Computational Argumentation for Responsible News Recommendation: connecting social science to NLP

Myrthe Reuver
Visiting Junior Researcher @ CSS, GESIS

Introduction talk, 26-03-2024

gesis
Leibniz-Institut
für Sozialwissenschaften

VU VRIJE UNIVERSITEIT AMSTERDAM

Who am I?

Myrthe Reuver, 4th year (!) PhD at Computational Linguistics & Text Mining (CLTL) lab, Vrije Universiteit Amsterdam.

supervisors: prof.dr. Antske Fokkens (VU), prof.dr.Suzan Verberne (Leiden University)

Computational linguist in an **interdisciplinary project** on diversity in news recommendation.

Fun facts: I used to be a local radio host in Almelo, and I love poffertjes.

Research interests

General: **argument mining**, **diversity, interdisciplinary/societal NLP**

are we measuring what we think we are measuring? 🤔

→ i.e.: careful evaluation, operationalization, and validation

Why do we do science this way, and how can we do it differently?

→ i.e. meta-scientific norms in NLP and beyond

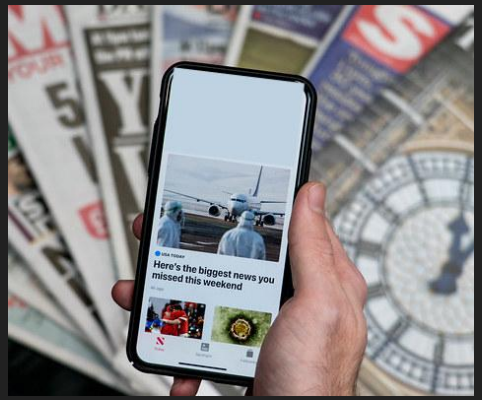
- *How can we combine theory, real-life context and use cases, and methods?*

Today, I briefly highlight work of 3 publications

OpenClipArt, Public domain

… And a brief discussion of my GESIS project!

1) News Recommendation and Diversity



- News recommendation: **more of the same**
- Why is this bad? Models of democracy
  - deliberative model
  - critical model

= citizens are required to see **diverse viewpoints on issues**

How to operationalize this?

- maybe **stance?**
- Stances are **positional claims about topics** (e.g. gun control, immigration, abortion). They indicate a position: **pro, against, or neutral.**

Reuver, M., Fokkens, A. & Verberne, S. (2021). **No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems.** In: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation (co-located at EACL 2021, online). Association of Computational Linguistics, p. 45–55.

Stance Detection

Stance detection, common definition: **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

1) Limitations of stance for viewpoint operationalization

In online news recommendation,

new topics and issues continuously appear online!

So:

How cross-topic robust are stances?

Joseph Mucira @ Pixabay, Simplified Pixabay License

Myrthe Reuver, Verberne, S., Morante, R., & Fokkens, A. (2021).
Is Stance Detection Topic-Independent and Cross-topic Generalizable?- A Reproduction Study.
In P*roceedings of the 8th Workshop on Argument Mining.*

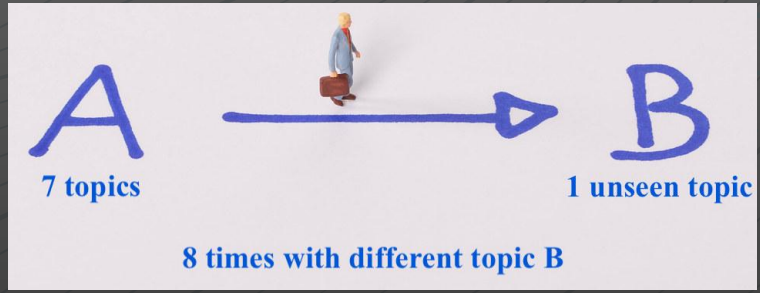
Cross-topic stance classification in Reimers (2019)

Train: 7 topics, test: 8th topic

Fine-tuning BERT (base & large)

Findings:

- avg. F1 (10 seeds) = 0.633
- +0.20 over reference model (LSTM)
- Results are *"very promising and stress the feasibility of the task"* (**Reimers et al. 2019**, p. 575)



Marco Verch @ Flickr, Creative Commons 2.0.
https://foto.wuestenigel.com/businessman-walking-from-a-to-b-point/

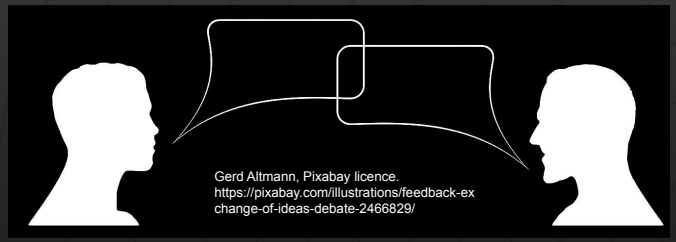
Dataset: UKP Dataset (Stab et. al., 2018)

25,492 arguments on 8 topics, in 3 classes:

- For or against "the use, adoption, or idea" of the topic, or no argument

- 8 controversial debate topics from internet forums:

  abortion, cloning, death penalty, gun control, marijuana legalization, minimum wage, nuclear energy and *school uniforms*.



Reproduction

- **Systematic reproduction:** 3 dimensions of reproduction (Cohen et. al.,2018): numeric values, findings, conclusions.

- Non-deterministic results of BERT:
  - **Standard deviation (SD) over seeds**;
  - value is reproduced if it falls **within 2 SDs.**

Cohen et. al. [2018]'s 3 dimensions of reproducibility

1. **(numeric) values:**

   Within 2 standard deviations

2. **findings** (relationship between variables, e.g. model & result):

   baseline < BERT-base < BERT-large,

3. **conclusion(s):**

   How feasible is cross-topic stance detection?

Mean (stdv) over 10 seeds	F1
Reimers et. al. (2019)	
LSTM (baseline)	**.424**
BERT-base	.613 (-)
BERT-large	.633 (-)
Reuver et. al (2021)	
SVM+tf-idf (baseline)	**.517**
Reproduction BERT-base	.617 (.006)
Reproduction BERT-large (all)	**.596 (.043)**
BERT-large - 5 good seeds	**.636 (.007)**

Results:

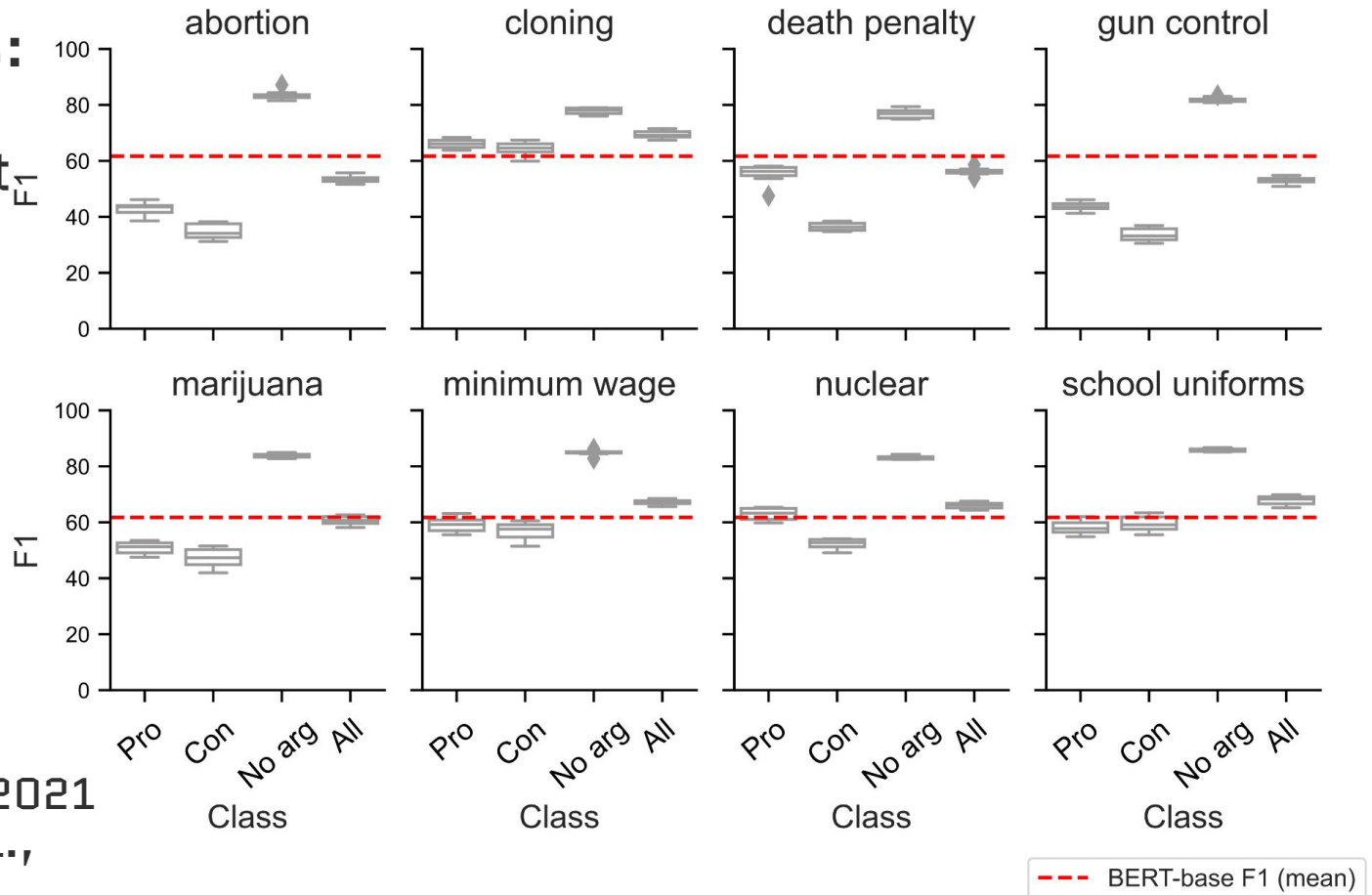
BERT-large under-performs in 50% of seeds

SVM+tf-idf model outperforms the LSTM reference model from the original study (F1 of .517 > .424)

Cohen et. al. [2018]'s 3 dimensions of reproducibility

1. **(numeric) values:**
   - ✓ Within 2 standard deviations (BERT-large = large SD)
2. **findings** (relationship between variables, e.g. model & result):
   - ✓ baseline < BERT-base < BERT-large,
   - ✗ .20 improvement over baseline is (much) smaller with SVM
3. **conclusion(s):**
   - ? How feasible is cross-topic? Let's investigate some more, especially on topics.

Crossing to other topics: difficult, inconsistent result



[Reuver et. al, 2021 of Reimers et al., 2019]

What does this mean?

Topic matters!

Stance not as topic-independent

- See also: Thorn Jakobsen et. al. (2021) >

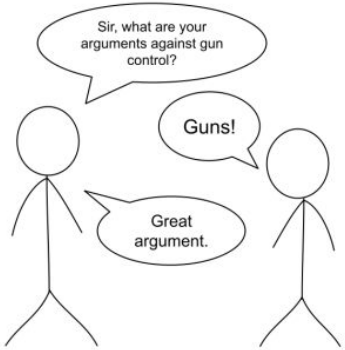
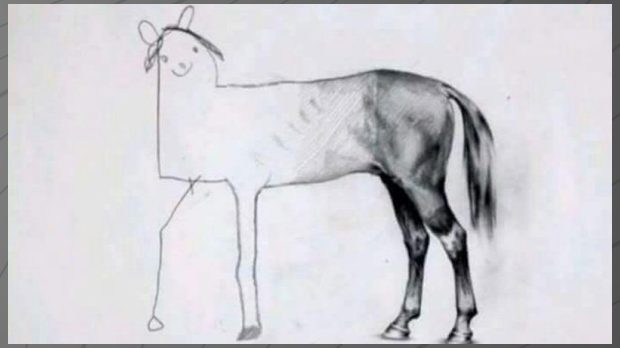


Figure 1: In human interaction, it is evident that relying on topic words for recognizing an argument is nonsensical. It is, nevertheless, what a BERT-based cross-topic argument mining model does.

3] Mixed Results in stance research

- What factors are **helping** in cross-topic stance?

  → What if people **only report what works**?

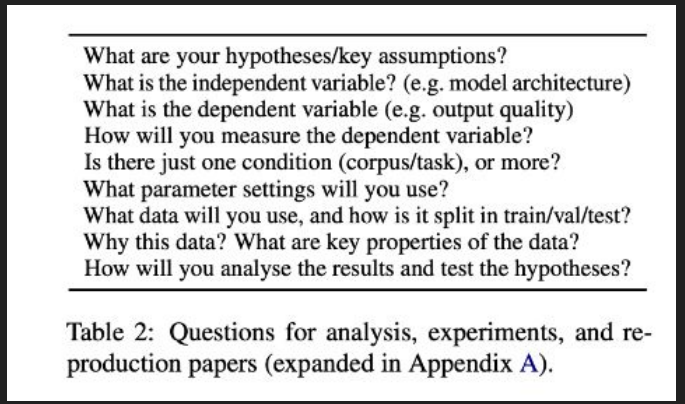


Myrthe Reuver, Suzan Verberne, Antske Fokkens (2023). **Investigating the Robustness of Modelling Decisions for Few-Shot Cross-Topic Stance Detection: A Preregistered Study**--> accepted to LREC-COLING 2024

Pre-registration

- 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?

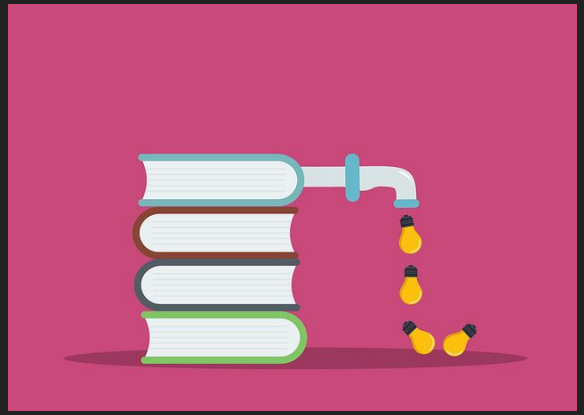
Table 2: Questions for analysis, experiments, and reproduction papers (expanded in Appendix A).

Why pre-registering stance?

Many papers in the few-shot, cross-topic stance field **claim exceptional progress while only testing one dataset,**

or only comparing one modelling choice.

- Positive results bias?
- Robust improvement?

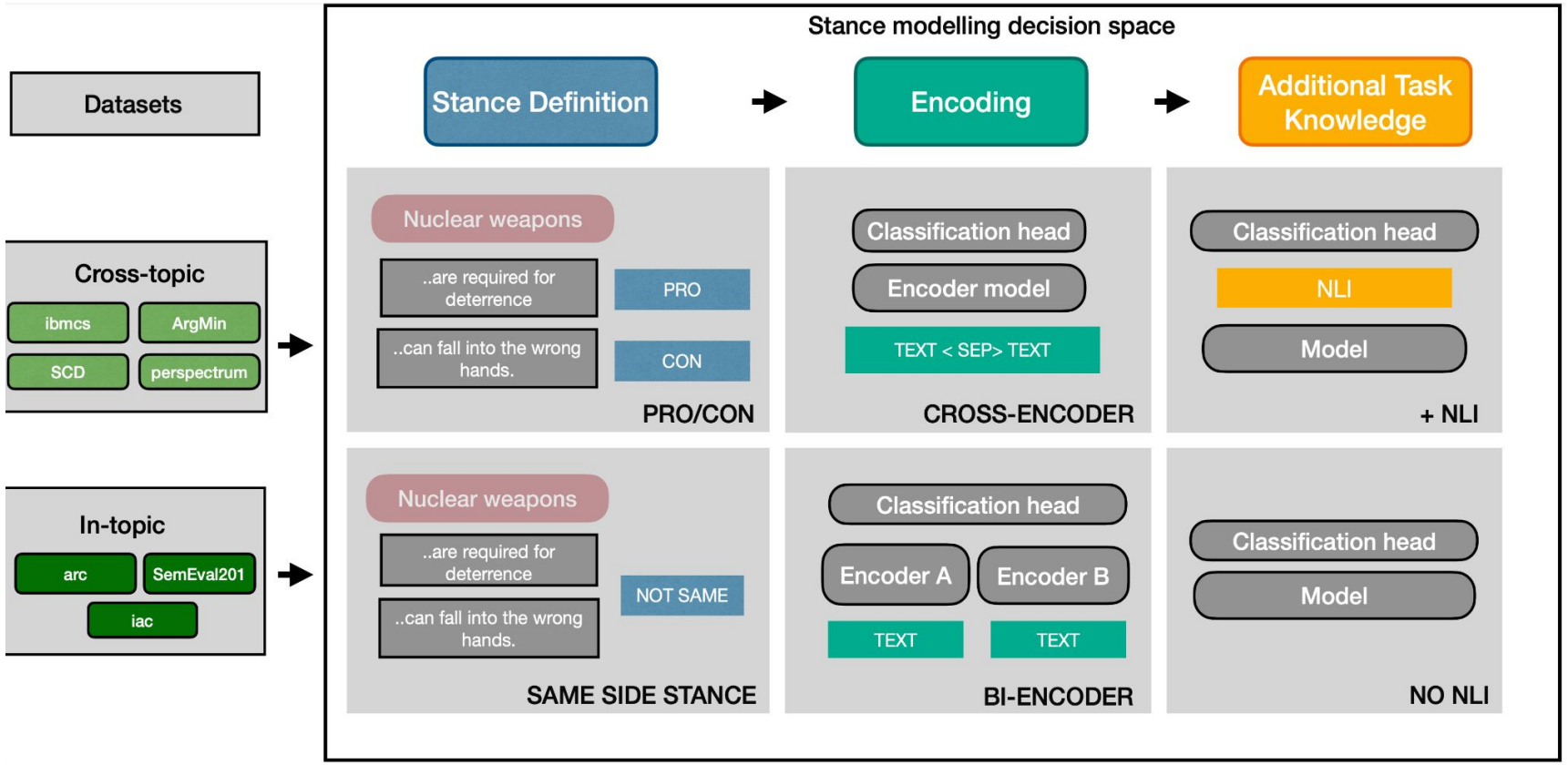


Systematic stance detection experiments

I **pre-registered** RQs, hypotheses and analysis plans.

> From AsPredicted.com: *“Would a reader wonder whether a given decision about analysis, data source or hypothesis was made after knowing the results?"*

- **What?** Testing claims on what is more topic-independent, specifically Same Side Stance (SSS) in a **pair-wise classification setting.**


Datasets
Cross-topic
ibmcs
ArgMin
SCD
perspectrum
In-topic
arc
SemEval201
iac
Stance modelling decision space
Stance Definition
Encoding
Additional Task Knowledge
Nuclear weapons
..are required for deterrence
..can fall into the wrong hands.
PRO
CON
PRO/CON
Classification head
Encoder model
TEXT < SEP> TEXT
CROSS-ENCODER
Classification head
NLI
Model
+ NLI
Nuclear weapons
..are required for deterrence
..can fall into the wrong hands.
NOT SAME
SAME SIDE STANCE
Classification head
Encoder A
Encoder B
TEXT
TEXT
BI-ENCODER
Classification head
Model
NO NLI

5 Hypotheses, 7 datasets, 100 shots from each dataset

- **Task definition:**

1.1: **SSSC definition** to be more cross-topic robust than the pro/con

1.2: **Size of the topics** in training/test splits does not relate with the classification performance in cross-topic pro/con stance classification.

- **Encoding Choices:**

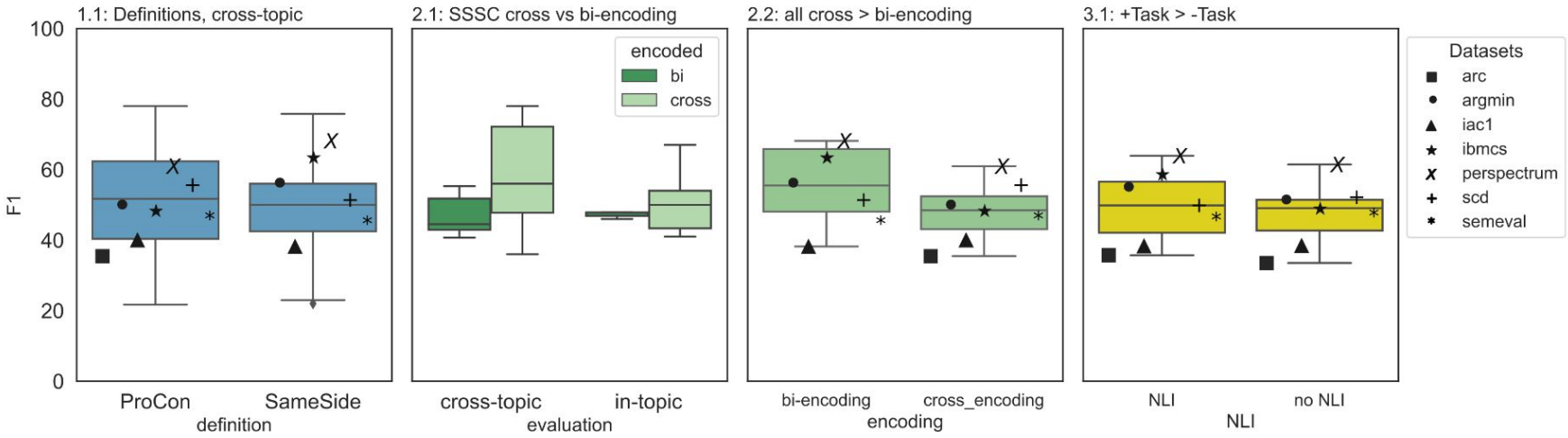
2.1: we expect **bi-encoding** to **fluctuate less** between in-topic to cross-topic performance, and improve cross-topic performance.

2.2: We expect **cross-encoding** to perform better in both cross-topic and in-topic

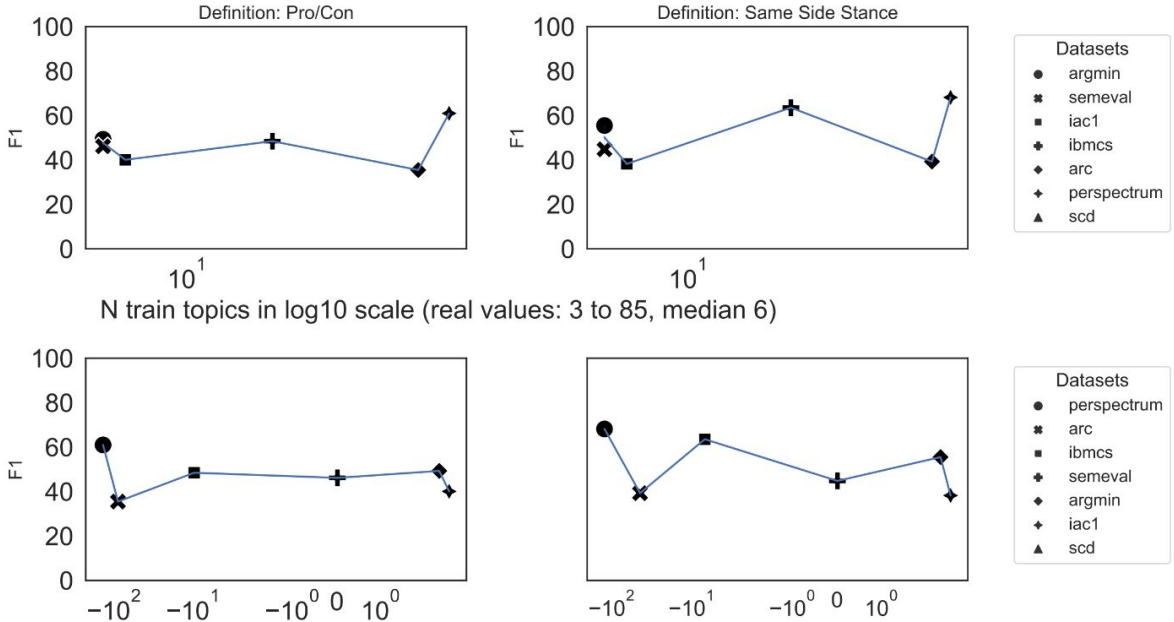
- **Task Knowledge**

3.1: adding **NLI training** to the model will lead to classification performance gains over models without NLI training

Results, per hypothesis



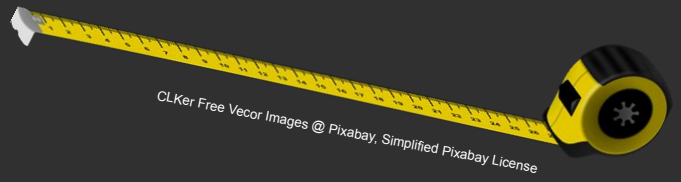
1.2: Influence of N Topics on Classification Performance



Preregistration of stance experiments shows:

- Properly measuring “this works better” only works when measuring **different modelling choices**, and different datasets;

- often, performance is more related to benchmark dataset choice than actual modelling choice.



GESIS Project: Instruction-tuned models and theory knowledge

Research Questions:

- are instruction-tuned generative models able to **detect complex theoretical constructs in texts,** and how can we **evaluate** whether models can?
- Can we **combine theoretical knowledge about the concept** with model evaluation?

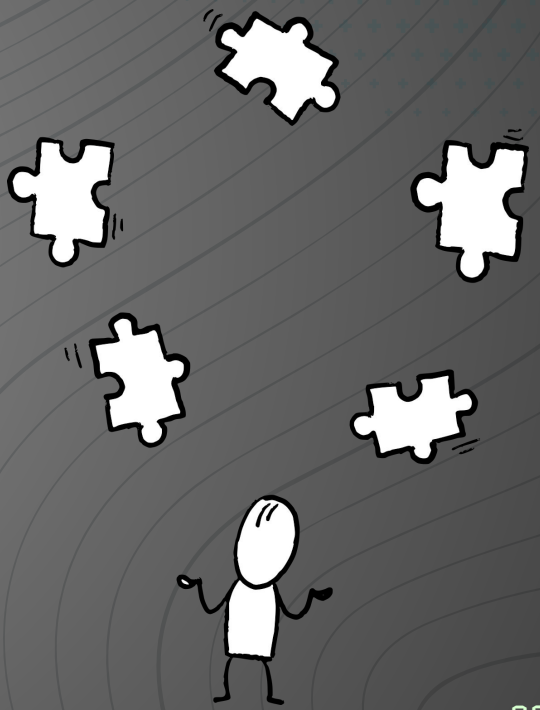
Method :

- ask experts on constructs on evaluation;
- carefully distinguish effects of (i) construct theory; and (ii) generating additional data

- What are effects on model accuracy and validity?

Overall, my research shows:

NLP in news recommendation means juggling different key decisions: theoretical concept (of viewpoint and of democracy), task, data, and evaluation. Also, input from different experts!



Thank you!

Myrthe Reuver, Vrije Universiteit Amsterdam

myrthe.reuver[at]vu.nl

@myrthereuver