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

VRIJE UNIVERSITEIT AMSTERDAM

What is (going on with) stance?

Stance detection, common definition: **classification task** with labels Pro, Con, Neutral towards an issue or topic

"Abortion is a sin, and should never be practiced."

Topic: Abortion, Stance: Con

societal challenges with (online) information:

diversifying stances in an online news rec

(Reuver et. al., 2021)

- New topics and issues continuously appear online!

Joseph Mucira @ Pixabay, Simplified Pixabay License

Cross-topic, cross-domain stance

Main question: can we detect stance (pro, con) on **topics or issues unseen** in training?

(1) Topic similarity

- Wei & Mao (2019), meta topics (e.g. feminism, abortion → "equality"), even earlier Somasundaran & Wiebe (2009)

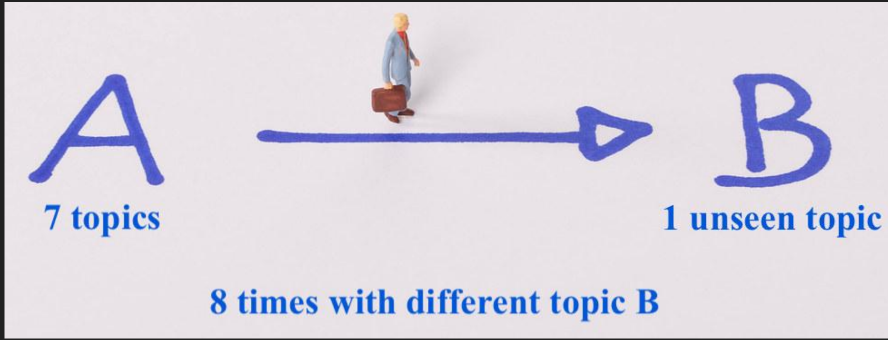
(2) topic-(in)dependent stance

Reimers et. al. (2019)

→ Train: 7 topics, test: 8th topic

→ Fine-tuning BERT (base & large)

→ Findings:



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

Reproduction

- **Important.**

- **Systematically** (with 3 dimensions).

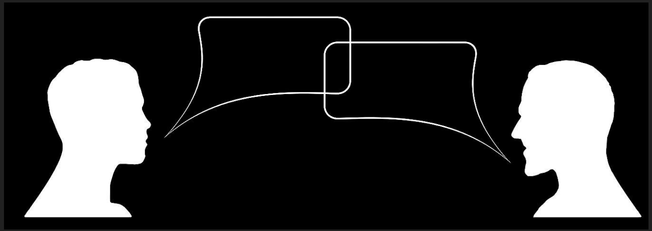
- Non-deterministic results of BERT:
  - **Standard deviation (SD) over seeds**;
  - value is reproduced if it falls **within 2 SDs.**

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



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

Results

Model	**UKP Dataset**				
mean (stdev) 10 seeds	F1	P pro	P con	R pro	R con
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)

Reimers et. al. (2019) provided **excellent preliminaries for reproducibility:** documented, shared, working code (through a GitHub repository) + available for questions.

Results: further details

Model	UKP Dataset				
mean (stdev) 10 seeds	F1	P pro	P con	R pro	R con
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**,

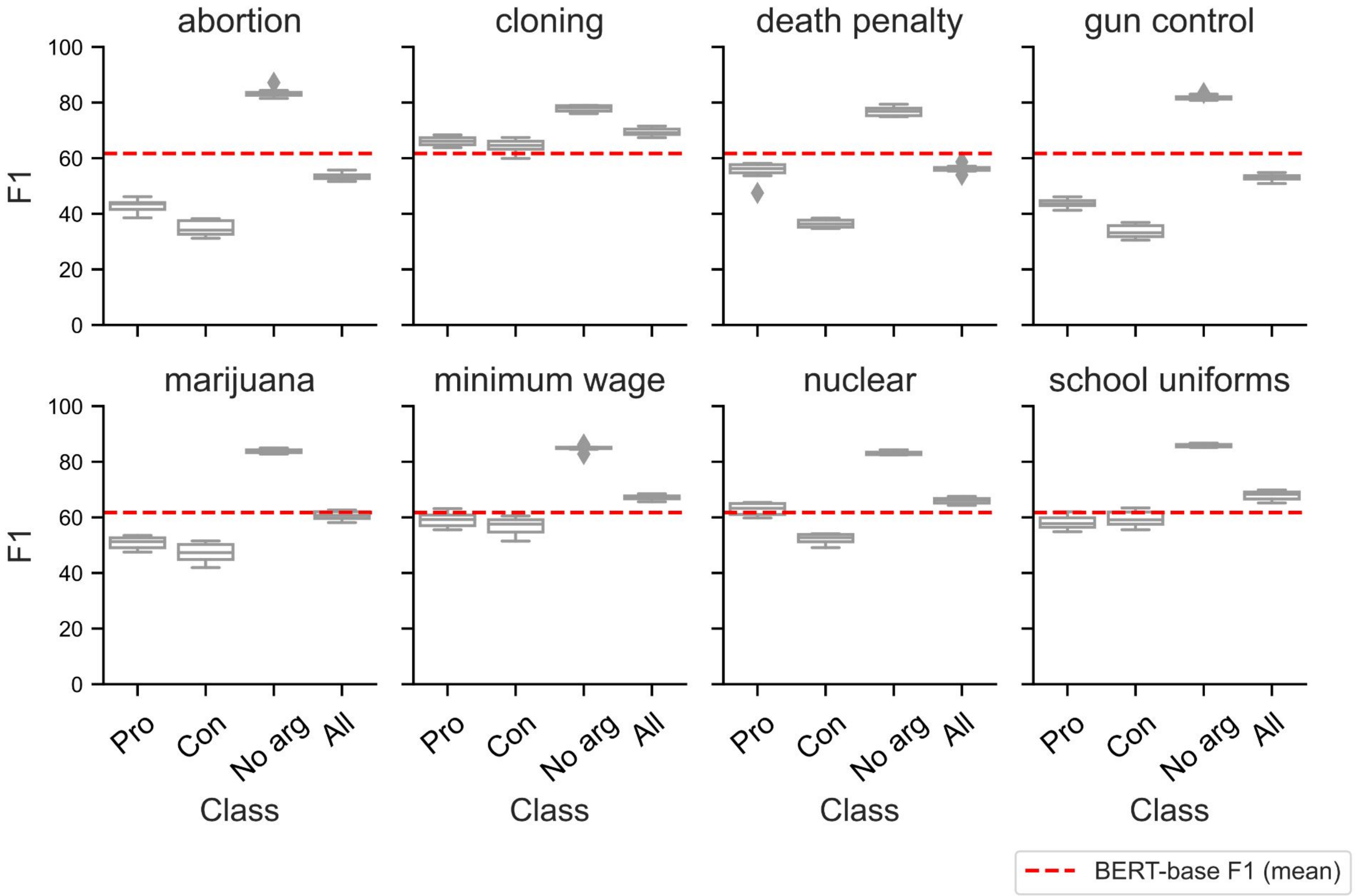
    ✗ .20 improvement over non-BERT model (LSTM < our SVM)

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


abortion
cloning
death penalty
gun control
marijuana
minimum wage
nuclear
school uniforms
F1
100
80
60
40
20
0
Pro
Con
No arg
All
Class
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 matters for BERT-large, & SVM is stronger reference.

- **Topic matters!** Stance not topic-independent → beyond one avg F1
    - See also: Thorn Jakobsen et. al. (2021)

OpenClipArt, Public domain

- A **class/topic interaction effect on performance**

- Time to (re)investigate **topic similarity**? **When can we cross to new topics?**

Thank you!

Myrthe Reuver, Free University of Amsterdam

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).