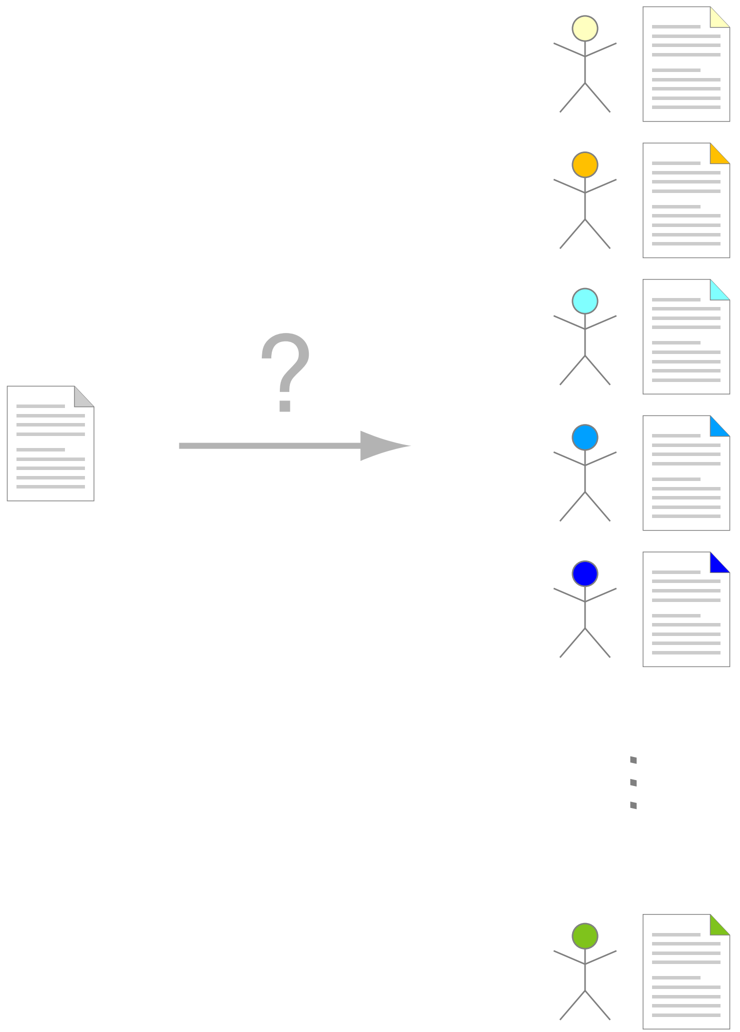
Overview of the Authorship Verification Task at PAN 2022

Efstathios Stamatatos, Mike Kestemont, Krzysztof Kredens, Piotr Pezik,
Annina Heini, **Janek Bevendorff**, Benno Stein, Martin Potthast

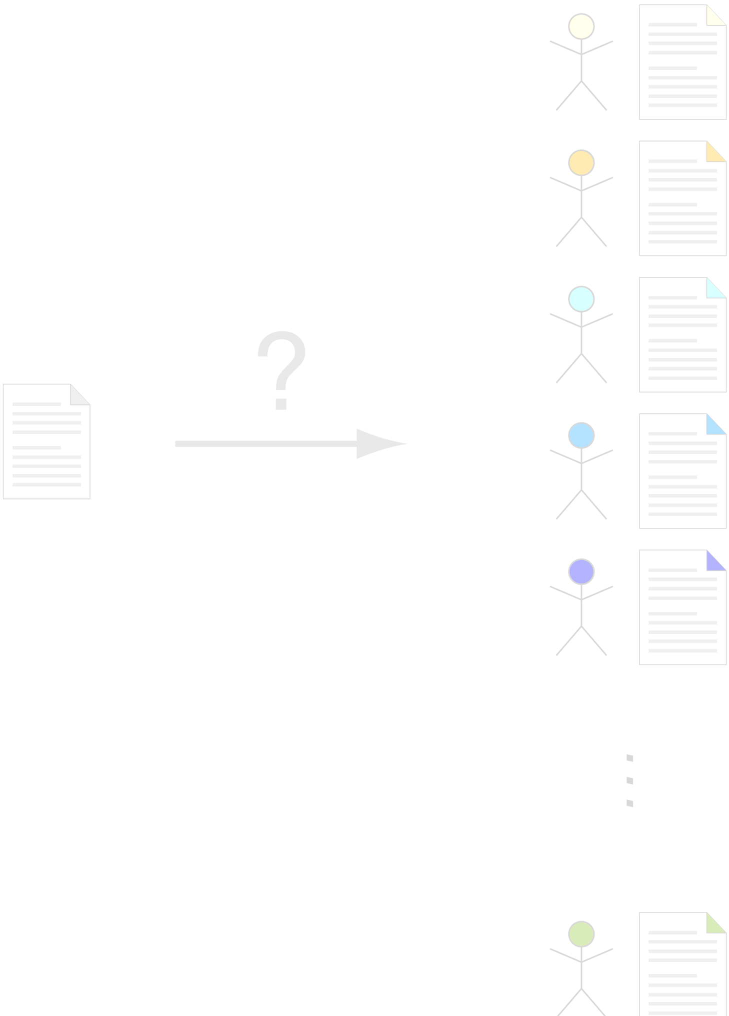
PAN @ CLEF 2022 – pan.webis.de

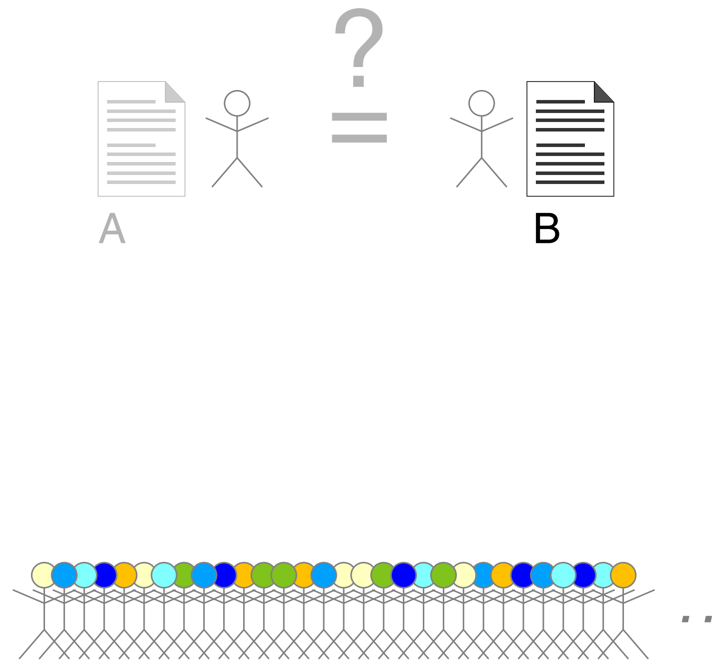
September 6, 2022, Bologna

Authorship Verification



Authorship Verification





PAN 2020–2022 Overview

1. PAN 2020:
   Closed-set verification on fanfiction texts

2. PAN 2021:
   Open-set verification on fanfiction texts

3. PAN 2022:
   “Surprise task”: cross-discourse type authorship verification

PAN 2020–2022 Overview

1. PAN 2020:
   Closed-set verification on fanfiction texts

2. PAN 2021:
   Open-set verification on fanfiction texts

3. PAN 2022:
   “Surprise task”: cross-discourse type authorship verification

The Data

The task's training and test data is based on the *Aston 100 Idiolects*[1] corpus:

- Text samples by 112 individuals using various discourse types.
- Authors have similar age characteristics.
- Authors are native speakers of English.
- Topic is unrestricted.

[1] Kredens, Heini, and Pezik; 2021

The Data

The task's training and test data is based on the *Aston 100 Idiolects*[1] corpus:

- Text samples by 112 individuals using various discourse types.
- Authors have similar age characteristics.
- Authors are native speakers of English.
- Topic is unrestricted.

Selected Discourse Types:

Essays, emails, business memos, text messages.

[1] Kredens, Heini, and Pezik; 2021

The Data (continued)

Subset	**Training**	**Test**
Author match	*Text pairs*	
Positive (same author)	6,132 (50.0%)	5,239 (50.0%)
Negative (different author)	6,132 (50.0%)	5,239 (50.0%)
Discourse type pairings	*Text pairs*	
Email–Text message	7,484 (61.0%)	6,092 (58.1%)
Essay–Email	1,618 (13.2%)	1,454 (13.9%)
Essay–Text message	1,182 (9.6%)	1,128 (10.8%)
Business memo–Email	1,014 (8.3%)	900 (8.6%)
Business memo–Text message	780 (6.4%)	718 (6.9%)
Essay–Business memo	186 (1.5%)	186 (1.8%)
Discourse type	*Text length (avg. chars)*	
Essay	11,098	10,117
Email	2,385	2,323
Business memo	1,255	1,042
Text message	611	601

The Data (continued)

Subset	Training	Test
Author match	*Text pairs*	
Positive (same author)	6,132 (50.0%)	5,239 (50.0%)
Negative (different author)	6,132 (50.0%)	5,239 (50.0%)
Discourse type pairings	*Text pairs*	
Email–Text message	7,484 (61.0%)	6,092 (58.1%)
Essay–Email	1,618 (13.2%)	1,454 (13.9%)
Essay–Text message	1,182 (9.6%)	1,128 (10.8%)
Business memo–Email	1,014 (8.3%)	900 (8.6%)
Business memo–Text message	780 (6.4%)	718 (6.9%)
Essay–Business memo	186 (1.5%)	186 (1.8%)
Discourse type	*Text length (avg. chars)*	
Essay	11,098	10,117
Email	2,385	2,323
Business memo	1,255	1,042
Text message	611	601

The Data (continued)

Subset	**Training**	**Test**
Author match	*Text pairs*	
Positive (same author)	6,132 (50.0%)	5,239 (50.0%)
Negative (different author)	6,132 (50.0%)	5,239 (50.0%)
Discourse type pairings	*Text pairs*	
Email–Text message	7,484 (61.0%)	6,092 (58.1%)
Essay–Email	1,618 (13.2%)	1,454 (13.9%)
Essay–Text message	1,182 (9.6%)	1,128 (10.8%)
Business memo–Email	1,014 (8.3%)	900 (8.6%)
Business memo–Text message	780 (6.4%)	718 (6.9%)
Essay–Business memo	186 (1.5%)	186 (1.8%)
Discourse type	*Text length (avg. chars)*	
Essay	11,098	10,117
Email	2,385	2,323
Business memo	1,255	1,042
Text message	611	601

The Data (continued)

Source Data:

`pairs.jsonl:`


```
{"id": "a09fdc6b-ed15-48c5-9d2e-572f989b9b45",
     "discourse_type": ["essay", "text_message"],
     "pair": ["Text 1...", "Text 2..."]}
...
```


`truth.jsonl:`


```
{"id": "a09fdc6b-ed15-48c5-9d2e-572f989b9b45",
     "same": false, "authors": ["en_110", "en_112"]}
...
```

The Data (continued)

Source Data:

`pairs.jsonl:`


```
{"id": "a09fdc6b-ed15-48c5-9d2e-572f989b9b45",
     "discourse_type": ["essay", "text_message"],
     "pair": ["Text 1...", "Text 2..."]}
...
```


`truth.jsonl:`


```
{"id": "a09fdc6b-ed15-48c5-9d2e-572f989b9b45",
     "same": false, "authors": ["en_110", "en_112"]}
...
```


Answer Submission:


```
{"id": "a09fdc6b-ed15-48c5-9d2e-572f989b9b45", "value": 0.4921}
...
```

Evaluation

Answers are in the range $[0, 1]$ indicating the *same author* class probability:

- > 0.5: most likely same author
- < 0.5: most likely different authors
- $= 0.5$: no answer commitment

Evaluation

Answers are in the range $[0, 1]$ indicating the *same author* class probability:

- > 0.5: most likely same author
- < 0.5: most likely different authors
- $= 0.5$: no answer commitment

Performance is assessed by five measures:

- AUROC: area under the ROC curve
- F_1: Harmonic mean of precision and recall for *same author* class
- $F_{0.5U}$: Precision-weighted F score which rewards non-answers
- C@1: Modified binary accuracy which rewards non-answers
- BRIER: Brier score complement (inverse binary quadratic loss)

Final score is calculated as the arithmetic mean of all five.

Baselines

- CNGDIST22: Distance-based character n-gram model: cosine similarity on most frequent 4-grams with two thresholds for classes or “undecided”.
- COMPRESSOR22: Compression-based model: logistic regression classifier trained on the PPM cross-entropy between texts, scores ≈ 0.5 are set to 0.5.

Baselines

- CNGDIST22: Distance-based character n-gram model: cosine similarity on most frequent 4-grams with two thresholds for classes or "undecided".
- COMPRESSOR22: Compression-based model: logistic regression classifier trained on the PPM cross-entropy between texts, scores ≈ 0.5 are set to 0.5.

Baseline Name	AUROC	C@1	F_1	$F_{0.5u}$	BRIER	MEAN
BASELINE-CNGDIST22	**0.546**	**0.496**	**0.669**	**0.542**	0.749	**0.600**
BASELINE-COMPRESSOR22	0.541	0.493	0.570	0.478	**0.750**	0.566

Submitted Systems

Seven participants handed in their models.

Models were evaluated (but not trained) on the Tira[1] platform.

[1] `https://www.tira.io`

Submitted Systems

Seven participants handed in their models.

Models were evaluated (but not trained) on the Tira[1] platform.

System	Representation	Architecture	Augm.
NAJAFI22	T5, word unigrams, POS,	CNN	No
	NEs, Punctuation		No
GALICIA22	graph-based, POS	Siamese network	Yes
JINLI22	MPNET		No
LEI22	BERT		No
YIHUIYE22	BERT	TextCNN	Yes
HUANG22	BERT		No
CRESPOSANCHEZ22	word unigrams, doc2vec		Yes
	(text and POS), SOM		Yes

[1] https://www.tira.io

Participant Results

System	AUROC	c@1	F_1	$F_{0.5u}$	BRIER	MEAN
BASELINE-CNGDIST22	0.546	0.496	0.669	0.542	0.749	**0.600**
NAJAFI22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
GALICIA22	0.512	0.499	0.628	0.544	0.741	0.585
JINLI22	0.577	0.557	0.581	0.563	0.589	0.573
BASELINE-COMPRESSOR22	0.541	0.493	0.570	0.478	**0.750**	0.566
LEI22	0.539	0.539	0.399	0.488	0.539	0.501
YIHUIYE22	0.542	0.526	0.398	0.461	0.565	0.499
HUANG22	0.519	0.519	0.196	0.328	0.519	0.416
CRESPOSANCHEZ22	0.500	0.500	0	0	0.748	0.350

Participant Results

System	AUROC	c@1	F_1	$F_{0.5u}$	BRIER	MEAN
BASELINE-CNGDIST22	0.546	0.496	0.669	0.542	0.749	**0.600**
NAJAFI22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
GALICIA22	0.512	0.499	0.628	0.544	0.741	0.585
JINLI22	0.577	0.557	0.581	0.563	0.589	0.573
BASELINE-COMPRESSOR22	0.541	0.493	0.570	0.478	**0.750**	0.566
LEI22	0.539	0.539	0.399	0.488	0.539	0.501
YIHUIYE22	0.542	0.526	0.398	0.461	0.565	0.499
HUANG22	0.519	0.519	0.196	0.328	0.519	0.416
CRESPOSANCHEZ22	0.500	0.500	0	0	0.748	0.350

Participant Results

System	AUROC	c@1	F_1	$F_{0.5u}$	BRIER	MEAN
BASELINE-CNGDIST22	0.546	0.496	0.669	0.542	0.749	**0.600**
NAJAFI22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
GALICIA22	0.512	0.499	0.628	0.544	0.741	0.585
JINLI22	0.577	0.557	0.581	0.563	0.589	0.573
BASELINE-COMPRESSOR22	0.541	0.493	0.570	0.478	**0.750**	0.566
LEI22	0.539	0.539	0.399	0.488	0.539	0.501
YIHUIYE22	0.542	0.526	0.398	0.461	0.565	0.499
HUANG22	0.519	0.519	0.196	0.328	0.519	0.416
CRESPOSANCHEZ22	0.500	0.500	0	0	0.748	0.350

Participant Results

System	AUROC	c@1	F_1	$F_{0.5u}$	BRIER	MEAN
BASELINE-CNGDIST22	0.546	0.496	0.669	0.542	0.749	**0.600**
NAJAFI22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
GALICIA22	0.512	0.499	0.628	0.544	0.741	0.585
JINLI22	0.577	0.557	0.581	0.563	0.589	0.573
BASELINE-COMPRESSOR22	0.541	0.493	0.570	0.478	**0.750**	0.566
LEI22	0.539	0.539	0.399	0.488	0.539	0.501
YIHUIYE22	0.542	0.526	0.398	0.461	0.565	0.499
HUANG22	0.519	0.519	0.196	0.328	0.519	0.416
CRESPOSANCHEZ22	0.500	0.500	0	0	0.748	0.350

Model Biases

System	Positive	Negative	Unanswered
NAJAFI22	5,355	5,083	40
GALICIA22	8,874	1,604	0
JINLI22	5,820	4,658	0
LEI22	2,805	7,673	0
YIHUIYE22	2,841	7,116	521
HUANG22	1,031	9,447	0
CRESPOSANCHEZ22	0	10,478	0

Baseline Name	Positive	Negative	Unanswered
BASELINE-CNGDIST22	9,199	17	1,262
BASELINE-COMPRESSOR22	3,927	3,268	3,283

Model Biases

System	Positive	Negative	Unanswered
NAJAFI22	5,355	5,083	40
GALICIA22	8,874	1,604	0
JINLI22	5,820	4,658	0
LEI22	2,805	7,673	0
YIHUIYE22	2,841	7,116	521
HUANG22	1,031	9,447	0
CRESPOSANCHEZ22	0	10,478	0

Baseline Name	Positive	Negative	Unanswered
BASELINE-CNGDIST22	9,199	17	1,262
BASELINE-COMPRESSOR22	3,927	3,268	3,283

Explanations?

- Models too complex for the data?
- Data lends itself to overfitting?
- Issues with the test split?
- Task too difficult?
- . . .

Lots of hypotheses to investigate.

Do Previous Systems Perform Better?

System	AUROC	c@1	F_1	$F_{0.5u}$	BRIER	MEAN
BASELINE-CNGDIST22	0.546	0.496	0.669	0.542	0.749	**0.600**
NAJAFI22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
GALICIA22	0.512	0.499	0.628	0.544	0.741	0.585
JINLI22	0.577	0.557	0.581	0.563	0.589	0.573
BASELINE-COMPRESSOR22	0.541	0.493	0.570	0.478	**0.750**	0.566
LEI22	0.539	0.539	0.399	0.488	0.539	0.501
YIHUIYE22	0.542	0.526	0.398	0.461	0.565	0.499
HUANG22	0.519	0.519	0.196	0.328	0.519	0.416
CRESPOSANCHEZ22	0.500	0.500	0	0	0.748	0.350

Do Previous Systems Perform Better?

Short answer: No.

First place of last year trails behind last place of this year.

System	AUROC	c@1	F_1	$F_{0.5u}$	BRIER	MEAN
BASELINE-CNGDIST22	0.546	0.496	0.669	0.542	0.749	**0.600**
NAJAFI22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
GALICIA22	0.512	0.499	0.628	0.544	0.741	0.585
JINLI22	0.577	0.557	0.581	0.563	0.589	0.573
BASELINE-COMPRESSOR22	0.541	0.493	0.570	0.478	**0.750**	0.566
LEI22	0.539	0.539	0.399	0.488	0.539	0.501
YIHUIYE22	0.542	0.526	0.398	0.461	0.565	0.499
HUANG22	0.519	0.519	0.196	0.328	0.519	0.416
EMBARCADERORUIZ21	0.538	0.502	0.063	0.116	0.581	0.360
CRESPOSANCHEZ22	0.500	0.500	0	0	0.748	0.350
BOENNINGHOFF21*	0.513	0.501	0.002	0.005	0.531	0.310
WEERASINGHE21	0.488	0.500	0.011	0.027	0.506	0.306

* Previous winner

Conclusion

- Authorship verification is *not* a solved task.
- Bigger models do not necessarily lead to better results.
- Cross-discourse-type verification may be particularly challenging.
- Systems are still failing to find a generalization of “style”.
- Previously successful systems to not transfer well to new task variants.

Conclusion

- Authorship verification is *not* a solved task.
- Bigger models do not necessarily lead to better results.
- Cross-discourse-type verification may be particularly challenging.
- Systems are still failing to find a generalization of “style”.
- Previously successful systems to not transfer well to new task variants.

Thanks