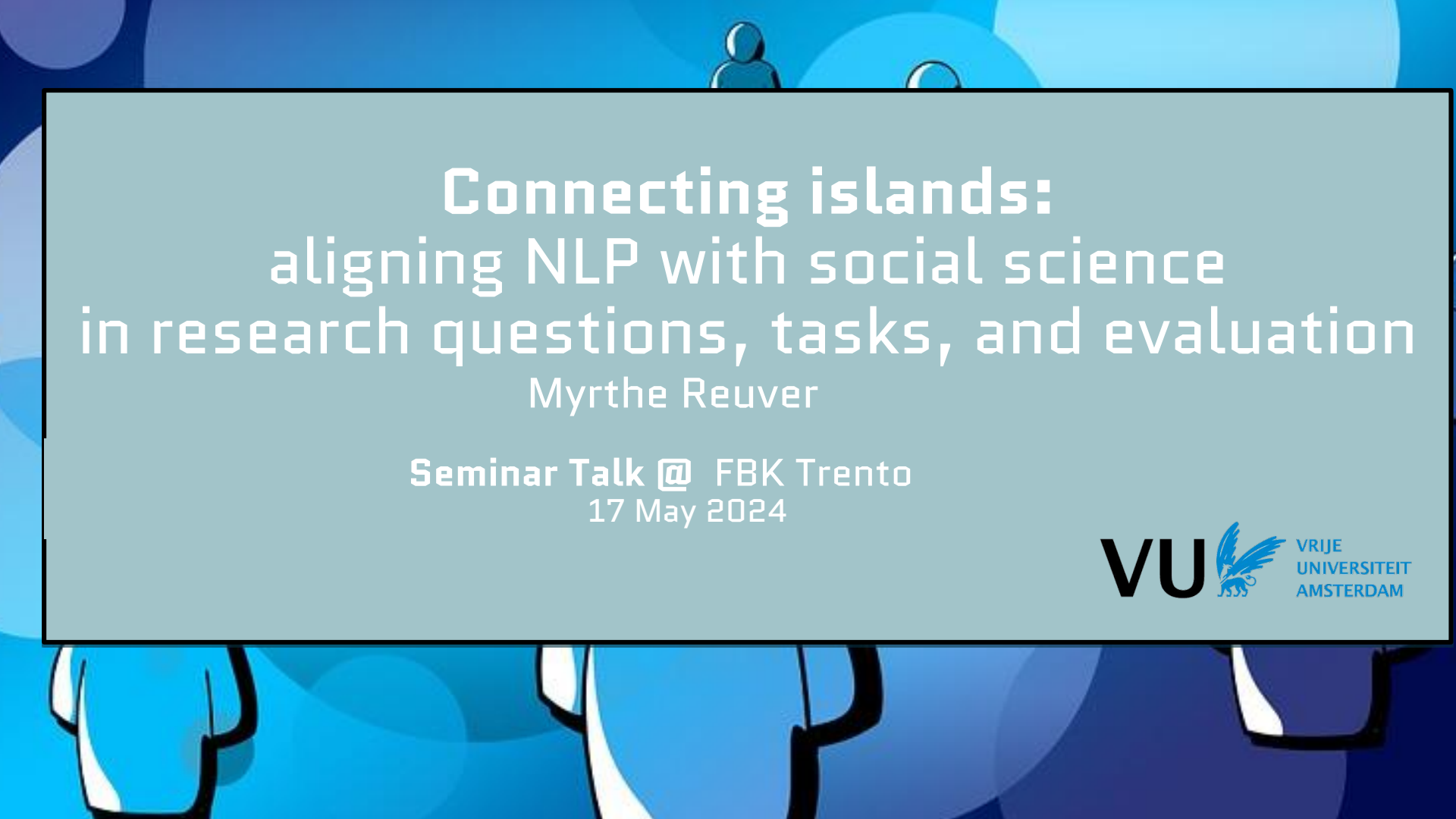


Connecting islands:
aligning NLP with social science in research questions, tasks, and evaluation

Myrthe Reuver

Seminar Talk @ FBK Trento
17 May 2024

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?*

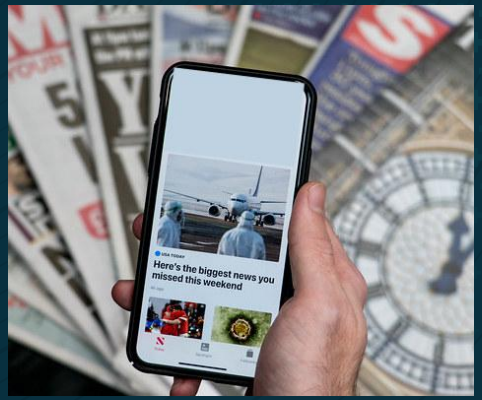
Bridging two islands

Today, I will highlight several projects, on:

- viewpoint diversity;
- stance detection; and
- hypocrisy accusation detection.

which contain **a connection** between social science and NLP;
but highlight **difficulties** of the bridge between the two islands.

1] News Recommendation and Diversity



News recommendation: more of the same,

→ filter bubbles

Why is this bad?

- Models of democracy
  - deliberative model
  - critical model

citizens are required to see diverse viewpoints on issues

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.

NLP task connection

Deliberative democracy requires rational debate , and for news recommendation recommending a diverse set of viewpoints.

OpenClipart Vectors @ Pixabay

How to operationalize this?

- our position paper discusses several established NLP Tasks (media frames, perspective detection, stance detection), and how they (do not) fit this goal.

- **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

2) 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.*

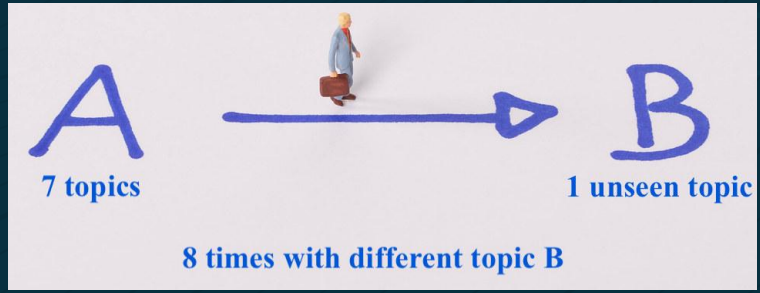
Cross-topic stance classification in Reimers (2019)

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/

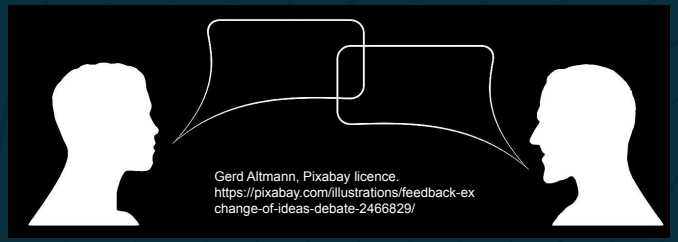
Dataset: UKP Dataset (Stab et. al., 2018)

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:

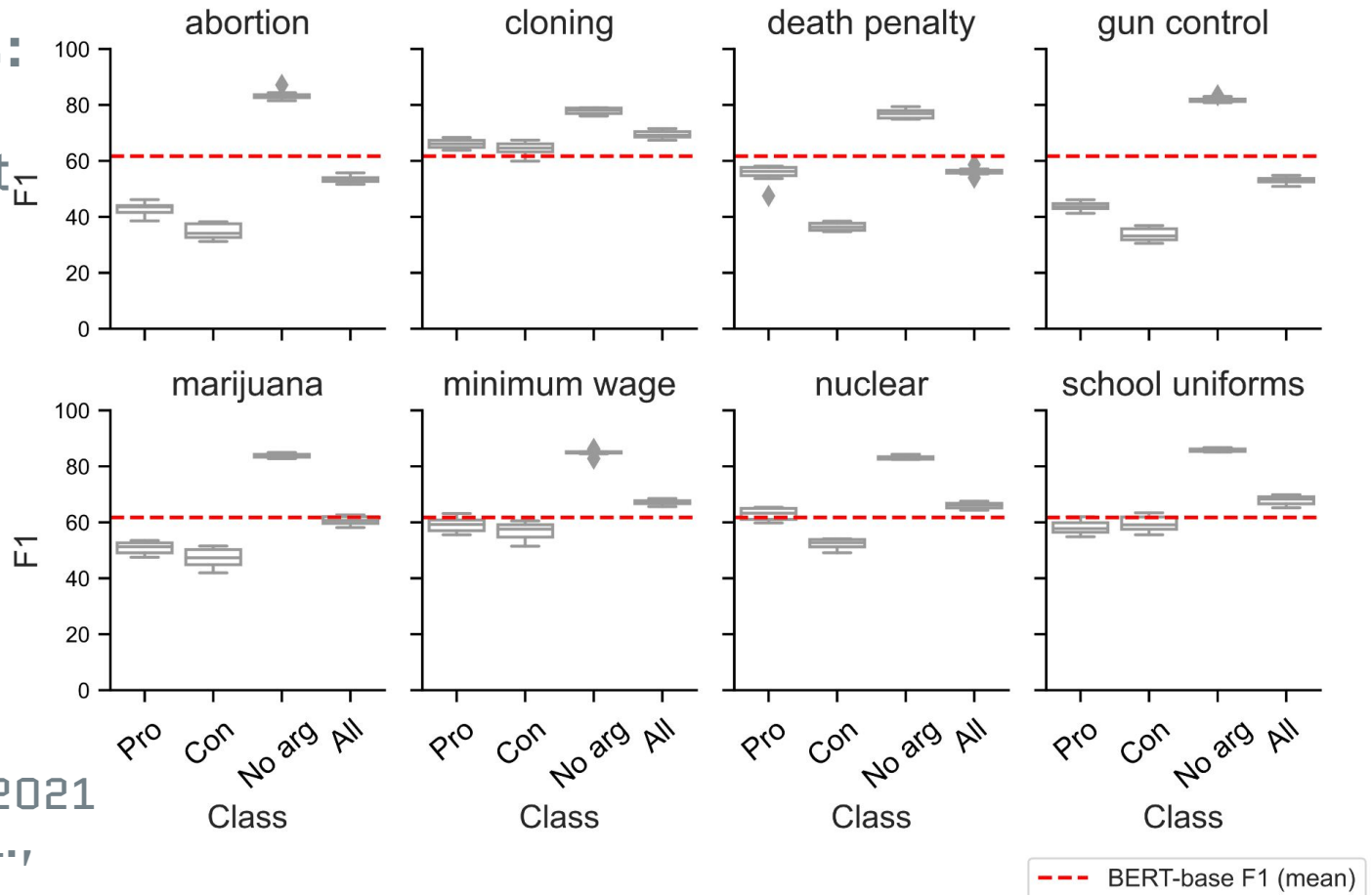
BERT-large under-performs in 50% of seeds

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) >

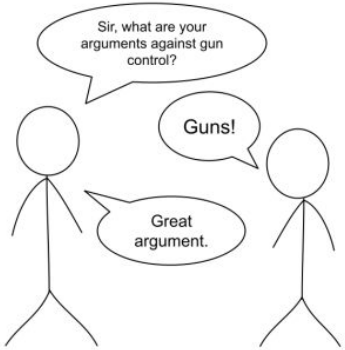


Figure 1: In human interaction, it is evident that relying on topic words for recognizing an argument is nonsensical. It is, nevertheless, what a BERT-based cross-topic argument mining model does.

2] Mixed Results in stance research

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

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

  → what if we use approaches against positive results bias from social science: **preregistration?**

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

Pre-registration

- 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.

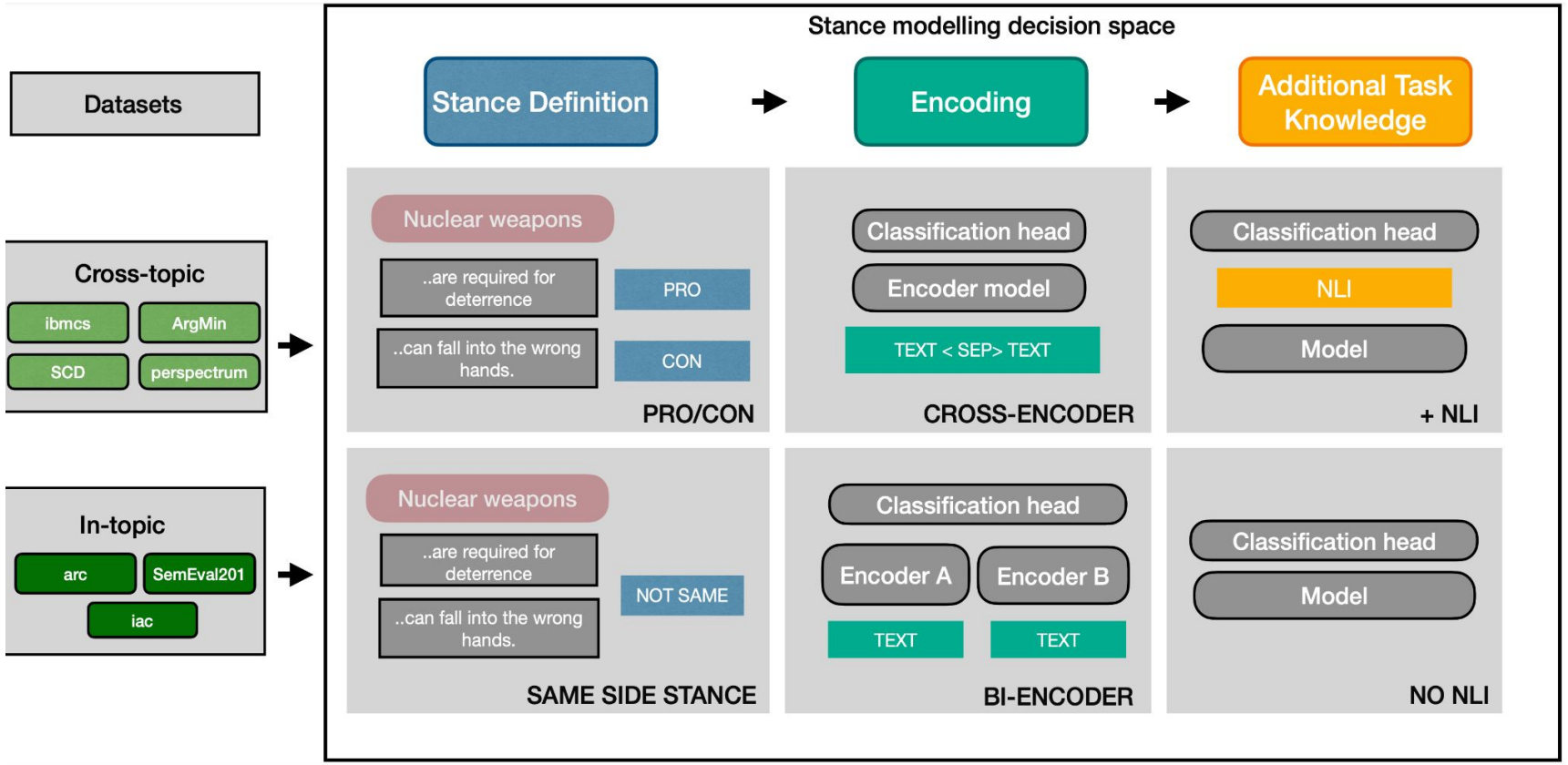
- Positive results bias?
- Robust improvement?

Systematic stance detection experiments

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:

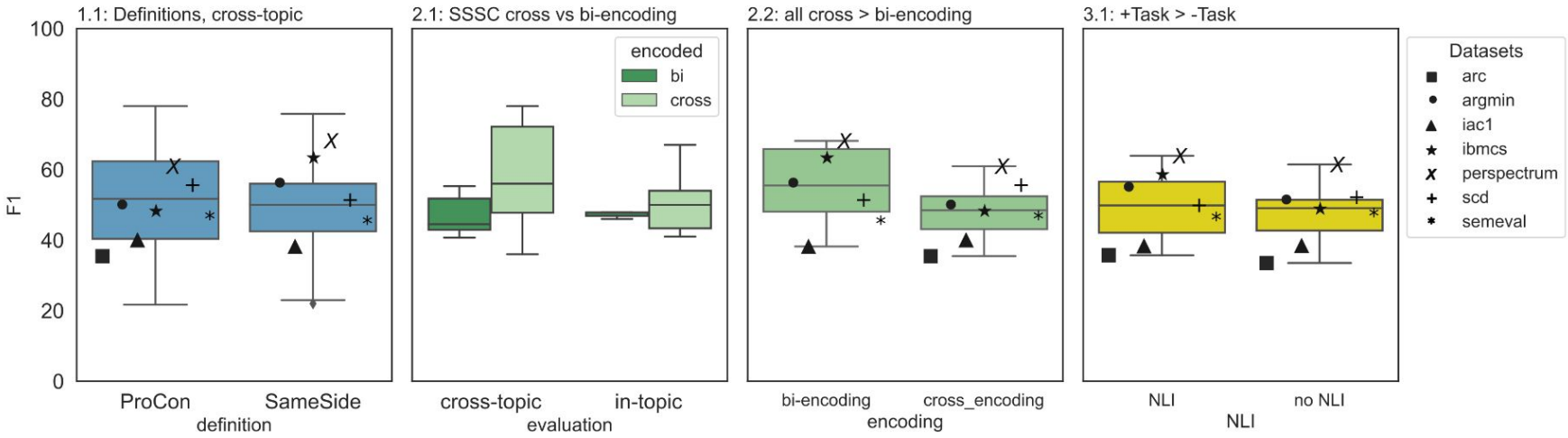
2.1: we expect bi-encoding to fluctuate less between in-topic to cross-topic performance, and improve cross-topic performance.

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

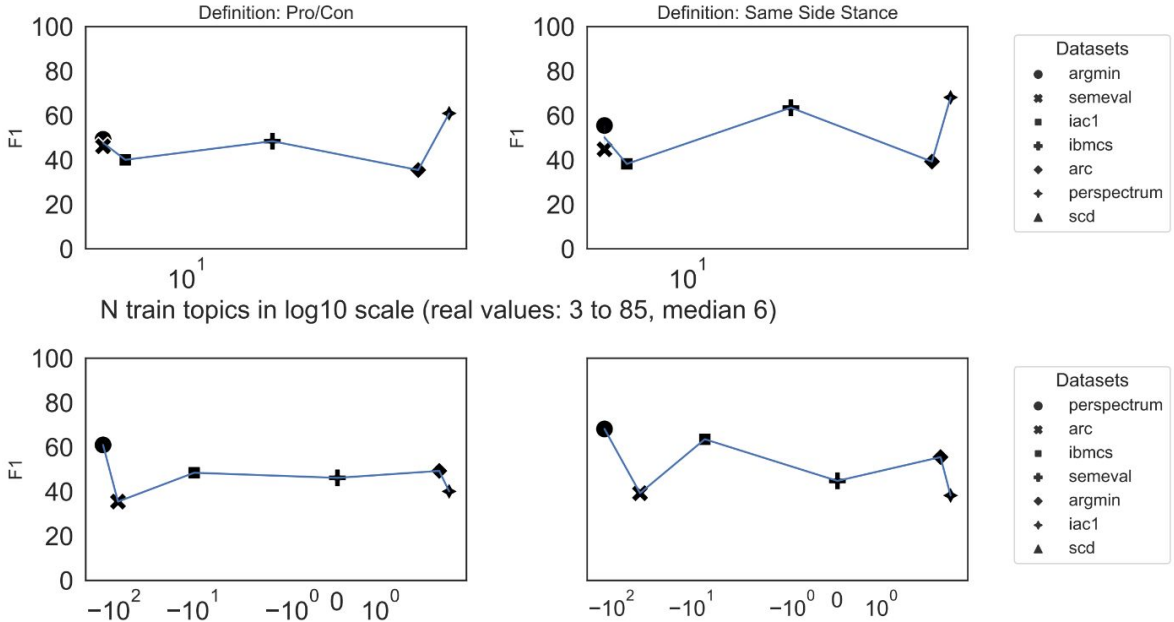
- Task Knowledge

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 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.

3] Hypocrisy accusation in online climate debate

Why hypocrisy? the Go-to Political Accusation

Hypocrisy **accusations** are abundant in politics

- Easy to make
- Effective
- In polarized polity - the only rhetorical tool available?

Image: Johan Eklund, Flickr

Politics: a “never ending fight to ferret out hypocrites” (Arendt, 2006, p. 93)

Paulina Garcia-Corral, Avishai Green, Hendrik Meyer, Myrthe Reuver,

Xiaoyue Yan, and Anke Stoll. CompText talk, 2024


#HACKICA23
TORONTO
This project: born at the ICA23 hackathon
NIKE
SPORTSWEA

Hypocrisy accusation: less attention in NLP tasks and low recall even with state-of-the-art models

Habernal et al. (2018): **sub-concept in fallacy detection**.

Instruction-tuned models (GPT and T-5) & climate debate by Alhindi et al. (2022):

- In their 5 fallacy datasets, only **one dataset** with hypocrisy-related category, “whataboutism”
- Training on the other 4 datasets, they detect **"whataboutism" with .44 accuracy,**
- adding a definition leads to a small reduction to .43.

Piskorski et al. (2023) have designed a multilingual dataset on online news with an annotated hypocrisy accusation concept, as part of a **"persuasian techniques"** task.

- They also introduce an XLM-RoBERTa model as baseline. One of the debates: on climate change
- Their appendix reports a performance of the Whataboutism concept of .25% precision, with **extremely low recall (.034%)** leading to an F1 of .06.
- This concept is **only .05% of their dataset**

Hypocrisy Accusation Detection with small training samples

Detecting hypocrisy accusations in online debates (Reddit) with few examples.

Challenging because

- Explicit vs. implicit
  - *"Exactly! Imagine the US with three times the CO2 per capita to ask China to reduce emissions… THAT is hypocrisy."*
  - *Yet when I see those who make money on fossil fuels brag how clean they are … seriously, how dare you?*

RQ: Is few-shot learning suitable to detect accusations of hypocrisy?

Initial Results

- SVM: unsuccessful/incapable of predicting minority-class positive cases.
- SETFIT outperforms SVM when around 50-100 shots.
- GPT 3.5 Turbo: Most promising. 42/45 comments correctly classified!
  - including: *"Like Zuckerberg, who bought his neighbors houses to protect his privacy, while making billions selling other people privacy. ~~Hypocrisy is a virtue for these people~~"*

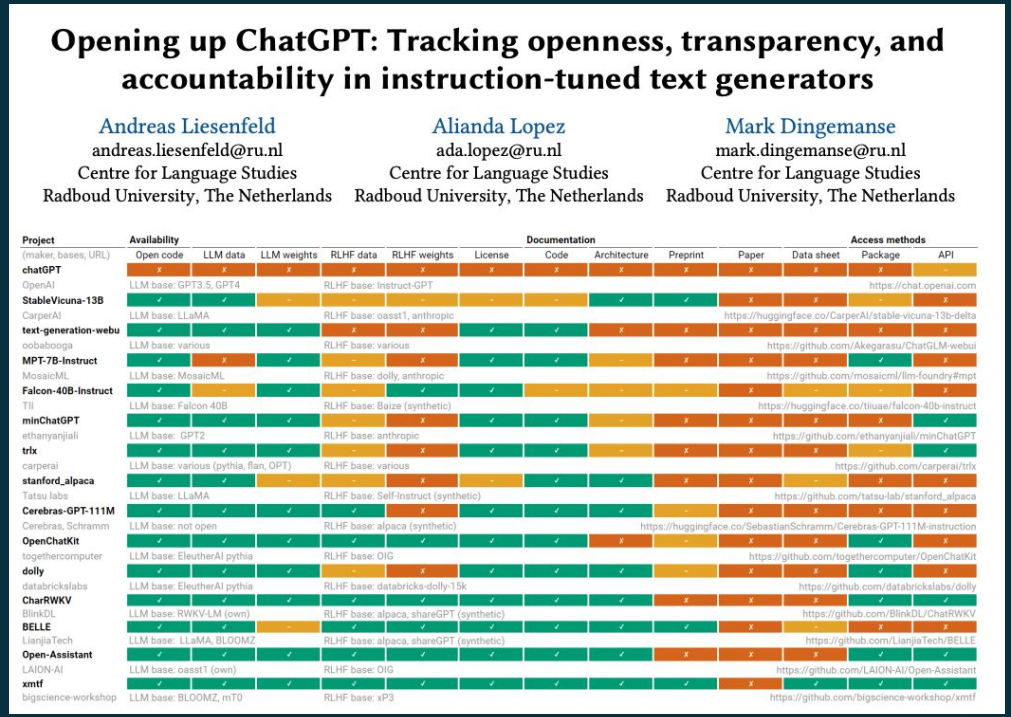
Conclusions:

- Task requires precise conceptualization of the complex concept + careful evaluation
- Few-shot classifying struggles with distinguishing implicit instances
- GPT looks promising…

Mixtral-8x7B-Instruct: more open model

Mistral AI, French start-up.
"Mixture of Experts", like GPT (is rumored to be)

- Access through huggingface API access: free hourly rate limit
  - Pro: Apache 2.0 license , free for academic and commercial usage, release papers about model development
- Con: not fully open, e.g. dataset information



Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators

Andreas Liesenfeld
andreas.liesenfeld@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Alianda Lopez
ada.lopez@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Mark Dingemanse
mark.dingemanse@ru.nl
Centre for Language Studies
Radboud University, The Netherlands

Project	Availability						Documentation					Access methods	
(maker, bases, URL)	Open code	LLM data	LLM weights	RLHF data	RLHF weights	License	Code	Architecture	Preprint	Paper	Data sheet	Package	API
chatGPT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	-
OpenAI	LLM base: GPT3.5, GPT4			RLHF base: Instruct-GPT									https://chat.openai.com
StableVicuna-13B	✓	✓	-	-	-	-	-	✓	✓	✗	✗	-	✗
CarperAI	LLM base: LLaMA			RLHF base: oasst1, anthropic									https://huggingface.co/CarperAI/stable-vicuna-13b-delta
text-generation-webu	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
oobabooga	LLM base: various			RLHF base: various									https://github.com/Akegarasu/ChatGLM-webui
MPT-7B-Instruct	✓	✗	✓	-	✗	✓	✓	-	✗	✗	✗	✓	✗
MosaicML	LLM base: MosaicML			RLHF base: dolly, anthropic									https://github.com/mosaicml/llm-foundry#mpt
Falcon-40B-Instruct	✓	-	✓	-	✓	✓	-	-	-	✗	-	-	✗
TII	LLM base: Falcon 40B			RLHF base: Baize (synthetic)									https://huggingface.co/tiiuae/falcon-40b-instruct
minChatGPT	✓	✓	✓	-	✗	✓	✓	-	✗	✗	✗	✗	✓
ethanyanjiali	LLM base: GPT2			RLHF base: anthropic									https://github.com/ethanyanjiali/minChatGPT
trlx	✓	✓	✓	-	✗	✓	✓	-	✗	✗	✗	-	✓
carperai	LLM base: various (pythia, flan, OPT)			RLHF base: various									https://github.com/carperai/trlx
stanford_alpaca	✓	✓	-	-	✗	-	✓	✓	✗	✗	-	✗	✗
Tatsu labs	LLM base: LLaMA			RLHF base: Self-Instruct (synthetic)									https://github.com/tatsu-lab/stanford_alpaca
Cerebras-GPT-111M	✓	✓	✓	✓	✗	✓	✓	✓	-	✗	✗	✗	✗
Cerebras, Schramm	LLM base: not open			RLHF base: alpaca (synthetic)									https://huggingface.co/SebastianSchramm/Cerebras-GPT-111M-instruction
OpenChatKit	✓	✓	✓	✓	✓	✓	✓	✗	-	✗	✗	✓	✗
togethercomputer	LLM base: EleutherAI pythia			RLHF base: OIG									https://github.com/togethercomputer/OpenChatKit
dolly	✓	✓	✓	-	✗	✓	✓	✓	-	✗	✗	✓	✗
databrickslabs	LLM base: EleutherAI pythia			RLHF base: databricks-dolly-15k									https://github.com/databrickslabs/dolly
CharRWKV	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
BlinkDL	LLM base: RWKV-LM (own)			RLHF base: alpaca, shareGPT (synthetic)									https://github.com/BlinkDL/ChatRWKV
BELLE	✓	✓	-	✓	✓	✓	✓	✓	✓	✗	-	✗	✗
LianjiaTech	LLM base: LLaMA, BLOOMZ			RLHF base: alpaca, shareGPT (synthetic)									https://github.com/LianjiaTech/BELLE
Open-Assistant	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
LAION-AI	LLM base: oasst1 (own)			RLHF base: OIG									https://github.com/LAION-AI/Open-Assistant
xmtf	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
bigscience-workshop	LLM base: BLOOMZ, mT0			RLHF base: xP3									https://github.com/bigscience-workshop/xmtf

Preliminary results - how reliable is our model?

Comparing **different prompt versions**, we found:

+ “reason about it”:

  Often does not lead to mistral adding reasons, but does seem to increase proficiency at task.

+ Examples:

  More is not always better

Conclusion: Islands of concepts, methods, and ideas

Social science:

- defining and analyzing (viewpoint, hypocrisy, etc);

NLP: developing tasks, model development and testing:

- tasks that are difficult are interesting (to test model limits);
- Difficult tasks, but: easy(ish) to annotate;
- less social science theories in task conceptualizations;
- fast development

Both:

- real-world application and impact is a wish but not always there

But the ✨**combination of both**✨ **may actually make the magic happen!**

Challenges in connecting the islands

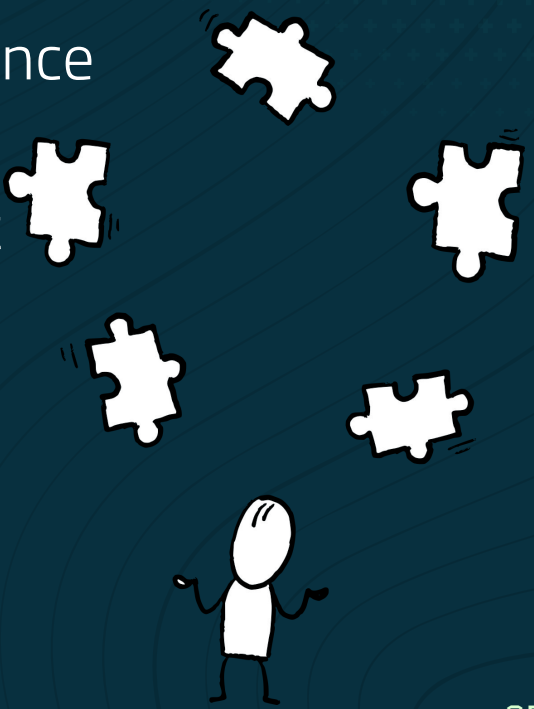
- **shared language and talking;**
- **balancing approaches to research,** e.g. preregistration vs fast development;
- Connecting **tasks to real-world** and social science research needs

In general:

Interdisciplinary research in NLP and social science means **juggling the different views**;

key decisions: theoretical concept (of viewpoint and of democracy), task, data, and evaluation.

Talking to each other is key!



Thank you!

Myrthe Reuver, Vrije Universiteit Amsterdam

myrthe.reuver[at]vu.nl

@myrthereuver