Overview of the Cross-Domain Authorship Verification Task at PAN 2020

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

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

DOI 10.5281/zenodo.3716403

Renewed strategy

- Long-running task on relevant problem, but:
  - Lack of diversity in submissions recently
  - Lack of large-scale resources in field in general
  - Lack of task realism (cf. court settings)
- 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

- Authorship verification (and not attribution, obfuscation, …)
- Calibration and 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)
- Only unseen test texts, but “closed” scenario: no new authors. All pairs are cross-fandom.

Benchmark dataset

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

Dataset size

(Largest resource in verification that we know of)

	Same-Author Pairs	Different-Author Pairs	# fandoms	SA authors	DA authors
Train (“large”)	148K	128K	1.6K	41K	250K
Train (“small”)	28K	25K	1.6K	25K	48.5K
Test (2020)	10K	6.9K	0.4K	3.5K	12K

Interesting differences: some only used subset of “small”, others enlarged “large” even further

Evaluation framework

- Varied set of 4 metrics, sensitive to different aspects:
  - AUC: conventional area-under-the-curve score
  - c@1: variant of F1, rewarding systems that leave difficult problems unanswered
  - F1: classic metric, but *not* taking into account non-answers
  - F0.5u: new measure, emphasis on deciding same-author cases correctly
- Combined score for final ranking

Two 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

[All code available from Github (https://github.com/pan-webis-de/pan-code/tree/master/clef20); all data from Zenodo (https://zenodo.org/record/3716403#.X2neLpMzZ25)]

Submissions

- 13 submissions from 10 teams
- Novelty: no calibration on Tira (only testing/deployment) for more flexibility
- 3 teams submitted “small” and “large” versions
  - Others used “small” (or subset!) apart from ordonez20
- Much more diverse array of methods, including use of e.g. siamese nets and fandom info

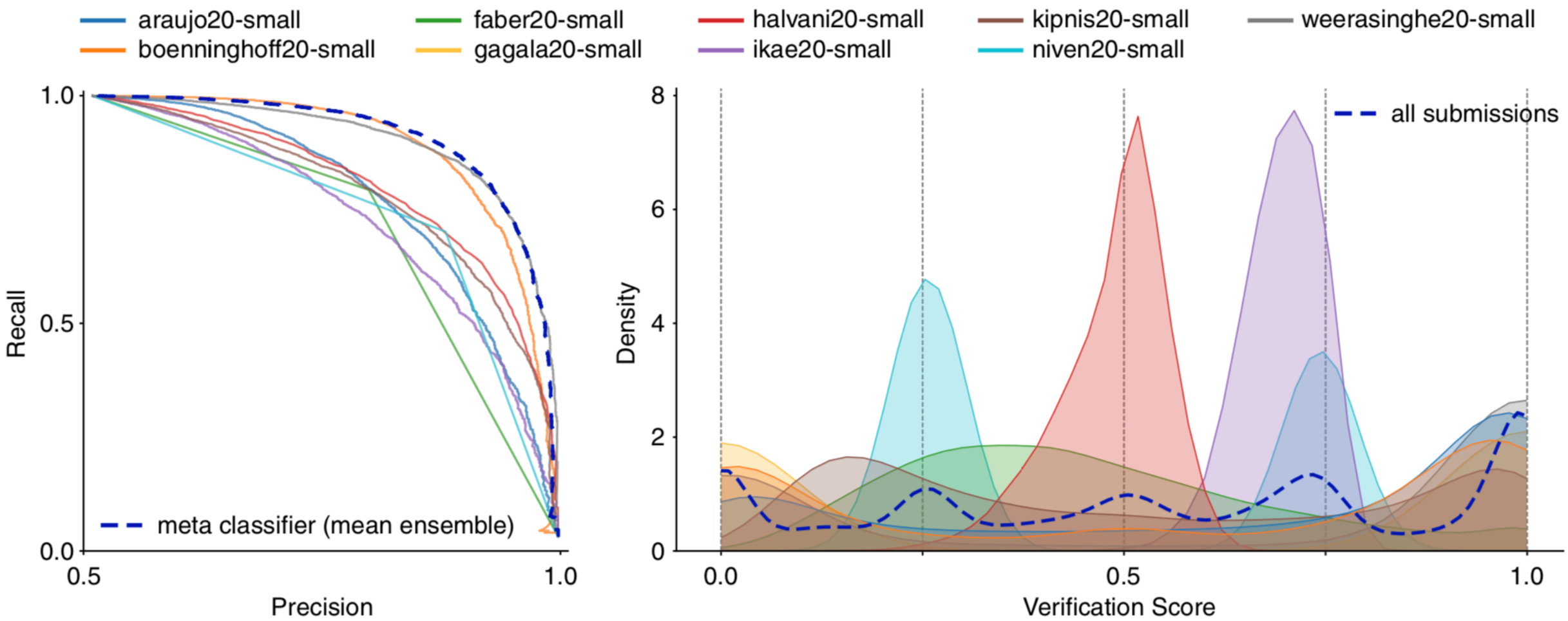
Results

Pair-wise differences mostly significant: indicative of diversity

Submission	AUC	c@1	F0.5u	F1	Overall
boenninghoff20-large	**0.969**	**0.928**	**0.907**	**0.936**	**0.935**
weerasinghe20-large	0.953	0.880	0.882	0.891	0.902
boenninghoff20-small	0.940	0.889	0.853	0.906	0.897
weerasinghe20-small	0.939	0.833	0.817	0.860	0.862
halvani20-small	0.878	0.796	0.819	0.807	0.825
kipnis20-small	0.866	0.801	0.815	0.809	0.823
araujo20-small	0.874	0.770	0.762	0.811	0.804
niven20-small	0.795	0.786	0.842	0.778	0.800
gagala20-small	0.786	0.786	0.809	0.800	0.796
araujo20-large	0.859	0.751	0.745	0.800	0.789
baseline (naive)	0.780	0.723	0.716	0.767	0.747
baseline (compression)	0.778	0.719	0.703	0.770	0.742
ordonez20-large	0.696	0.640	0.655	0.748	0.685
ikae20-small	0.840	0.544	0.704	0.598	0.672
faber20-small	0.293	0.331	0.314	0.262	0.300

Analysis (1): “small” distributions

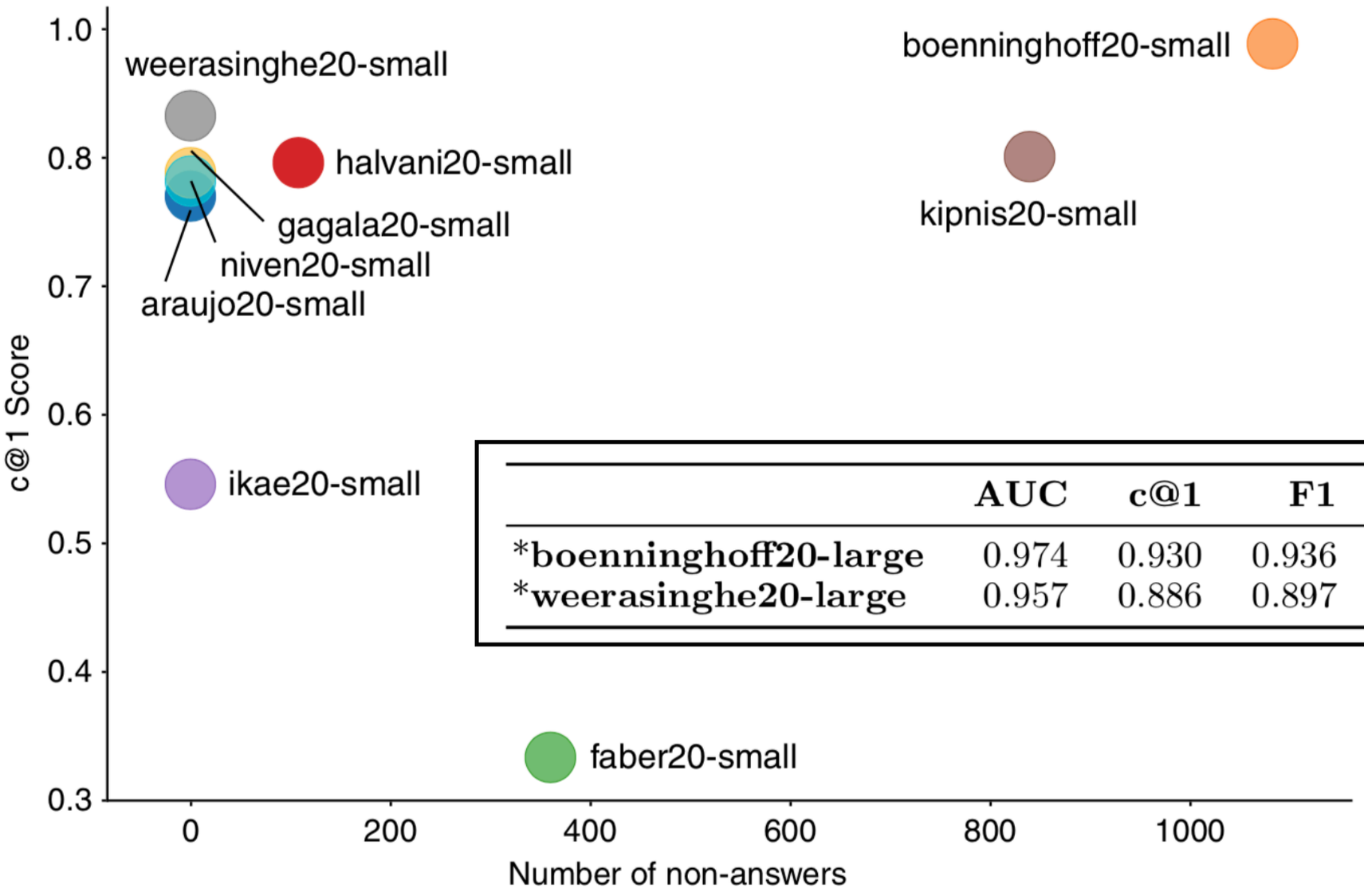
Number heaping but strong metaclassifier



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

Analysis (2): Non-answers

boeninghoff20 surprisingly solid non-response without compromizing score

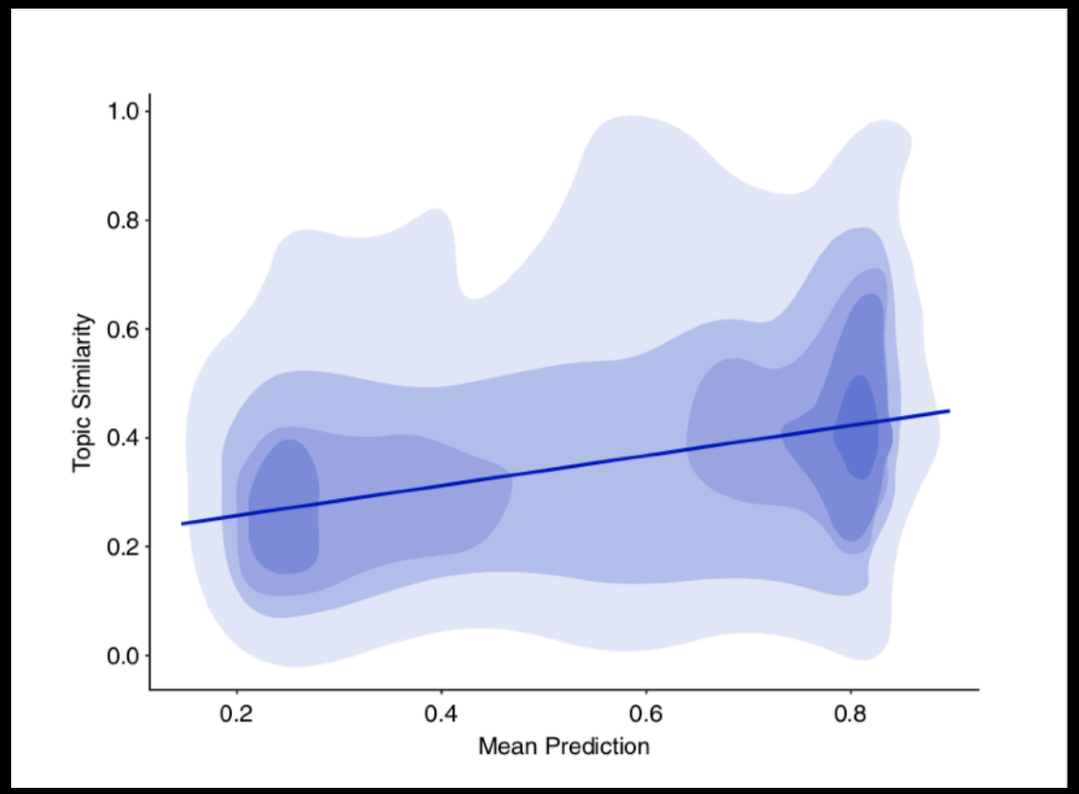


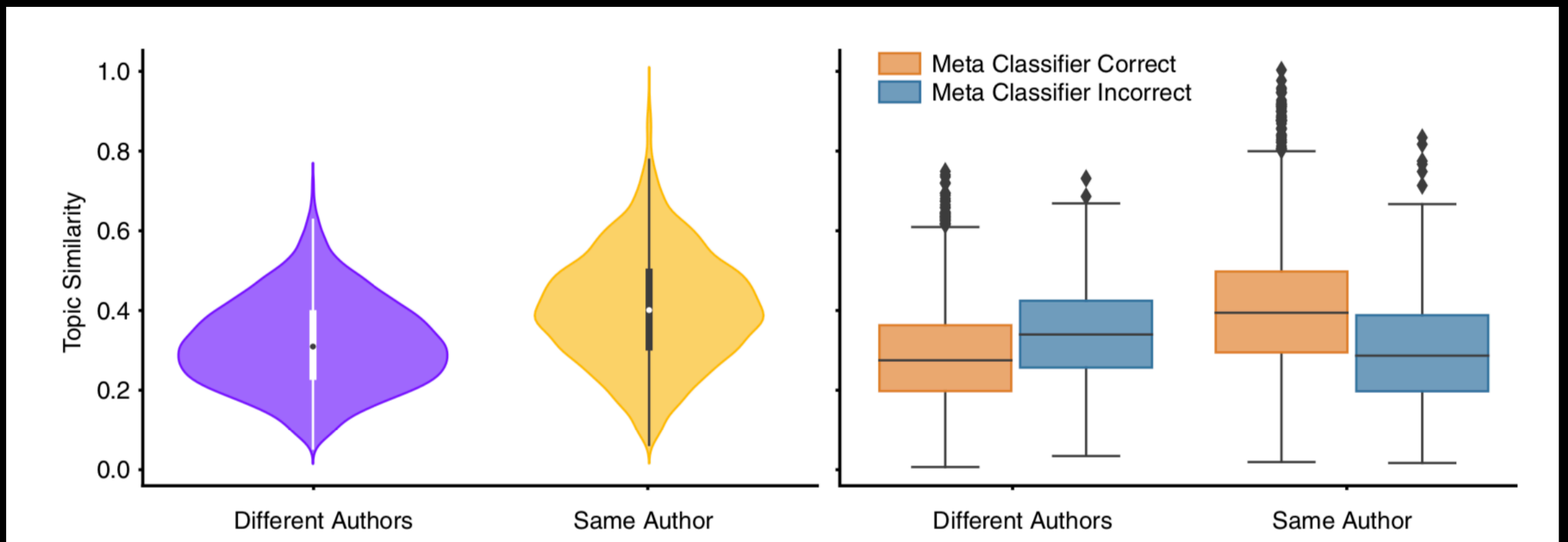
	AUC	c@1	F1	F0.5u	Overall
***boenninghoff20-large**	0.974	0.930	0.936	0.934	0.943
***weerasinghe20-large**	0.957	0.886	0.897	0.888	0.907

Analysis (3): topic model

NMF on TFIDF | 150 dims | top 5K tokens

Analysis (4): Topic effect is real





All score: “small” meta-classifier on test set

Conclusions

- Higher diversity in submissions lead to interesting edition
- Reliably established that scale and size matter
- Promising new neural approaches (but brittle, cf. ordonez20)
- Closing in on solution for in-domain authorship attribution
- Topic-author orthogonality remains holy grail (next year!)
- Next year:
  - same training data
  - but much (!) more challenging test dataset

Many thanks to the AV@PAN team,
but especially the participants!

Check out the task overview paper for
more info and see you next year!