets

The collection of input instances for a method under which the method is evaluated is called a dataset. A dataset may or may not include the true/optimum/best answer for a given input. Also, a dataset may be comprehensive; that is, it has all the possible input instances, or it may be partial. In the latter case, the dataset must be sampled such that it represents all possible input instances. For instance, for a method that analyses the sentiment of an input sentence, a comprehensive dataset includes all meaningful sentences in a language. However, curating such a dataset is not practically possible. As a result, we have to *fairly* sample the meaningful sentences of *reasonable lengths, different genres,* representing *all likely sentiments* (positive, negative, neutral). A *biased* or *cooked* dataset is skewed toward or favours a specific method and is *not* accepted. For instance, a dataset that includes only short sentences*!*

Synthesis vs. Real-world Datasets. To fairly sample the possible inputs for a method, we can use random procedures to generate the inputs, e.g, generating a random graph to evaluate graph clustering methods. Such a dataset is called a synthesis dataset. In contrast, we may use the information available in the real-world such as the social network of people on Twitter or Facebook. Evaluation strategy can have both synthesis and real-world datasets. In some research areas, synthesis datasets may not be easy, like in sentiment analysis (e.g., generating meaningful sample sentences via random procedures!)

Gold Standards. For a reliable evaluation strategy, researchers try to provide *standard* datasets. These datasets have been investigated carefully and tried to remove any possible biases. Therefore, such datasets offer a fair comparison testbed for all methods—for instance, MS MARCO[3] in Web Search or RTE datasets[4] in Text Matching. Indeed, curating Gold Standards are an area of research and some research projects are defined and developed solely for this purpose.

If you choose a dataset that is not a standard dataset, you have to justify your dataset: Why choose the dataset? What are the dataset statistics (to show the fair sample/representation of input instances)? Is it publically available, and where? (If it is private or not available, it may raise concerns)

(2) Metrics

Metrics are the *quantitative* values to indicate the *quality* of the methods. For instance, to show that a method is faster than the competitors, wallclock elapsed time is a metric; the lower, the better. To show that a mathematical proof is shorter, the number of steps is a metric; the lower, the better. To show that a sentiment analysis method finds the positive sentiments better, recall[5] is a metric; the higher, the better. A metric, however, represents the

[3] https://microsoft.github.io/msmarco/
[4] https://aclweb.org/aclwiki/Textual_Entailment_Resource_Pool#RTE_data_sets
[5] https://en.wikipedia.org/wiki/Precision_and_recall

quality from a very limited perspective. For example, a method may be fast (lower wallclock time) but whose output solutions may not be close to the true answer (low recall). Therefore, an evaluation strategy usually includes more than one metric to showcase the power of methods from different perspectives, esp., where trade-offs are happening, such as accuracy-speed.

Like Gold Standards, there are standard metrics in most research areas: perplexity in language modelling, AUC in classification, nDCG in information retrieval, modularity in graph partitioning, etc. If you decide to use unknown, new, or not standard metrics, you must justify them.

As part of the evaluation strategy, comparing the metric values is not enough to claim a *significant* improvement! If you claim this, you have to do statistical significance tests[6] considering the assumptions of these tests.

(3) Baselines

Your research should improve state of the art in one or more aspects, including time and space (storage) complexity, and accuracy. In this regard, you have to compare your method with the state of the arts or related work, called baselines in general. In the literature review milestone, you already explained related work and the state of the arts. Among the related work, you should select those that are so-called *strong*; meaning they are

- highly-cited,
- made a significant contribution,
- peer-reviewed in world-class venues,
- recent

Additionally, it is common that the proposed method, as well as the baselines, have different configurations, and a search is needed to find the best running settings or configuration. You can add each configuration or each running setting as a baseline. For instance, if your method proposes an embedding method, the embedding size would be an essential parameter to study. In the result part, you could show the metric values for a range of sizes, e.g., [10, 20, …, 100].

A challenge for selecting the baselines is whether the implementation is publicly available. Fortunately, most strong baselines have their code publicly available. Otherwise, you can communicate with the authors and ask for the implementation. In the worst case, you have to implement the baseline*!* The latter case may introduce some concern in the evaluation strategy as your implementation may not be the same as the original baseline and miss some essential parts. Therefore, you have to try your best to find the original implementation.

An advantage of having Gold Standard datasets and metrics is that you do not need to implement the baselines[7]. Why?

(4) Results

The last but not the least part of an evaluation strategy is the *comparative* result presentation. In this part, you have to use tables and charts to showcase i) metric values, ii) on the datasets, iii) for the baselines. This section should be organized into subsections based on the aspects you claimed, each of which explains i) the result and ii) the intuition why we see such results to support your claims. You should neither put all the results in a single figure nor leave the results without explanations*!*

Submission Guidelines

- Submission must be written in English, in the current ACM two-column conference format in LaTeX. Overleaf templates are available from the ACM Website (use the "sigconf" proceedings template).

[6] https://en.wikipedia.org/wiki/Statistical_hypothesis_testing
[7] For instance, look at the MS MARCO's Document Ranking Leaderboard at https://microsoft.github.io/msmarco/

- Submission must be 2 pages (4 columns) in length, no more, no less, including figures, tables, authored by the team members, and 1 column for references.

- The implementations (code) should be available in an online repo (preferably Github), and the link should be mentioned as a footnote to the report's title. See the example below.

- Submission must be in one single zip file with Evaluation_firsname1_firstname2.zip, including the LaTeX files and the pdf file

A sample submission has been attached to this manual in Blackboard, also available online here.

In summary, your submission has (%marking schema for this milestone):
1.1) (20%) Datasets
1.2) (10%) Metrics
1.3) (20%) Baselines
1.4) (20%) Results
1.5) (30%) Evaluation Code (FAIR-compliant)