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?

Formulated as decision problem:

**Problem.** AV FIND

**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

<table>
<thead>
<tr>
<th>Impurity assessment</th>
<th>Decomposition strategy</th>
<th>Style model construction</th>
<th>Classification</th>
<th>Post-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document length analysis</td>
<td>Uniform length</td>
<td>Formatting</td>
<td>Two-class discriminant analysis</td>
<td>Citation analysis</td>
</tr>
<tr>
<td>Genre Analysis</td>
<td>Structural boundaries</td>
<td>Surface analysis</td>
<td>One-class classifier: density estimation</td>
<td>Unmasking</td>
</tr>
<tr>
<td>Analysis of issuing institution</td>
<td>Text element boundaries</td>
<td>Structure analysis</td>
<td>One-class classifier: boundary estimation</td>
<td>Batch means</td>
</tr>
<tr>
<td></td>
<td>Topical boundaries</td>
<td>Complexity measures</td>
<td>One-class classifier: reconstruction</td>
<td>Human inspection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n$-gram analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Language modeling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dialectic analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
Abstract The paper in hand presents a Web-based application for the analysis of text documents with respect to plagiarism. Aside from reporting experiences with standard algorithms, a new method for plagiarism analysis is introduced. Since well-known algorithms for plagiarism detection assume the existence of a candidate document collection against which a suspicious document can be compared, they are unsuited to spot potentially copied passages using only the input document. This kind of plagiarism remains undetected e.g. when paragraphs are copied from sources that are not available electronically. Our method is able to detect a change in writing style, and consequently to identify suspicious passages within a single document. Apart from contributing to solve the outlined problem, the presented method can also be used to focus a search for potentially original documents.

Key words: plagiarism analysis, style analysis, focused search, chunking, Kullback-Leibler divergence

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 $s = \begin{pmatrix} \sigma_1(s) \\ \vdots \\ \sigma_m(s) \end{pmatrix} \in \mathbb{R}^m$
   - compute relative deviations $s_\Delta = \begin{pmatrix} \frac{\sigma_1(s) - \sigma_1(d)}{\sigma_1(d)} \\ \vdots \\ \frac{\sigma_m(s) - \sigma_m(d)}{\sigma_m(d)} \end{pmatrix} \in \mathbb{R}^m$

3. Learn an outlier hypothesis $h$ from a sample $\{(s_\Delta, c(s))\}$, $c(s) \in \{0, 1\}$.
The unsatisfying precision is rooted in the class imbalance.

The Gretchenfrage: Are parts of \( d \) plagiarized, if we find an outlier?
The unsatisfying precision is rooted in the class imbalance.

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

<table>
<thead>
<tr>
<th># Outliers</th>
<th>Strategy</th>
<th>→ Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>minimum risk</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>1</td>
<td>minimum risk</td>
<td>plagiarized</td>
</tr>
<tr>
<td>2</td>
<td>minimum risk</td>
<td>plagiarized</td>
</tr>
<tr>
<td>3</td>
<td>minimum risk</td>
<td>plagiarized</td>
</tr>
</tbody>
</table>
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?

<table>
<thead>
<tr>
<th># Outliers</th>
<th>Strategy</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>minimum risk</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>1</td>
<td>minimum risk</td>
<td>plagiarized</td>
</tr>
<tr>
<td>2</td>
<td>minimum risk</td>
<td>plagiarized</td>
</tr>
<tr>
<td>3</td>
<td>minimum risk</td>
<td>plagiarized</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>post-processing</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>post-processing</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>post-processing</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>post-processing</td>
<td>plagiarized</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th># Outliers</th>
<th>Strategy</th>
<th>→</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>minimum risk, post-processing</td>
<td>→</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>2</td>
<td>minimum risk</td>
<td>→</td>
<td>plagiarized</td>
</tr>
<tr>
<td>2</td>
<td>post-processing</td>
<td>→</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>4</td>
<td>minimum risk, post-processing</td>
<td>→</td>
<td>plagiarized</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th># Outliers</th>
<th>Strategy</th>
<th>→</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>minimum risk, post-processing</td>
<td>→</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>2</td>
<td>minimum risk</td>
<td>→</td>
<td>plagiarized</td>
</tr>
<tr>
<td>2</td>
<td>post-processing</td>
<td>→</td>
<td>not plagiarized</td>
</tr>
<tr>
<td>4</td>
<td>minimum risk, post-processing</td>
<td>→</td>
<td>plagiarized</td>
</tr>
</tbody>
</table>

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:

![Graph showing % correct classifications vs. # eliminated features with lines for different authors and same author.]

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

Strategy Overview

1. Solve AV_{OUTLIER} with one-class classifier. For borderline situations:
2. Construct AV_{BATCH} from the classified target and outlier sections.
3. Apply Unmasking to solve AV_{BATCH}.
Post-Processing with Unmasking

Evaluation: Artificial Data

<table>
<thead>
<tr>
<th>Impurity</th>
<th>Classification</th>
<th>Post-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AVOUTLIER</td>
<td>AVBATCH</td>
</tr>
<tr>
<td></td>
<td>Minimum risk</td>
<td>Majority</td>
</tr>
<tr>
<td>$\theta$</td>
<td>prec rec $F$</td>
<td>prec rec $F$</td>
</tr>
<tr>
<td>0.20</td>
<td>0.12 1.00 0.56</td>
<td>0.71 0.83 0.77</td>
</tr>
<tr>
<td>0.30</td>
<td>0.20 1.00 0.60</td>
<td>1.00 0.56 0.78</td>
</tr>
<tr>
<td>0.40</td>
<td>0.18 1.00 0.59</td>
<td>1.00 0.83 0.92</td>
</tr>
</tbody>
</table>
Post-Processing with Unmasking

Evaluation: Artificial Data

<table>
<thead>
<tr>
<th>Impurity</th>
<th>Classification</th>
<th>Post-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AVOUTLIER</td>
<td>AVBATCH</td>
</tr>
<tr>
<td></td>
<td>Minimum risk</td>
<td>Majority</td>
</tr>
<tr>
<td></td>
<td>( \theta )</td>
<td>( \theta )</td>
</tr>
<tr>
<td></td>
<td>( \text{prec} )</td>
<td>( \text{rec} )</td>
</tr>
<tr>
<td>0.20</td>
<td>0.12</td>
<td>1.00</td>
</tr>
<tr>
<td>0.30</td>
<td>0.20</td>
<td>1.00</td>
</tr>
<tr>
<td>0.40</td>
<td>0.18</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

Main contribution:
A post-processing strategy for borderline situations, based on unmasking.
Thank you!