

Stance Detection: A Task Definition, Use-Case, and Recent Research

Myrthe Reuver

Course: MSc Text Mining (Leiden University)
9 November 2022

Who am I?



Myrthe Reuver, PhD candidate at CLTL at VU Amsterdam.

→ Supervisors: Antske Fokkens (CLTL @ VU), Suzan Verberne (LIACS @ Leiden).



My Research is on Text Mining in an **interdisciplinary project** on diversity in news recommendation.

with: social scientists, philosophers, and RecSys/computer scientists.

Argument Mining and Stance

- **Argument Mining** is a sub-field of NLP dealing with argumentative texts and debates - for instance in online debate portals, essays, or news texts.
- Human debate is full of **stances**:
People expressing whether they agree or disagree with arguments and topics.



What is stance detection?

- **Stance Detection:** a classification task classifying **texts** (tweets, comments, reviews..)
- Modelling the **stance relationship between such a text and a target:**
 - ..a topic/issue/question, OR;
 - ..a second text/headline/news article.
- Common labels:
 - **Pro** (text¹ agrees with text²/topic);
 - **Con** (text¹ disagrees with text²/topic);
 - **Neutral** (text¹ does not agree but also not disagree with text²/topic);
 - Sometimes: a **questioning/discussing** label: text¹ asks a question about text²/topic

Example (not necessarily my own stance):

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

Topic: Abortion, Stance: Con

Current methods

- **Classification method:** Pre-trained Large Language Models such as BERT and RoBERTa
- **Stance Benchmark** (Schiller et al., 2021) combines 10 different stance datasets:

Table 2 All datasets, grouped by domain and with examples

Dataset	Domain	Topic	Comment	Stance
ibmcs	Encyclopedia	[...] atheism is the only way	Atheism is a superior basis for ethics	PRO
semeval2019t7	Social media	(Charlie Hebdo)	"[...] #CharlieHebdo gunmen have been killed" yayyy [...]	Support
semeval2016t6		Feminist Movement	[...] every women should have their own rights!! #SemST	Favor
fnc1	News	Hugh Hefner Dead?	Hugh Hefner has denied reports that he is dead [...]	Disagree
snopes	Debating forums	Farmers feed their cattle candy [...]	[...] padding out cow feed with waste candy is nothing new.	Agree
scd		(Obama)	I think Obama has been a great President. [...]	For
perspectrum		School Day Should Be Extended	So much easier for parents!	Support
iac1	Web search	existence of god	[...] the Bible tells me that Jesus existed [...]	Pro
arc		Salt should have a place at the table	[...] the iodine in salt is necessary to prevent goiter. [...]	Agree
argmin		school uniforms	We believe in freedom of choice.	CON

Topics in parentheses signal implicit information

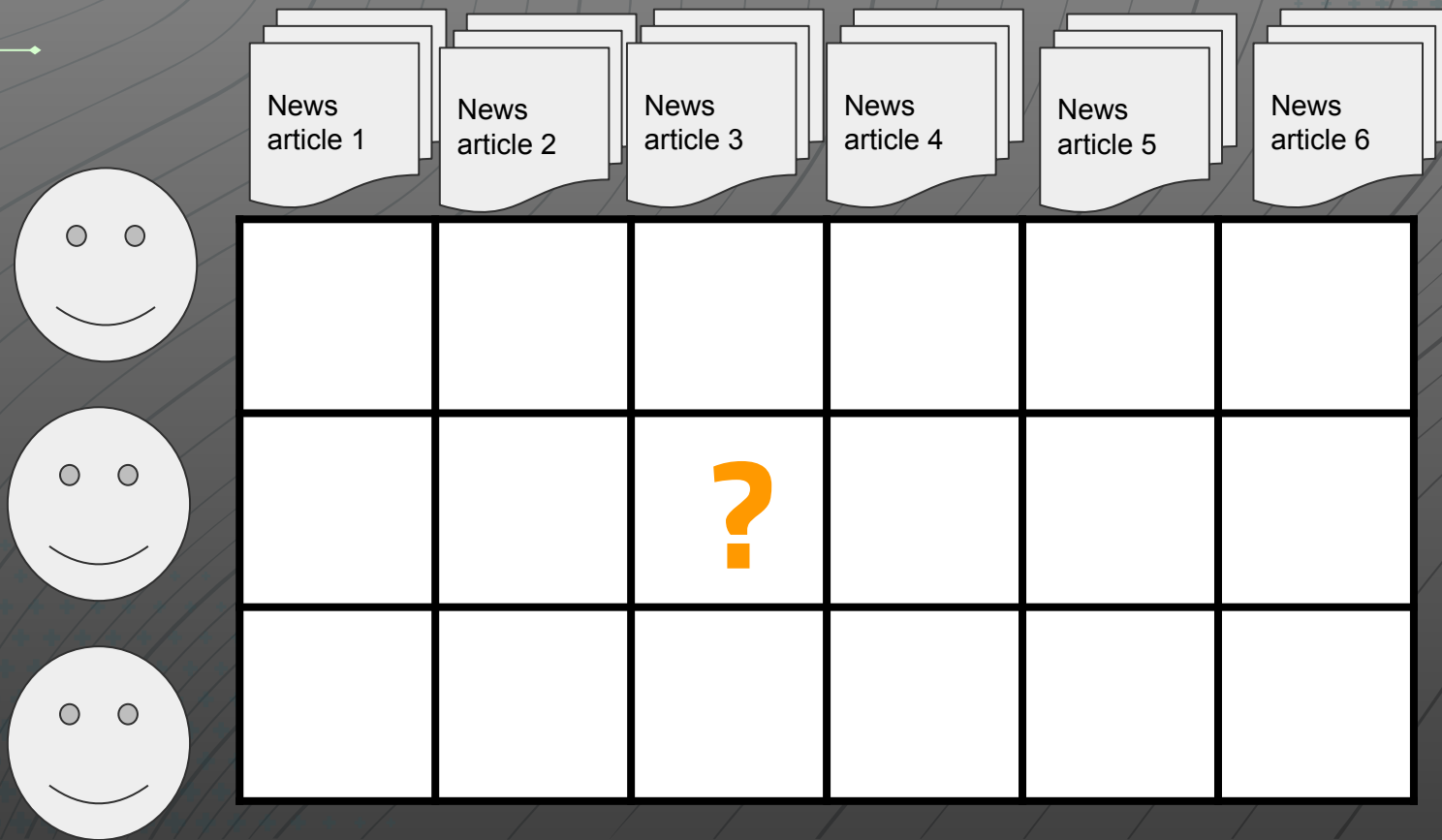
What could be the role of **stance detection models** in news recommender systems?

My own use case: **diversity of viewpoints** in news recommendation

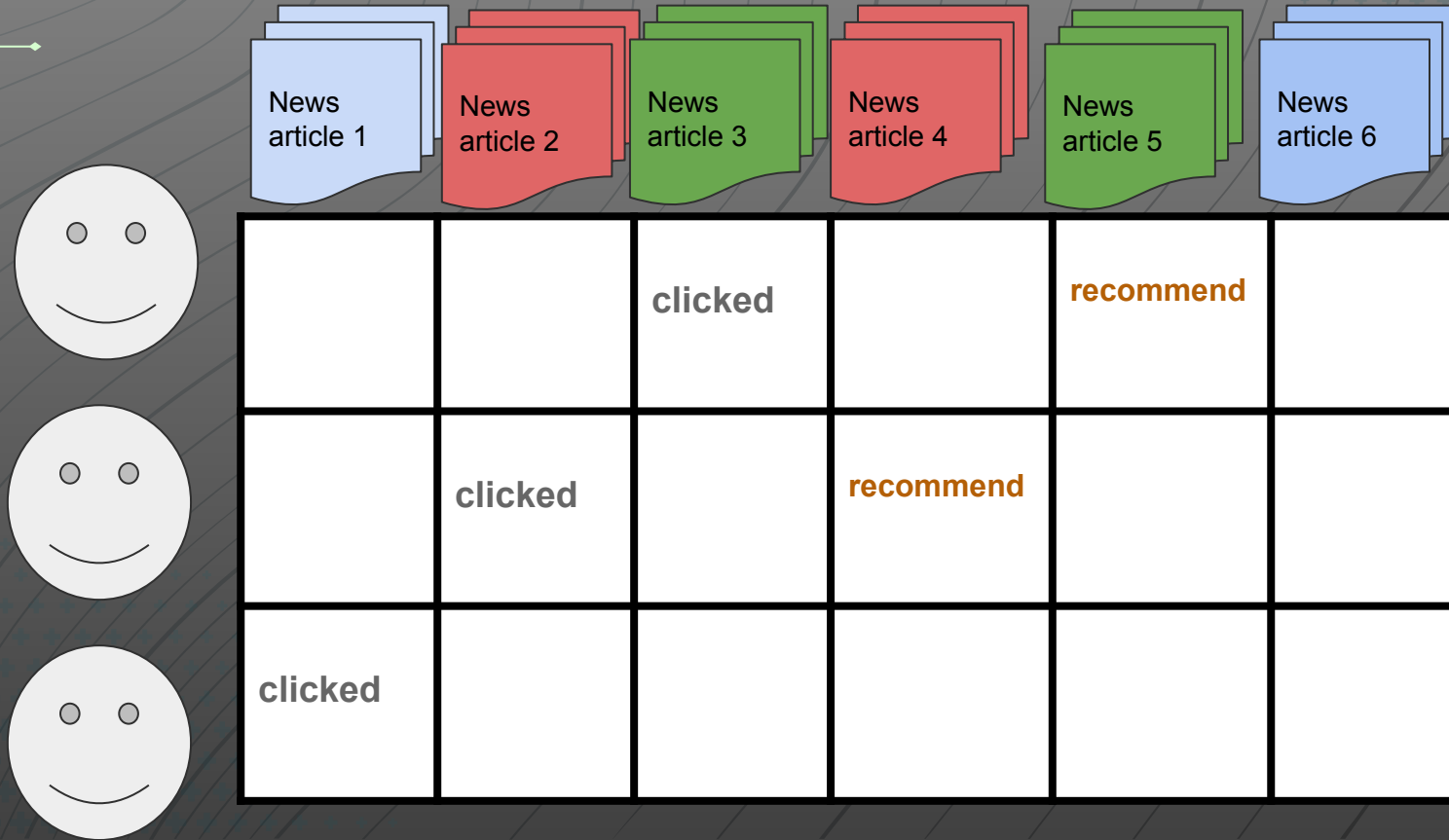


CartridgeSave Images, Attribution 2.0 Generic
(CC BY 2.0)

What is a [news] recommender system?



What is a [news] recommender system?





Optimizing in News Recommendation

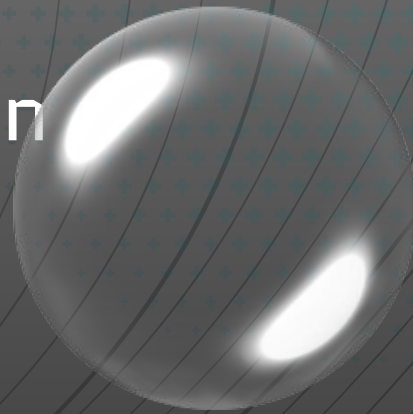
RecSys: **click-accuracy** (as proxy for user interest).

- **Predicting clicks** means showing users more of the same,
- More of **what they already agree with.**

Could lead to 'filter bubbles';

- problematic for democracy and public debate;
- **if you always see the same, how do you know other ideas exist?**

In my PhD project, we work with social scientists, political scientists, and computer scientists to try to optimize for **different viewpoints** in **recommendation**



OpenClipart Vectors @
Pixabay

Using stances to diversify news recommendations

Stance in news articles towards topics:

Dutch stance dataset on sentences from news texts on the 2020 Dutch elections

Stances in the news on four Issues: *Immigration, Climate measures, taxes, and European Union membership.*

Aim: diversity of stances, actors, issues in news recommendation

VVD komt in
opstand tegen
stikstofplannen
eigen minister

Beyond Gun Control: Creating a Dutch Stance Dataset for Diversity in News Recommendation. Myrthe Reuver, Kasper Welbers, Wouter van Atteveldt, Antske Fokkens, Mariken van der Velden and Felicia Locherbach. CLIN32 (2022)

Stance in news articles towards questions:

Alam, M., Iana, A., Grote, A., Ludwig, K., Müller, P., & Paulheim, H. (2022). Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection. *2nd International Workshop on Knowledge Graphs for Online Discourse Analysis (KnOD 2022) collocated with The Web Conference 2022*.

Table 2: Questions for the question-news article pairs.

German Question	English Translation (for understandability)
(Q1) Befürworten Sie, dass Flüchtlinge nach Deutschland kommen?	Are you in favor of refugees coming to Germany?
(Q2) Befürworten Sie, dass Flüchtlinge in Deutschland leben?	Are you in favor of refugees living in Germany?
(Q3) Befürworten Sie, dass Flüchtlinge in Deutschland arbeiten?	Are you in favor of refugees working in Germany?
(Q4) Sollte Deutschland Flüchtlinge aufnehmen?	Should Germany take in refugees?
(Q5) Sollte Deutschland Flüchtlingen helfen?	Should Germany help refugees?

How do we actually **develop** and **evaluate**
stance detection models?

Is Stance Detection Topic-Independent and Cross-topic Generalizable? - A Reproduction Study

Myrthe Reuver*, Suzan Verberne#, Roser Morante*, Antske Fokkens*^

*Computational Linguistics and Text Mining Lab, Vrije Universiteit Amsterdam

^ Dept. of Mathematics and Computer Science, Eindhoven University of Technology

#Leiden Institute of Advanced Computer Science, Leiden University

Cross-topic, cross-domain stance

Main question of **cross-topic** stance detection:

can we detect stance (pro, con)

on **topics or issues not seen** in training?



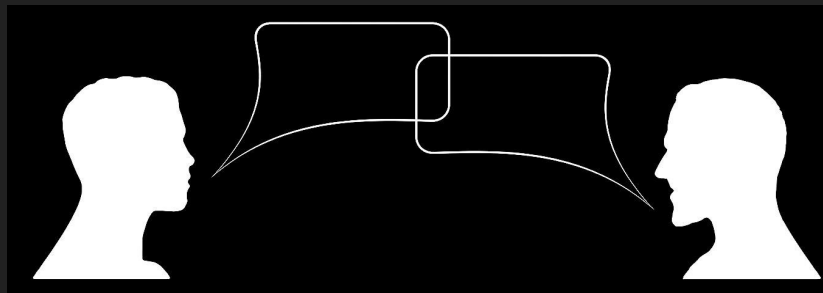
OpenClipart Vectors @ Pixabay

(The news always has new topics coming up!)

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**



Gerd Altmann, Pixabay licence.
<https://pixabay.com/illustrations/feedback-exchange-of-ideas-debate-2466829/>

- **8 controversial debate topics** from the internet: *abortion, cloning, death penalty, gun control, marijuana legalization, minimum wage, nuclear energy and school uniforms.*

Reimers et. al. (2019)

Experimental set-up:

- Training on 7 topics, testing on 8th topic
- Fine-tuning BERT



Reimers et. al. (2019) results

- avg. F1 (over 10 seeds) = .633
- +.20 improvement over reference model (LSTM)
- Results are “***very promising and stress the feasibility of the task***” (Reimers et al. 2019, p. 575)

Reproduction

Reproducibility Crisis in social science since 2016, now broader in all fields.

Following the ACM (Association for Computing Machinery):

“An experimental result is **not fully established** unless it can be independently reproduced.”

ACM Terminology

Repeatability (Same team, same experimental setup)

→ can you find your own result again with your own hardware, code, and data?

Reproducibility (Different team, same experimental setup)

→ same artifact (code, data, experimental set-up) as the original researchers.

Replicability (Different team, different experimental setup)

→ someone else can find the same results (e.g. “Transformers are better for this problem than SVM!”) with their own code.

Reproduction of results: why do we do it?

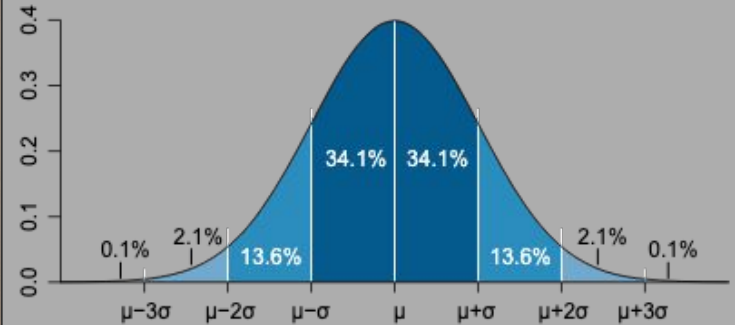
- **Important for science.** One result could be accident, fluke, or not reliable.
- **Non-deterministic results** of Transformers:
 - seeds are random factor & can widely vary the performance
- How to minimize that factor, and deal with it in reproduction:
 - **Standard deviation (SD) over seeds;**
 - value is reproduced if it falls **within 2 SDs.**

Standard Deviation & What it Tells Us About Data



total ; 36 kilo
 $36 / 4 = 9$ kilo average (mean)
 $SD = 4.06$

In a normal distribution: probability of an item **from the same population** > 2 SD from the mean: very low.



Results of reproduction

Model	UKP Dataset				
	F1	P pro	P con	R pro	R con
mean (stdev) 10 seeds					
Reimers et al. (2019) bilstm+BERT	.424	.267	.389	.281	.403
Reimers et al. (2019) BERT base	.613 (-)	.505 (-)	.531 (-)	.470 (-)	.576 (-)
Reimers et al. (2019) BERT large	.633 (-)	.554 (-)	.584 (-)	.505 (-)	.560 (-)
SVM+tf-idf					
Reproduction BERT-base					
Repr. BERT-large - all seeds					
Repr. BERT-large - 5 evenly performing seeds					

> What do you think will happen here? Results within two standard deviations?

> And: which BASELINE is stronger, SVM or LSTM?

Results of reproduction

Model	UKP Dataset				
	F1	P pro	P con	R pro	R con
mean (stdev) 10 seeds	.424	.267	.389	.281	.403
Reimers et al. (2019) biclestm+BERT	.613 (-)	.505 (-)	.531 (-)	.470 (-)	.576 (-)
Reimers et al. (2019) BERT base	.633 (-)	.554 (-)	.584 (-)	.505 (-)	.560 (-)
Reimers et al. (2019) BERT large					
SVM+tf-idf	.517	.418	.460	.414	.423
Reproduction BERT-base	.617 (.006)	.519 (.011)	.538 (.007)	.464 (.029)	.581 (.019)
Repr. BERT-large - all seeds	.596 (.043)	.483 (.057)	.527 (.057)	.464 (.058)	.516 (.063)
Repr. BERT-large - 5 evenly performing seeds	.636 (.007)	.532 (.014)	.578 (.016)	.515 (.016)	.567 (.022)



Difference with original results **within two standard deviations**

Results of reproduction

Model	UKP Dataset				
	F1	P pro	P con	R pro	R con
mean (stdev) 10 seeds					
Reimers et al. (2019) biclstm+BERT	.424	.267	.389	.281	.403
Reimers et al. (2019) BERT base	.613 (-)	.505 (-)	.531 (-)	.470 (-)	.576 (-)
Reimers et al. (2019) BERT large	.633 (-)	.554 (-)	.584 (-)	.505 (-)	.560 (-)
SVM+tf-idf	.517	.418	.460	.414	.423
Reproduction BERT-base	.617 (.006)	.519 (.011)	.538 (.007)	.464 (.029)	.581 (.019)
Repr. BERT-large - all seeds	.596 (.043)	.483 (.057)	.527 (.057)	.464 (.058)	.516 (.063)
Repr. BERT-large - 5 evenly performing seeds	.636 (.007)	.532 (.014)	.578 (.016)	.515 (.016)	.567 (.022)

- BERT-large under-performs in 50% of seeds
- SVM+tf-idf model

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 non-BERT model (LSTM) does not work for other model (SVM+tf-idf);

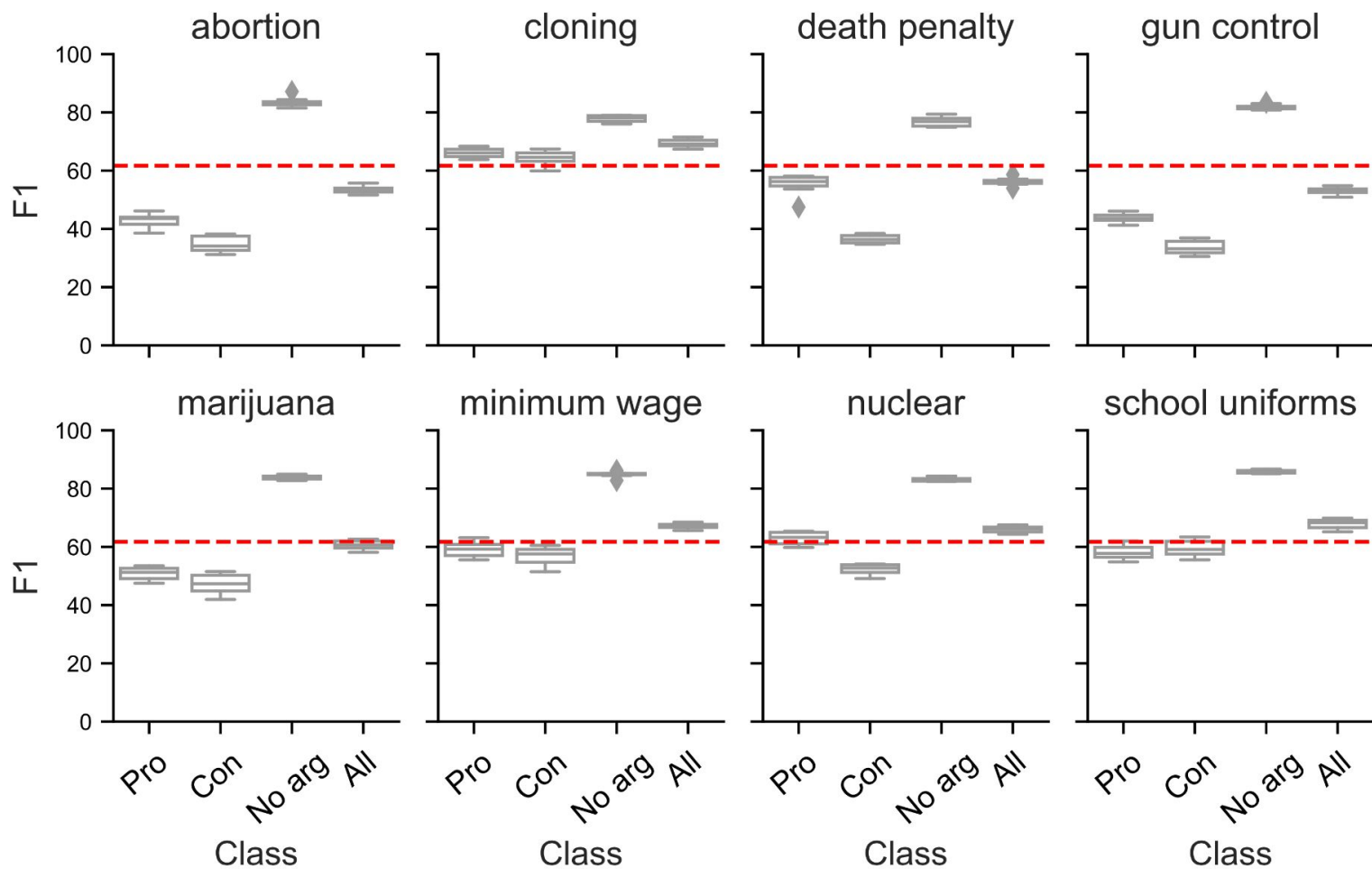
3. conclusion(s):

❓ How feasible is cross-topic? Let's investigate some more, especially on topics.

What about different topics?

held-out topic	abortion	cloning	death penalty	gun control	marijuana legalization	minimum wage	nuclear energy	school uniform
SVM+tf-idf	.463	.585	.482	.515	.323	.615	.598	.576
BERT-base	.533 (.011)	.693 (.013)	.562 (.012)	.530 (.013)	.607 (.016)	.670 (.009)	.660 (.011)	.678 (.016)
diff.	+.070	+.108	+.080	+.028	+.283	+.055	+.0850	+.102

- Some topics (*abortion, death penalty*) perform near baseline (SVM F1 = .517 average over all topics)
- Others (*minimum wage, cloning, gun control*) perform markedly higher (F1 > .670).



--- BERT-base F1 (mean)

Some examples of difficult arguments

“The second amendment protects the right to possess a firearm”

Topic: **gun control**, True: **Con**, Predicted (7/10 seeds): **Pro**

“The fetus is not a person, which makes abortion morally permissable”

Topic: **abortion**, True: **Pro**, Predicted (5/10 seeds): **Con**

“People were freed from death row because they were later found to be innocent”

Topic: **death penalty**, True: **Con**, Predicted (9/10 seeds): **Pro**

What does this mean? Take home messages

- **Successful reproduction** cross-topic stance (Reimers et. al., 2019), but random seed does matter for BERT-large.
- **Topic matters!** Stance not as topic-independent as seems with one averaged F1 metric reported.
 - See also: Thorn Jakobsen et. al. (2021)
- **A class/topic interaction effect in SOTA stance detection**
- Time to (re)investigate **topic similarity?** **When can we cross to new topics?**



OpenClipArt, Public domain

Thank you!

Myrthe Reuver, Free University of Amsterdam



[myrthe.reuver\[at\]vu.nl](mailto:myrthe.reuver[at]vu.nl)



@myrthereuver

References

- Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., & Gurevych, I. (2019, July). Classification and Clustering of Arguments with Contextualized Word Embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 567-578).
- Schiller, B., Daxenberger, J., & Gurevych, I. (2021). Stance detection benchmark: How robust is your stance detection?. KI-Künstliche Intelligenz, 1-13.
- Du Bois, J. W. (2007). The stance triangle. Stancetaking in discourse: Subjectivity, evaluation, interaction, 164(3), 139-182.
- Küçük, D., & Can, F. (2020). Stance detection: A survey. ACM Computing Surveys (CSUR), 53(1), 1-37.
- Jakobsen, T. S. T., Barrett, M., & Søgaard, A. Spurious Correlations in Cross-Topic Argument Mining. Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics.
- Allaway, E., Srikanth, M., & McKeown, K. (2021, June). Adversarial Learning for Zero-Shot Stance Detection on Social Media. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4756-4767).
- Wei, P., & Mao, W. (2019, July). Modeling transferable topics for cross-target stance detection. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1173-1176).

After this lecture

- You can define stance detection
- You can explain the purpose of stance detection for diverse news recommendation
- You can explain the importance of reproducibility in NLP
- You can explain the challenges of cross-topic model learning