

A benchmark dataset for detecting frames in multi-topical news content: Online Appendix

A benchmark dataset for detecting frames in multi-topical news content: Online Appendix

Regression coefficients

Table 1

# Fixed Effects

Parameter	Median	89% CI	pd	Rhat	ESS
(Intercept)	0.47	(0.38, 0.56)	99.92%	1.005	1091.00
method_typeSemisupervised	-0.11	(-0.22, 5.98e-03)	93.88%	1.005	1020.00
method_typeAutomatic	-0.15	(-0.26, -0.05)	98.02%	1.005	1037.00

Table 2

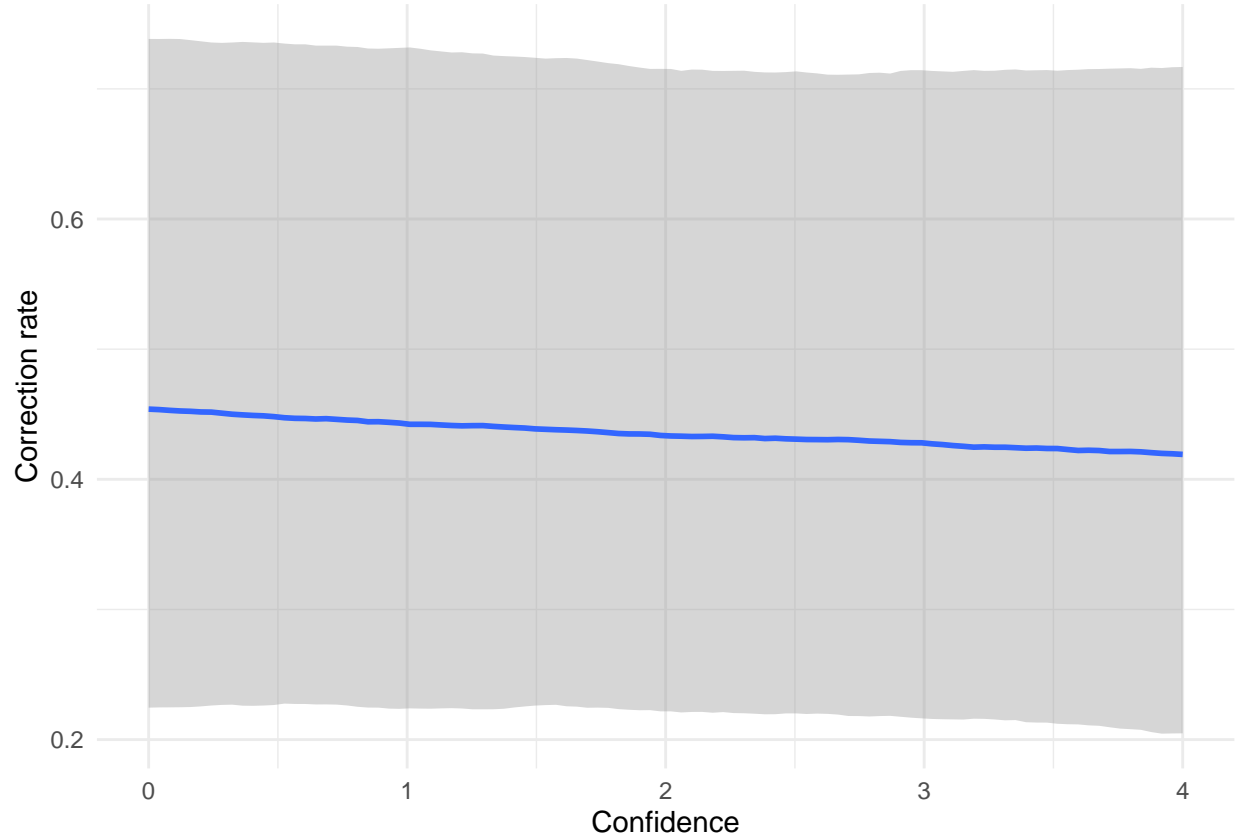
# Sigma

Parameter	Median	89% CI	pd	Rhat	ESS
sigma	0.05	(0.05, 0.05)	100%	0.999	3924.00

$R^2 = 0.1840042$

Comparing confidence level of correct and incorrect expert coding

We modeled the correctness of expert coding (“F1” is equal to the ground truth) and confidence level (“F2”), while adjusting for individual differences between the two experts using Bayesian multilevel logistic regression analysis. The following is the robust conditional effect plot. There is no evidence to suggest that there is a trend. Therefore, experts can either confidently give correct and incorrect coding.



*Figure 1.* Robust conditional effects from the Bayesian model on the relationship between correction rate of expert coding and confidence at 89% credibility

### Simulation of increasing sample size

In this analysis, we simulated the possible outcome of increasing the sample size on the multiverse analysis.

From our 100 articles, we created further synthetic articles following the principle of bootstrapping. We synthesized new article based on the following algorithm:

1. Randomly select one article
2. Tokenize this article into  $n$  sentences
3. From these  $n$  sentences, randomly draw  $n$  sentences from these sentences with replacement. Therefore, one sentence can appear more than once.

4. Concatenate these randomly drawn  $n$  sentences into a synthetic article, assign this article with the same topic and frame as the original article

We repeat the above process for 500, 1000, and 2000 times to generate four different corpora. This approach is compatible with the bag-of-words representation used in all unsupervised and semi-supervised methods. Also in step 3, topical and frame clues have the same natural chance of being selected. To put this simulation in another perspective, it simulates whether frames, rather than topics, are more likely to be picked up by these unsupervised and semi-supervised methods when the sample size increases.

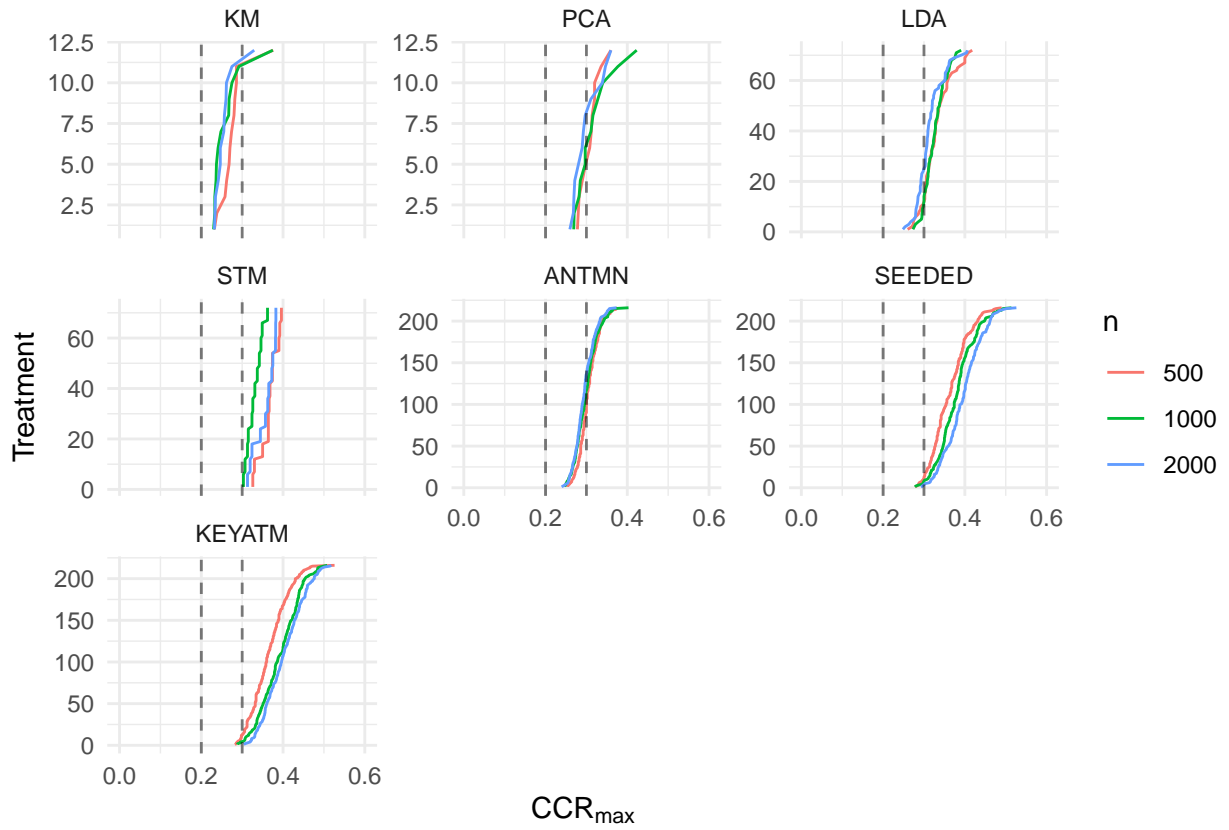


Figure 2. Multiverse analyses with different sample sizes: 500, 1000, 2000

From this simulation, there is highly unlikely that increasing the sample size would increase the correct classification rate for all unsupervised methods. Instead, all methods, except STM, perform more like the de-facto null (0.3) with the increasing sample size.

Therefore, these methods appear to be more keen to pick up **topics**, rather than frames, when sample size increases. It reflects by the performance curve rescinding towards the null value with the highest sample size. For the two semi-supervised methods, increasing the sample size appears to provide better resistance against the rescinding towards the null value.

We also performed the same Bayesian analysis using the simulated data. By increasing the sample size, it appears that both H1 and H2 would become more significant, instead of the other way around.

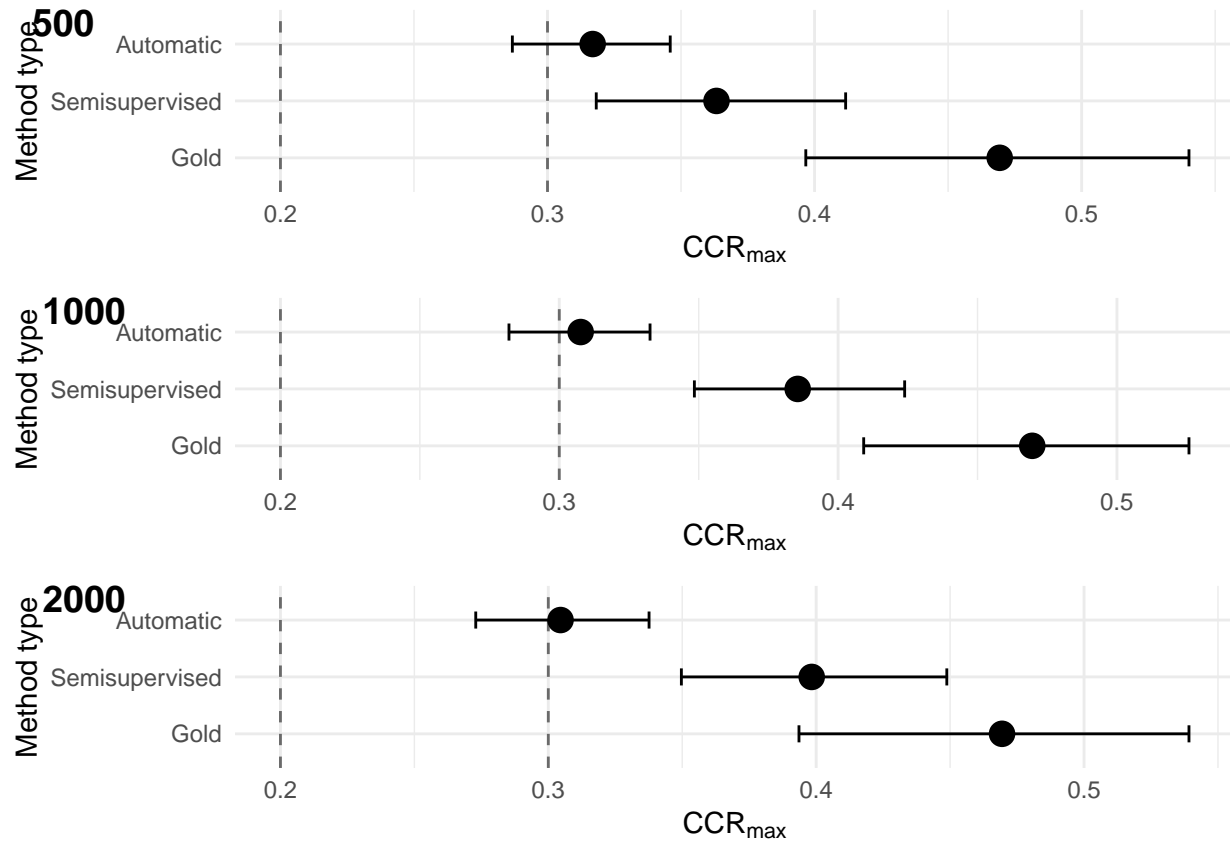


Figure 3. Robust conditional effects from the Bayesian model of the “gold standard”, semisupervised and automatic methods at 89% credibility with different simulated sample sizes