Q&A Session Answers

- *Course*: Big Data and Language Technologies
- *Date*: 20.06.2022

Questions on Organizational Matters

Q: Can we change the title?
A: Changing the specific title/topic of the project is possible before exposé submission; after exposé submission, please consult us regarding title/topic changes.

Q: Overlapping event on Mondays?
A: Only one person per group is required to attend, but we encourage attention.

Q: Are remote presentations for the exposés okay?
A: Yes.

Q: Sending in Slides?
A: Yes. We will compile a master slide deck beforehand. See deadline for slides submission. If you have content that does not fit a static PDF document, get in touch with us beforehand.

Q: Plan for next week?
A: Individual feedback on exposés by the topic supervisors. We will contact each group regarding the timeslot for each. The idea for those discussions is to refine your project plan; please consider integrating the suggested change into your upcoming presentations.

Questions on Formal Requirements

Q: Full text or keywords in exposés?
A: Full text preferred, but keywords are okay if it works better in the writing context.

Q: Template for exposé?
A: No template. Only requirement is 1(ish; a bit more is okay if needed) page A4 in PDF format. Feel free to style it however you want.

Q: Deliverables in the exposé?
A: The deliverables noted in the exposé instructions are guidelines; not all might apply to your case, but specifying them already in the exposé helps us understand your project plan better. Internally, it could help to attach a timeframe to each deliverable, but these are not needed in the exposé.

Q: Handing in code?
A: Your final submission needs to include a git project with any code you might created as part of the project. This does not need to be part of the exposé.

Q: Model cards? Datasheets? What?
A: The statement that you plan on creating model cards/datasheets can be included in the exposé (if applicable), but until your final submission these do not need to be fully specified.

Q: How detailed should the expose be?
A: The more detailed you make the exposé, the easier it is for us to provide valuable feedback. The exposé is not graded, but writing it is likely to make your project work much easier. Having a plan before starting to work is always good. So, keeping it general is okay, but details are appreciated.

Q: Is it okay to deviate from the exposé plan later throughout the project?
A: Yes, but please consult with us first.

Q: Do group roles need to be specified at all?
A: Yes, please indicate them in the exposé. Also, a contribution statement should be part of the final report.

Q: Does the distribution of work regarding the writing of the final report need to be specified in the exposé (i.e. a preliminary contribution statement)?
A: No, this does not need to be part of the exposé.

Q: Do organisation-related tasks for group roles (i.e. "setting up git repository", "proofreading") have to be specified as well?
A: No. The task overview should be kept general and succinct.

Q: Personal pronouns in exposé and report?
A: If it is the common or easier way of expressing a sentence, using pronouns is totally fine.

Q: What does "Are there baseline datasets?" mean?
A: Baseline datasets are datasets that exist in literature and that can be used to quantify the performance of your approach on established, external data.

Q: What is the difference between test data and baseline data?
A: Baseline dataset are usually used by most relevant papers in the field. Testing on them allows to compare to others' work. This adds value beyond just testing on a split of the training data in the same domain.

Questions on Research Questions

Q: How precisely does the research question need to formulated in the exposé?
A: As precisely as you can at this point. The more precise, the better feedback we can give. Also, precise RQ's are generally easier to answer than broad claims.

Q: Does the research question itself need to be BDLT-related, or only the method?
A: Only the method needs to be BDLT-related. Feel free to propose research on whatever you find interesting.

Q: Is a dataset in itself a deliverable?
A: Yes, of course. Annotations are a core part of many NLP-related tasks, and are a task in their own right.

Q: Focus only on a subset of the goals as indicated in the project fair?
A: Of course. Keep in mind though that if you narrow down your scope to a specific goal, the depth in which it would need to be worked on is deeper (to keep the same amount of work).

Q: What are regular expressions that could be used for explicit sentiment analysis?
A: The examples given in the slides are only suggestions. Coming up with other regexes is very much dependent on your research question (or are a research question in itself). The specific regexes are only a method, not an end goal; i.e. the language patterns defined by a regex are only representations of an underlying concept (that you want to classify). Also, using regexes is not a requirement, feel free to come up with other solutions.

Q: What is the overarching course topic of the seminar?
A: The focus is on modern (i.e. deep-learning) NLP techniques, which are commonly enabled by large amounts of text data. Alternatively, the topic of the course can be characterized by everything we did in lecture and lab.

Questions on Data

Q: Example dataset derived from the Web Archive pipeline?
A: For an example, see here: https://webis.de/data/webis-web-errors-19.html

Q: Annotated gold standard data – how big? (In this case, future prediction)
A: This depends if you intend to test, or also train the data. Make an educated guess (and see what similar datasets are like). Also take your available time for annotation into consideration.

Q: Publishing e.g. annotated data derived from the Web Archive?
A: We are open to publishing course results. Yet, these would be held to a higher standard than just course submissions and need to be conducted and written with a higher rigor.

Q: Model finetuned on A, retrained on B (Web-Archive-based), measured performance on C, where C is more similar to C than A. Makes sense?
A: Yes. The evaluation data should be from the same or a similar domain as the data you later want to apply the model to.

Questions on Experiments

Q: In which extend do we have to mention train-test leakage? We didn't feel like it is a big problem when having a clear train-test split.
A: It is good to think about this as potential problem. But of course, there are settings where leakage is improbable; yet with larger amounts of training data from e.g. the web (especially LLMs like GPT), the likelihood of leakage increases and should be kept in mind.

Questions on Group Work

Q: Is the task-to-person assignment fixed?
A: The final contribution statement in the report is what matters for us.

Questions on Core Project Work

Q: How is the work ideally distributed time-wise (sub-tasks)?
A: The total amount is roughly indicated by the ECTS equivalent. The distribution of that time is up to you and very dependent on the specific project.

Q: Consultations during the semester holidays?
A: Supervision is generally possible. Ask us for appointments at any time.

Questions on Presentations

Q: What is the point of having a presentation?
A: The idea is that other groups get an impression of what you're working on. Thus, the presentations should be approachable to people without prior knowledge on your specific topic.

Questions on Final Report

Q: What is a contribution statement?
A: See https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement for guidance.

Questions on Scientific Writing

Q: Is the use of personal pronouns discouraged in scientific writing?
A: If it is the common or easier way of expressing a sentence, using pronouns (e.g. "We developed a model...") is totally fine.