Guidelines for the Final Course Report

Profilierungsmodul Computerlinguistik I – Trustworthy Data-centric AI

The overall goal of the final course report is to reflect on the theme of Trustworthy Data-centric AI. That is, to provide you an opportunity to apply a topic of interest (that is covered in or related to topics of this class) and a tool to reflect upon or constructively critique a research paper that can benefit from such a discussion. Specifically, your task is to identify gaps, suggest, outline, and implement improvements, and reflect on the paper's implications in the context of trustworthy AI. **Your final report has to be done individually.**

Guidelines

Select a Research Paper

Choose a peer-reviewed research paper from a major NLP, ML or HCI venue. You can search for papers that interest you, e.g., in recent proceedings of

- *ACL conferences
- COLM
- NeurIPS, ICML and ICLR
- CHI

The paper should align with the topics covered in the course within the broad theme of Trustworthy Data-centric AI. The paper should at least have one major aspect that you can perform major constructive critique on and suggest and implement improvements. For example, if a paper lacks proper evaluation, then you can perform an evaluation; if a paper is missing a proper or major probing study, then you can implement one.

Report Structure & Style

The final report should use the **MaiNLP LaTex template** (see Moodle for the template) and (at least), include the following sections, and be **4-6 pages** of content, plus references (and optionally, an appendix):

1. **Introduction**: briefly introduce the research paper you selected and explain why you choose this paper and its relevance to the course
2. **Summary of the Paper**: provide a concise summary of the paper's key contributions, methodologies, and implications
3. **Critical Analysis**:

- briefly analyze the strengths of the paper in relation to trustworthy data-centric AI
- identify limitations or areas where the paper falls short (e.g. no discussion of severe data biases that implicate the results; lack of interoperability, limited evaluation protocols etc.) that you plan to improve on
- discuss whether the methods used align with the principles of trustworthy AI and suggest alternative approaches or potential ways to improve

4. **Improvements**:
   - select one of the major aspects identified above and then propose specific improvements or alternative strategies / approaches that could address the limitations or concerns
   - discuss new results or insights obtained from the proposed method or implementation
5. **Implementation**:
   - put into action your suggested improvement, but on a limited scale:
     i. If you suggested a user study, you could e.g. run it as a "pilot study" on one user (e.g. one of the other students in the course or a friend of yours).
     ii. If you suggested a model evaluation, you could run it on a smaller model (Llama2 can be run for free on Google Colab https://colab.research.google.com/drive/1X1z9Q6domMKl2CnEM0QGHNwidLfR4dW2?usp=sharing).
6. **Reflection**:
   - discuss your own implementation including your findings and your limitations
   - discuss the broader impact or implications of the paper's findings and/or your own findings related to trustworthy AI in practice
   - try to connect these findings with what's covered in the course (e.g., evaluation protocols, case studies, tools) to support your recommendations
7. **Conclusion**: summarize your critique and reflection and offer final thoughts on what could contribute to or hinder progress in trustworthy data-centric AI that you think are crucial but has yet to gain any attention

Below we provide two examples on how the critique/improvement/implementation could look like below. See the Examples for Final Reports section.

Timeline / Milestones

date (s)	action item	notes
08 January, 2025 (or before if you want to start early) **Hard deadline** (no extension)	**Step 1: Submit your topic proposal. Example topic proposals are provided. Requirement for exam.**	via Moodle
15 + 22 January	Support and feedback sessions (between step 1 and 2)	during lecture time (Slack also possible during that time)
14 February, 2025, time: 23:59 **Hard deadline** (no extension)	**Step 2:** Submit your **final course report and code/data**.	via Moodle

Key Considerations

1. **Relevance to Course Topics**: the report should demonstrate the ability to apply concepts from the course, including specific tools or frameworks etc. discussed in class
2. **Depth of Critique and Analysis**: the critique should go beyond surface-level observations; one should engage with the material in a way that reflects deep understanding
3. **Implementation of improvements**: the proposed improvement should be sound, realistic, actionable, and reproducible.

AI Policy

When utilizing generative AI tools, you must adhere to the following guidelines:

- **Transparency**. Explicitly disclose any use of AI tools in your work. This should be included in the last section of your report called **Ethical considerations**. This disclosure applies to all AI models, including coding assistants like Copilot. Note that **responsible** use of AI will **not** result in decreased grades. **When in doubt, ask and double check with the instructors!**
- **Accountability.** The accountability for any submitted work is your own only. You are thus responsible for:
    - The **correctness** of your own work, irrespective of the use of any tool. This includes, for instance, any factual statement or citation in the report. Be aware that generative AI can generate outputs including inaccurate and completely made up information, often referred to as “hallucinations".
    - The **originality** of your work. All material that is directly or indirectly attributable to papers, books, tutorials, or online content must be correctly referenced. Failure to do so will be considered plagiarism. Note that AI outputs may inadvertently contain plagiarized content, often without appropriate sourcing.

You will be **fully responsible** for the content of your report, including any portions produced by AI tools, and are liable for any breaches that may result from the use of such content.

Examples for Final Reports

*Example **proposal** and example **structure** for a report on the paper. Note that the actual report would obviously be a full text and not just these short notes. The papers and their corresponding specifics provided below should NOT be used for your own final report.*

Example A

Example Proposal

The proposal (which you have to submit on Moodle) could look like this:

Paper: Label-Descriptive Patterns and Their Application to Characterizing Classification Errors ICML 2022
One sentence summary: Global explanation method that helps users understand why their classifier failed.
Criticism: Authors claim that their method finds a more succinct and less redundant set of patterns. But they do the judging of what succinct means themselves and not by an external, blind evaluation.
Planned Implementation: Doing a pilot study with one user showing the users the raw outputs of the proposed and the baseline methods and asking users to identify reasons for misclassification

Example Report Structure

- Providing a summary of the paper:
    - Method to find global explanations for why a classifier misclassified using data mining technique. Studied on synthetic data with known ground truth and two use cases for VQA and NER.
- Possible criticisms:
    - Connecting to the question of actionability of explainability methods Possible criticisms you could point out:
    - Authors claim that “patterns” are easy to understand. But they do not verify with actual users.
    - Authors claim that their method finds a more succinct and less redundant set of patterns. But they do the judging of what succinct means themselves and not by an external, blind evaluation.
    - The experiment on actionability for NER is based on sampling data and not on actually fixing the identified issues (like preprocessing issues)

    - Bringing the work into wider context:
        - Connecting to more recent LLM explainability work
        - Connecting to HCI literature on global explanations (e.g. https://dl.acm.org/doi/10.1145/3542921 mentioned in the lecture)
- Possible future work to address criticisms: A report could contain a description of one of these studies (research question, how you would run the study, results you might expect)
    - Idea 1
        - Research question: Can users understand the patterns?
        - How: Semi-structured interview study showing users the patterns and asking them what these mean/what the reasons for misclassification with the classifier are
    - Idea 2
        - Research question: Does the proposed method provide more helpful insights than the baselines?
        - How: User study showing the users the raw outputs of the proposed and the baseline methods (which were published in the supplementary material) and asking users to identify reasons for misclassification
    - Idea 3
        - Research question: Are these insights actionable? Can an ML engineer improve a classifier based on these insights?
        - How: User study with ML engineers, letting them fix the NER classifier (either with the support of the proposed method or without) and measuring performance differences. “Think aloud” part of the study for qualitative evaluation.
- Implementation
    - Doing a pilot study of one of the suggested studies with just one participant.
- Discussions of your own suggested work and results including limitations.
- Reflection & Conclusion

Example B

Example Proposal

Paper: BAE: BERT-based Adversarial Examples for Text Classification, EMNLP 2020

One sentence summary: A method leveraging BERT embeddings to generate adversarial examples for text classification tasks by identifying key tokens and replacing them in a context-aware manner.
Criticism: While the method shows promising results in generating adversarial examples, it remains constrained to single-token replacements, which may limit the complexity and realism of the adversarial texts.
Planned Implementation: Reproduction of the methodology on the same benchmarks, extending it to multi-token replacement with a more in-depth error analysis comparing the previous single-token vs the new multi-token replacement strategy.

Example Report Structure

(the actual report would obviously be a full text and not just these short notes)

- **Summary of the paper**
  Black-box adversarial attack method. Proposes to use contextualized embeddings (BERT) to generate adversarial attacks on input texts. Prior work mostly used static embeddings and/or rule-based approaches.
  Method: find the most important tokens (by deletion of every token and inspecting drop in correct classification probability) and filter top-k predicted BERT alternatives by sentence embedding similarity (to avoid replacing "good" with "bad"). Insertions are also explored.
  Evaluation on 4 text classification tasks (Amazon, Yelp, IMDB, MR). Compared to a synonym-replacement strategy.
- **Possible criticisms**
  Identify multiple sources of criticism. You can go into more details on the points strictly related to your proposed future works and/or implementation.
  - Practical relevance/impact
    - Evaluation on topic classification tasks, less of societal relevance (e.g., hate speech)
  - Methodology / technical aspects
    - Devising alternative ways to judge token importance and determine replacement.
    - Currently restricted to single token replacement.

    - …
- **Possible future work to address limitations and criticisms**
  Identify multiple possible future works. You can go into more details on one specific aspect, designing a research plan to implement such future work
    - Related to practical relevance/impact
        - How does the method perform out-of-the-box on other tasks (e.g. hate speech detection) and/or languages?
    - Related to methodology / technical aspects
        - Improvement of existing method
            - E.g., better token replacement strategy, e.g., more linguistically informed
            - E.g., authors only used encoder-based small pre-trained LMs. Is there a connection to LLMs (beyond token replacement)?
    - Related to interpretability
        - Connecting to model internal interpretability methods
        - Connection to behavioral testing framework
    - Related to evaluation
        - Are the generated examples actually indistinguishable from the original ones? Would humans find them natural? Would humans change the label?
- **Implementation**
  The following are alternatives. Focus on one direction only
    - Replicate result of paper (code / TextAttack is available), improve it in **one** direction (relevance/impact, methodology, interpretability, evaluation)
    - Compare their baseline with new way of generating adversarial attacks via a human evaluation study (e.g., 1 student judging the resulting adversarial attacks)
    - Detailed error analysis
- **Discussions** of your own suggested work and results including **limitations**.
- **Reflection & Conclusion**