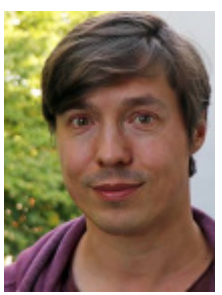
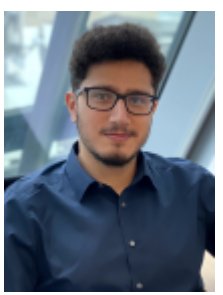
# On Classifying whether Two Texts are on the Same Side of an Argument

## 8th International Workshop on Argument Mining @ EMNLP 2021

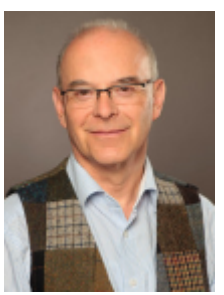

**Erik Körner**[1]



**Gregor Wiedemann**[2]



Ahmad Dawar Hakimi[3]



Gerhard Heyer[3]

Martin Potthast[4]

[1,3,4] UNIVERSITÄT LEIPZIG

[2] LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG HANS-BREDOW-INSTITUT

[1] Bauhaus-Universität Weimar

[1,4] Webis

# On Classifying whether Two Texts are on the Same Side of an Argument

## Same Side Stance Classification

- Shared task at 6th Workshop on Argument Mining [1] [Stein et al. 2021]

  *Identifying (classifying) the stance of an argument towards a particular topic is a fundamental task in computational argumentation. The stance of an argument as considered here is a two-valued function: it can either be "pro" a topic (= yes, I agree), or "con" a topic (= no, I do not agree).*

  *With the new task "same side (stance) classification" we address a simpler variant of this problem: Given two arguments regarding a certain topic, the task is to decide whether or not the two arguments have the same stance.*

- Two topics: *gay marriage* and *abortion*
- Two tasks: **within**, same set of topics for training and test (*abortion* and *gay marriage*); **cross**, training set for topic *abortion*, and test set with arguments related to another set of topics

[1] https://sameside.webis.de/, https://webis.de/events/argmining-19/

Motivation

- ease the difficulty of argument stance classification
- only **argument similarity** within stances needs to be learned
- in contrast to actual stance classification which requires a substantial amount of **domain knowledge** to identify whether an argument is in favor or against a certain issue

# On Classifying whether Two Texts are on the Same Side of an Argument

## Examples [Stein et al. 2021]

Arguments on the topic “gay marriage”:

**Argument 1.** Marriage is a commitment to love and care for your spouse till death. This is what is heard in all wedding vows. Gays can clearly qualify for marriage according to these vows, and any definition of marriage deduced from these vows.

**Argument 2.** Gay Marriage should be legalized since denying some people the option to marry is discriminatory and creates a second class of citizens.

**Argument 3.** Marriage is the institution that forms and upholds for society, its values and symbols are related to procreation. To change the definition of marriage to include same-sex couples would destroy its function, because it could no longer represent the inherently procreative relationship of opposite-sex pair-bonding.

## Motivation

- Participation in S3C shared task in 2019,
  achieving 1st place in *within* and 2nd place in *cross* task
- Noticed certain properties in the **official dataset**
  - **overlap** of single argument stances between *train* and *test*
  - great **variety of sizes** for single debates from which pairs are sampled

  ➜ results may be **unrealistically optimistic**

# On Classifying whether Two Texts are on the Same Side of an Argument

## Goals

1. **Improving on the state of the art** using recent transformer-based approaches
2. Renewed **assessment of the original S3C shared task dataset** & Compilation of **new training and test sets** that enable a more realistic evaluation scenario
3. Compilation of a hand-crafted **test set consisting of adversarial cases**
   ➜ Investigate the hypothesis underlying S3C in particular
4. Improve the (training) data scarcity by utilizing **cross-domain dataset**

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 1: Optimization

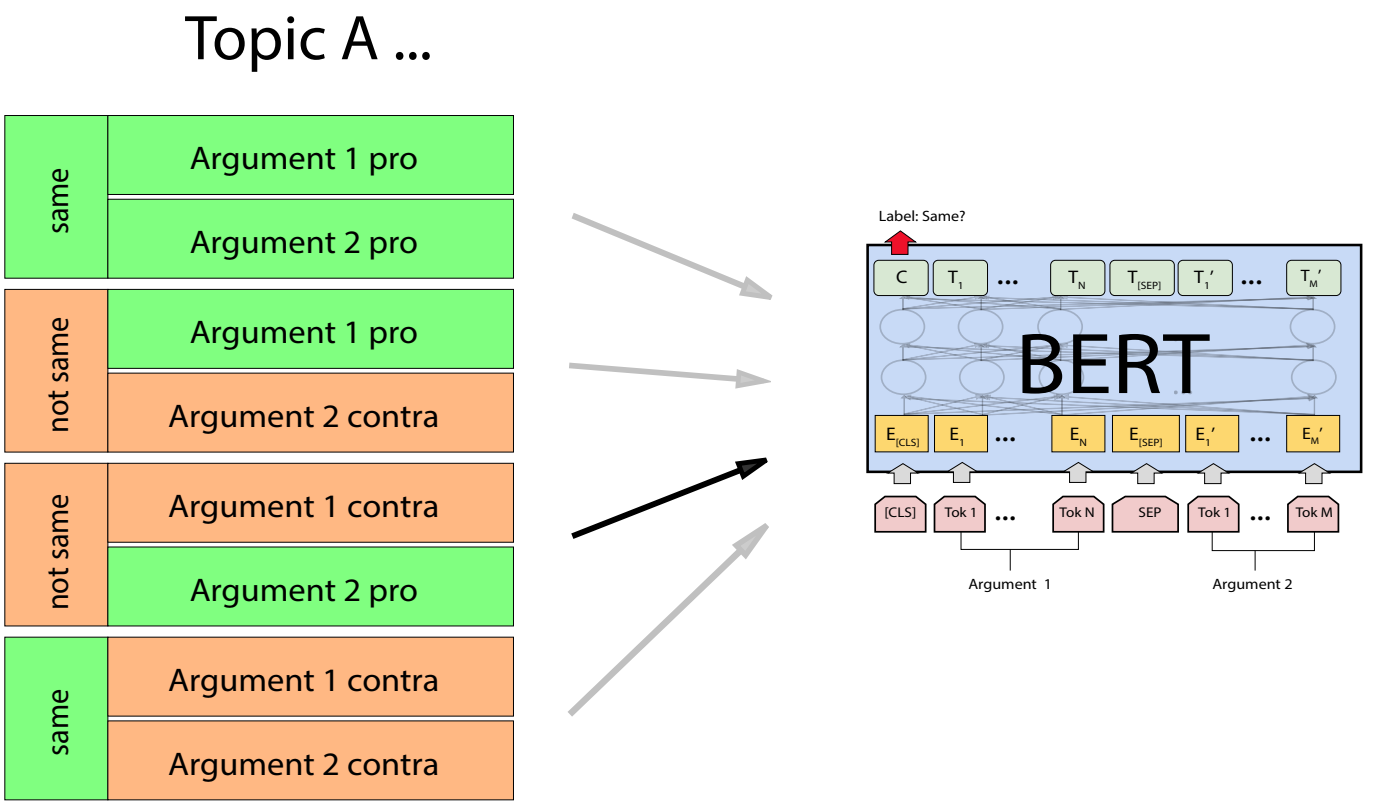
- Reproduce the shared task in its original form & best-performing approach at the S3C shared task by [Ollinger et al. 2021]
- Approach:
  - English pre-trained **BERT** [Devlin et al. 2019] model for sequence pair classification
  - Fine-tuning for 3 epochs with binary cross-entropy loss
  - Standard hyper-parameters values

## Experiment 1: Optimization Experiment 1: Optimization

- Reproduce the shared task in its original form & best-performing approach at the S3C shared task by [Ollinger et al. 2021]
- Approach:
  - English pre-trained **BERT** [Devlin et al. 2019] model for sequence pair classification
  - Fine-tuning for 3 epochs with binary cross-entropy loss
  - Standard hyper-parameters values
- Newer transformer-based pre-trained networks:
  - **RoBERTa** [Liu et al. 2019]: BERT with larger and cleaner datasets for pre-training
  - **XLNet** [Yang et al. 2019]: employs autoregressive pre-training
  - **DistilBERT** [Sanh et al. 2019]: knowledge distillation during pre-training
  - **ALBERT** [Lan et al. 2020], embedding matrix compression and sentence order prediction as a pre-training task
  - . . .

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 1: Optimization



# On Classifying whether Two Texts are on the Same Side of an Argument

| Model | Cross | | Within | |
|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 |
| – sequence length: 128 | | | | |
| bert-base-uncased | **60.33** | 57.35 | 77.59 | 74.23 |
| albert-base-v2 | 59.25 | **58.65** | **80.79** | **80.38** |
| – sequence length: 256 | | | | |
| bert-base-uncased | 60.72 | 58.27 | 85.45 | 86.02 |
| bert-base-cased | 63.23 | 65.16 | **86.47** | **87.01** |
| roberta-base | 60.31 | 54.59 | 76.19 | 71.85 |
| distilbert-base-cased | 59.08 | 56.91 | 67.91 | 63.74 |
| distilroberta-base | 59.07 | 54.80 | 75.95 | 73.15 |
| xlnet-base-cased | 61.62 | 63.63 | 82.35 | 80.30 |
| albert-base-v1 | 63.93 | 66.51 | 83.76 | 84.09 |
| albert-base-v2 | **64.55** | **67.29** | 84.81 | 85.57 |
| electra-small-discriminator | 59.88 | 55.94 | 65.48 | 63.92 |
| electra-base-discriminator | 59.71 | 60.81 | 82.29 | 81.52 |
| sent.-transf.-stsb-dist. | 59.93 | 58.80 | 74.32 | 70.85 |
| queezebert-uncased | 61.86 | 59.96 | 82.96 | 82.28 |
| – sequence length: 512 | | | | |
| bert-base-uncased | 64.77 | 65.94 | 86.26 | 86.28 |
| bert-base-cased | 63.54 | 65.64 | 87.31 | 87.62 |
| roberta-base | 61.55 | 55.38 | 82.21 | 79.99 |
| distilbert-base-cased | 58.77 | 54.87 | 82.35 | 80.44 |
| distilroberta-base | 60.10 | 55.69 | 82.23 | 80.51 |
| xlnet-base-cased | 59.84 | 57.91 | 85.32 | 86.62 |
| **albert-base-v2** | **66.19** | **68.95** | **88.81** | **89.30** |
| electra-small-discriminator | 59.61 | 60.61 | 76.81 | 73.41 |
| electra-base-discriminator | 59.45 | 60.68 | 82.04 | 80.42 |
| sent.-transf.-stsb-dist. | 51.47 | 46.44 | 81.16 | 79.26 |
| queezebert-uncased | 64.25 | 66.32 | 84.46 | 83.98 |

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 1: Optimization

| **Task:** | **Cross** | | **Within** | |
|---|---|---|---|---|
| **Model** | Acc. | F1 | Acc. | F1 |
| BERT base | 63.6 | 66.0 | 86.8 | 87.2 |
| RoBERTa base | 60.5 | 55.2 | 82.3 | 80.3 |
| DistilBERT base | 59.1 | 56.0 | 82.3 | 80.5 |
| XLNet base | 61.0 | 60.7 | 84.2 | 84.2 |
| **ALBERT base v2** | **66.2** | **68.9** | **88.4** | **89.1** |
| Ollinger et al. (2021) | 73.0 | 72.0 | **77.0** | **74.3** |
| ALBERT base v2 | **74.2** | **73.7** | 73.8 | 72.0 |

- length of 512 tokens (3 runs) on our recompiled test set
- state of the art by Ollinger et al. 2021 (baseline)
- our best model evaluated on the shared task test set (bottom)

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 1: Optimization

- Recreated evaluation scenario equivalent to official S3C shared task
- Surprisingly, RoBERTa and XLNet, which commonly improve results upon the standard BERT model, do not perform better for S3C
- Only *ALBERT base v2* model slightly outperforms the baseline of the previous state of the art

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 2: Bias Control

- Sampling of the official dataset may lead to **unrealistically optimistic results** ➜ non-overlapping pairs but overlap of single arguments between *train* and *test*, debates of greatly varying *sizes*.
- We sample 3 new roughly equal-sized dataset splits with **varying degrees of overlap of single arguments**:

- **random**: replicate sampling process of S3C task
- **disjoint**: no single argument from *train* in *test*; split across debates (*cross*) or topic (*within*)
- **single**: only one *single* argument from each pair is also contained in *train*

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 2: Bias Control

| S3C Scenario | Accuracy | F1 |
|---|---|---|
| **Majority baseline** | 53.4 | 34.8 |
| **random** | 86.6 ($\pm$ 0.73) | 86.6 ($\pm$ 0.74) |
| **disjoint** | | |
| – within | 61.7 ($\pm$ 1.64) | 61.4 ($\pm$ 1.46) |
| – cross (A $\rightarrow$ G) | 62.4 | 62.3 |
| – cross (G $\rightarrow$ A) | 61.2 | 61.0 |
| **single** | 67.0 | 64.5 |

- Model: ALBERT base v2
- Scenario *disjoint-cross* reverses the topics *abortion* (A) and *gay marriage* (G) for training and testing
- Random selection for splitting strategies *random* and *disjoint*

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 2: Bias Control

- All scenarios surpass majority baseline
  ➜ model actually learns to recognize (dis-)agreement of arguments
- S3C works accurately (86.6% F1) for the randomly composed test set
- Performance drops severely (ca. 62% F1) for *disjoint* datasets with no overlap of individual arguments
  - Performance for *within* does not even surpass *cross* which is trained on a completely different topic!
- Low performance (65% F1) for *single* scenario, where one argument of a test pair has been seen during training

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 3: Adversarial Examples

- Why? ➜ Reveal the ability of our best model to solve different types of *"adversarial"* cases for same stance prediction more systematically
- Artificial dataset based on **25 distinct arguments** from *gay marriage* topic (short and express their stance clearly)
- Construct new arguments of **four distinct types** to obtain two pairs, one with *same stance*, and one with *opposing stance*

- **Negation**: simple negation of the argument
- **Paraphrase**: alters important words from the argument to synonymous expressions with the same stance
- **Argument**: uses an argument from the same topic and stance, but semantically completely different regarding the first one
- **Citation**: repeats or summarizes the first argument, expresses agreement or rejection (a case frequently occurring in the dataset)

➜ test set with **175 cases**

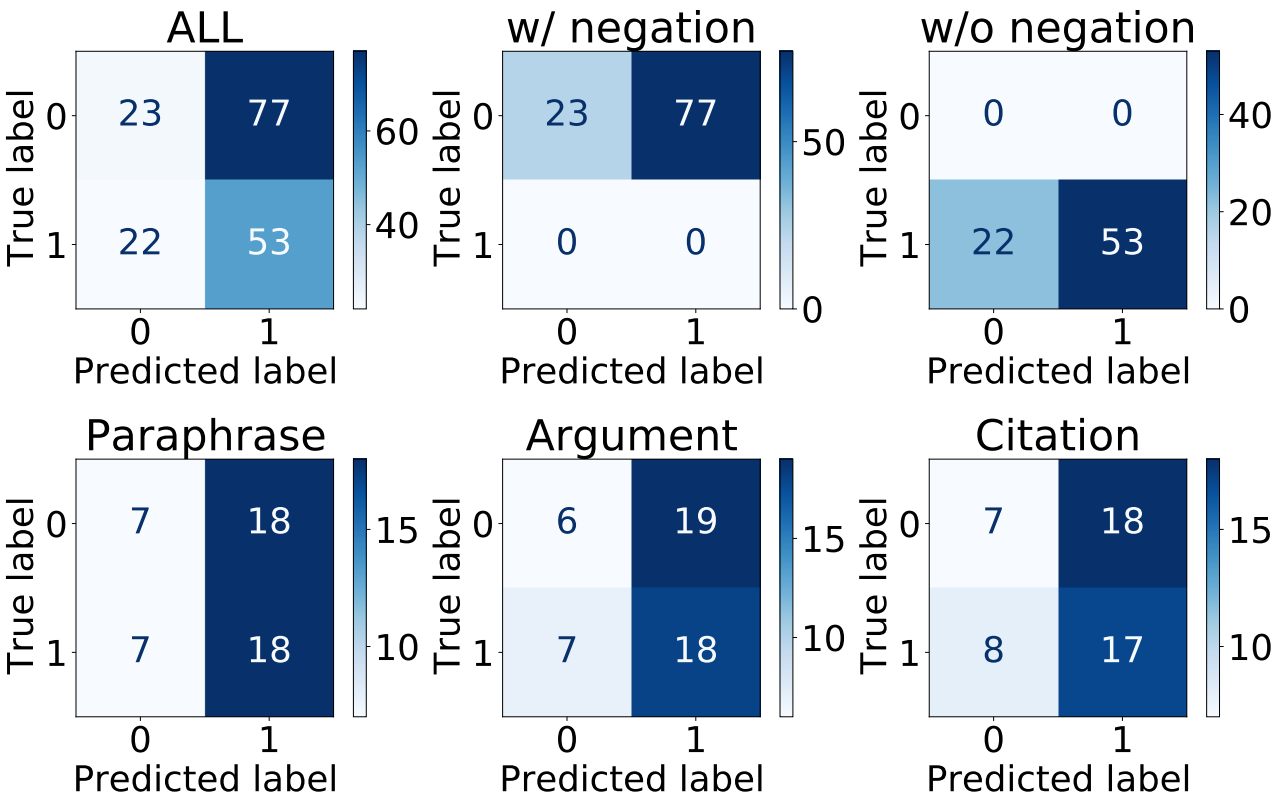
# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 3: Adversarial Examples - Example

| **Claim:** | **The gay marriage ban goes against human rights.** | **Same side?** |
|---|---|---|
| **Negation:** | Banning gay marriage is not a violation of the human rights. | false |
| **Paraphrase:** | Basic rights, including the right to marry, apply to homosexual couples, too. | true |
| **Paraphrase-Negation:** | Denying gays the right to marry does not violate their human rights. | false |
| **Argument:** | Denying gays the right to adopt children violates their human rights. | true |
| **Argument-Negation:** | Denying gays the right to adopt children does not violate their human rights. | false |
| **Citation:** | Some say banning gay marriage goes against their human rights. And it sure is. | true |
| **Citation-Negation:** | Some say banning gay marriage goes against their human rights. But it is not. | false |

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 3: Adversarial Examples



# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiment 3: Adversarial Examples

- For adversarial cases, even our best model only achieves 43.4% Accuracy (41.7% F1-score)
- Model is able to capture **shallow semantic similarity** between arguments (*paraphrase*)
- Not capable to predict the **semantically more challenging types** (*argument* and *citation*)
- *Negation*, leading to opposing stance, is **completely overlooked**

# On Classifying whether Two Texts are on the Same Side of an Argument

## Conclusion

- Recent transformer models **improve over the state of the art** in the recent S3C shared task, ALBERT base v2 with best performance (73.7% F1-score)
- S3C shared task's experimental setup suffers from **overfitting**, yielding overly optimistic results ➜ all models fail on **adversarial cases** involving **negation** and **citation** of opposing arguments
  - More realistic evaluation scenario: training and test set pairs sampled from **distinct sets of arguments**
  - Training set with re-occurring arguments in different pairings: pay particular attention to **measures against overfitting**
  - E.g., don't randomly sample validation set from the training set
- Our best models struggle to accurately predict the cross-topic scenario, or **complex cases involving different arguments expressing the same stance**.
  ➜ **Topic-specific knowledge** and a **deeper semantic representation** of individual arguments than those encoded by current transformer models

## Outlook - Improving results

- Same Side Stance Classification main problem
  ➜ data scarcity

- Our idea:
  ➜ **Distant-Supervision Learning** / additional **pretraining** using **data-rich domains** with **similar semantics**
  ➜ Sentiment datasets

- ***Same Sentiment Classification Problem***
  - "new" problem variant of sentiment analysis
  - analogous to Same Side Stance Classification:
    *"given a pair of texts, determine if they have the **same sentiment**, disregarding the actual sentiment polarity"*

Requirements:

- Texts with clear stances or sentiments
- Both **multiple positive and negative samples** about the same topic (e.g. product, movie, business, ...)
- **Multiple topics** with enough samples for **cross-topic comparisons**

# On Classifying whether Two Texts are on the Same Side of an Argument

## Same Sentiment Data

Requirements:

- Texts with clear stances or sentiments
- Both **multiple positive and negative samples** about the same topic (e.g. product, movie, business, ...)
- **Multiple topics** with enough samples for **cross-topic comparisons**

➜ We chose yelp Open Dataset.[2] (business reviews)

- 6,685,900 user reviews
- 192,127 businesses
- 22 main categories

Other options:
Amazon product reviews, IMDb movie reviews

[2] https://www.yelp.com/dataset

# On Classifying whether Two Texts are on the Same Side of an Argument

## Same Sentiment Data

Training data generation:

- Translate the star rating of 1 to 5 to **binary labels**, *good* or *bad* (*good* if the star rating is above 3 stars)
- Filter out businesses that have less than 8 positive and negative reviews
- Randomly combine pairs of reviews about the **same business** per pair type
- **4 sentiment pairs** each for *good-good*, *good-bad*, *bad-bad*, and *bad-good*.

Final dataset:

- 175,940 samples for each pair-type; **703,760** total

# On Classifying whether Two Texts are on the Same Side of an Argument

## Baselines

- **Count**- and **TFIDF**-Vectorizer for feature vector with various classifiers (SVM, logistic regression, SGD)
  ➜ not much better than random baseline
- **Doc2Vec** DBOW embeddings & different embedding-pair pooling & different classifiers
  ➜ slightly better but only around 57%
- Both are no good baselines!

# On Classifying whether Two Texts are on the Same Side of an Argument

## Baselines

- **Count**- and **TFIDF**-Vectorizer for feature vector with various classifiers (SVM, logistic regression, SGD)
  ➜ not much better than random baseline
- **Doc2Vec** DBOW embeddings & different embedding-pair pooling & different classifiers
  ➜ slightly better but only around 57%
- Both are no good baselines!

- **Siamese network**: 50-dim GloVe embeddings + 50 LSTM + 50 hidden units
  [Neculoiu et al. 2016], [Mueller and Thyagarajan 2016]
  ➜ strong baseline, 15 epochs with 83% Accuracy

# On Classifying whether Two Texts are on the Same Side of an Argument

## Transformer Model

- **Standard BERT-base** model [Devlin et al. 2019] for sequence pair classification, default hyper-parameters values
- sequence length of 128 to max. 512 tokens
- fine-tuning for 3 epochs
- *gradient accumulation* to batch small batches (2–6 samples $\rightarrow$ 64) at 512 sequence length

- more recent transformers: **DistilBERT** [Sanh et al. 2019], **ALBERT** [Lan et al. 2020]

# On Classifying whether Two Texts are on the Same Side of an Argument

## Experiments

- **Overall**
  - Random split for train/valid/test (80/10/10%), 5 epochs, sequence lengths (SL) 128 – 512, samples per pair-type 2 – 4
  - 81.3% – 82.0% Acc. for SL 128, 89.1% Acc. for SL 512
- **Per-Major Category**
  - Evaluate on single categories
  - 84% to 95% Acc.
- **Cross-Category**
  - 4-fold cross-validation of random main category splits
  - Evaluation on other fold (79.4% – 92.3% Acc.), single categories (71.5% – 95.3%), rest (83.4% – 85.2%)

➜ Performance as expected. **Slightly** better compared to S3C.

# On Classifying whether Two Texts are on the Same Side of an Argument

## S3C Prediction with Sentiment Pre-Training

| **Setup:** | **Only S3C** | | **+ Yelp Pretraining** | |
|---|---|---|---|---|
| **S3C Train size** | Acc. | F1 | Acc. | F1 |
| **within** | | | | |
| – 51.760 | **87.44** | **88.15** | 86.72 | 87.37 |
| – 5.000 | 60.77 | **63.49** | **61.77** | 61.35 |
| – 500 | **55.44** | 60.65 | 54.36 | **68.63** |
| **cross** | | | | |
| – 54.943 | **64.28** | **67.18** | 63.84 | 65.58 |
| – 5.000 | **58.80** | **54.68** | 58.50 | 51.79 |
| – 500 | 53.01 | 54.64 | **57.80** | **70.74** |

- Model: ALBERT-base-v2, 256 SeqLen, 3 Epochs fine-tuning on S3C train
- Pre-training with Yelp sentiment pair dataset, 1 Epoch on 359k samples
  $\sim$ 84.79% Acc. (84.75% F1)

## Conclusion

- Introduction of new perspective on sentiment analysis
- Initial results (on *same sentiment*) promising
  ➜ Application on different domains like *same stance argument classification* still unsolved
- Hope to find some common features for *“sameness”* to support and improve existing models

## Links

- Contact us:
  erik.koerner@uni-leipzig.de
  g.wiedemann@leibniz-hbi.de
- Code and Data:
  https://github.com/webis-de/EMNLP-21
- Adversarial Test Cases Dataset:
  https://webis.de/data.html#webis-sameside-21

Thank you for listening.