Cross-domain Authorship Attribution

Overview of the Author Identification Task at PAN-2018
PAN@CLEF2018, Avignon, 11 September 2018

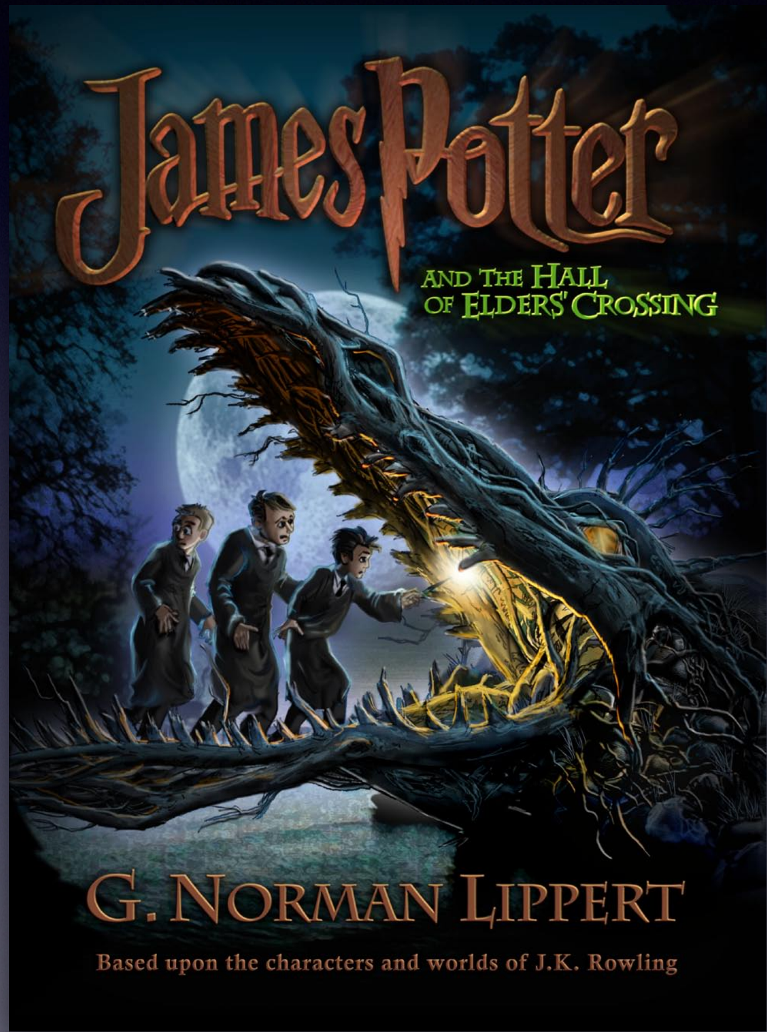
Mike Kestemont, Efstathios Stamatatos, Walter Daelemans, Benno Stein, Martin Potthast

Authorship attribution

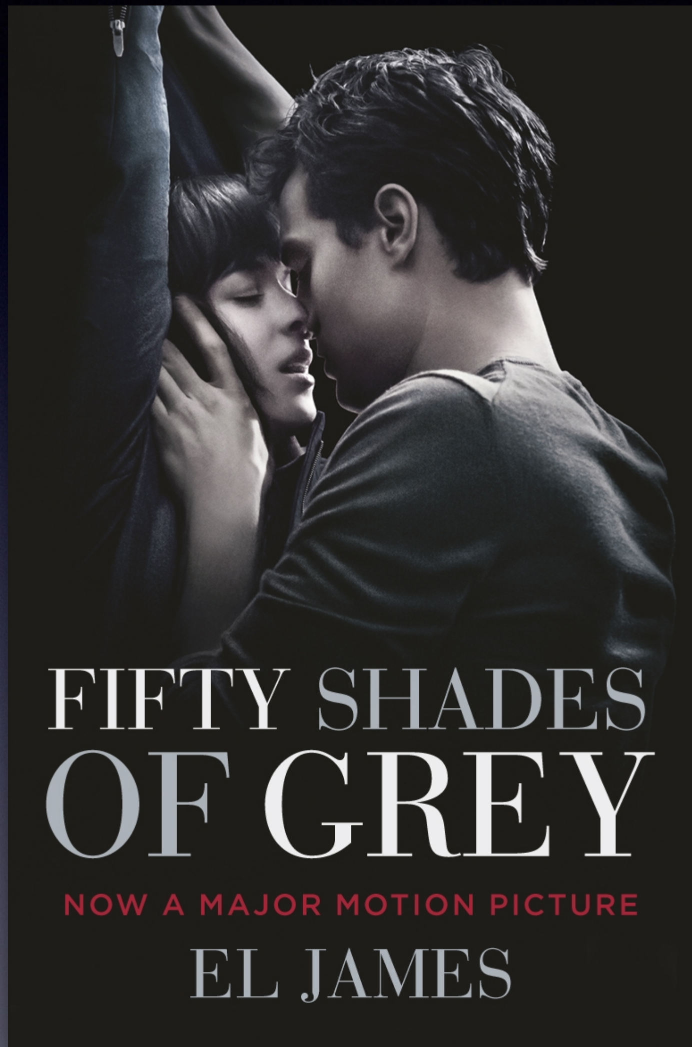
- Closed-set: assign anonymous text to one author from set of candidate authors (classification problem)
- Importance *and* difficulty of benchmarking: need for
    - Large but varied corpora
    - Accessible data (free of rights)
    - Control over topic and genre (domain)
    - Multilingual, yet comparable datasets

What is fan fiction?

- Fiction produced by non-professional authors
- that explicitly builds on previously published fiction (characters, themes, settings, etc.)



Canon



Fandom

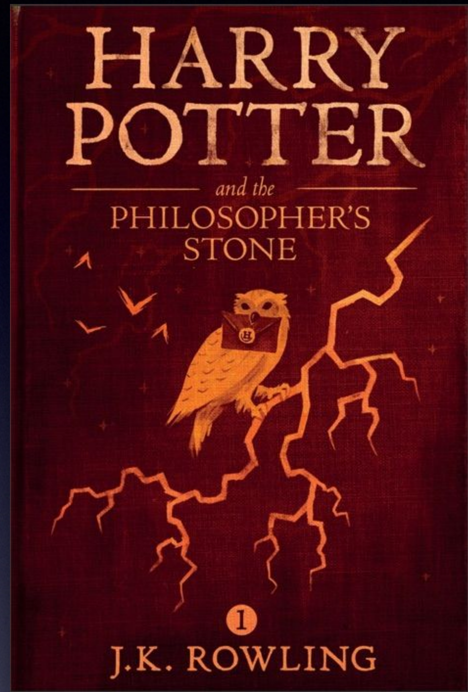
Attractive?

Characteristic	Advantage
Online, open platforms	Digitally accessible
Unmediated	No editorial interference
Explicit about canon	Rich metadata
Global phenomenon	Language-independent

Balanced cross-domain design

Table 1. The cross-domain authorship attribution corpus.

	Language	Problems	Authors (subsets size)	Texts per author: training	Texts per author: test	Text length (avg. words)
Development	English	2	5,20	7	1-22	795
	French	2	5,20	7	1-10	796
	Italian	2	5,20	7	1-17	795
	Polish	2	5,20	7	1-21	800
	Spanish	2	5,20	7	1-21	832
Evaluation	English	4	5,10,15,20	7	1-17	820
	French	4	5,10,15,20	7	1-20	782
	Italian	4	5,10,15,20	7	1-29	802
	Polish	4	5,10,15,20	7	1-42	802
	Spanish	4	5,10,15,20	7	1-24	829



All test texts, across 5 languages (!), from target fandom (Harry Potter) not represented in the training data. Each author: 7+ training texts

Submissions

Compared to a SVM char 3gram baseline

Table 4. Authorship attribution evaluation results (macro F1) per language.

Submission	Overall	English	French	Italian	Polish	Spanish
Custódio and Paraboni	**0.685**	0.744	**0.668**	0.676	0.482	**0.856**
Murauer et al.	0.643	**0.762**	0.607	0.663	0.450	0.734
Halvani and Graner	0.629	0.679	0.536	**0.752**	0.426	0.751
Mosavat	0.613	0.685	0.615	0.601	0.435	0.731
Yigal et al.	0.598	0.672	0.609	0.642	0.431	0.636
Martín dCR et al.	0.588	0.601	0.510	0.571	**0.556**	0.705
PAN18-BASELINE	0.584	0.697	0.585	0.605	0.419	0.615
Miller et al.	0.582	0.573	0.611	0.670	0.421	0.637
Schaetti	0.387	0.538	0.332	0.337	0.388	0.343
Gagala	0.267	0.376	0.215	0.248	0.216	0.280
López-Anguita et al.	0.139	0.190	0.065	0.161	0.128	0.153
Tabealhoje	0.028	0.037	0.048	0.014	0.024	0.018

Effect of number of authors

Table 5. Performance (macro F1) of the cross-domain authorship attribution submissions per candidate set size.

Submission	20 Authors	15 Authors	10 Authors	5 Authors
Custódio and Paraboni	**0.648**	**0.676**	**0.739**	**0.677**
Murauer et al.	0.609	0.642	0.680	0.642
Halvani and Graner	0.609	0.605	0.665	0.636
Mosavat	0.569	0.575	0.653	0.656
Yigal et al.	0.570	0.566	0.649	0.607
Martín dCR et al.	0.556	0.556	0.660	0.582
PAN18-BASELINE	0.546	0.532	0.595	0.663
Miller et al.	0.556	0.550	0.671	0.552
Schaetti	0.282	0.352	0.378	0.538
Gagala	0.204	0.240	0.285	0.339
López-Anguita et al.	0.064	0.065	0.195	0.233
Tabealhoje	0.012	0.015	0.030	0.056

Significance

Table 6. Significance of pairwise differences in output between submissions, across all problems.

	Murauer et al.	Halvani and Graner	Mosavat	Yigal et al.	Martín dCR et al.	Miller et al.	PAN18-BASELINE	Schaetti	Gagala	López-Anguita et al.	Tabealhoje
Custódio and Paraboni	=	***	***	***	***	***	***	***	***	***	***
Murauer et al.		**	***	**	***	***	***	***	***	***	***
Halvani and Graner			=	=	=	=	=	***	***	***	***
Mosavat				=	=	=	=	***	***	***	***
Yigal et al.					=	=	=	***	***	***	***
Martín dCR et al.						=	=	***	***	***	***
Miller et al.							=	***	***	***	***
PAN18-BASELINE								***	***	***	***
Schaetti									***	***	***
Gagala										***	***
López-Anguita et al.											***

Model criticism

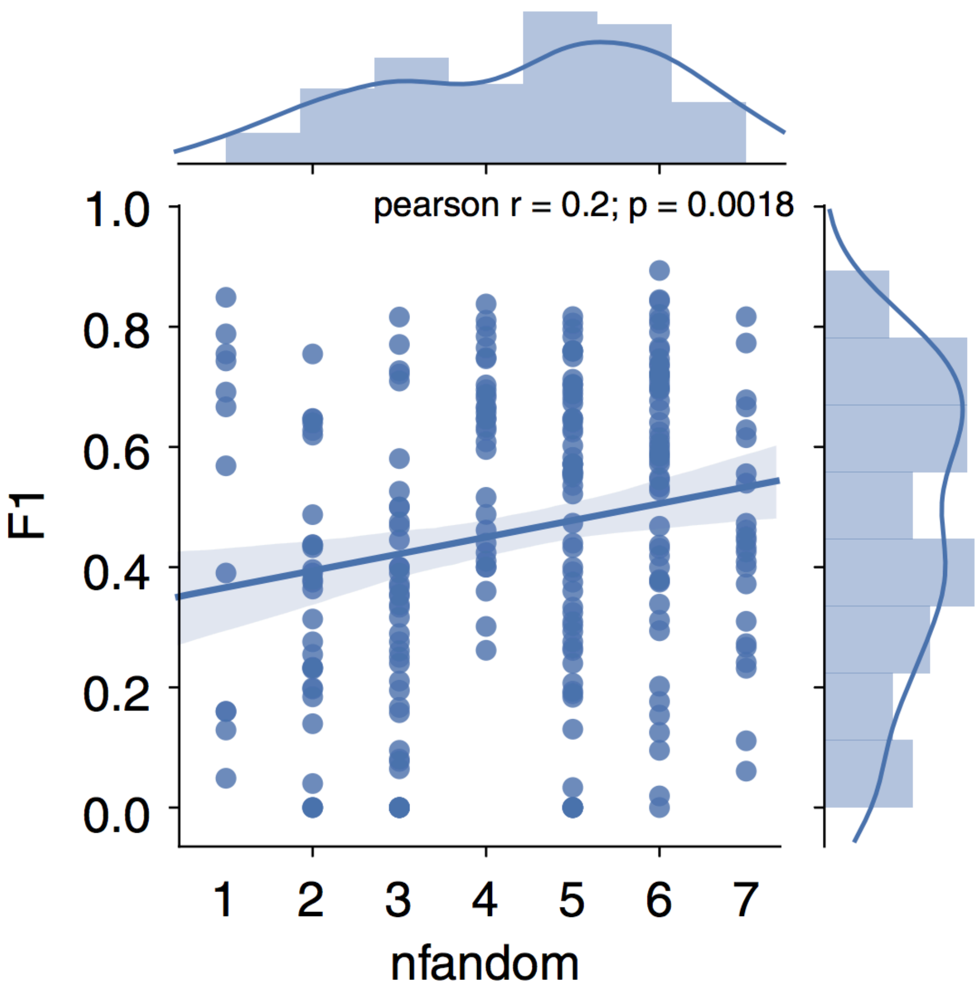
Dominance of ngrams (TF-IDF), instance-based, SVMs

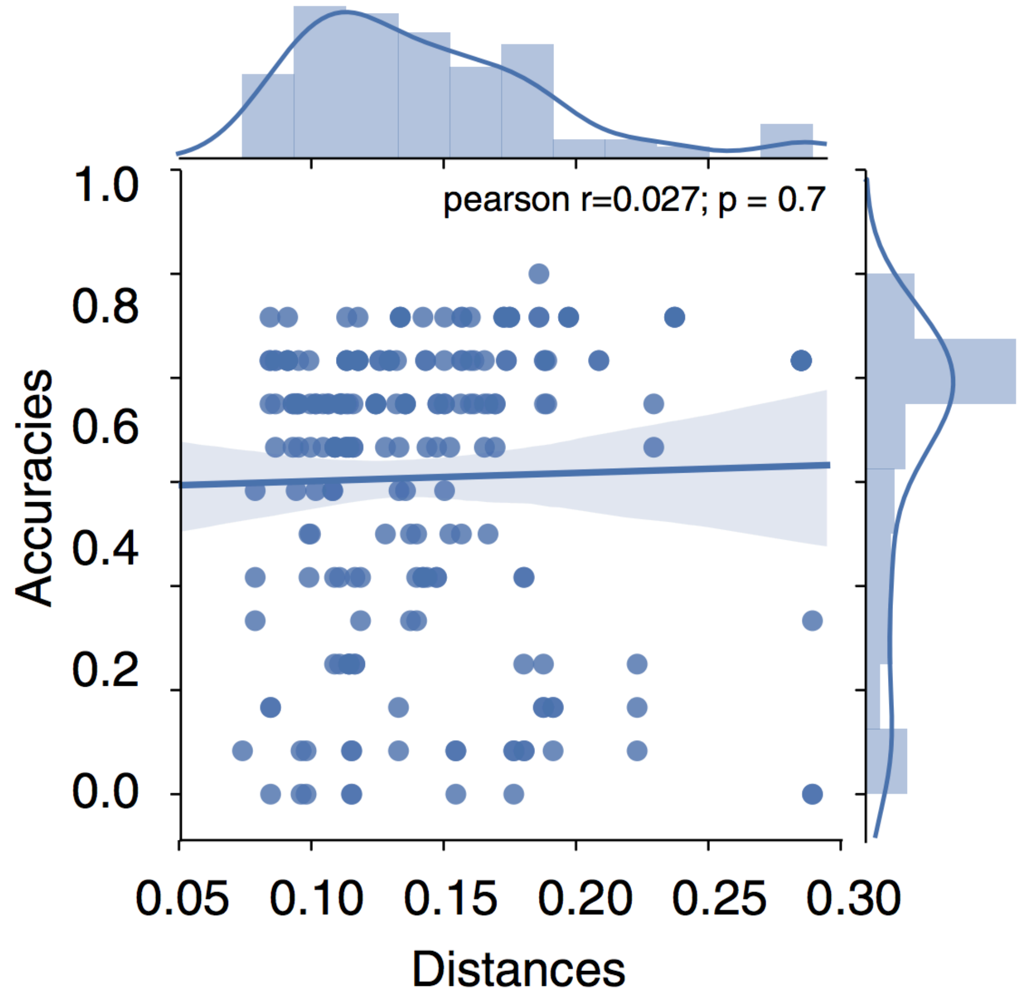
Submission		Features	Weighting / Normalization	Paradigm	Classifier	Parameter settings
Team	Reference					
Custódio and Paraboni	[6]	char & word n-grams	TF-IDF	i-b	ensemble	global
Gagala	[8]	various n-grams	none	i-b	NN	global
Halvani and Graner	[14]	compression	none	p-b	similarity	global
López-Anguita et al.	[25]	complexity	L2-norm.	i-b	SVM	l-s
Martín dCR et al.	[4]	various n-grams	log-entropy	i-b	SVM	l-s
Miller et al.	[13]	various n-grams & stylistic	TF-IDF & TF	i-b	SVM	global
Murauer et al.	[28]	char n-grams	TF-IDF	i-b	SVM	local
PAN18-BASELINE		char n-grams	TF	i-b	SVM	global
Schaetti	[41]	tokens	embeddings	i-b	ESN	local
Yigal et al.	[13]	various n-grams & stylistic	TF-IDF & TF	i-b	SVM	global

Submissions without a working notes paper: Saeed Mosavat; Hadi Tabealhojeh

Post-hoc analyses

More varied training data helps (cf. Sapkota 2014) — influence of original author is not a major factor





Observations

- Fanfiction validated: feasible, but not easy, so room for progress
- (Stylistic) influence of canon author not an issue? Focus on (semantic) domain
- Some stagnation in the field, both in feature extraction and classification
- (Where is deep learning? Cf. Bagnall@PAN2016)

Stay tuned

- Next year at PAN 2019 (Lugano)
- Focus on open-set attribution in fan fiction
  - No longer a single target fandom: more “adversarial” set up
  - Less restricted design: larger, more complex problems to push innovation

References

- Douglas Bagnall. Authorship Clustering Using Multi-headed Recurrent Neural Networks—Notebook for PAN at CLEF 2016.
- Kestemont at al. Overview of the Author Identification Task at PAN-2018 Cross-domain Authorship Attribution and Style Change Detection. PAN 2018.
- Hellekson, K., Busse, K. (eds.): The Fan Fiction Studies Reader. University of Iowa Press (2014).
- Sapkota, U. et al. Not all character n-grams are created equal: A study in authorship attribution. COLING 2014.
- Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60, 538–556 (2009)