Erik Körner † **Gregor Wiedemann** ‡ Ahmad Dawar Hakimi † Gerhard Heyer † Martin Potthast †

†Leipzig University ‡Leibniz-Institute for Media Research - Hans-Bredow-Institut

†erik.koerner@uni-leipzig.de ‡g.wiedemann@leibniz-hbi.de

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

Same Side Stance Classification

- S3C shared task introduced at 6th Workshop on Argument Mining [Stein et al., 2021]
- to ease the difficulty of argument stance classification
- only *argument similarity* within stances needs to be learned to successfully solve the task
- 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

Contributions

- Reproduction of shared task in its original form as well as the best-performing approach at the S3C shared task by [Ollinger et al., 2021]. Development of new transformer-based approaches which improve upon the state of the art.
- Renewed assessment of the original S3C shared task dataset, and compilation of new training and test sets that enable a more realistic evaluation scenario.
- Compilation of an additional, hand-crafted test set consisting of adversarial cases, such as negations and references to contrary positions within single arguments, to investigate the hypothesis underlying S3C in particular.

Code and data: `github.com/webis-de/EMNLP-21`

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

We recreated an evaluation scenario equivalent to the official S3C shared task. Surprisingly, some newer models, such as 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. [Ollinger et al., 2021]
Our *within* test set can be predicted significantly more accurate than the original S3C test set, the *cross* set performs significantly worse.

Experiment 2: Bias Control

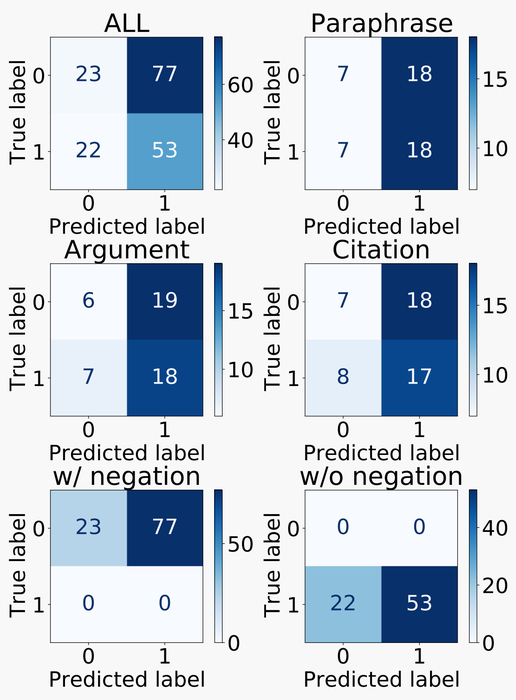
S3C Scenario	Accuracy	F1
Majority baseline	53.4	34.8
random	86.6 (± 0.73)	86.6 (± 0.74)
disjoint		
– within	61.7 (± 1.64)	61.4 (± 1.46)
– cross (A → G)	62.4	62.3
– cross (G → A)	61.2	61.0
single	67.0	64.5

Sampling of the official dataset may lead to unrealistically optimistic results
→ non-overlapping pairs but overlap of single arguments between *train* and *test*, varying *debate sizes*.
We sample 3 new roughly equal-sized dataset splits with varying degrees of overlap of single arguments:
random: replicate sampling of S3C task
disjoint: no single argument from *train* in *test*; split across debates (*cross*) or topic (*within*) (abortion (A) and gay marriage (G))
single: only one *single* argument from each pair is also contained in *train*

Experiment 3: Adversarial Test Cases

- Manually crafted hard adversarial test set
- 175 cases total, based on 25 short, distinct arguments from the “gay marriage” topic, express their stance clearly
- Construct new arguments of distinct types to obtain two pairs (same & 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)



Misclassified pairs from previous disjoint test set experiment reveal typical cases which require certain logical inference capabilities.

For adversarial cases, even our best model only achieves 43.4% Accuracy (41.7% F1-score):

- Successfully captures shallow semantic similarity between arguments (*Paraphrase*)
- Not capable to predict the semantically more challenging types (*Argument* and *Citation*)
- Completely overlooks *Negation*, leading to opposing stance

`webis.de/data.html#webis-sameside-21`

Conclusions

- Recent transformer models improve over the state of the art in the S3C shared task. With 73.7% F1-score, the best performance is achieved by the *ALBERT base v2* model.
- S3C shared task’s experimental setup suffers from overfitting, yielding overly optimistic results. A manually crafted test set of adversarial cases shows that all models fail on adversarial cases involving negation and citation of opposing arguments.
- For a more realistic evaluation scenario, training and test set pairs should be sampled from distinct sets of arguments.
- When the training set involves re-occurring arguments in different pairings, machine learning models should pay particular attention to measures against overfitting. For instance, a validation set should not be randomly sampled 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. For such cases, topic-specific knowledge and a deeper semantic representation of individual arguments than those encoded by current transformer models would be needed.