Who Wrote the Web?

Revisiting Influential Author Identification Research Applicable to Information Retrieval

Bauhaus-Universität Weimar	Martin Potthast, Benno Stein, Matthias Hagen
University of the Aegean	Efstathios Stamatatos
Technische Universität München	Sarah Braun, Sebastian Wilhelm
Technical University of Berlin	Tolga Buz, Maike Elisa Müller
RWTH Aachen University	Fabian Duffhauss
Heidelberg University	Florian Friedrich, Lucas Rettenmeier
University of Konstanz	Jörg Marvin Gülzow
Free University of Berlin	Jakob Köhler
Chemnitz University of Technology	Winfried Lötzsch
Karlsruhe University of Applied Sciences	Fabian Müller
University of Bonn	Robert Paßmann, Bernhard Reinke
University of Michigan	Thomas Rometsch
Hamburg University of Technology	Timo Sommer
University of Bamberg	Michael Träger

supported by

Studienstiftung
des deutschen Volkes

[www.webis.de]

Who Wrote the Web?

Who Wrote the Web?

- Humans possess individual writing style
- Since 1890, style has been quantified to identify authors
- With machine learning, $>$500 words per author seem to suffice
- Every text on the web written by a human encodes its author’s style
- Writing style allows to infer author traits, such as gender, age, etc.

➜ Style IR would be ripe for the taking, if ...

Who Wrote the Web?

- ❑ Humans possess individual writing style
- ❑ Since 1890, style has been quantified to identify authors
- ❑ With machine learning, >500 words per author seem to suffice
- ❑ Every text on the web written by a human encodes its author's style
- ❑ Writing style allows to infer author traits, such as gender, age, etc.

➜ Style IR would be ripe for the taking, if ...

- ❑ ... authorship technology scaled to the web

Author Identification [details]

- ❑ NLP research field
- ❑ Hundreds of papers in the past two decades
- ❑ Little intersection with IR to date, but accessible with IR background

- ❑ Estimated 50-60 reasonably different approaches
- ❑ No standardized evaluation frameworks until PAN 2011

➜ How to "quickly" get to grips with a research field?

How to “quickly” get to grips with a research field?

How to “quickly” get to grips with a research field?

Heuristics (ordered from simple to laborious)

- Citations: highly cited papers might have some merits
- Topics: identify papers specifically on your problem of interest
- Surveys: read literature reviews, systematic reviews, and meta studies
- Authority: follow leading research (groups)
- Experts: seek advise or buy consulting
- Benchmarks: identify best-performing approaches
- Usage: “*commonly used baselines are what actually works*” [citation needed]
- Libraries: hunt down and use published code
- DIY: reimplement and evaluate selected approaches

How to “quickly” get to grips with a research field?

Heuristics (ordered from simple to laborious)

- ❑ Citations: highly cited papers might have some merits
- ❑ Topics: identify papers specifically on your problem of interest
- ❑ Surveys: read literature reviews, systematic reviews, and meta studies
- ❑ Authority: follow leading research (groups)
- ❑ Experts: seek advise or buy consulting
- ❑ Benchmarks: identify best-performing approaches
- ❑ Usage: “*commonly used baselines are what actually works*” [citation needed]
- ❑ Libraries: hunt down and use published code
- ❑ DIY: reimplement and evaluate selected approaches

➜ Reimplementing approaches necessarily includes reproducing research

DIY ~ Do it Yourself

How to “quickly” get to grips with a research field?

Heuristics (ordered from simple to laborious)

- Citations: highly cited papers might have some merits
- Topics: identify papers specifically on your problem of interest
- Surveys: read literature reviews, systematic reviews, and meta studies
- Authority: follow leading research (groups)
- Experts: seek advise or buy consulting
- Benchmarks: identify best-performing approaches
- Usage: “*commonly used baselines are what actually works*” [citation needed]
- Libraries: hunt down and use published code
- DIY: reimplement and evaluate selected approaches

→ Reimplementing approaches necessarily includes reproducing research

Scaling DIY ... as in Don’t DIY

- Hire engineers
- Recruit students
- Crowdsourcing

DIY ~ Do it Yourself

Contributions

Contributions

to author identification and information retrieval alike

- ❑ Open source reimplementations of 15 of the most influential approaches
- ❑ First comparative evaluation of these approaches on standardized datasets
- ❑ Lowering the bar for newcomers to get started

Contributions

to author identification and information retrieval alike

- Open source reimplementations of 15 of the most influential approaches
- First comparative evaluation of these approaches on standardized datasets
- Lowering the bar for newcomers to get started

to computer science reproducibility

- first-time large-scale reproduction in human language technologies
- proof-of-concept on employing undergrad students in reproducibility studies

Contributions

to author identification and information retrieval alike

- Open source reimplementations of 15 of the most influential approaches
- First comparative evaluation of these approaches on standardized datasets
- Lowering the bar for newcomers to get started

to computer science reproducibility

- first-time large-scale reproduction in human language technologies
- proof-of-concept on employing undergrad students in reproducibility studies

A Reproducibility Study in 7 Steps

1. Paper selection
2. Student recruitment
3. Paper assignment and instruction
4. Implementation and experimentation
5. Auditing
6. Publication
7. Post-publication rebuttal

A Reproducibility Study in 7 Steps

1. Paper selection [details]
   - ❑ From the top: select “influential” papers
   - ❑ Coverage: represent different paradigms
   - ❑ 30 papers selected, 15 reimplemented due to limited human resources
2. Student recruitment
3. Paper assignment and instruction
4. Implementation and experimentation
5. Auditing
6. Publication
7. Post-publication rebuttal

A Reproducibility Study in 7 Steps

1. Paper selection
2. Student recruitment
   - ❑ Motivation via, e.g., a graded course, extracurricular activity, payment
   - ❑ Summer academy of the German National Academic Foundation
   - ❑ 16 students: computer science (5), engineering (4), physics (3), maths (4)
3. Paper assignment and instruction
4. Implementation and experimentation
5. Auditing
6. Publication
7. Post-publication rebuttal

A Reproducibility Study in 7 Steps

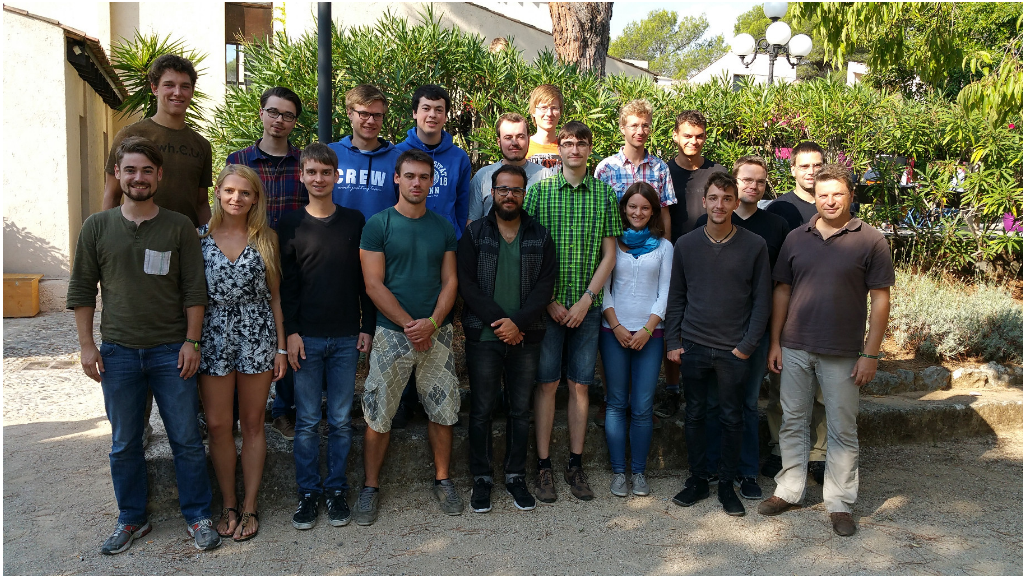
1. Paper selection
2. Student recruitment
3. Paper assignment and instruction
   - ❑ Interviews to learn students’ background and skills
   - ❑ Matching students with papers based on skills and paper complexity
   - ❑ Instructions
     - – Study the main algorithmic contribution for author identification
     - – Implement the approach in a programming language of your choice
     - – Replicate at least one of the experiments described
4. Implementation and experimentation
5. Auditing
6. Publication
7. Post-publication rebuttal

A Reproducibility Study in 7 Steps

1. Paper selection
2. Student recruitment
3. Paper assignment and instruction
4. Implementation and experimentation
   - ❑ Students worked on their own
   - ❑ Questions were answered (mostly pertaining to implementation basics)
   - ❑ Problem: procrastination
5. Auditing
6. Publication
7. Post-publication rebuttal

A Reproducibility Study in 7 Steps

1. Paper selection
2. Student recruitment
3. Paper assignment and instruction
4. Implementation and experimentation
5. Auditing
   - ❑ 2-week workshop in La Colle-sur-Loup, France
   - ❑ Students gave talks, demos, and were quizzed
   - ❑ Hackathon to finalize and fix implementations based on feedback
6. Publication
7. Post-publication rebuttal



A Reproducibility Study in 7 Steps

1. Paper selection
2. Student recruitment
3. Paper assignment and instruction
4. Implementation and experimentation
5. Auditing
6. Publication
   - ❑ Code published at GitHub: www.github.com/pan-webis-de
   - ❑ Permissible licensing
   - ❑ Report published here, at ECIR :-)
7. Post-publication rebuttal

A Reproducibility Study in 7 Steps

1. Paper selection
2. Student recruitment
3. Paper assignment and instruction
4. Implementation and experimentation
5. Auditing
6. Publication
7. Post-publication rebuttal
   - ❑ Authors were unaware of our study to avoid outside influence
   - ❑ After publication, authors were notified of the results
   - ❑ Invitation to feedback and rebuttal

John Burrows: *I congratulate you on a project that is so much to our communal advantage and I am delighted that Delta has a place there.*

Moshe Koppel: *Awesome project! We're very flattered to have been included.*

Shlomo Argamon: *This is a wonderful project - we are honored to be included!*

David Harper: *My congratulations of this useful work. [...] I am well-pleased with the result of the study.*

William Teahan: *Congratulations on the excellent work!*

Hugo Jair Escalante: *I feel honored for the inclusion of our paper in our study. I think this type of studies will pave the way for a radical change in reproducibility of research.*

Reproducibility Report

Reproducibility Report

Criterion	Publication														
	[4]	[5]	[7]	[10]	[12]	[22]	[23]	[24]	[25]	[29]	[32]	[33]	[34]	[35]	[41]
(1) *Approach clarity*															
Code available	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Description sound	●	●	◐	◐	◐	●	●	●	●	●	●	●	●	●	●
Details sufficient	●	●	◐	◐	◐	●	●	●	●	◐	◐	◐	●	●	●
Paper self-contained	◐	○	●	◐	●	●	◐	●	●	●	○	◐	●	●	●
Preprocessing	○	●	●	●	–	–	–	◐	–	○	○	●	●	–	–
Parameter settings	–	◐	●	◐	●	●	–	●	●	●	●	○	●	●	○
Library versions	–	–	–	○	◐	–	–	◐	–	–	○	○	○	–	–
Reimplementation															
Language	`Py`	`Py`	`Py`	`C++`	`J`	`Py`	`C++`	`Py`	`Py`	`C#`	`C++`	`J`	`Py`	`Py`	`Py`
(2) *Experiment clarity / soundness*															
Setup clear	◐	●	◐	◐	●	●	●	●	●	●	●	◐	●	●	●
Exhaustiveness	◐	○	◐	○	◐	◐	○	◐	●	●	●	◐	●	●	○
Compared to others	○	○	○	●	●	◐	●	●	○	●	●	●	○	◐	●
Result replicated	◐	◐	○	◐	◐	◐	●	○	◐	●	○	◐	●	●	●
(3) *Dataset reconstructability / availability*															
Text length	L	L	M	S	M	M	M	M	L	M	L	S	M	M	M
Candidate set	M	M	M	S	M	M	L	L	M	M	S	L	M	L	M
Origin given	●	●	◐	○	●	◐	●	●	●	●	●	●	●	●	○
Corpora available	○	○	○	○	●	◐	◐	○	○	●	●	○	●	◐	●
(4) *Overall assessment*															
Replicability	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Reproducibility	●	◐	●	◐	◐	●	●	●	●	●	○	●	●	●	●
Simplifiability	●	●	●	○	○	○	○	○	○	○	○	○	○	○	●
Improvability	●	●	●	○	○	○	●	○	○	○	○	●	○	○	○

Reproducibility Report

Criterion	Publication														
	[4]	[5]	[7]	[10]	[12]	[22]	[23]	[24]	[25]	[29]	[32]	[33]	[34]	[35]	[41]
(1) *Approach clarity*															
Code available	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Description sound	●	●	◐	◐	◐	●	●	●	●	●	●	●	●	●	●
Details sufficient	●	●	◐	◐	◐	●	●	●	●	◐	◐	◐	●	●	●
Paper self-contained	◐	○	●	◐	●	●	◐	●	●	●	○	◐	●	●	●
Preprocessing	○	●	●	●	–	–	–	◐	–	○	○	●	●	–	–
Parameter settings	–	◐	●	◐	●	●	–	●	●	●	●	○	●	●	○
Library versions	–	–	–	○	◐	–	–	◐	–	–	○	○	○	–	–
Reimplementation															
Language	Py	Py	Py	C++	J	Py	C++	Py	Py	C#	C++	J	Py	Py	Py

→ lack of formal, mathematical rigor; vague descriptions

→ references to other papers for details, and missing references

→ important processing steps, parameters, and libraries employed missing

● Sufficient reproducibility or information

◐ Partial reproducibility or information

○ Lack of reproducibility or information

– Criterion does not apply

Reproducibility Report

Criterion	Publication														
	[4]	[5]	[7]	[10]	[12]	[22]	[23]	[24]	[25]	[29]	[32]	[33]	[34]	[35]	[41]
(2) *Experiment clarity / soundness*															
Setup clear	◐	●	◐	◐	●	●	●	●	●	●	●	◐	●	●	●
Exhaustiveness	◐	○	◐	○	◐	◐	○	◐	●	●	●	◐	●	●	○
Compared to others	○	○	○	●	●	◐	●	●	○	●	●	●	○	◐	●
Result replicated	◐	◐	○	◐	◐	◐	●	○	◐	●	○	◐	●	●	●

→ unclear training-test-split

→ simple baselines, small-scale experiments, or no comparative evaluation

→ given missing details on approach and setup, replication not always possible

● Sufficient reproducibility or information

◐ Partial reproducibility or information

○ Lack of reproducibility or information

– Criterion does not apply

Reproducibility Report

Criterion	Publication														
	[4]	[5]	[7]	[10]	[12]	[22]	[23]	[24]	[25]	[29]	[32]	[33]	[34]	[35]	[41]
(3) *Dataset reconstructability / availability*															
Text length	L	L	M	S	M	M	M	M	L	M	L	S	M	M	M
Candidate set	M	M	M	S	M	M	L	L	M	M	S	L	M	L	M
Origin given	●	●	◐	○	●	◐	●	●	●	●	●	●	●	●	○
Corpora available	○	○	○	○	●	◐	◐	○	○	●	●	○	●	◐	●

➜ long texts and small candidate sets easier than short/large ones

➜ naming the source of data important for reconstruction

➜ few authors share their data up front

● Sufficient reproducibility or information

○ Lack of reproducibility or information

◐ Partial reproducibility or information

– Criterion does not apply

Reproducibility Report

Criterion	Publication														
	[4]	[5]	[7]	[10]	[12]	[22]	[23]	[24]	[25]	[29]	[32]	[33]	[34]	[35]	[41]
(4) *Overall assessment*															
Replicability	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Reproducibility	●	◐	●	◐	◐	●	●	●	●	●	○	●	●	●	●
Simplifiability	●	●	●	○	○	○	○	○	○	○	○	○	○	○	●
Improvability	●	●	●	○	○	○	●	○	○	○	○	●	○	○	○

→ none of the papers replicable

→ all except one reproducible

→ some approaches simplifiable, and some even improvable (e.g., runtime)

●	Sufficient reproducibility or information	○	Lack of reproducibility or information
◐	Partial reproducibility or information	–	Criterion does not apply

Evaluation

Evaluation

Corpus								Publication							
	[4]	[5]	[7]	[10]	[12]	[22]	[23]	[24]	[25]	[29]	[32]	[33]	[34]	[35]	[41]	BR
C10	9.0	72.8	59.8	50.2	75.4	71.0	77.2	22.4	72.0	76.6	–	29.8	73.8	70.8	76.6	**86.4**
PAN11	0.1	29.6	5.4	13.5	43.1	1.8	32.8	n/a	20.2	46.2	–	n/a	7.6	34.5	65.0	**65.8**
PAN12	85.7	71.4	92.9	28.6	28.6	71.4	n/a	78.6	78.6	57.1	–	n/a	7.1	85.7	64.3	**92.9**

- C10. English newswire stories from the CCAT topic of the Reuters Corpus Volume 1 for 10 candidate authors with 100 texts each
- PAN11. English emails from the Enron corpus for 72 candidate authors with imbalanced distribution of texts
- PAN12. English novels for 14 candidate authors with three texts each

- Performance scores indicate classification accuracy
- BR = best result from the literature; n/a cases due to runtime complexity

Evaluation

Corpus	Publication															
	[4]	[5]	[7]	[10]	[12]	[22]	[23]	[24]	[25]	[29]	[32]	[33]	[34]	[35]	[41]	BR
C10	9.0	72.8	59.8	50.2	75.4	71.0	77.2	22.4	72.0	76.6	–	29.8	73.8	70.8	76.6	**86.4**
PAN11	0.1	29.6	5.4	13.5	43.1	1.8	32.8	n/a	20.2	46.2	–	n/a	7.6	34.5	65.0	**65.8**
PAN12	85.7	71.4	92.9	28.6	28.6	71.4	n/a	78.6	78.6	57.1	–	n/a	7.1	85.7	64.3	**92.9**

- C10. English newswire stories from the CCAT topic of the Reuters Corpus Volume 1 for 10 candidate authors with 100 texts each
- PAN11. English emails from the Enron corpus for 72 candidate authors with imbalanced distribution of texts
- PAN12. English novels for 14 candidate authors with three texts each

→ some approaches compete with the state of the art

→ some approaches have stable performance across two corpora

→ one approach has stable performance across all corpora

- Performance scores indicate classification accuracy
- BR = best result from the literature; n/a cases due to runtime complexity

The Fallacy of Reproducibility Perfection

The Fallacy of Reproducibility Perfection

Why computer science papers can’t be written to be perfectly reproducible?

- ❑ Papers are written sometimes after, sometimes before experiments
- ❑ Papers are not a documentation of “How to (re)construct this software?”
- ❑ Papers are optimized for (reviewer) readability
- ❑ Paper authors deliberately abstract over a subject matter
- ❑ Paper authors straighten story and reasoning with hindsight
- ❑ Paper authors did not necessarily write the software
- ❑ Paper authors deliberately omit implementation details

➜ A computer science paper is not a software documentation

The Fallacy of Reproducibility Perfection

Why computer science papers can’t be written to be perfectly reproducible?

- ❑ Papers are written sometimes after, sometimes before experiments
- ❑ Papers are not a documentation of “How to (re)construct this software?”
- ❑ Papers are optimized for (reviewer) readability
- ❑ Paper authors deliberately abstract over a subject matter
- ❑ Paper authors straighten story and reasoning with hindsight
- ❑ Paper authors did not necessarily write the software
- ❑ Paper authors deliberately omit implementation details

➜ A computer science paper is not a software documentation

Why failure to reproduce doesn’t invalidate a paper?

- ❑ Because the paper might have misrepresented its approach
- ❑ Because we cannot know what the original software did
- ❑ Because we might have made an honest mistake

➜ A reproducibility study is not a final verdict about a paper’s merits

Conclusion and Future Work

Summary

- Reimplementation of influential author identification approaches
- Main goal: publication of working code to lower the bar of entry
- Laying the groundwork to answering the question “Who wrote the web”?

Conclusion and Future Work

Summary

- Reimplementation of influential author identification approaches
- Main goal: publication of working code to lower the bar of entry
- Laying the groundwork to answering the question “Who wrote the web”?

Take-away messages

- Reimplementation can be outsourced to undergrad students
- Some “old” author identification approaches are still competitive
- Sharing code is essential to improve computer science replicability
- Reimplementing other people’s software entitles you to citations

Conclusion and Future Work

Summary

- Reimplementation of influential author identification approaches
- Main goal: publication of working code to lower the bar of entry
- Laying the groundwork to answering the question “Who wrote the web”?

Take-away messages

- Reimplementation can be outsourced to undergrad students
- Some “old” author identification approaches are still competitive
- Sharing code is essential to improve computer science replicability
- Reimplementing other people’s software entitles you to citations

Open questions

- How to scale “student reproducibility studies”?
- Can we reproduce ALL (algorithmic) computer science papers?

Conclusion and Future Work

Summary

- Reimplementation of influential author identification approaches
- Main goal: publication of working code to lower the bar of entry
- Laying the groundwork to answering the question “Who wrote the web”?

Take-away messages

- Reimplementation can be outsourced to undergrad students
- Some “old” author identification approaches are still competitive
- Sharing code is essential to improve computer science replicability
- Reimplementing other people’s software entitles you to citations

Open questions

- How to scale “student reproducibility studies”?
- Can we reproduce ALL (algorithmic) computer science papers?

Code available at GitHub: www.github.com/pan-webis-de

Conclusion and Future Work

Summary

- Reimplementation of influential author identification approaches
- Main goal: publication of working code to lower the bar of entry
- Laying the groundwork to answering the question “Who wrote the web”?

Take-away messages

- Reimplementation can be outsourced to undergrad students
- Some “old” author identification approaches are still competitive
- Sharing code is essential to improve computer science replicability
- Reimplementing other people’s software entitles you to citations

Open questions

- How to scale “student reproducibility studies”?
- Can we reproduce ALL (algorithmic) computer science papers?

Code available at GitHub: www.github.com/pan-webis-de

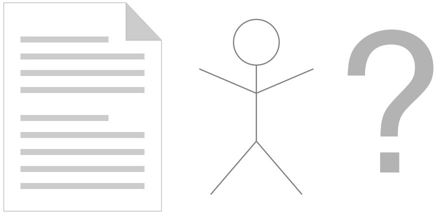
Thank you for your attention!

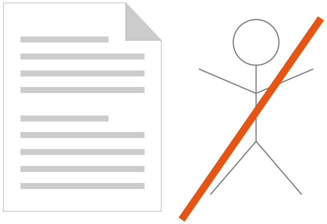
Author Identification

Author Identification

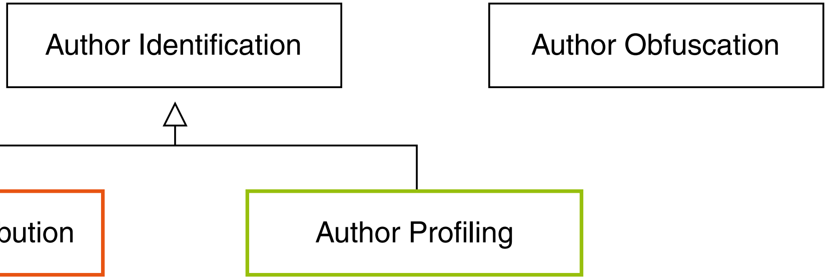
Author Identification

Author Obfuscation





Author Identification

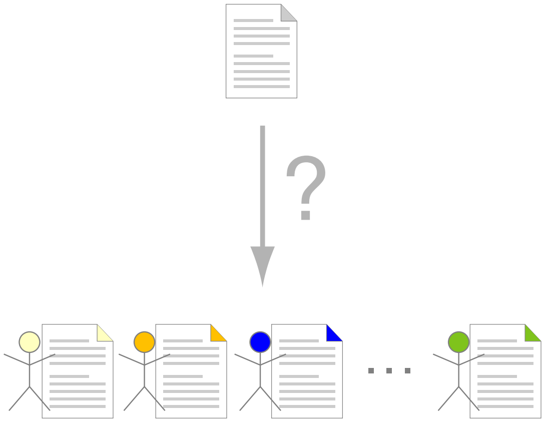


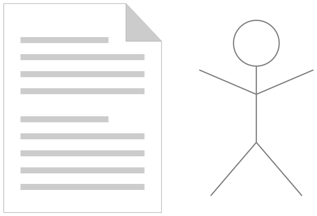
Author Identification

Author Obfuscation

Authorship Attribution

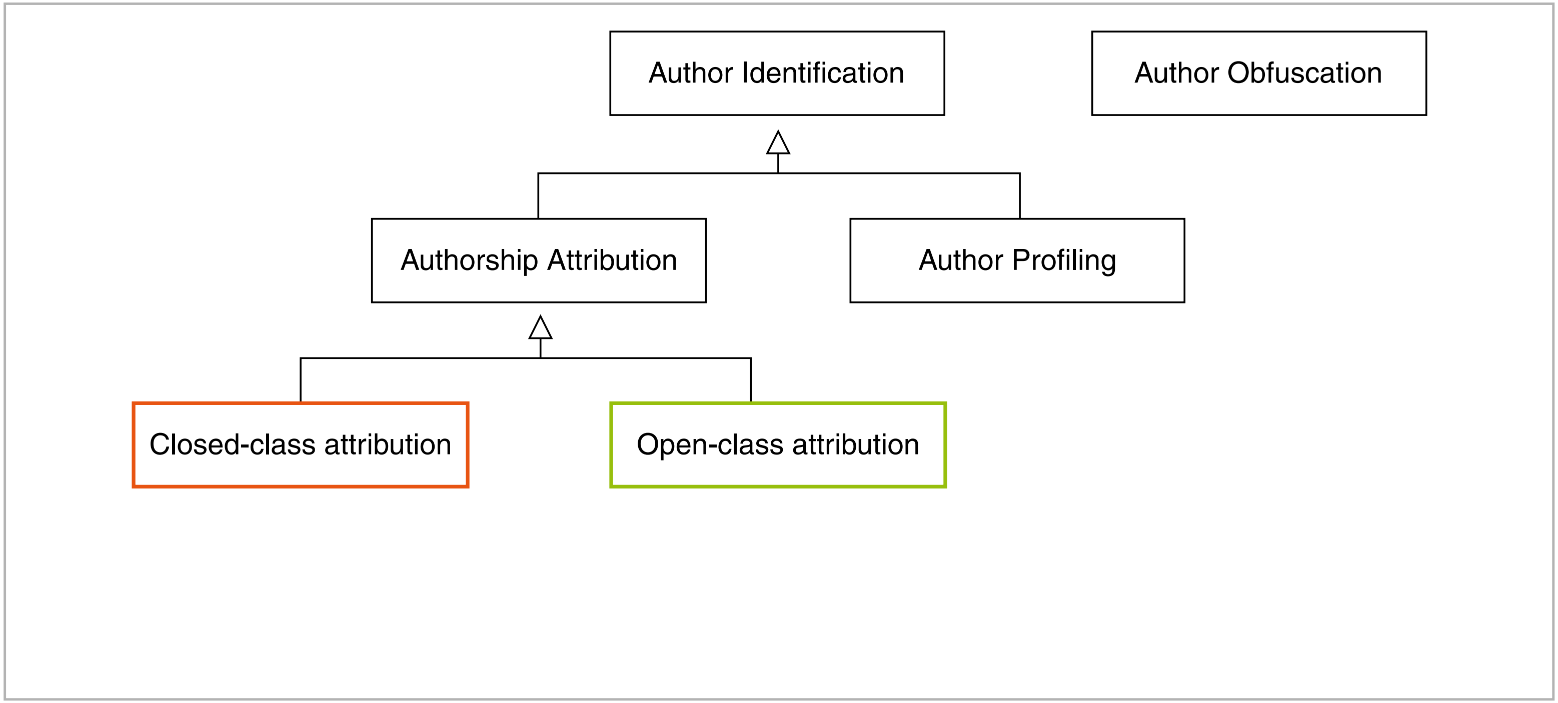
Author Profiling





Gender?
Age?
Mother tongue?
Education?
Personality?
⋮

Author Identification



Author Identification

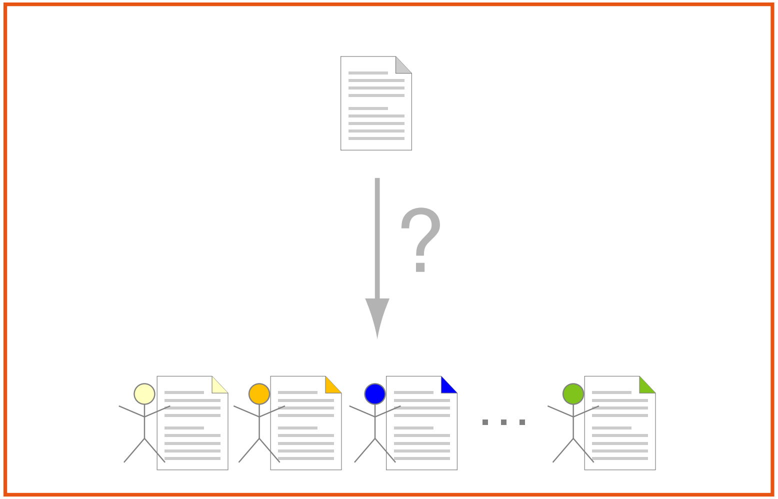
Author Obfuscation

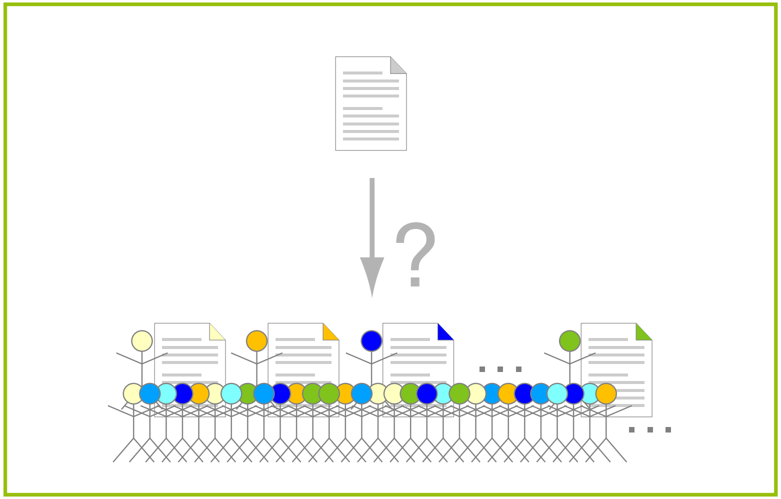
Authorship Attribution

Author Profiling

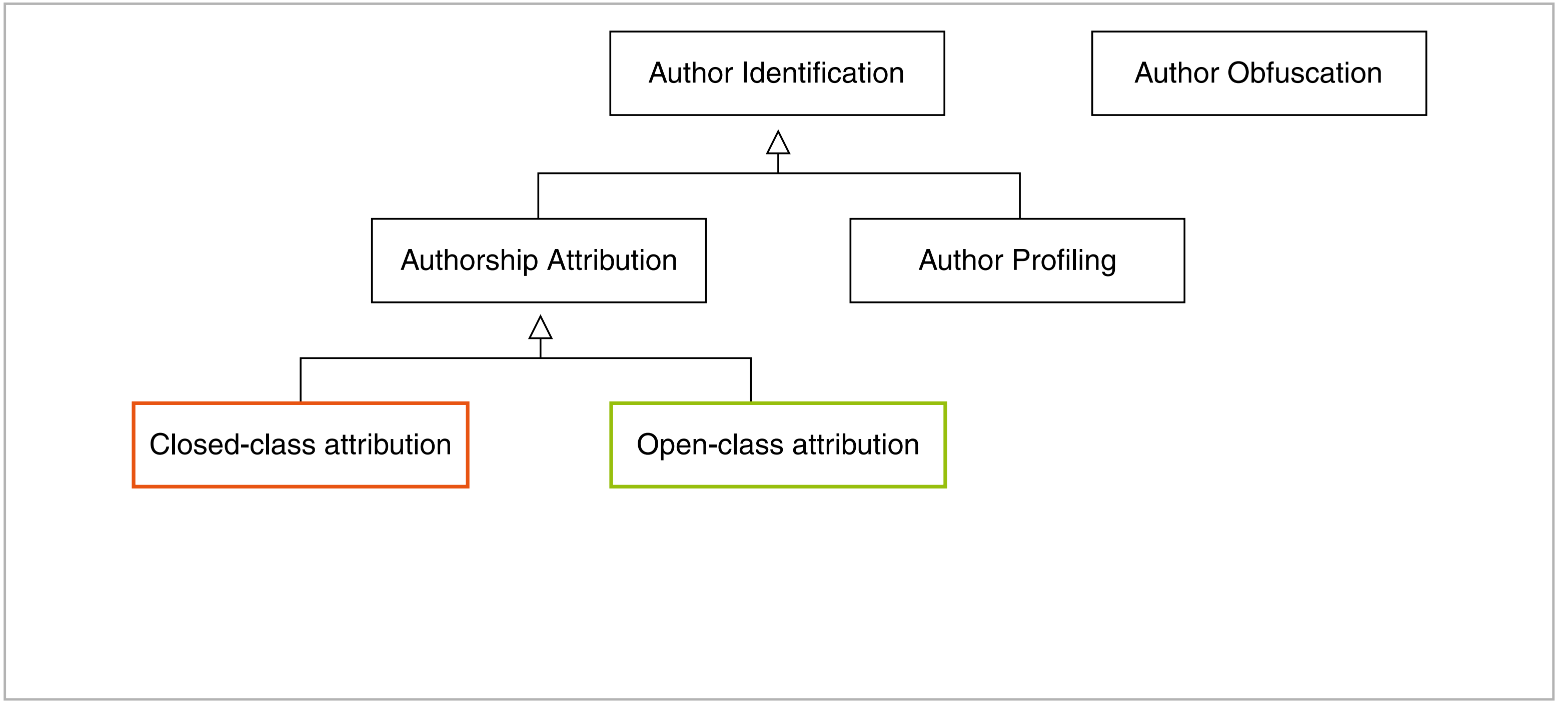
Closed-class attribution

Open-class attribution





Author Identification


Author Identification

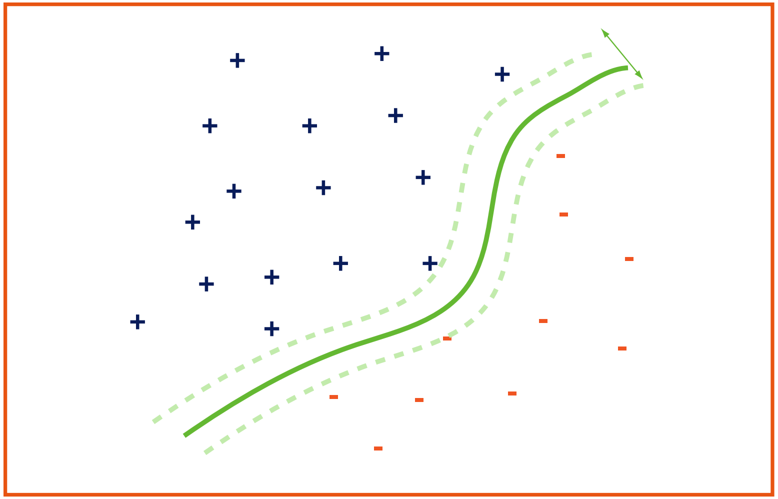
Author Obfuscation

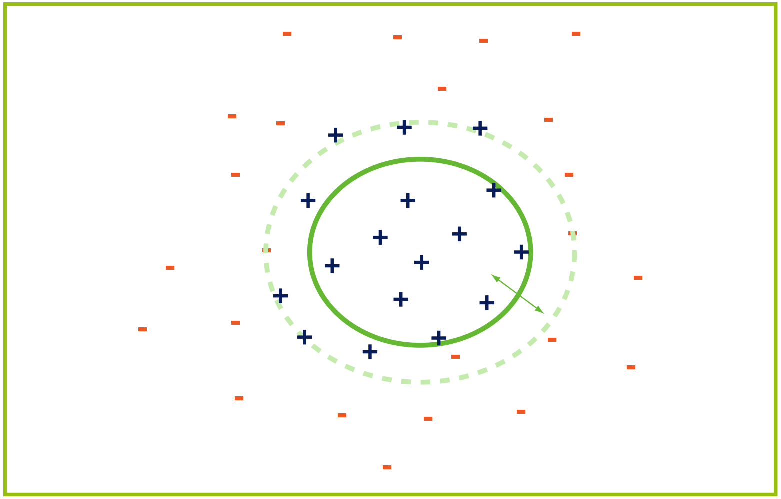
Authorship Attribution

Author Profiling

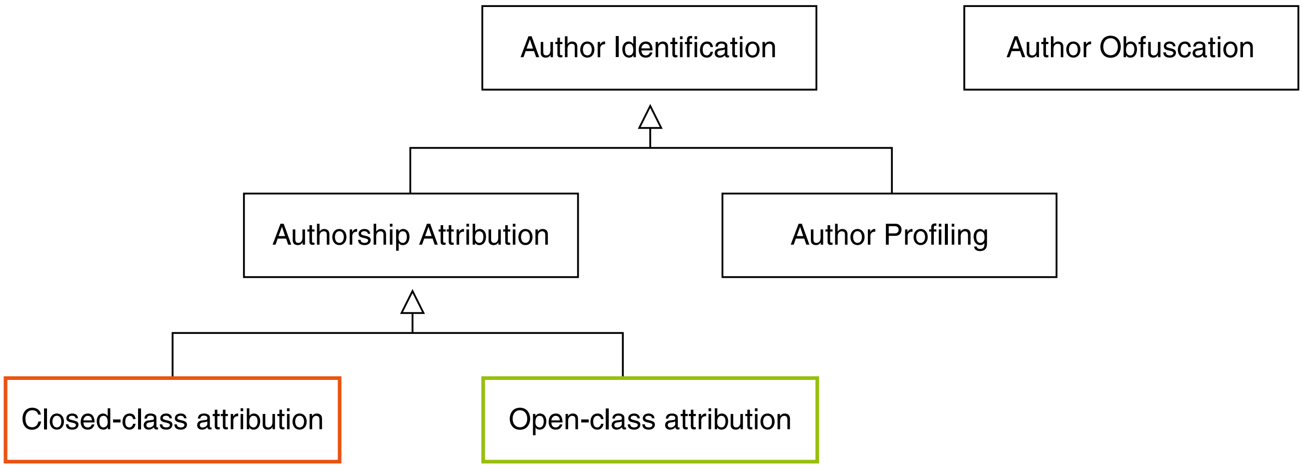
Closed-class attribution

Open-class attribution





Author Identification



Author Identification

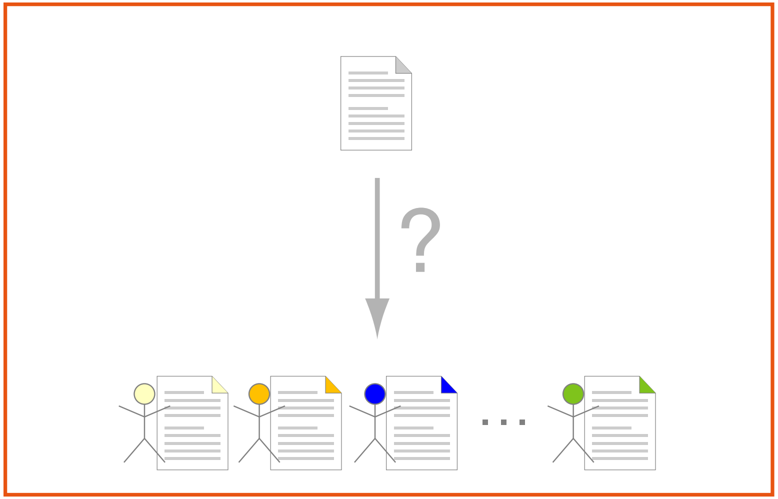
Author Obfuscation

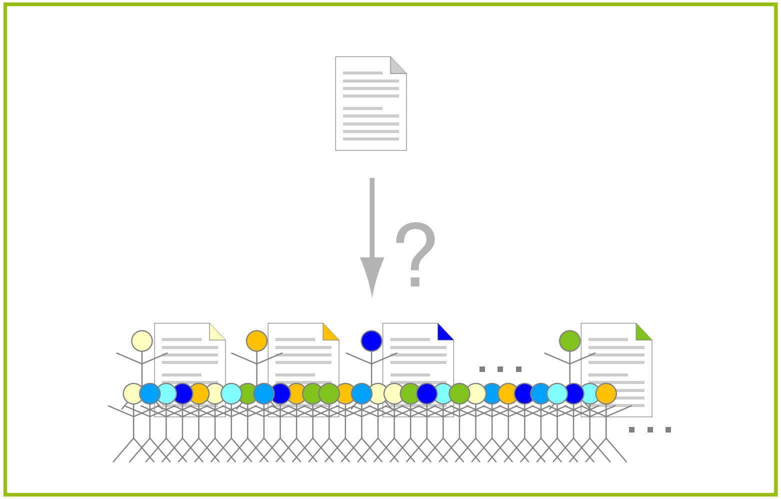
Authorship Attribution

Author Profiling

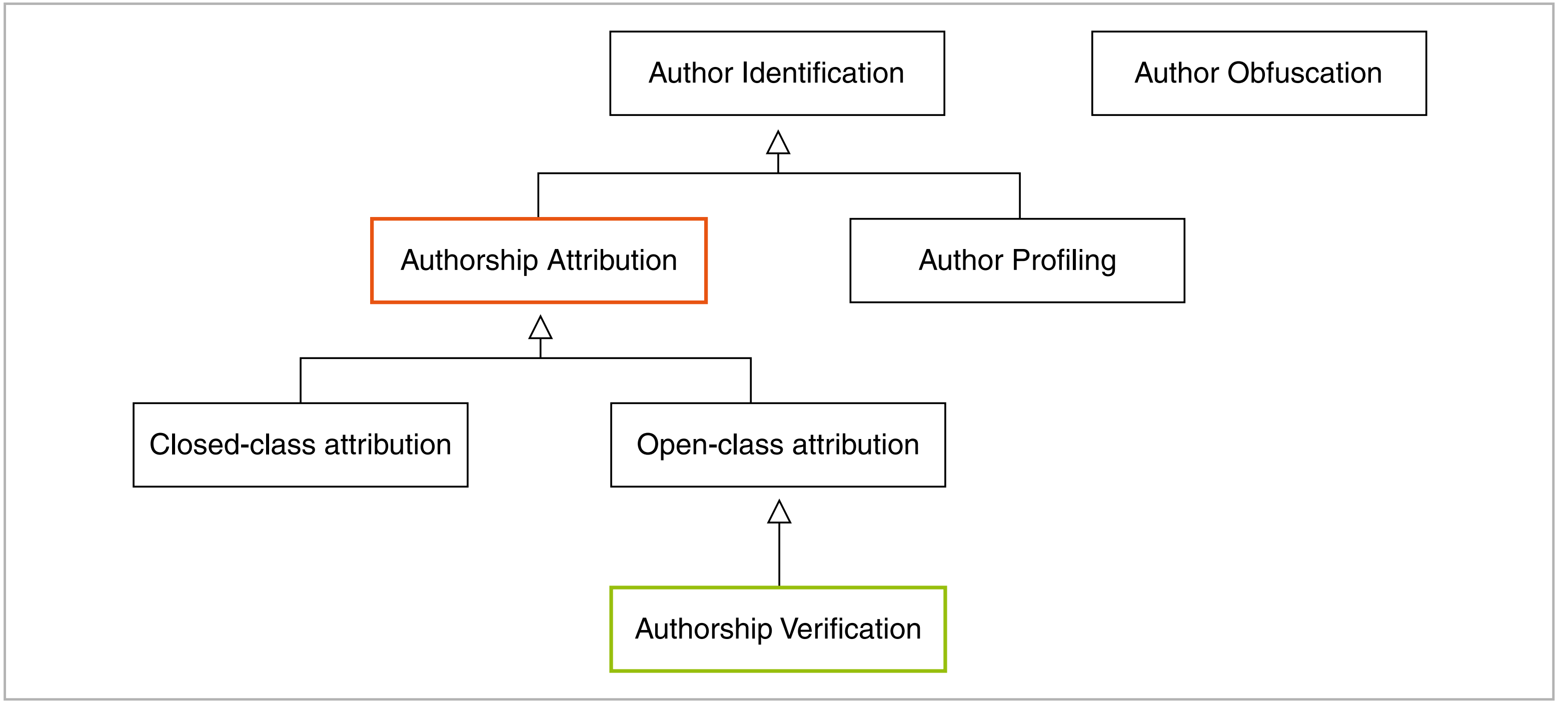
Closed-class attribution

Open-class attribution





Author Identification



Author Identification

Author Obfuscation

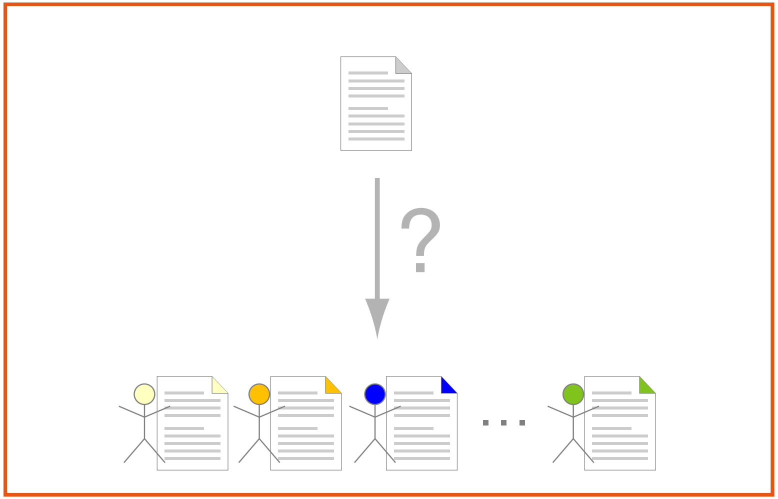
Authorship Attribution

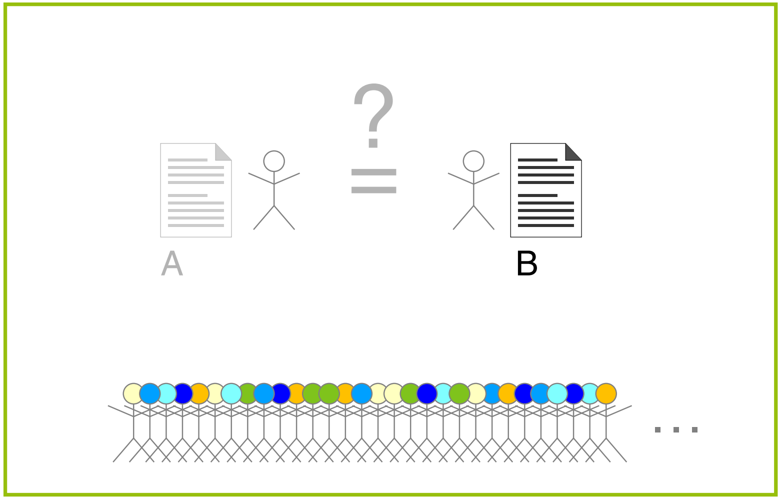
Author Profiling

Closed-class attribution

Open-class attribution

Authorship Verification



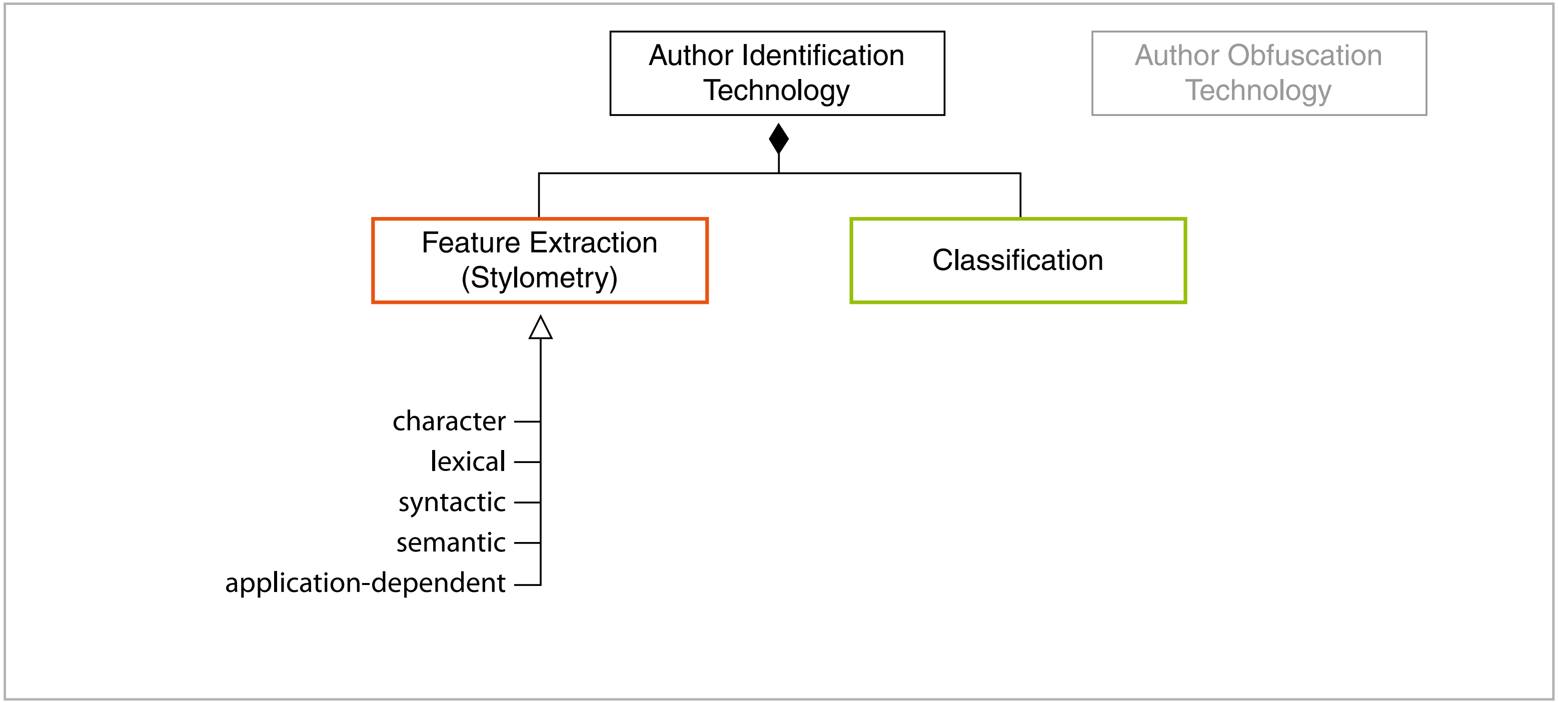


Author Identification

Author Identification Technology

Author Obfuscation Technology

Author Identification



Author Identification Technology

Author Obfuscation Technology

Feature Extraction (Stylometry)

Classification

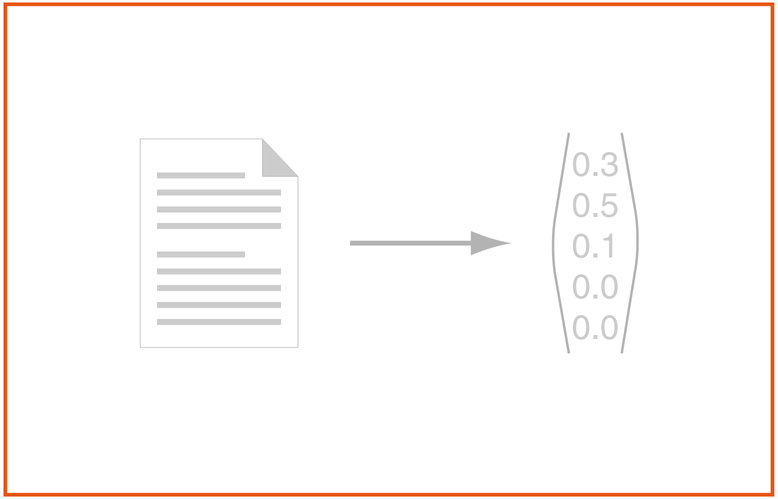
character

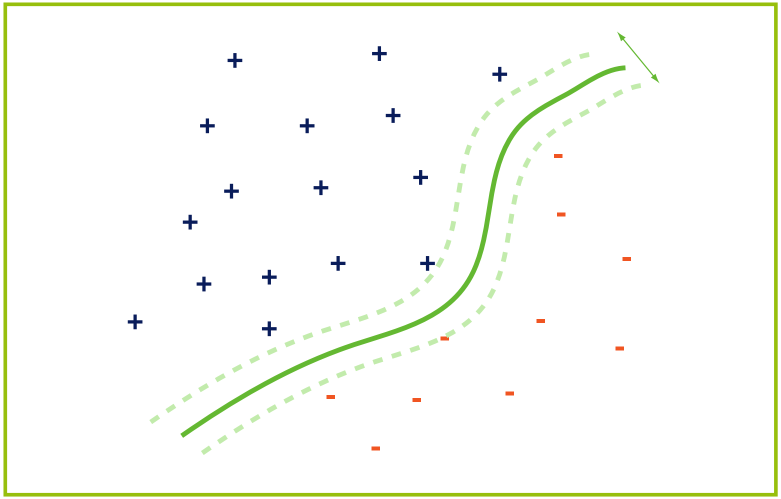
lexical

syntactic

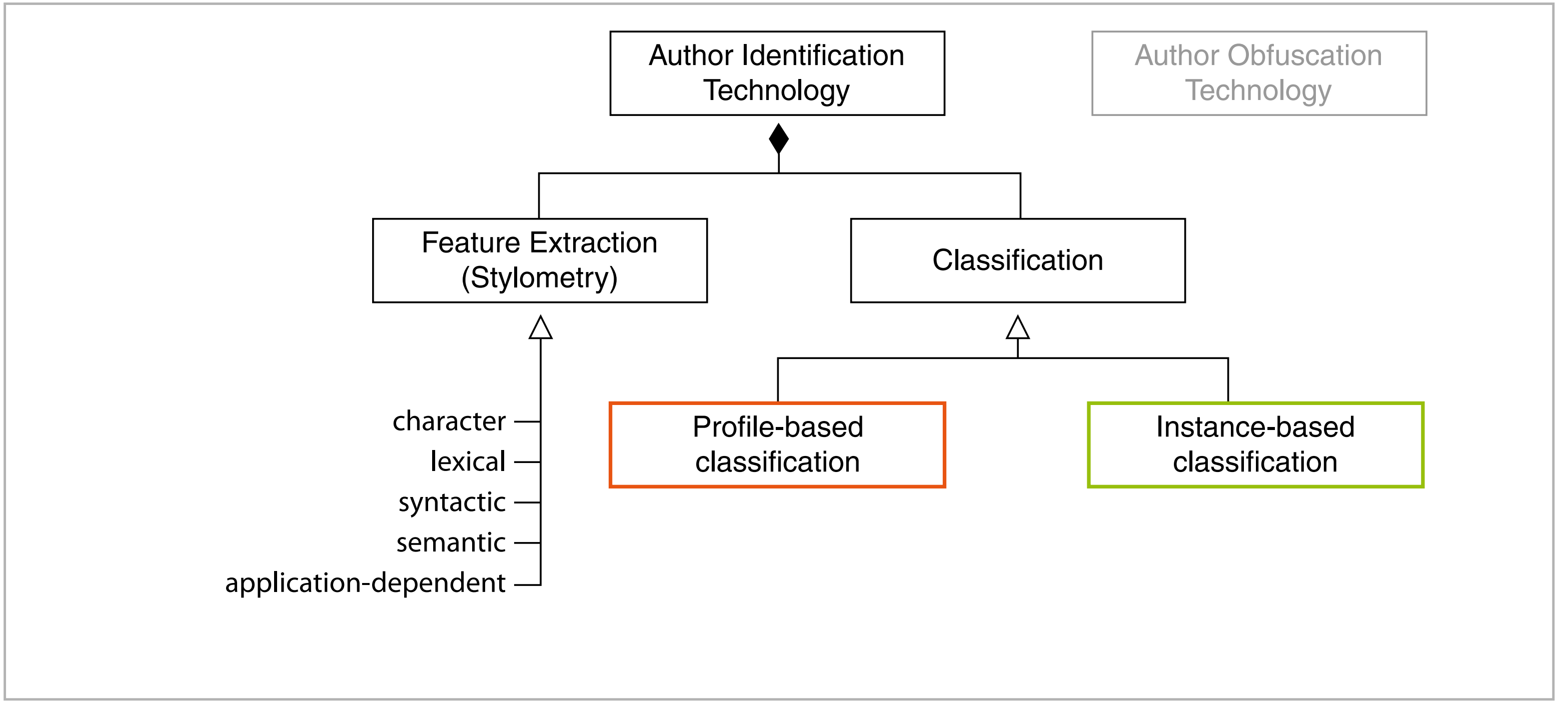
semantic

application-dependent

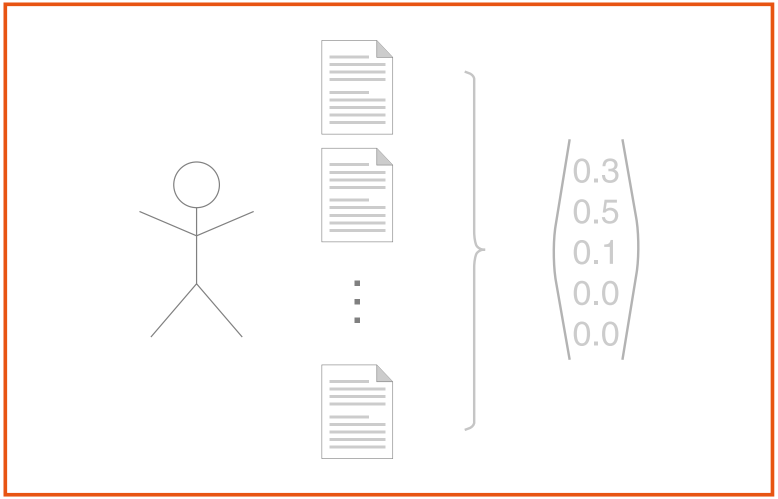


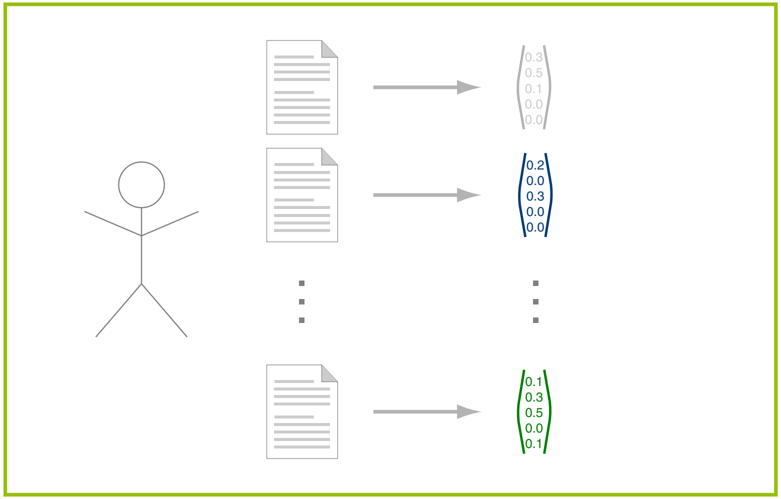


Author Identification



Author Identification Technology
Author Obfuscation Technology
Feature Extraction (Stylometry)
Classification
character
lexical
syntactic
semantic
application-dependent
Profile-based classification
Instance-based classification





Paper Selection

Influential Authorship Attribution Papers

	Publication														
	[4]	[5]	[7]	[10]	[12]	[22]	[23]	[24]	[25]	[29]	[32]	[33]	[34]	[35]	[41]
Task	cA	cA	cA	cA	cA	cA	cA	V	oA	cA	cA	cA	cA	cA	cA
Features	lex	chr	lex	mix	chr	chr	chr	lex	chr	mix	lex	syn	lex	chr	chr
Paradigm	p	i	i	i	i	p	p	i	p	p	i	i	i	p	p
Complexity	**	*	*	*	***	*	**	**	*	**	***	**	*	*	**
Citations	14	377	213	366	41	267	60	75	89	201	17	44	26	43	80
Year	09	02	02	01	11	03	03	07	11	04	12	14	06	07	03

- Influentiality judged by domain expert
- Tasks: closed-class and open-class attribution (cA, oA), verification (V)
- Style features encode character (chr), lexical (lex), syntactical (syn) information, or mixtures (mix) thereof
- Representation paradigms are profile-based (p) and instance-based (i)
- Complexity ranges from string processing to statistical topic modeling

- Publication references correspond to the paper