Meta Analysis within Author Verification

Benno Stein   Nedim Lipka   Sven Meyer zu Eissen

webis.de

Bauhaus-Universität Weimar

Outline

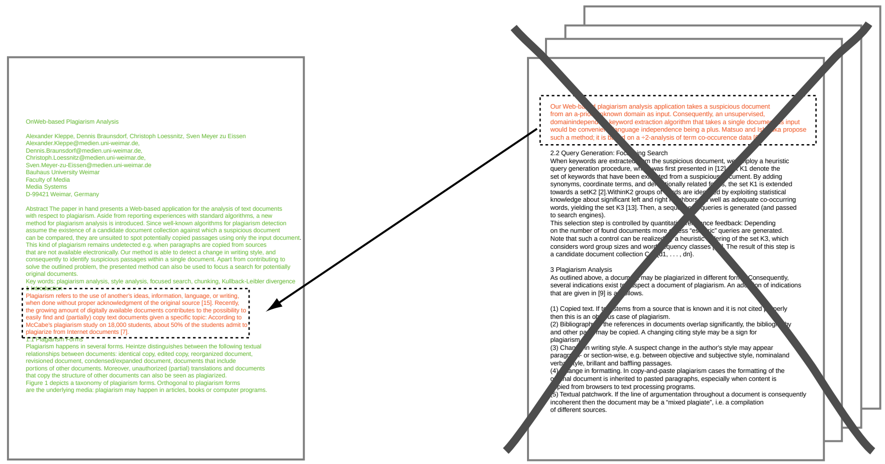
- Intrinsic Plagiarism Analysis and Authorship Verification
- Post-Processing with Unmasking

Intrinsic Analysis and Authorship Verification

Intrinsic Analysis and Authorship Verification

Problem Setting

How to find a plagiarized section / foreign authorship without a reference corpus?



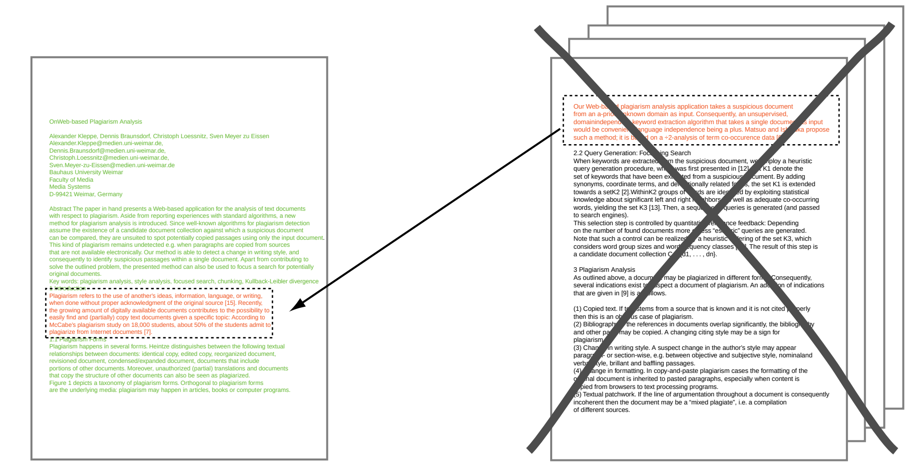
suspicious document

corpus documents

Intrinsic Analysis and Authorship Verification

Problem Setting

How to find a plagiarized section / foreign authorship without a reference corpus?



suspicious document   corpus documents

Formulated as decision problem:

Problem.	AVFIND
Given.	A text d, allegedly written by author A.
Question.	Does d contain sections written by an author B, $B \neq A$?

Intrinsic plagiarism analysis and authorship verification (AV) are two sides of the same coin.

Intrinsic Analysis and Authorship Verification

Building Blocks for Authorship Verification

Pre-analysis			Classification	Post-processing	
Impurity assessment	Decomposition strategy	**Style model construction**	**Style outlier identification**	Improvement at section level	**Improvement at document level**
Document length analysis	Uniform length	Formatting	**Two-class discriminant analysis**	Citation analysis	**Confidence-based majority decision**
Genre Analysis	Structural boundaries	Surface analysis	**One-class classifier: density estimation**		**Unmasking**
Analysis of issuing institution	Text element boundaries	Structure analysis	One-class classifier: boundary estimation		Batch means
	Topical boundaries	**Complexity measures**	One-class classifier: reconstruction		Human inspection
		n-gram analysis			
		Language modeling			
		Dialectic analysis			

Intrinsic Analysis and Authorship Verification

Style Model Construction: Starting Points

Selected quantifiable feature classes (from easy to difficult):

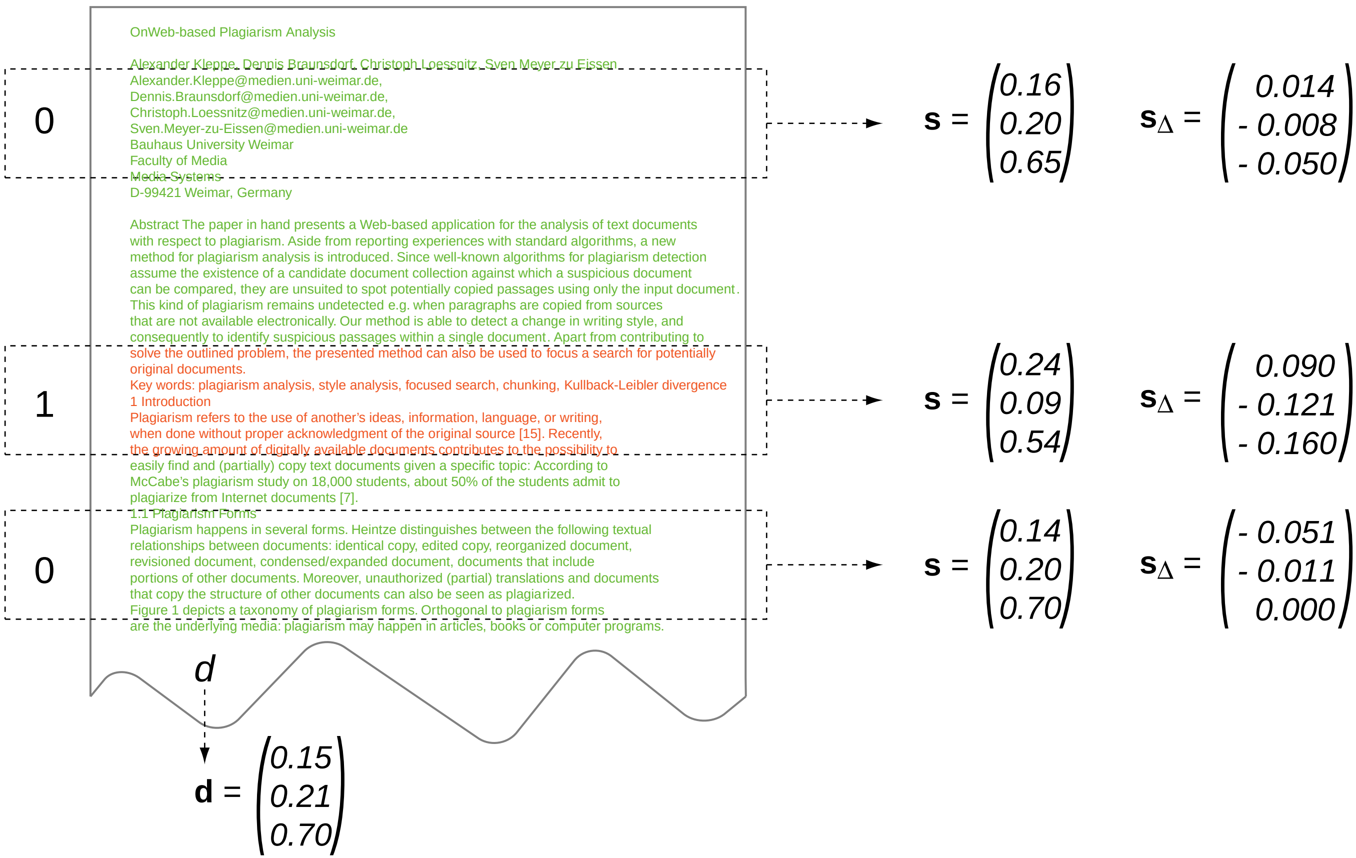
- surface features
- structure and organization
- complexity measures
  - readability
  - writing complexity
  - vocabulary richness, diction
- dialectic power
  - argumentation consistency
  - argumentation strategy

For a machine-based identification, features have to be developed and operationalized within a style model $\mathcal{R}$.

Intrinsic Analysis and Authorship Verification

Style Model Construction: Language Modeling

Style Outlier Identification



Supervised learning situation: given are sections s_i from both the target class (author A), where $c(s) = 0$, and the outlier class (other authors), where $c(s) = 1$.

Intrinsic Analysis and Authorship Verification

Style Outlier Identification

Compute for each section the relative differences between section-specific style feature values and document-specific style feature values.

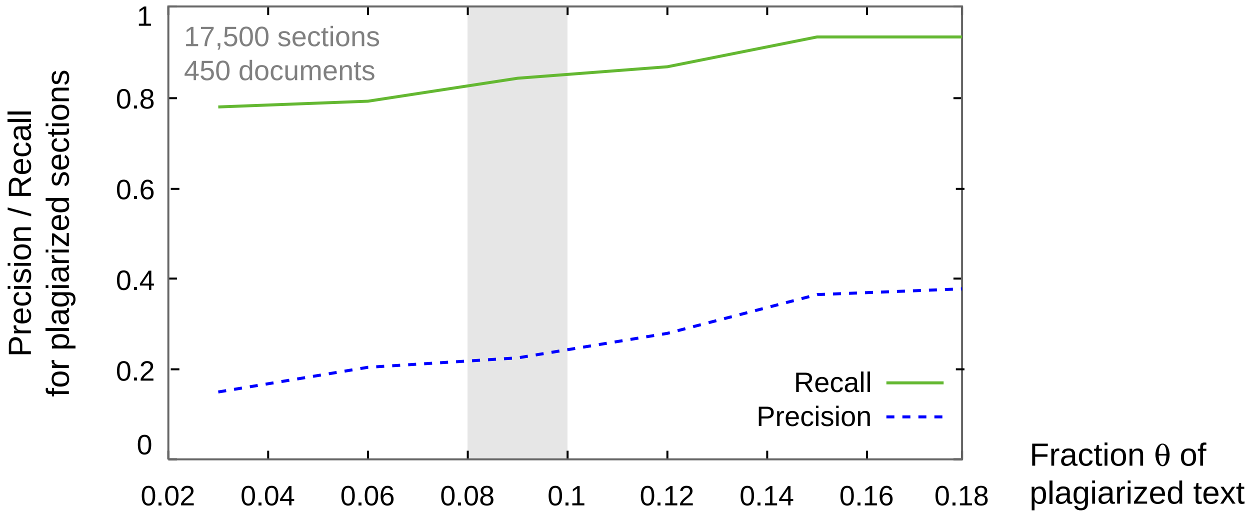
1. Let $\sigma_1, \ldots, \sigma_m$ denote style feature functions.

2. For each section $s \subseteq d$:
   - compute style model $\mathbf{s} = \begin{pmatrix} \sigma_1(s) \\ \vdots \\ \sigma_m(s) \end{pmatrix} \in \mathbf{R}^m$
   - compute relative deviations $\mathbf{s}_\Delta = \begin{pmatrix} \frac{\sigma_1(s) - \sigma_1(d)}{\sigma_1(d)} \\ \vdots \\ \frac{\sigma_m(s) - \sigma_m(d)}{\sigma_m(d)} \end{pmatrix} \in \mathbf{R}^m$

3. Learn an outlier hypothesis h from a sample $\{(\mathbf{s}_\Delta, c(s))\},\ c(s) \in \{0, 1\}$.

Intrinsic Analysis and Authorship Verification

Evaluation: Style Model Performance

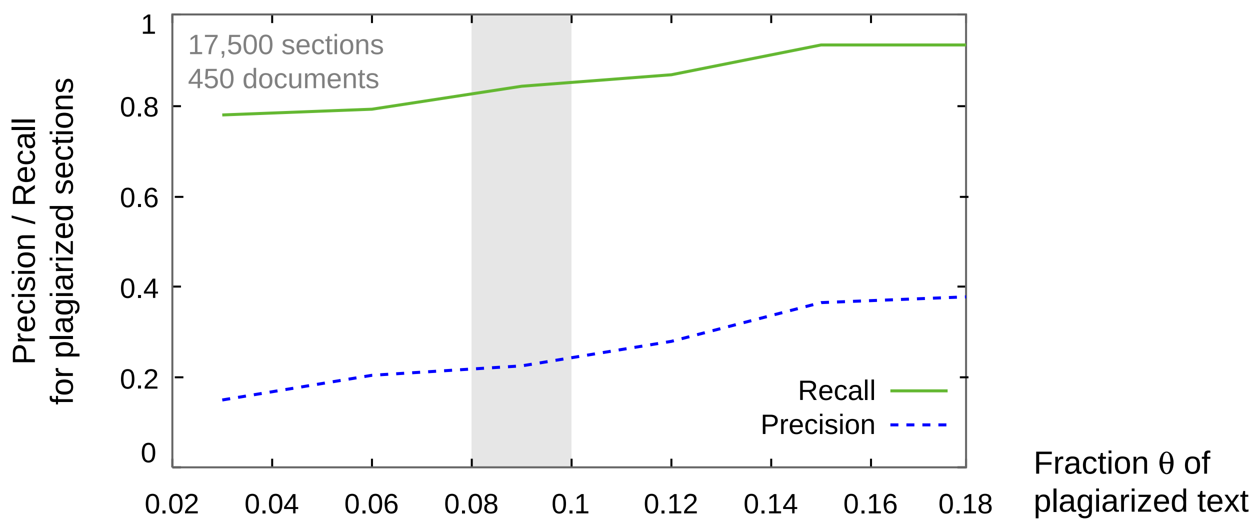


The unsatisfying precision is rooted in the class imbalance.

The Gretchenfrage: Are parts of d plagiarized, if we find an outlier?

Intrinsic Analysis and Authorship Verification

Evaluation: Style Model Performance

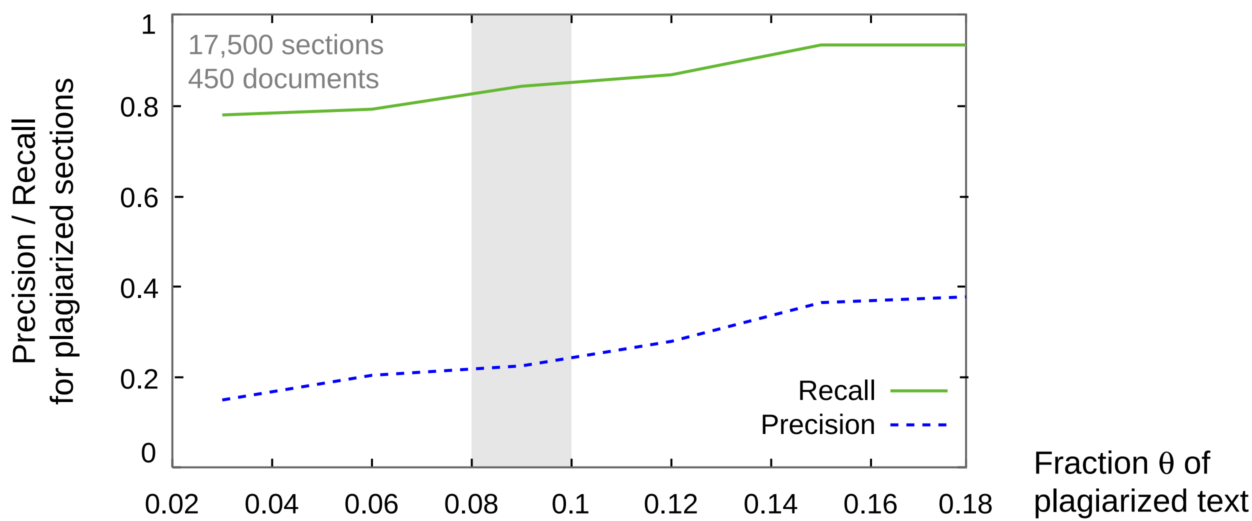


The unsatisfying precision is rooted in the class imbalance.

The Gretchenfrage: Are parts of d plagiarized, if we find an outlier?

# Outliers	Strategy	$\rightarrow$	Hypothesis
0	minimum risk	$\rightarrow$	not plagiarized
1	minimum risk	$\rightarrow$	plagiarized
2	minimum risk	$\rightarrow$	plagiarized
3	minimum risk	$\rightarrow$	plagiarized

Evaluation: Style Model Performance



The unsatisfying precision is rooted in the class imbalance.

The Gretchenfrage: Are parts of d plagiarized, if we find an outlier?

# Outliers	Strategy	→	Hypothesis	Strategy	→	Hypothesis
0	minimum risk	→	not plagiarized	post-processing	→	not plagiarized
1	minimum risk	→	plagiarized	post-processing	→	not plagiarized
2	minimum risk	→	plagiarized	post-processing	→	not plagiarized
3	minimum risk	→	plagiarized	post-processing	→	plagiarized

Post-Processing with Unmasking [Building Blocks]

Post-Processing with Unmasking

Reliable Interpretation of Outliers

Problem. AVOUTLIER (an easier variant of AVFIND)

Given. A set of texts $D = \{d_1, \ldots, d_n\}$, allegedly written by author A.

Question. Does D contain texts written by an author B, $B \neq A$?

Post-Processing with Unmasking

Reliable Interpretation of Outliers

Problem. AVOUTLIER (an easier variant of AVFIND)

Given. A set of texts $D = \{d_1, \ldots, d_n\}$, allegedly written by author A.

Question. Does D contain texts written by an author B, $B \neq A$?

The belief into an answer depends on the number of found outliers:

# Outliers	Strategy	$\rightarrow$	Hypothesis
0	minimum risk, post-processing	$\rightarrow$	not plagiarized
2	minimum risk	$\rightarrow$	plagiarized
2	post-processing	$\rightarrow$	not plagiarized
4	minimum risk, post-processing	$\rightarrow$	plagiarized

Post-Processing with Unmasking

Reliable Interpretation of Outliers

Problem. AVOUTLIER (an easier variant of AVFIND)
Given. A set of texts $D = \{d_1, \ldots, d_n\}$, allegedly written by author A.
Question. Does D contain texts written by an author B, $B \neq A$?

The belief into an answer depends on the number of found outliers:

# Outliers	Strategy	$\rightarrow$	Hypothesis
0	minimum risk, post-processing	$\rightarrow$	not plagiarized
2	minimum risk	$\rightarrow$	plagiarized
2	post-processing	$\rightarrow$	not plagiarized
4	minimum risk, post-processing	$\rightarrow$	plagiarized

Post-process borderline situations to gain further evidence for accepting or rejecting a hypothesis.

Idea: Interpret AVOUTLIER results under the Unmasking framework.

Post-Processing with Unmasking

Unmasking for Authorship Verification [Koppel/Schler 2004]

Problem. AV
Given. Two documents d_1, d_2.

Question. Are d_1 and d_2 written by the same author?

Procedure Unmasking:

1. *Chunking.*

2. *Model Fitting.*

3. *Impairing.*

4. Goto Step 2 until the feature space is sufficiently reduced.

Post-Processing with Unmasking

Unmasking for Authorship Verification [Koppel/Schler 2004]

Problem. AV

Given. Two documents d_1, d_2.

Question. Are d_1 and d_2 written by the same author?

Procedure Unmasking:

1. *Chunking.* Decompose d_1, d_2 into two sets of sections, D_1, D_2.

2. *Model Fitting.* With the 250 most frequent words in d_1, d_2 build a VSM for each s in D_1, D_2. Learn a classifier that discriminates between D_1, D_2.

3. *Impairing.* Drop the 3 most discriminating features from the VSMs.

4. Goto Step 2 until the feature space is sufficiently reduced.

Post-Processing with Unmasking

Unmasking for Authorship Verification [Koppel/Schler 2004]

Problem. AV

Given. Two documents d_1, d_2.

Question. Are d_1 and d_2 written by the same author?

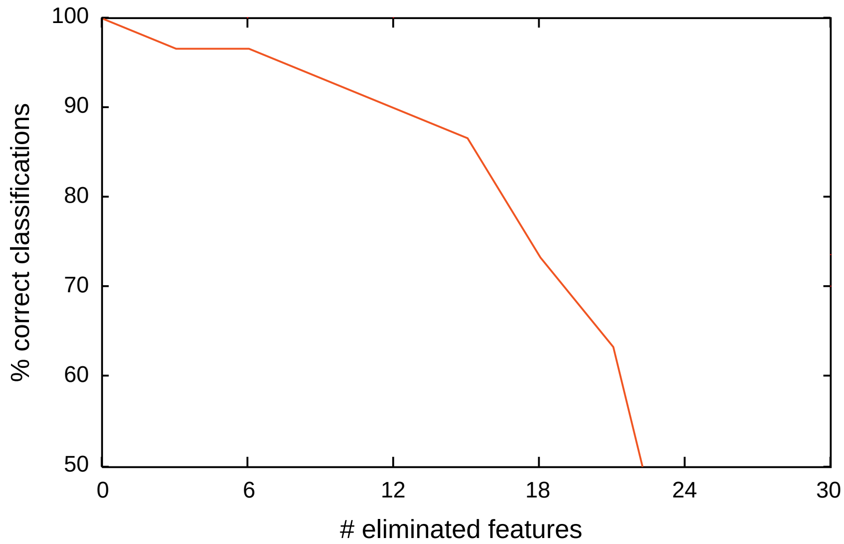
Procedure Unmasking:

1. *Chunking.* Decompose d_1, d_2 into two sets of sections, D_1, D_2.
2. *Model Fitting.* With the 250 most frequent words in d_1, d_2 build a VSM for each s in D_1, D_2. Learn a classifier that discriminates between D_1, D_2.
3. *Impairing.* Drop the 3 most discriminating features from the VSMs.
4. Goto Step 2 until the feature space is sufficiently reduced.
5. *Meta Learning.* Analyze the degradation in the quality of the model fitting.

Post-Processing with Unmasking

Unmasking for Authorship Verification

Characteristic of a typical outcome:



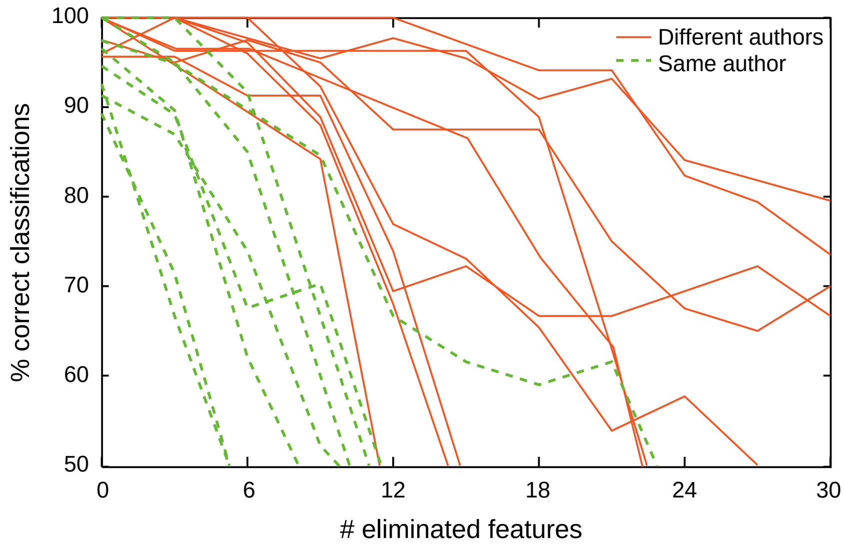
Rationale:

- A large fraction of the 250 words are function words and stop words.
- Only few of the words are related to topic.
- Only few words do the discrimination job—the topic words for a large part.
- Different authors can be distinguished by their use of function words.

Post-Processing with Unmasking

Unmasking for Authorship Verification

Characteristic of a typical outcome:



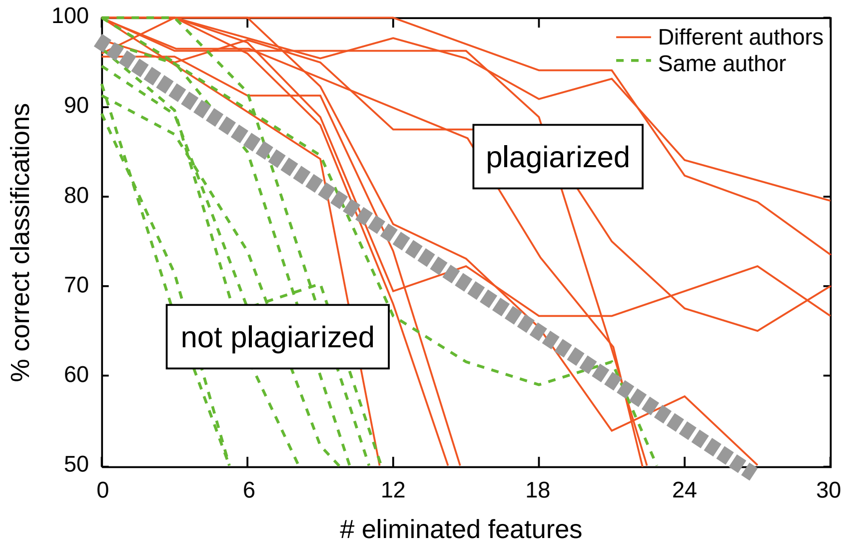
Rationale:

- A large fraction of the 250 words are function words and stop words.
- Only few of the words are related to topic.
- Only few words do the discrimination job—the topic words for a large part.
- Different authors can be distinguished by their use of function words.

Post-Processing with Unmasking

Unmasking for Authorship Verification

Characteristic of a typical outcome:



Rationale:

- A large fraction of the 250 words are function words and stop words.
- Only few of the words are related to topic.
- Only few words do the discrimination job—the topic words for a large part.
- Different authors can be distinguished by their use of function words.

Post-Processing with Unmasking [Results]

Strategy Overview

1. Solve AVOUTLIER with one-class classifier. For borderline situations:
2. Construct AVBATCH from the classified target and outlier sections.
3. Apply Unmasking to solve AVBATCH.

Post-Processing with Unmasking [^]

Evaluation: Artificial Data

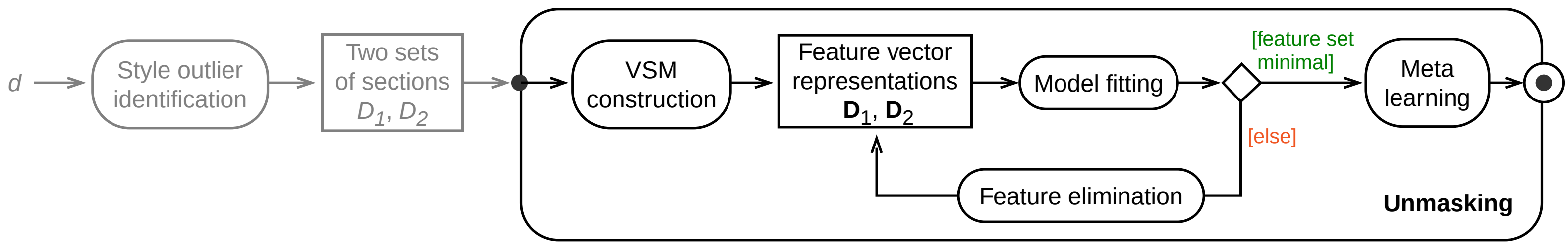
Impurity	Classification			Post-processing					
	AVOUTLIER Minimum risk			AVBATCH Majority			AVBATCH Unmasking		
θ	*prec*	*rec*	F	*prec*	*rec*	F	*prec*	*rec*	F
0.20	0.12	1.00	0.56	0.71	0.83	0.77	0.73	0.90	0.82
0.30	0.20	1.00	0.60	1.00	0.56	0.78	1.00	0.93	0.97
0.40	0.18	1.00	0.59	1.00	0.83	0.92	1.00	0.87	0.94

Post-Processing with Unmasking [^]

Evaluation: Artificial Data

Impurity	Classification			Post-processing					
	AVOUTLIER Minimum risk			AVBATCH Majority			AVBATCH Unmasking		
θ	*prec*	*rec*	F	*prec*	*rec*	F	*prec*	*rec*	F
0.20	0.12	1.00	0.56	0.71	0.83	0.77	0.73	0.90	0.82
0.30	0.20	1.00	0.60	1.00	0.56	0.78	1.00	0.93	0.97
0.40	0.18	1.00	0.59	1.00	0.83	0.92	1.00	0.87	0.94

Strategy overview:



Summary

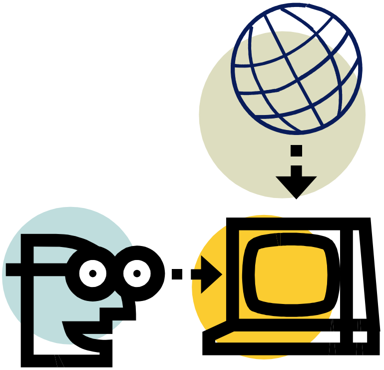
Summary

Authorship verification happens within three steps:

1. Pre-processing. Text decomposition + style model construction
2. Classification. Style outlier identification / one-class classification
3. Post-processing. Improve reliability of the classification step.

Main contribution:

A post-processing strategy for borderline situations, based on unmasking.



Thank you!