Investigating the Robustness of Modelling Decisions for Few-Shot Cross-Topic Stance Detection: A Preregistered Study

Myrthe Reuver*, Suzan Verberne^ , Antske Fokkens*§

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)

* Computational Linguistics & Text Mining Lab, Vrije Universiteit Amsterdam

^ Text Mining and Retrieval Leiden, Universiteit Leiden

§ Eindhoven University of Technology

Our project: diversity in news recommendations

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

Consequence: Showing users more of the same.

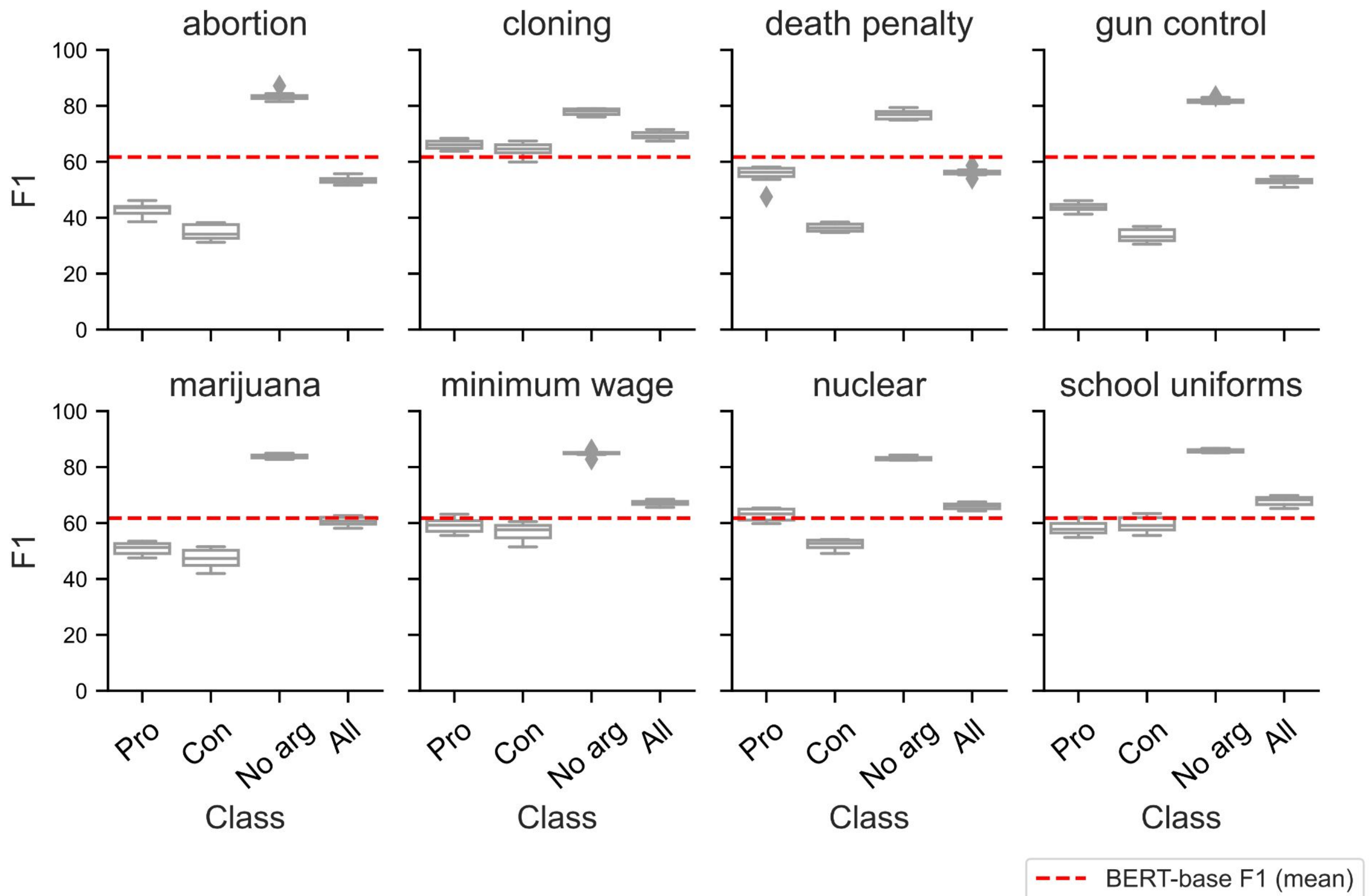
→ Filter bubbles and echo chambers.

Why is this problematic for democracy and society?

Theoretic models of **democracy** (Vrijenhoek et. al., 2021):

Needed: **diverse viewpoints on issues**

Previous work: crossing to other topics is difficult, inconsistent results



(Reuver et. al, 2021 replication of Reimers et al., 2019)

Same Side Stance

Task operationalization: **“Same Side Stance Detection” (SSSC)** (Stein et. al., 2020)

Training to classify whether two arguments on an issue have **the same** or a **different stance.** Aim: reducing the model’s leaning on topic-specific pro- and con-vocabulary

Possibility: **bi-encoding and immediately measuring the similarity between a pair of stances** (e.g. a read article vs a new article)

vectorportal.com

Example SSSC

Topic: '[This house believes] all nations have a right to nuclear weapons'

Are these arguments on the same side?

"Nuclear weapons may lessen a state's reliance on allies for security, thus preventing allies from dragging each other into wars" (**used to be PRO**)

“Nuclear holocaust could result in an end to human life” (**used to be CON**)

Same side stance label: FALSE

Pre-registration

Van Miltenburg et. al. (2021) identified how to **pre register** in NLP experiments.

Preregistration: deciding on experiments, comparisons, and datasets before running them, since experimental conditions and hypotheses are often **implicit** in NLP work (assumptions about what will work better etc.)

Goal: making these **explicit,** and being **transparent about choices** in research design.

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?

Main motivations for pre-registration

Registering: expectations of models + datasets, in **explicit hypotheses**

Papers could claim exceptional progress while only testing one dataset, or only comparing one modelling choice, and not reporting what does not work.

We wanted to:

- **systematically** comparing modelling choices
- Also reporting **negative or mixed results**

Hypotheses, example:

Hypothesis: *based on Shnarch et. al. (2022)'s experimental results on topic-dependent versus topic-independent tasks and pre-fine-tuning clustering,*

- Grounding in literature and/or earlier experiments;

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,

- Grounding in literature and/or earlier experiments;
- Expectation;

Hypotheses, example:

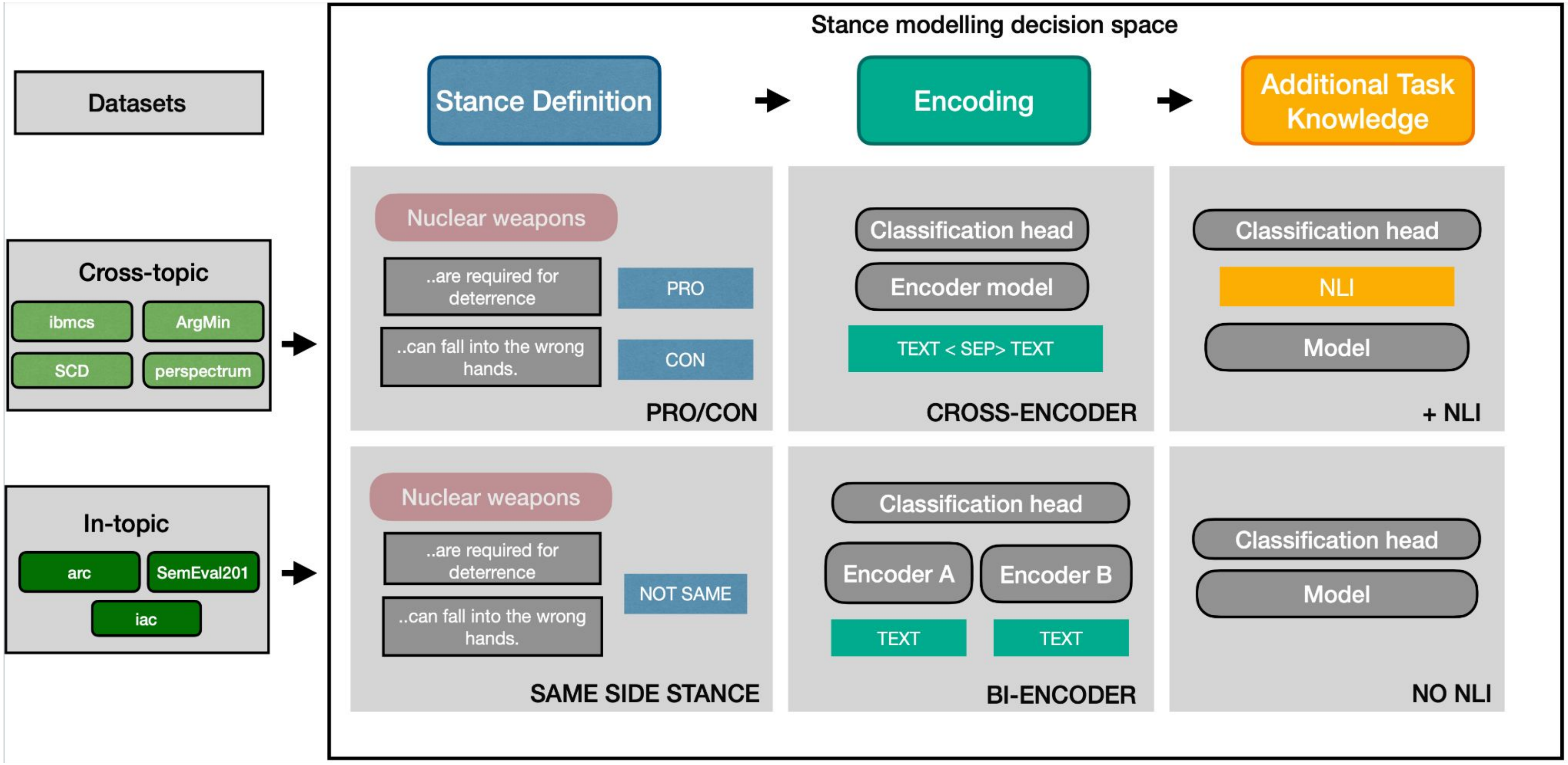
Hypothesis: *based on Shnarch et. al. (2022)'s experimental results on topic-dependent versus topic-independent tasks and pre-fine-tuning clustering,* we expect that SSSC models + pre-fine-tune clustering approach improve significantly over SSSC models without the pre-fine-tuning approach, since we consider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

Main research questions

1. How do **different modelling choices** (task definitions and architecture differences) affect **few-shot classification performance** on different stance datasets?

2. To what extent do these modelling choices affect few-shot **cross-topic** robustness?


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
PRO
..can fall into the wrong hands.
CON
PRO/CON
Classification head
Encoder model
TEXT < SEP> TEXT
CROSS-ENCODER
Classification head
NLI
Model
+ NLI
Nuclear weapons
..are required for deterrence
NOT SAME
..can fall into the wrong hands.
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:**

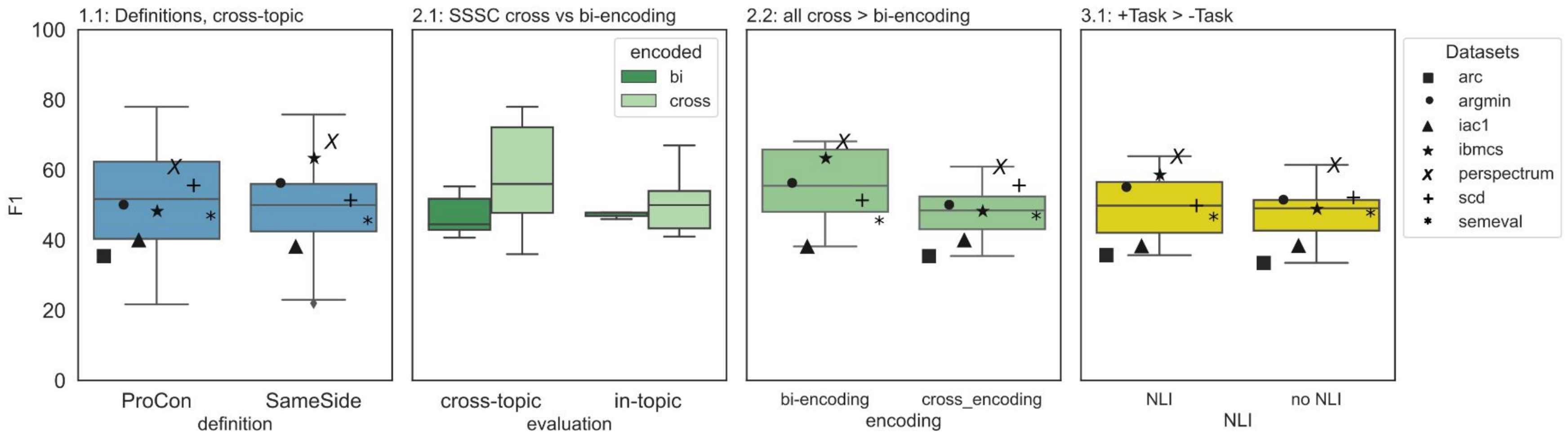
2.1: we expect **bi-encoding to fluctuate less** between in-topic to cross-topic performance, and improve cross-topic performance.

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

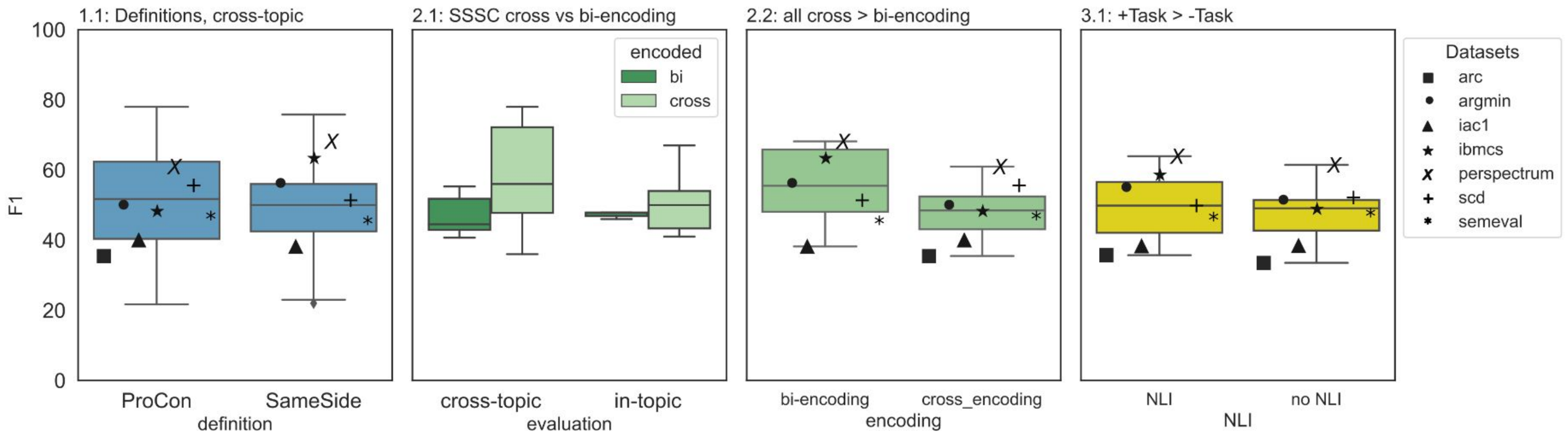
- **Task Knowledge**

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

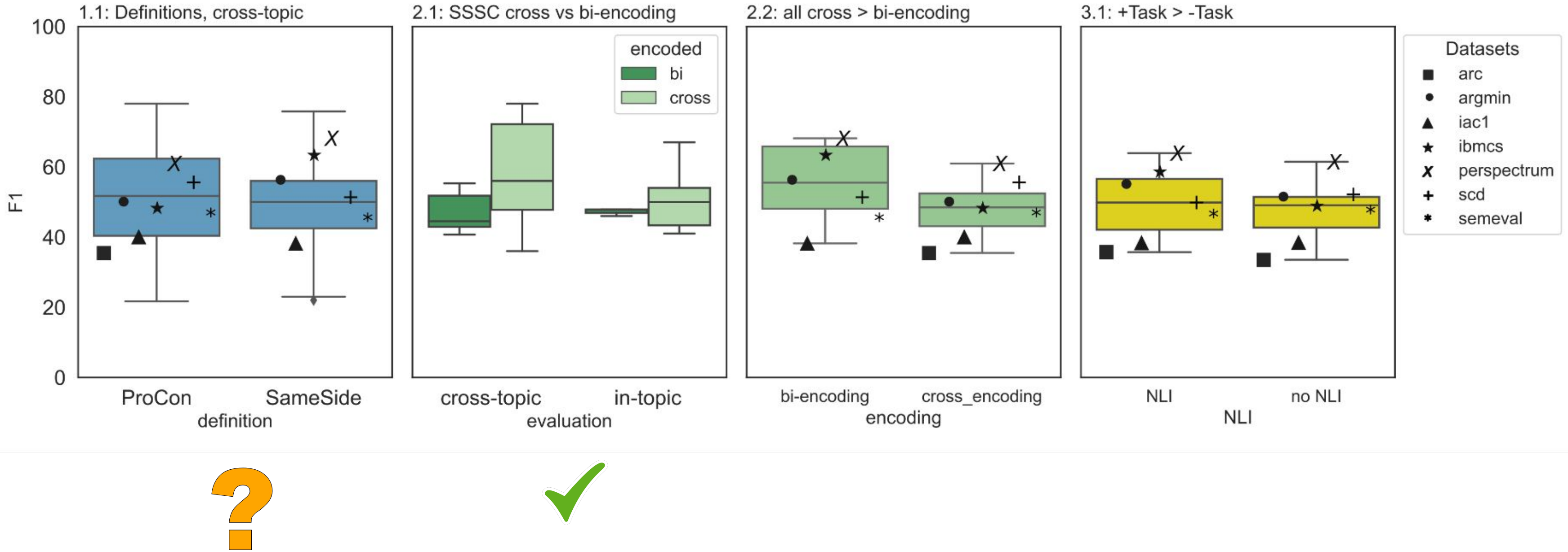
Results, per hypothesis:



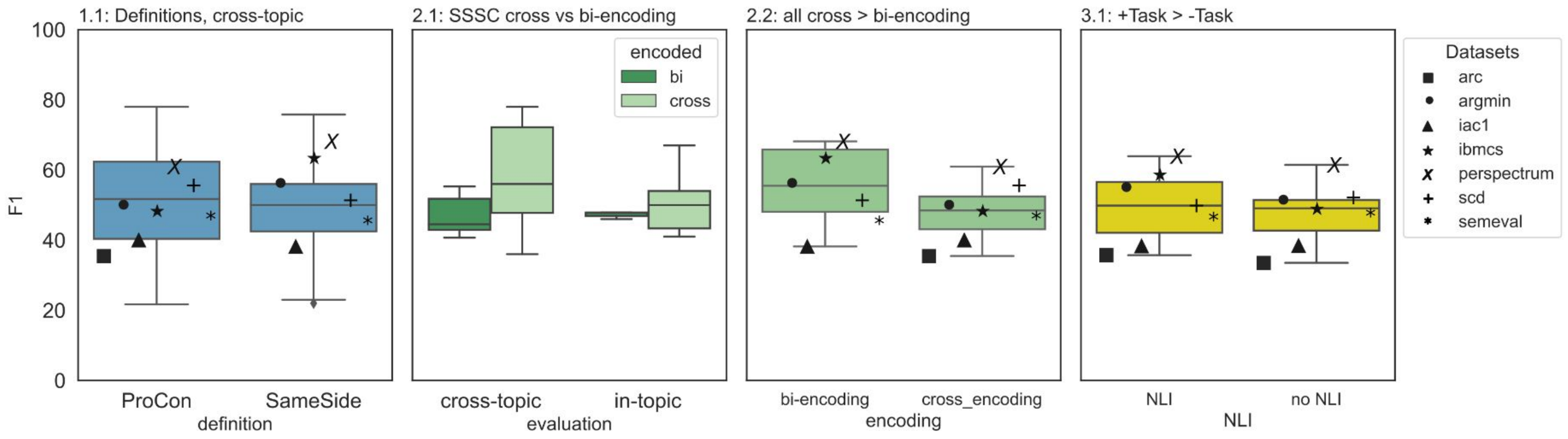
Results, per hypothesis:



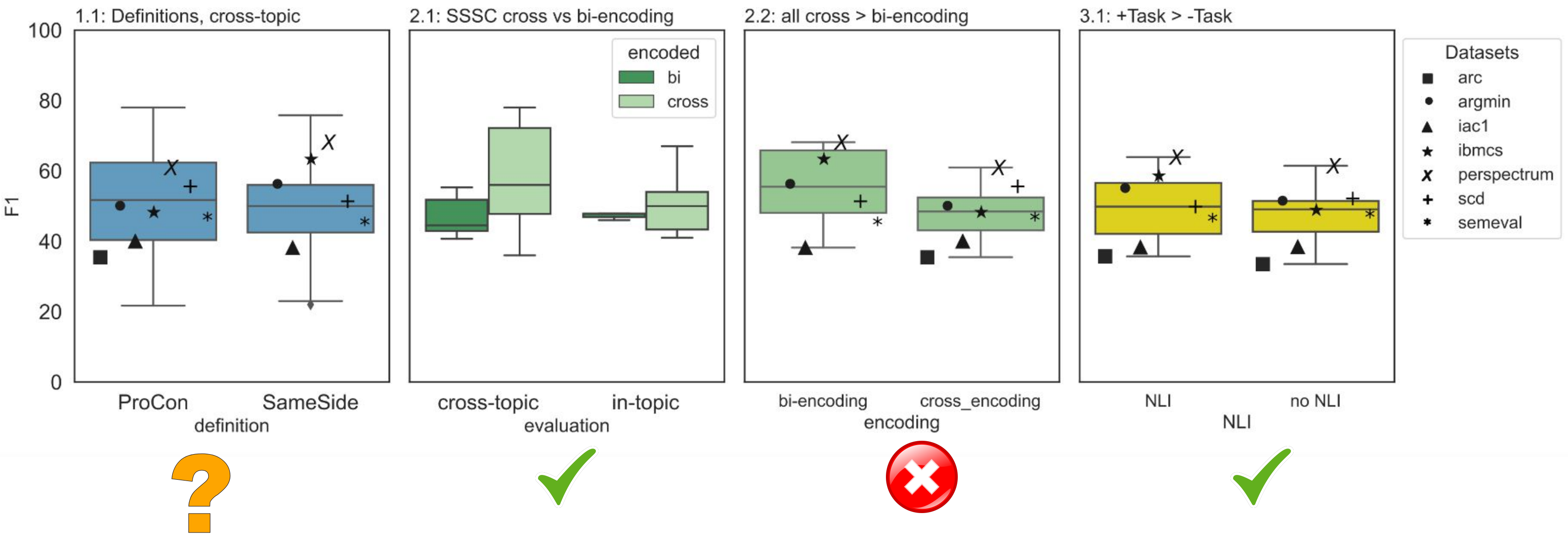
Results, per hypothesis:

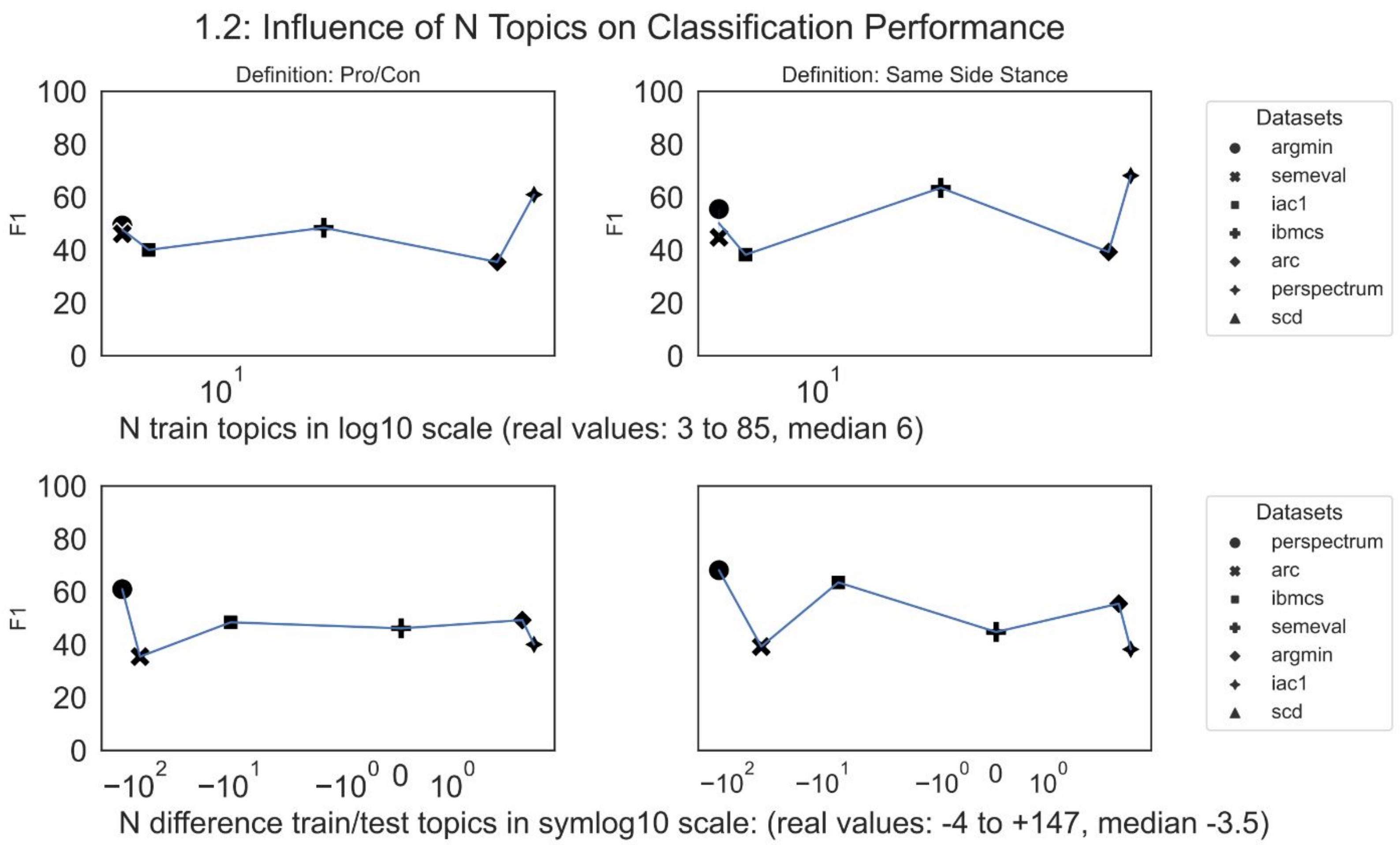


Results, per hypothesis:



Results, per hypothesis




1.2: Influence of N Topics on Classification Performance
Definition: Pro/Con
Definition: Same Side Stance
F1
100
80
60
40
20
0
10^1
N train topics in log10 scale (real values: 3 to 85, median 6)
Datasets
argmin
semeval
iac1
ibmcs
arc
perspectrum
scd
-10^2
-10^1
-10^0
0
10^0
N difference train/test topics in symlog10 scale: (real values: -4 to +147, median -3.5)
Datasets
perspectrum
arc
ibmcs
semeval
argmin
iac1
scd

Discussion

- It appears that **stance datasets with the highest performance** contain texts from websites specifically aimed at debating (e.g. *perspectrum*).

  Other recent work explores different modelling decisions for stance:

- Arakelyan et al. (2023) **optimizing data seems similar to optimizing modelling choices.**
- Recently, Waldis et al. (2024) **differently pre-trained models for cross-topic stance detection**: diverse pre-training objectives allow for better cross-topic stance capabilities.

Conclusion(s): stance dataset require different mixes of modelling choices

Same Side Stance definition on performance **differs per dataset and other modelling choice**, and also the relation between cross and bi-encoding is not the same for every dataset.

We found **no clear relationship** between number of training topics and performance.

Adding **NLI training to our models gives considerable improvement** for most datasets, but inconsistent results for others.