Overview of the Cross-Domain Authorship Verification Task at PAN 2021

Mike Kestemont, Enrique Manjavacas, Ilia Markov,
Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos,
Benno Stein and Martin Potthast

pan@webis.de https://pan.webis.de

DOI 10.5281/zenodo.5106099

Context

- Long-running task on relevant problem, but
  - Lack of large-scale resources in field in general
  - Lack of task realism
- Renewed 3-year strategy, increasing difficulty, scope and realism
  - Year 1 (2020): increased size
  - Year 2 (2021): increased difficulty
  - Year 3 (2022): “mystery task”

Task (= 2020)

- Authorship verification (not attribution, obfuscation, …)
- Test set consist of series of “problems”:
  - Given a pair of texts, assign a verification score [0, 1]
  - < 0.5 (different-author: DA) or > 0.5 (same-author: SA)
  - Exactly 0.5: non-response (for “difficult” pairs)
- Reference set of pairs available to calibrate systems

Dataset

- Fanfiction dataset (from fanfiction.net): non-professional authors expanding “canons” of well-known works and authors (“fandoms”)
  - Fandom information as a proxy for “domain”
  - English-language (but global phenomenon)
  - Huge scale (and no moderation)
  - User-provided metadata

Novelty

- Test set: 19,999 problems. Still cross-fandom but:
- Last year: closed set scenario (no new authors in test set)
  - (Without participants knowing!)
  - Could be reformulated as attribution task
- This year: fully disjunct test set (open set scenario)
  - Only unseen authors, only in new domains (fandoms)
  - Supposedly much more difficult (!)

Dataset sizes (approx.)

(Largest resource in verification that we know of)

	Same-Author Pairs	Different-Author Pairs
Calibration (“large”)	148K	128K
Calibration (“small”)	28K	25K
Test (2020)	10K	6.9K
Test (2021)	10K	10K

Evaluation framework

- Varied set of 5 metrics, sensitive to different aspects, with new addition:

    1. AUC: conventional area-under-the-curve score
    2. F1: classic metric, but *not* taking into account non-answers
    3. c@1: F1 variant: rewards systems that leave difficult problems unanswered
    4. F0.5u: new measure, emphasis on deciding same-author cases correctly
    5. Brier: complement Brier score loss (kudos F. Sebastiani)

- Combined score for final ranking

2+1 baselines

Straightforward but competitive

Calibrated on “small” set only (give “large” systems edge):

1. Cosine similarity between TF-IDF BOW of 4-grams (with naive “hack” to shift scores)

2. Text compression method, based on cross-entropy for “text2” using Prediction by Partial Matching

3. [Post-hoc] Short-text unmasking, Bevendorff et al. (2019) based on Koppel and Schler (2004)

Submissions

- 13 submissions from 10 teams (similar to last year)
- Again: no calibration on Tira (only testing/deployment) for more flexibility
- 3 teams submitted “small” and “large” versions
  - Others used “small” or “large” version
- Diverse array of methods, including representation learning

Results

Most participants above baselines (baselines remarkably similar)

Team	**Dataset**	AUC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
boenninghoff21	large	**0.9869**	**0.9502**	**0.9524**	**0.9378**	**0.9452**	**0.9545**
embarcaderoruiz21	large	0.9697	0.9306	0.9342	0.9147	0.9305	0.9359
weerasinghe21	large	0.9719	0.9172	0.9159	0.9245	0.9340	0.9327
weerasinghe21	small	0.9666	0.9103	0.9071	0.9270	0.9290	0.9280
menta21	large	0.9635	0.9024	0.8990	0.9186	0.9155	0.9198
peng21	small	0.9172	0.9172	0.9167	0.9200	0.9172	0.9177
embarcaderoruiz21	small	0.9470	0.8982	0.9040	0.8785	0.9072	0.9070
menta21	small	0.9385	0.8662	0.8620	0.8787	0.8762	0.8843
rabinovits21	small	0.8129	0.8129	0.8094	0.8186	0.8129	0.8133
ikae21	small	0.9041	0.7586	0.8145	0.7233	0.8247	0.8050
unmasking21	small	0.8298	0.7707	0.7803	0.7466	0.7904	0.7836
tyo21	large	0.8275	0.7594	0.7911	0.7257	0.8123	0.7832
naive21	small	0.7956	0.7320	0.7856	0.6998	0.7867	0.7600
compressor21	small	0.7896	0.7282	0.7609	0.7027	0.8094	0.7581
futrzynski21	large	0.7982	0.6632	0.8324	0.6682	0.7957	0.7516
liaozhihao21	small	0.4962	0.4962	0.0067	0.0161	0.4962	0.3023

Significance

Approximate randomization testing (F1 as reference)

	embarcaderoruiz21-large	weerasinghe21-large	weerasinghe21-small	menta21-large	peng21-small	embarcaderoruiz21-small
boenninghoff21-large	***	***	***	***	***	***
embarcaderoruiz21-large		*	=	***	**	***
weerasinghe21-large			***	***	=	***
weerasinghe21-small				***	***	***
menta21-large					***	***
peng21-small						***

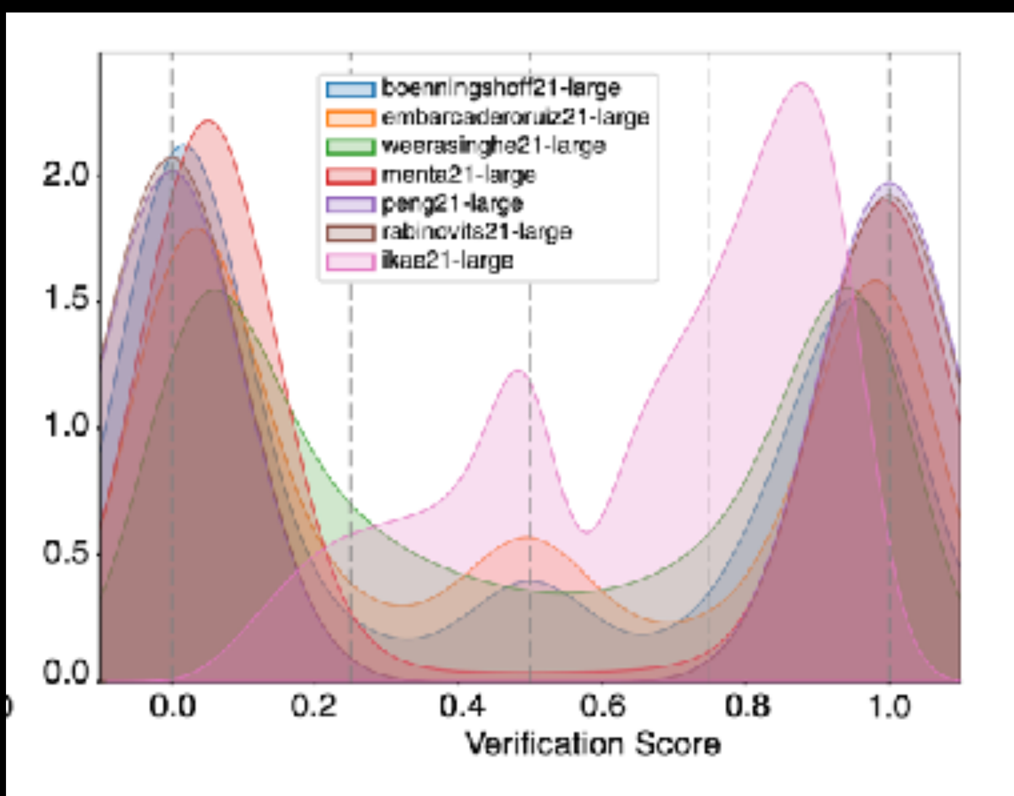
Table 2
Significance of pairwise differences in F_1 scores between submissions. Notation: '=' (not significant: $p \geq 0.05$), '*' (significant with $p < 0.05$), '**' (significant with $p < 0.01$), '***' (significant with $p < 0.001$).

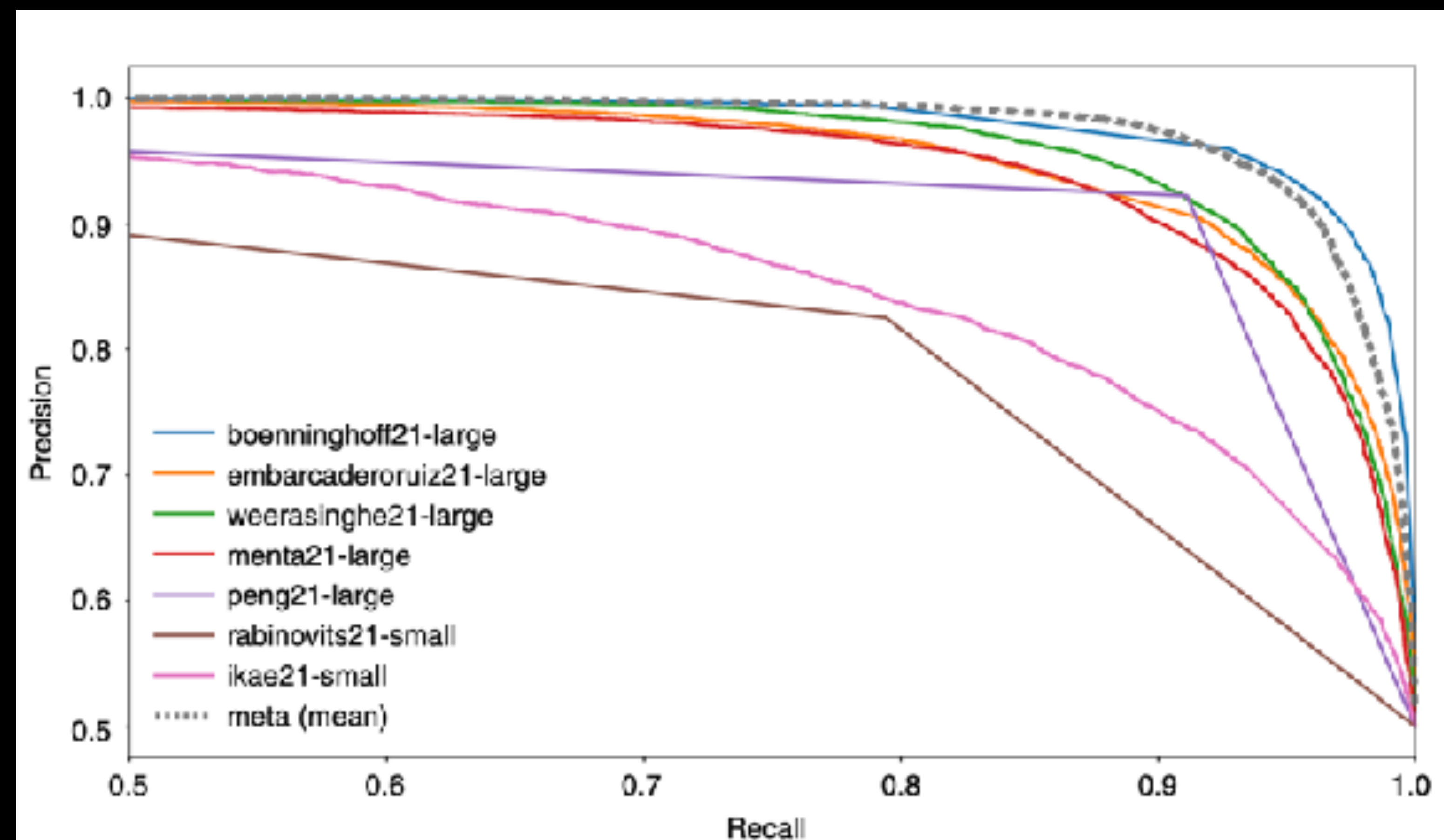
Evolution

Team	2020 System		2021 System	
	2020 Data	2021 Data	2020 Data	2021 Data
niven	0.786	–	–	–
araujo	0.770	0.81	–	–
boenninghoff	0.928	–	0.917	0.950
weerasinghe	0.880	0.913	0.885	0.917
ordonez	0.640	–	–	–
faber	0.331	–	–	–
ikae	0.544	0.503	0.742	0.758
kipnis	0.801	0.815	–	–
gagala	0.786	0.804	–	–
halvani	0.796	0.822	–	–
embarcaderoruiz	–	–	0.914	0.930
menta	–	–	0.878	0.902
peng	–	–	–	0.917
rabinovits	–	–	0.795	0.812
tyo	–	–	–	0.759
futrzynski	–	–	0.662	0.663
liaozhihao	–	–	–	0.496

Score distributions

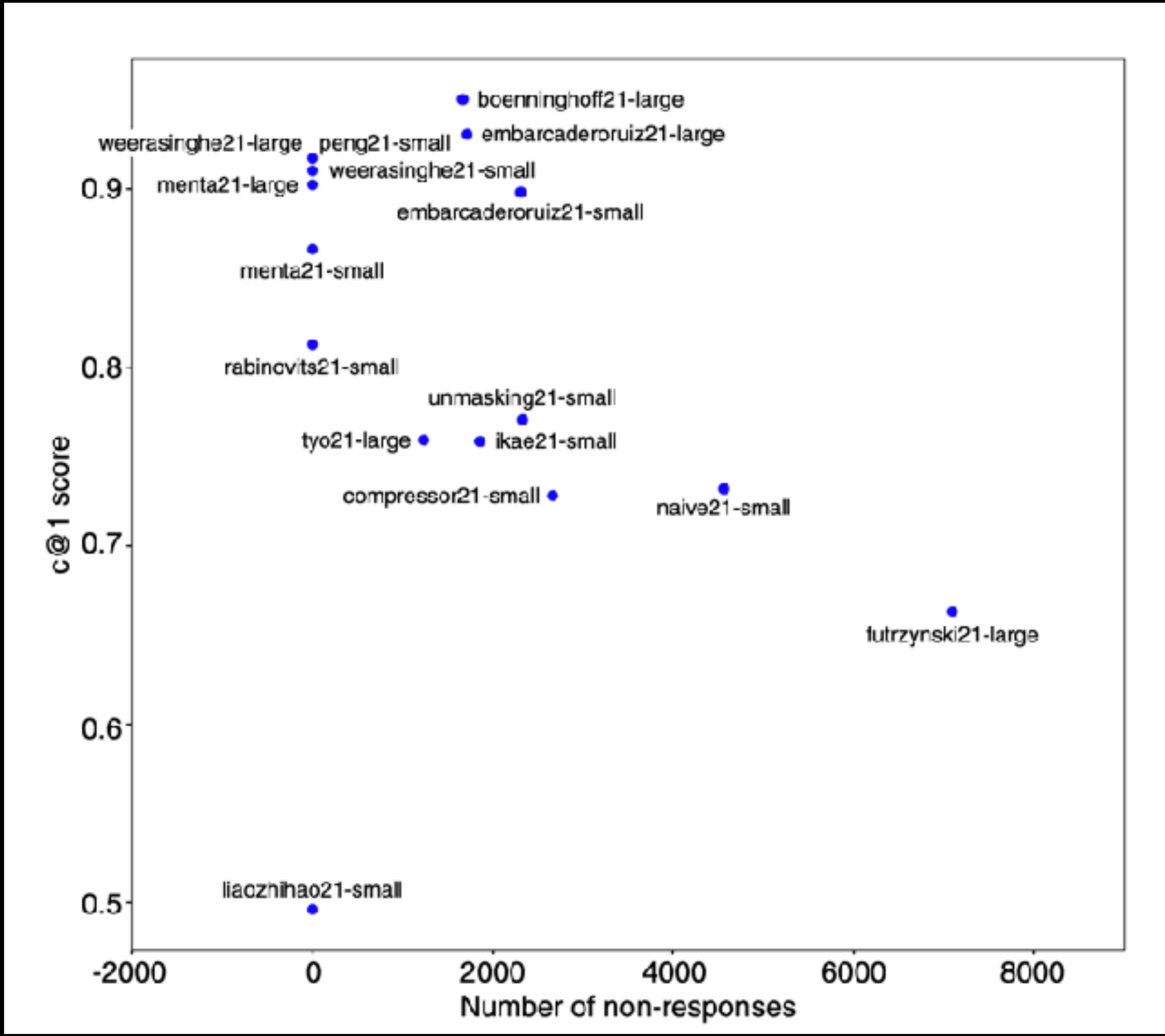
Number heaping but strong metaclassifier





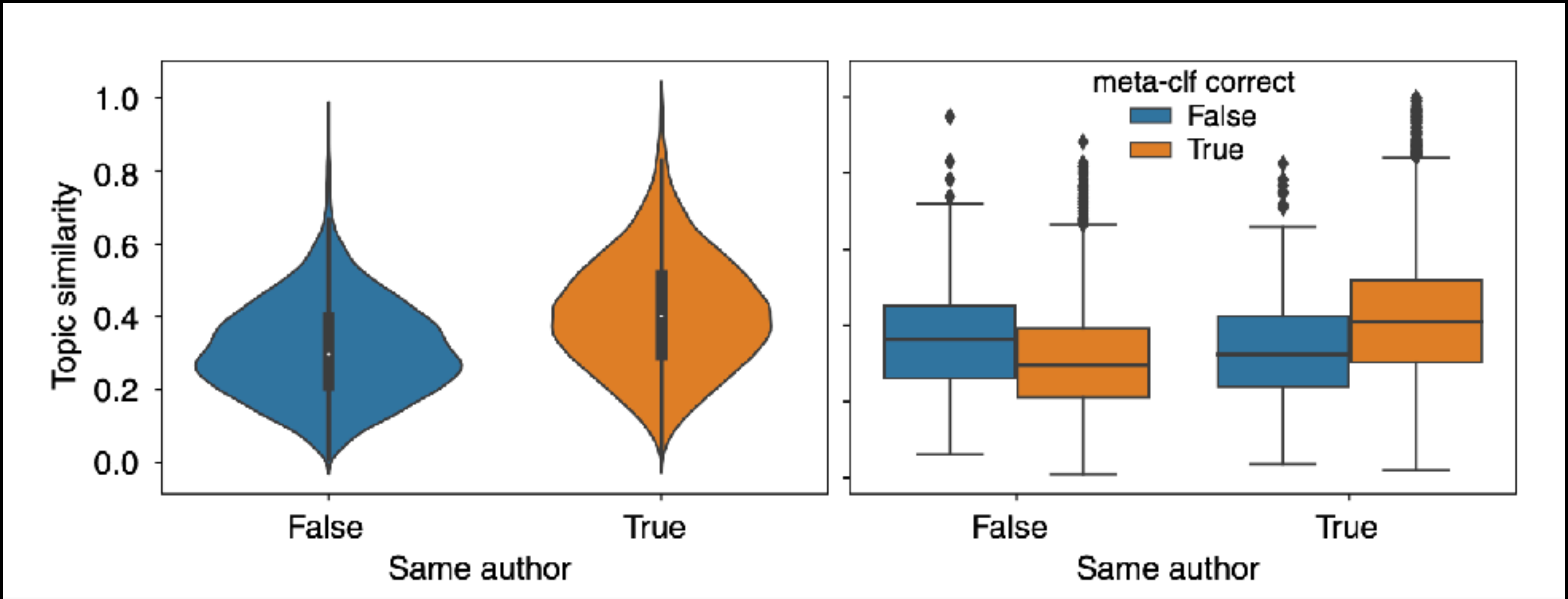
[Last year, metaclassifier did not outperform strongest participant…]

Non-response



c@1 as a function of absolute number of non-answers

Topical similarity (*cont.*)



Topical similarity is useful cue for authorship, but can be misleading

Conclusions

- Last year as turning point? Consolidated this year
- At least within this domain (but how representative?):
  - Large-scale authorship verification feasible
  - Open-set did not degrade results (counter-intuitive)
- Thanks to team and participants and see you next year!