

# Intrinsic Plagiarism Analysis with Meta Learning

Benno Stein

Faculty of Media, Media Systems  
Bauhaus University Weimar  
99421 Weimar, Germany  
benno.stein@  
medien.uni-weimar.de

Sven Meyer zu Eissen

Faculty of Media, Media Systems  
Bauhaus University Weimar  
99421 Weimar, Germany  
sven.meyer-zu-eissen@  
medien.uni-weimar.de

## ABSTRACT

In intrinsic plagiarism analysis we are given a document, allegedly written by a single author, and the task is to find sufficient evidence either to accept or to reject this hypothesis. Existing research to intrinsic plagiarism analysis tries to quantify changes in the writing style by analyzing the distributions of particular style markers. This way, acceptable detection rates can be achieved if the portion of plagiarized sections is known a-priori and if the document is of a single genre. However, both assumptions may not be fulfilled in practice.

In [6] Koppel and Schler propose a new approach to the authorship verification problem, where the task is to determine whether two texts are written by the same author. Their approach is ingenious in that it provides a means to detect relatively shallow differences in writing style while being independent of language, period, and genre. Since the approach requires two (relatively large) samples of text to be compared to each other it cannot be applied directly to the intrinsic plagiarism analysis problem.

Main contribution of our paper is the idea to address the shortcomings of existing approaches to intrinsic plagiarism analysis with the technology presented in [6]. We propose a hybrid approach that employs style marker analysis for the purpose of hypotheses generation which then are accepted or rejected by an authorship verification analysis. A second contribution of our paper is the evaluation of style markers for German text and their application to a real-world plagiarism case.

## Keywords

intrinsic plagiarism analysis, one-class classification, meta learning

## 1. INTRODUCTION

Intrinsic plagiarism analysis is characterized as follows. We are given a document  $d$ , allegedly written by a single author, and we want to identify sections in  $d$  which stem from another author and which are not labeled as such, e.g. by proper citation.<sup>1</sup> Intrinsic

<sup>1</sup>The intrinsic plagiarism analysis problem becomes harder if  $d$  is declared as a multi author document.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection.*

---

$D$	collection of real-world documents
$d \in D$	real-world document
$\mathbf{d}$	vector space representation of document $d$
$\mathbf{D}$	collection of vector space representations of $d \in D$
$s \subseteq d$	section of a real-world document
$n$	number of sections in which $d$ is decomposed
$\sigma : s \mapsto \sigma(s) \in \mathbf{R}$	scalar style marker or style feature of a section $s$
$\mathbf{s}$	style model representation of a section $s$ = vector of style markers
$m$	length of a vector $\mathbf{s}$ of style markers
$\theta$	portion of $d$ that is plagiarized

---

Table 1: Notation used in this paper.

insic plagiarism analysis is a one-class classification problem. The salient property of such classification problems is that information of only one class is available. This class is called the target class, all other objects are comprised in the so-called outlier class.

In the context of intrinsic plagiarism analysis all documents or document parts of the pretended author form the target class, and all documents or document parts of an arbitrary other author form the outlier class. Note that the document  $d$  is the only source to formulate a writing style model for objects in the target class, whereas the formulation of this model is impeded to the extent at which  $d$  is plagiarized. Also note that the documents in the outlier class are so abundant that neither a representative sampling nor the formulation of a writing style model for this class is possible.

One-class classification problems, and hence the intrinsic plagiarism analysis problem, must be solved on the basis of examples from the target class. Tax distinguishes following methods to solve one-class classification problems [10]:

1. *Outlier Detection Methods.* These methods are further distinguished with respect to the detection strategy:
  - (a) Methods that rely on standard classification and learning technology. Outliers are generated artificially, and a standard classification approach is applied to separate outliers from the target class.
  - (b) Modified methods from the field of classification or regression problem solving. Instead of using the most probable feature weights  $\mathbf{w}$  in a classifier, which aims at the minimization of the classification error given a training set, the classifier utilizes the probability of the correctness of  $\mathbf{w}$ .

- (c) Density methods, which directly estimate the probability distributions of features for the target class. Outliers are assumed to be uniformly distributed, and Bayes rule can be applied to separate outliers from the target class.
2. *Reconstruction Methods.* If we are given both an object’s feature vector (which is a style model representation  $\mathbf{s}$  here) as well as the original object (which is the document  $d$  or its VSM representation  $\mathbf{d}$  here), we may be able to reconstruct  $\mathbf{s}$  from  $d$  as  $\alpha(d)$  as well as to measure the reconstruction error  $\alpha(d) \ominus \mathbf{s}$ . It is assumed that  $\alpha$  captures the domain theory underlying the target class, and the smaller the reconstruction error is the more likely  $\mathbf{s}$  belongs to the target class.
3. *Boundary Methods.* These methods avoid the estimation of the multi-dimensional density and focus on the definition of a boundary around the set of target objects. The computation of the boundary is based on the distances between the objects in the target set.

## 1.1 Contributions

The contributions of this paper are as follows. Section 2 outlines existing as well as, up to now, not applied technology to solve the problem of intrinsic plagiarism detection. The two presented methods rely on a style marker analysis and can be regarded as specific variants of what Tax terms “outlier detection methods” [10]. A weakness of the presented plagiarism analysis methods is that they require meta knowledge about the amount and the distribution of the plagiarized text in a document  $d$  in order to achieve acceptable values for precision and recall.

To improve the classification performance and to become more independent of a-priori knowledge we propose to verify the classification results obtained by a style marker analysis with the meta learning approach developed by Koppel and Schler [6]. Section 3 outlines their approach and its application to the intrinsic plagiarism analysis problem. Section 4 presents first results based on both artificial data and a real plagiarism case.

Table 1 compiles the notation that is used throughout the paper.

## 2. INTRINSIC PLAGIARISM ANALYSIS

Intrinsic plagiarism analysis deals with the detection of plagiarized sections within a document  $d$ , without comparing  $d$  to extraneous sources [8]. To solve this ambitious task the writing style of individual sections has to be analyzed in order to spot those sections whose style differs significantly from the rest. There are several subproblems that arise in this connection, including the smart decomposition of  $d$ , the identification of features that capture style information, the detection of stylistic anomalies or changes in style, or the construction of a corpus with positive and negative examples for plagiarism.

Writing style aspects can be quantified with style markers: Let  $s_1, \dots, s_n$  be a decomposition of a document  $d$  into  $n$  contiguous, non-overlapping sections. Moreover, let  $\sigma_1, \dots, \sigma_m$  denote a set of style markers, each of which assigning a real value to a section  $s \subseteq d$  in order to quantify a certain style aspect of the writing. The style model representation  $\mathbf{s}$  of a section  $s$  is an  $m$ -dimensional vector, comparable to an instance of the vector space model or a genre retrieval model:

$$\mathbf{s} = \begin{pmatrix} \sigma_1(s) \\ \vdots \\ \sigma_m(s) \end{pmatrix}, s \subseteq d$$

When a section  $s^- \subset d$  is plagiarized, the assumption is that its style model representation,  $\mathbf{s}^-$ , differs significantly from other representations  $\mathbf{s}^+$  that belong to non-plagiarized sections  $s^+ \subset d$ . Using an outlier detection method,  $\mathbf{s}^-$  may be distinguished from  $\mathbf{s}^+$  with acceptable reliability.

In [8] Meyer zu Eissen and Stein proposed and analyzed an outlier detection method of Type (1a). They developed a “factory” corpus for plagiarism analysis, and generated test corpora with several thousand positive and negative training examples. Based on these corpora different classifiers were constructed, using discriminant analysis and SVM training among others. Input for the training are the relative deviations of 10 carefully selected style markers and about 10 part-of-speech features, whereas for each section  $s \in \{s_1, \dots, s_n\}$  the vector  $\mathbf{s}_\Delta$  of relative deviations of its style marker values from the document mean is computed:

$$\mathbf{s}_\Delta = \begin{pmatrix} \frac{\sigma_1(s) - \sigma_1(d)}{\sigma_1(d)} \\ \vdots \\ \frac{\sigma_m(s) - \sigma_m(d)}{\sigma_m(d)} \end{pmatrix}, s \subseteq d$$

Meyer zu Eissen and Stein reported precision and recall values of about 80% provided that meta knowledge about the plagiarized portion  $\theta$  of  $d$  is given. In particular, they distinguished for  $\theta$  the values  $0.03 \cdot i, i = 1, \dots, 6$ .

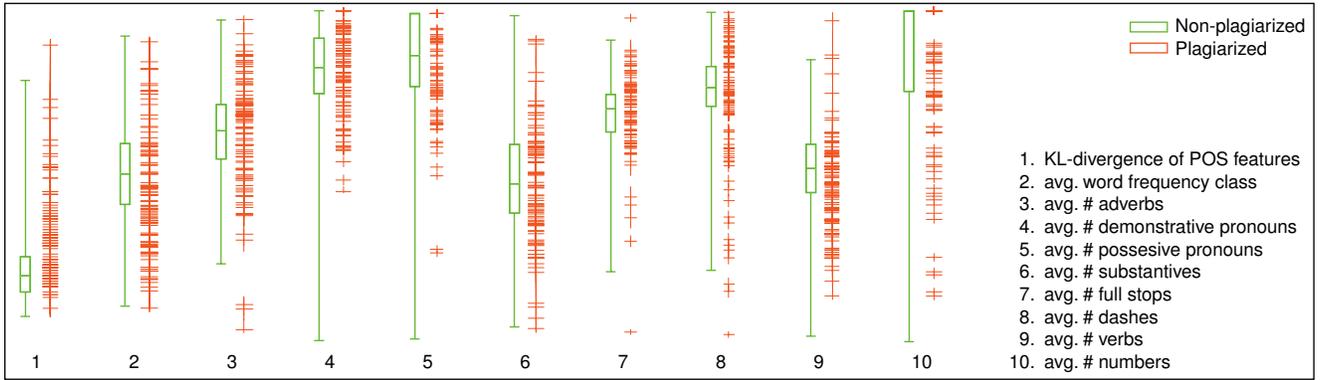
Main contribution of [8] is the analysis of style markers with respect to their robustness, and the identification of a new class of robust style markers. In this connection, robustness pertains to the sensitivity  $\zeta$  of a style marker  $\sigma(s)$  with respect to the length  $|s|$  of a section:  $\zeta(\sigma(s), |s|)$  of a robust style marker has a small variance.

### 2.1 Improved Style Marker Analysis

The most severe deficiency of outlier detection methods of Type (1a) roots in their dependency on the dimensionality of  $\mathbf{s}$ : the number of examples must grow exponentially in the number of *relevant* features, in order to apply a machine learning approach without bad conscience. This fact is sometimes termed as “curse of dimensionality”. The second-worst deficiency relates to the artificiality of the generated examples: the less we know about the stylistic impacts of plagiarism and the possible means to model these impacts the more unrepresentative the examples will be. It is in the nature of one-class classification problems that we have only very restricted knowledge and very few examples to model the outlier class, which are the plagiarized sections here.

By directly modeling the target objects, outlier detection methods of Type (1c) provide a way out for the mentioned problems. In this connection it is reasonable to presume the style markers in the objects of the target group being Gaussian distributed, while being uniformly distributed in the outlier group. Let  $S^+$  denote the event that a section  $s \in \{s_1, \dots, s_n\}$  belongs to the target group (= not plagiarized); likewise, let  $S^-$  denote the event that an  $s$  belongs to the outlier group (= plagiarized). Given a suspicious document  $d$  and a single style marker  $\sigma$  the acceptance or rejection of the hypothesis whether a paragraph  $s \subset d$  is plagiarized happens in five steps:

1. Hypothesizing an a-priori probability,  $P(S^-) = \theta$ , that some section  $s \subset d$  is plagiarized;  $P(S^+) = 1 - P(S^-)$ .
2. Depending on  $P(S^+)$ , decomposition of  $d$  into sections  $s_1, \dots, s_n$ . Note that  $P(S^+)$  provides valuable meta knowledge for the estimation of reasonable values for the section lengths  $|s_i|$ .



**Figure 1: Distribution of 10 style markers in 16,000 non-plagiarized (green) and 1,500 plagiarized (red) sections. The sections have a length of about 400 words and result from an equidistant partitioning of 900 plagiarized documents. The plagiarized portion,  $\theta$ , of a document ranges between 0.05 and 0.5.**

3. Estimation of  $\sigma$ 's expectation value and variance with respect to  $s_1, \dots, s_n$ .
4. Provided an equidistant segmentation of  $\sigma$ 's domain, computation of the conditional probabilities  $P(\sigma(s) | S^+)$  and  $P(\sigma(s) | S^-)$ , assuming a Gaussian and a uniform distribution respectively.
5. Application of Bayes rule and determination of the maximum a-posteriori hypothesis:

$$P(S^+ | \sigma(s)) = \frac{P(\sigma(s) | S^+) \cdot P(S^+)}{P(\sigma(s))} \quad \text{and}$$

$$P(S^- | \sigma(s)) = \frac{P(\sigma(s) | S^-) \cdot P(S^-)}{P(\sigma(s))}, \quad \text{with}$$

$$P(\sigma(s)) = P(\sigma(s) | S^+) \cdot P(S^+) + P(\sigma(s) | S^-) \cdot P(S^-).$$

The above decision procedure is formulated for a single style marker. Multiple style markers  $\sigma_1, \dots, \sigma_m$  require the accounting of multiple conditional probabilities. Under the conditional independence assumption the naive Bayes approach can be applied; the accepted a-posteriori hypothesis then computes as

$$\operatorname{argmax}_{S \in \{S^+, S^-\}} P(S) \cdot \prod_{i=1}^m P(\sigma_i(s) | S).$$

An alternative—and, dependent on the training corpus—more powerful approach is the construction of a Gaussian mixture for the  $\sigma_1, \dots, \sigma_m$ . The respective weights,  $\mathbf{w}$ , can be estimated by the linear model of a discriminant analysis, similar to the construction of a classifier when pursuing an outlier detection method of Type (1a).

The question that remains to be answered is which style markers qualify for intrinsic plagiarism analysis?

## 2.2 Style Markers

Quantifying the writing style of text is an active field of research since the 1940s [11, 3]. Several style markers have been proposed to measure writer-specific style aspects like vocabulary richness [4, 11] or text complexity and understandability [3], as well as to determine reader-specific requirements that are necessary to understand a text, like grading levels [2, 5, 1]. These style markers have been

developed to judge longer texts ranging from a few pages up to book size.

Since plagiarizers often copy sections that are shorter than a page [7], the section decomposition  $\{s_1, \dots, s_n\}$  of a document must not be too coarse, and, it is questionable which of the style markers will work for shorter sections. It should be clear that style markers that employ measures like average paragraph length are not reliable for shorter sections that consist of one or two paragraphs.

The work in [9] investigates the robustness of the vocabulary richness measures Yule's  $K$ , Honore's  $R$ , and the average word frequency class. The outcome is that only the average word frequency class can be called robust: it provides reliable results even for short sections, which can be explained with its word-based granularity. To get an idea of the usability of different style markers, Figure 1 contrasts their distribution in both original (shown green) and plagiarized (shown red) sections in a collection of 1000 documents.

## 3. COUPLING STYLE MARKER ANALYSIS AND META LEARNING

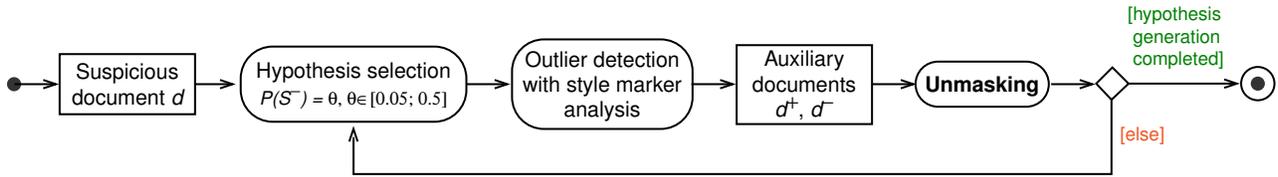
With the methods presented in the former section, we are able to identify possibly plagiarized sections in a document  $d$ . Let  $d^+ \subseteq d$  and  $d^- \subseteq d$  denote two auxiliary documents constructed from  $d$ , where  $d^+$  is comprised of all allegedly non-plagiarized sections in  $d$ , while  $d^-$  is comprised of all allegedly plagiarized sections in  $d$ . In particular we claim that  $d^+ \cup d^- = d$ .

Note that, based on the decomposition  $s_1, \dots, s_n$  of  $d$  and the quality of the detection approach,  $d^- \subseteq d$  may contain non-plagiarized sections, say, its precision is  $< 1$ . Likewise,  $d^+$  may not be complete, say, the recall of the plagiarized sections is  $< 1$ . Moreover, different a-priori probabilities  $P(S^-)$  will result in different documents  $d^-$  to be synthesized.

Given  $d^+$  and  $d^-$  our objective now is to find further evidence whether  $d$  contains plagiarized sections at all. I.e., we will not try to verify whether a single section  $s \subset d$  is plagiarized—instead we try to answer the following relaxed decision problem:

*“Is  $d$  written by a single author?”*

For this purpose we employ the unmasking approach of Koppel and Schler, originally developed to solve the authorship verification problem [6]. Unmasking is a special meta learning approach, where two documents  $d_1$  and  $d_2$  (likewise  $d^+$  and  $d^-$ ) are incrementally reduced towards author-specific writing style essentials. If  $d^+$  and  $d^-$  in fact stem from different authors, unmasking is a



**Figure 2: UML activity diagram of a new hybrid approach to intrinsic plagiarism analysis: (a) selection of a hypothesis for the plagiarized portion  $\theta$  of  $d$ , (b) generation of two auxiliary documents  $d^+ \subseteq d$  and  $d^- \subseteq d$  with style marker analysis, (c) authorship verification with unmasking. See Figure 3 for a detailed description of the unmasking step.**

powerful method to discover this fact. Figure 2 shows, in the form of a UML activity diagram, the combination of style marker analysis with subsequent unmasking.

### 3.1 Authorship Verification with Unmasking

In the authorship verification problem, one is given examples  $d_{1_1}, \dots, d_{1_n}$  of the writing of a single author, and one is asked to determine if a given document,  $d_2$ , were or were not written by this author.

For universal applicability we consider the examples  $d_{1_1}, \dots, d_{1_n}$  being combined into a single document  $d_1$ . The basic technology of unmasking is captured in the following procedure (cf. Figure 3):

0. *Chunking and Collection Construction.* Decomposition of  $d_1$  and  $d_2$  into a number of chunks. In [6] Koppel and Schler report on approximately 100 chunks of at least 500 words without breaking up paragraphs. The result of this step are two “collections” of chunks,  $D_1, D_2$ , generated from  $d_1$  and  $d_2$  respectively. The sets  $D_1$  and  $D_2$  are represented under a reduced vector space model, designated as  $\mathbf{D}_1$  and  $\mathbf{D}_2$ . As an initial feature set the 250 words with the highest (relative) frequency in  $D_1 \cup D_2$  are chosen.
1. *Model Fitting.* Training of a classifier that is able to separate  $\mathbf{D}_1$  from  $\mathbf{D}_2$ . Koppel and Schler implement a ten-fold cross-validation experiment using an SVM with a linear kernel to determine the achievable accuracy. Within our analyses logistic regression is applied.
2. *Impairing.* Elimination of the most discriminative features with regard to the model obtained in Step 1, and construction of new collections  $\mathbf{D}_1, \mathbf{D}_2$  which now contain the impaired representations of the chunks. Koppel and Schler achieved convincing results by eliminating the three most strongly-weighted positive features and most strongly-weighted negative features. Note, however, this heuristic depends on the

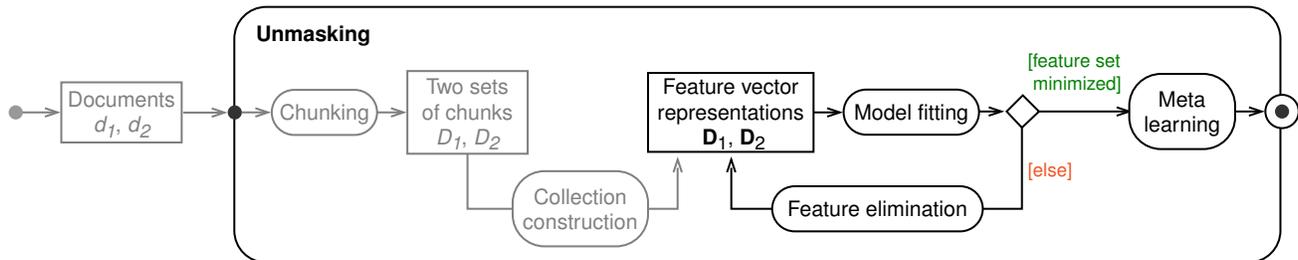
section length which in turn depends on the length of  $d_1$  and  $d_2$ .

3. Go to Step 1 until the feature set is sufficiently reduced. Typically about 5-10 iterations are necessary.
4. *Meta Learning.* Analyze the degradation in the quality of the model fitting process: if after the last impairing step the sets  $\mathbf{D}_1, \mathbf{D}_2$  can still be separated with a small error, assume that  $d_1$  and  $d_2$  stem from different authors.

Unmasking operationalizes following observation: two sets of chunks,  $D_1, D_2$ , constructed from two different documents  $d_1$  and  $d_2$  of the same author can be told apart easily if a vector space model (VSM) representation for the chunks in  $D_1 \cup D_2$  is chosen. The VSM representation considers all words in  $d_1 \cup d_2$ , and hence it includes all kinds of open class and closed class word sets. If only the 250 most-frequent words are selected, a large fraction of them will be function words and stop words.<sup>2</sup> Among these 250 most-frequent words a small number does the major part of the discrimination job. These words may capture topical differences, differences that result from genre or purpose, and the like. By eliminating them we approach step by step the distinctive and subconscious manifestation of an author’s writing style. After several iterations the remaining features are not powerful enough to discriminate two documents of the same author. By contrast, if  $d_1$  and  $d_2$  stem from two different authors, the remaining features will still quantify significant differences between the impaired representations  $\mathbf{D}_1$  and  $\mathbf{D}_2$  of the two chunk sets  $D_1$  and  $D_2$ .

*Remarks.* At heart, unmasking is a representative of what Tax terms “reconstruction methods” in his taxonomy [10]. Unmasking measures the increase of a sequence of reconstruction errors, starting with a good reconstruction which then is more and more

<sup>2</sup>Function words and stop words are not disjoint sets: most function words in fact are stop words; however, the converse does not hold.



**Figure 3: UML activity diagram of the unmasking technology from [6]. Input are two sufficiently large documents,  $d_1, d_2$ , from which two collections  $D_1$  and  $D_2$  are constructed. Basic idea is a meta learning analysis, which quantifies the separability of  $D_1$  and  $D_2$  when the feature representation of the chunks in  $D_1 \cup D_2$  is increasingly impaired.**

impaired. For two documents from the same author the reconstruction error develops differently compared to two documents from two different authors. In their paper Koppel and Schler present also a meta learning procedure to automatically identify the same-author curves, given a large set of unmasking experiments.

### 3.2 Rationale of the Hybrid Approach

Authorship verification and intrinsic plagiarism analysis represent two sides of the same coin. This subsection discusses the similarities and differences and gives the rationale of our hybrid approach.

In an authorship verification problem the interesting document  $d_2$  with the unsettled authorship is explicitly given, and,  $d_2$  is large enough to be analyzed with unmasking. In an intrinsic plagiarism analysis problem the sections in  $d$  for which the authorship is unsettled are unknown. In principle, unmasking could be applied to the decomposition  $s_1, \dots, s_n$  of  $d$ , taking each  $s_i$  in the role of  $d^-$  and the remaining  $d \setminus \{s_i\}$  in the role of  $d^+$ . However, in most cases a single section  $s_i$  is too small to be analyzed with unmasking, and our style marker analysis serves the purpose to construct a  $d^-$  of maximum length.

In this sense the style marker analysis is a heuristic filter (or generator) function that identifies both potentially plagiarized and sufficiently long auxiliary documents  $d^-$ . The underlying search space is the set of all subsets of a document  $d$ . Let  $k, k < n$ , denote the minimum number of sections that must be chosen from a decomposition  $s_1, \dots, s_n$  of  $d$  in order to obtain an auxiliary documents  $d^-$  of sufficient length. With  $\theta$  as the plagiarized portion of  $d$ ,  $k' = \lceil \theta \cdot n \rceil$  defines an upper bound for the number of sections that can be plagiarized at all. Hence, a brute-force analysis of  $d$  had to investigate  $r$  auxiliary documents, with

$$r = \binom{n}{k} + \dots + \binom{n}{k'}, \quad k < k'$$

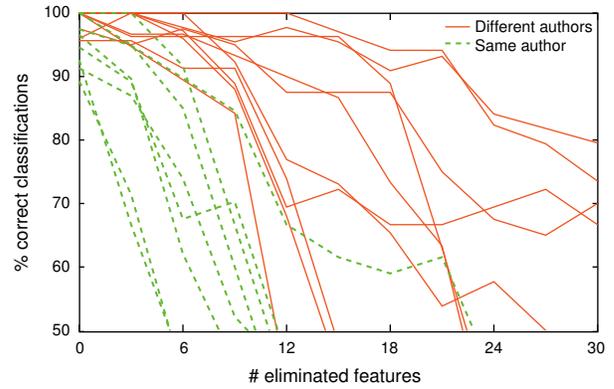
An unmasking analysis of  $r$  document pairs will not be tractable in most cases, which shows the finesse of the hybrid approach: the preceding style marker analysis enables us to concentrate on a very small number of auxiliary documents  $d^-$ .

A further important difference between authorship verification and intrinsic plagiarism analysis relates to impurity. In an authorship verification problem a model of the target class can be learned from the examples  $d_{1_1}, \dots, d_{1_n}$ , each of which belonging definitely to the target class. In an intrinsic plagiarism analysis problem a model of the target class has to be learned from the examples  $s_1, \dots, s_n$  (= document sections), from which only the—a-priori unknown—portion  $1 - \theta$  belongs to the target class.

Note that, from a statistical viewpoint, the reliability of the unmasking analysis depends not only on the length of an auxiliary document  $d^-$  but also on its “purity”, i. e., the precision of the retrieved plagiarized sections. Like before, without a style marker analysis this problem had to be addressed by a complete but intractable brute-force search.

*Related Questions.* Koppel and Schler evaluate their method with twenty-one 19<sup>th</sup> century English books written by ten authors, and they obtain convincing results. However, against the background of intrinsic plagiarism detection several questions arise with respect to the flexibility of the unmasking approach:

1. Does unmasking work for technical and scientific texts or is it primarily suited for novels?
2. What are minimum section lengths in the chunking step?



**Figure 4: Authorship verification with unmasking for short documents of 4-8 pages. Each line corresponds to a comparison of two papers, where each solid red (dashed green) line results from the analysis of papers from two different authors (the same author).**

3. Are the initial feature set and the number of eliminated features in the impairing step independent of document lengths and section lengths?
4. Within the model fitting step a model for the target class is learned. Within an intrinsic plagiarism analysis problem the model fitting for the target class relies on  $d^+$ ; likewise the model fitting for the unknown (outlier or target) class relies on  $d^-$ . How large is the impact of precision and recall that was achieved by the style marker analysis on the model fitting step?

We analyzed these questions within our experiments; one result is shown in Figure 4. Here, short scientific computer science texts formed the analysis base; the average section length in the chunking step was 500 words.

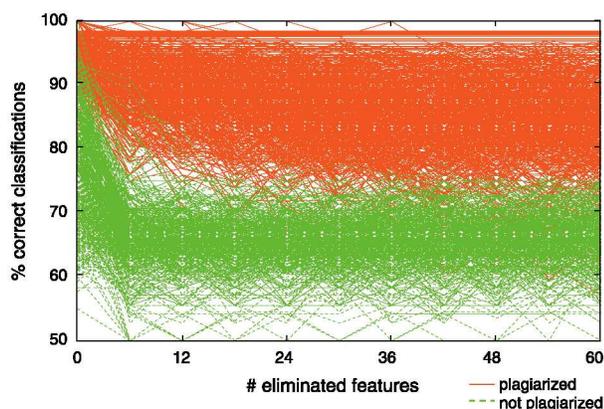
## 4. ANALYSIS

The analysis presented here relates to documents written in German. In the next subsection an analysis of the intrinsic approach according to the outlier detection method of Type (1a) on artificial plagiarism cases is presented, and its results are further refined using a meta learning approach. The next but one subsection reports on a real-world plagiarism case.

### 4.1 Artificial Data

We compiled a corpus of 50 scientific documents from several domains that were downloaded from German universities. Each of these documents (written in German by a single author,) was cut down to 12-15 pages. We plagiarized the documents by hand with up to five sections from other authors. A resulting document with  $k$  plagiarized passages served as a template document from which  $2^k$  instance documents were generated, depending on which of the  $k$  plagiarized passages were actually included in the instance. The resulting instance documents are plagiarized at a portion  $\theta \in [0.05; 0.5]$ .

The first experiment with this corpus analyzes the power of the unmasking technology, illustrated in Figure 5: Each of the red lines shows a learning curve of the plagiarized sections,  $d^-$ , against the remaining document,  $d^+$ . Likewise, a dashed green line shows a learning curve of randomly drawn sections from  $d^+$  against the rest from  $d^+$ .



**Figure 5: Unmasking applied to artificial data.** Each red line shows a learning curve when separating the plagiarized parts of a document ( $= d^-$ ) from the non-plagiarized part ( $= d^+$ ). A dashed green line shows a learning curve when two different non-plagiarized parts of the same document are to be distinguished. The document lengths varied between 10-20 pages.

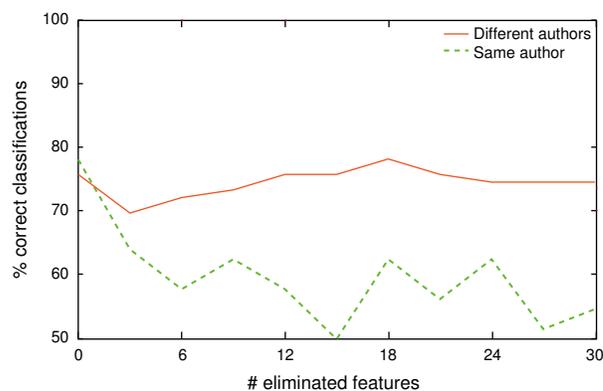
In a second experiment the intrinsic analysis as described in Section 2 was analyzed. For this purpose a classifier based on 20 part-of-speech features and 9 style markers was trained, including simple markers like average sentence length, average syllables per word, average stopword number, as well as specially crafted indexes like the Wiener Sachtextformel index, Amdahl's index, Honore's  $R$ , the Smog index, average German word frequency class, and the Kullback-Leibler divergence of the POS feature distribution. Altogether, the instance documents gave 16,000 vectors for non-plagiarized sections, and 1500 vectors for the plagiarized passages.

The classifier, based on a discriminant analysis, performed acceptably well: the precision and recall values for the non-plagiarized sections were between 80-90%, depending on the portion  $\theta$  of plagiarized passages. The recall of the plagiarized sections was about 70%, having a precision of 55% given that an a priori probability of 50% for the plagiarized and non-plagiarized sections is assumed. Note that these results for an imbalanced set of feature vectors correspond to a realistic setting in which only a fraction  $\theta$  of a suspicious document is plagiarized.

## 4.2 A Real-World Case

Given was a plagiarized postdoctoral thesis from the 1980s. The thesis was scanned, converted to plain text using OCR technology, and decomposed into 138 "natural" sections. The classifier that was outlined in the previous section was applied to generate a  $d^-$ , resulting in 13 suspicious sections. Three of these sections are known to be plagiarized from other textbooks from the 1980s, while the remaining 10 suspicious sections may or may not be plagiarized. Two more passages that are known to be partly plagiarized have not been detected by the classifier; an analysis has shown that the reason for missing these surrounding sections lies in the decomposition, which was too coarse for this purpose.

Figure 6 shows two learning curves for the plagiarism case. The red curve shows the classification rate when the 13 suspicious sections from  $d^-$  are learned against the rest of the thesis,  $d^+$ . The green dashed curve shows the classification rate when the original parts from  $d^+$  are trained against 13 randomly drawn sections from  $d^+$ . The allegedly plagiarized parts can be distinguished from the original parts even when dropping the most important features. According to [6] this is a strong indication for different authors.



**Figure 6: Analysis of a possibly plagiarized habilitation.** The red line shows the learning curve when separating the 13 suspicious sections ( $= d^-$ ) from the rest of the thesis ( $= d^+$ ). The dashed green line shows the learning curve when 13 randomly drawn sections from  $d^+$  are to be distinguished from the rest of  $d^+$ .

## 5. REFERENCES

- [1] J. Chall and E. Dale. *Readability Revisited: The new Dale-Chall Readability Formula*. Brookline Books, 1995.
- [2] E. Dale and J. Chall. A formula for predicting readability. *Educ. Res. Bull.*, 27, 1948.
- [3] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 1948.
- [4] A. Honore. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177, 1979.
- [5] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8Ú75 Millington TN: Naval Technical Training US Naval Air Station, 1975.
- [6] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004. ACM Press.
- [7] J. Mansfield. Textbook plagiarism in psy101 general psychology: incidence and prevention. In *Proceedings of the 18th Annual Conference on Undergraduate teaching of psychology: ideas and innovations*, SUNY Farmingdale, New York, USA, 2004.
- [8] S. Meyer zu Eissen and B. Stein. Intrinsic plagiarism detection. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsirikika, and A. Yavlinisky, editors, *Proceedings of the European Conference on Information Retrieval (ECIR 2006)*, volume 3936 of *Lecture Notes in Computer Science*, pages 565–569. Springer, 2006.
- [9] S. Meyer zu Eissen, B. Stein, and M. Kulig. Plagiarism Detection without Reference Collections. In R. Decker and H. Lenz, editors, *Advances in Data Analysis*, pages 359–366. Springer, 2007.
- [10] D. Tax. *One-Class Classification*. PhD thesis, Technische Universiteit Delft, 2001.
- [11] G. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.