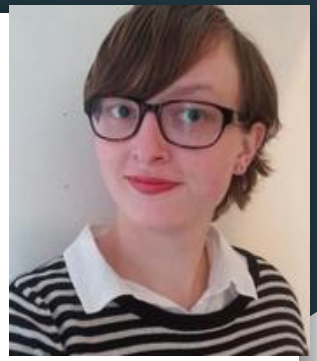
Preregistering NLP Experiments

From optimism to pitfalls

Myrthe Reuver, Open Science Coffee Leiden
24 Nov 2022

First: Who am I?



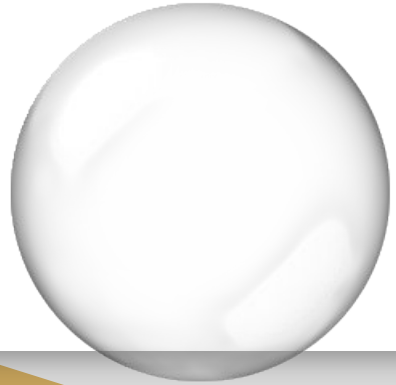
PhD Candidate @ CLTL, in computational linguistics
(Natural Language Processing - NLP).
Supervisors: Antske Fokkens (CLTL @ VU), Suzan Verberne (LIACS @ Leiden).

NLP is “Teaching computers how to deal with language”, related to Computer Science & AI.

- Spell checking, web searches, auto-fill… all use NLP.
- Language, and working with language data, is very complex.
  → nuances in meaning, pragmatics, normalization..

My work is on **viewpoint diversity in news recommendation** → specifically: detecting different opinions, arguments, and ideas in text with NLP → pop filter bubbles

I also like discussions about **responsible science** and **meta-science**



How does NLP research work?

- I work in the VU Humanities department, but my work is related to Computer Science for many reasons.
  - working with code, math, and sometimes more of an engineering approach ("*how can we build a system that does X*?") rather than hypothesis-driven research.

- Quite a young field (1950s), and explosive growth the past few years (2010s).
- The field goes fast since the 2010s and rise of neural models: many researchers want to "beat" other systems at SOTA (state of the art) at one or more NLP task(s) (detecting Named Entities, summarizing texts, etc).
- This is done on **benchmarks** (test or "exam" datasets on which systems are scored). **Methods from August 2022 can already be old!**

- Recent **meta-scientific community discussions** within NLP:
  - Involvement of **large private tech companies** → ethics, research directions, datasets and code.
  - Do we want "slow science" rather than "flag-planting"?

What to register

- Van Miltenburg et. al. (2021) identified how to preregister in NLP experiments
- They mention experimental conditions and hypotheses are often **implicit** in NLP work (assumptions about what will work better etc.)
- Neurips2021 had a **preregistration workshop** with acceptance of preregs: https://preregister.science/

What are your hypotheses/key assumptions?
What is the independent variable? (e.g. model architecture)
What is the dependent variable (e.g. output quality)
How will you measure the dependent variable?
Is there just one condition (corpus/task), or more?
What parameter settings will you use?
What data will you use, and how is it split in train/val/test?
Why this data? What are key properties of the data?
How will you analyse the results and test the hypotheses?

Table 2: Questions for analysis, experiments, and reproduction papers (expanded in Appendix A).

Questionable Research Practices in NLP

- **seed hacking**

  → recently dominant Transformer models are initialized with a random seed. This is supposed to be random, but optimizing for this factor can make your system appear better than it is.

  → Solutions: reproduction (my own stance reproduction paper found 50% of seeds underperformed), publishing 5 to 10 seeds with SD over seeds.

- **publishing only positive results**/well-performing systems

  → recent ‘medicine’ against this: the **negative results workshop** at main conferences by Anna Rogers and others.

  → maybe pre-registration can help this? → My work on **stance detection preregistration**

What is (preregistering) stance?

Stance detection: **classification task** (on texts, often tweets) with labels Pro, Con, Neutral towards an issue or topic;

“Abortion is a sin, and should never be practiced.”

Topic: Abortion, Stance: Con

- Registering: my expectations of models + datasets, in **explicit hypotheses**
- Already writing out literature review and results table, “just filling in”.

Hypotheses, example:

Hypothesis: *based on Shnarch et. al. (2022)'s experimental results on topic-dependent versus topic-independent tasks and pre-fine-tuning clustering,* we expect that SSSC models + pre-fine-tune clustering approach improve significantly over SSSC models without the pre-fine-tuning approach, since we consider stance classification a topic-dependent task and topic-dependent tasks responded well to this pre-fine-tuning task.

- Grounding in literature and/or earlier experiments;
- expectation;
- reasons

Pitfalls (mostly foreseen)

Pitfall 1: the **fast-changing methods;**

→ preregistering some models or parameters is risky when every month or week a new method comes out

Pitfall 2: NLP being about **building systems**

→ usually, strong publications are ones that “beat” others. This requires continuous changing and improving of plans, quite the opposite of preregistration

Pitfall 3: very **new and unestablished in NLP**

→ convincing reviewers is difficult, also publishing may be difficult when your system is not-SOTA.

→ Open Science platforms are not built for NLP research (some questions obsolete etc).

Unforeseen pitfalls

- I found NLP is relatively creative (solving puzzles/problems; connecting to recent work): having a preregistration felt **constricting** and led to me progress less

- Van Miltenburg's recommendation of writing before you do experiments is not the way many NLP researchers work, and **working with collaborators this way proved challenging.**

- Is NLP **a science** (with careful hypotheses and experiments) or more close to **engineering** or even **art**? In practice, many work less on hypothesis-based research. Ideals (of transparent science, clear hypotheses and results) are not aligned with everyday reality of how the field and research in it works.