Language Resources and Evaluation
DOI: 10.1007/s10579-010-9115-y

Intrinsic Plagiarism Analysis

Benno Stein · Nedim Lipka · Peter Prettenhofer

Published: January 20, 2010

Abstract Research in automatic text plagiarism detection focuses on algorithms that compare suspicious documents against a collection of reference documents. Recent approaches perform well in identifying copied or modified foreign sections, but they assume a closed world where a reference collection is given. This article investigates the question whether plagiarism can be detected by a computer program if no reference can be provided, e.g., if the foreign sections stem from a book that is not available in digital form. We call this problem class *intrinsic plagiarism analysis*; it is closely related to the problem of authorship verification.

Our contributions are threefold. (*i*) We organize the algorithmic building blocks for intrinsic plagiarism analysis and authorship verification and survey the state of the art. (*ii*) We show how the meta learning approach of Koppel and Schler, termed "unmasking", can be employed to post-process unreliable stylometric analysis results. (*iii*) We operationalize and evaluate an analysis chain that combines document chunking, style model computation, one-class classification, and meta learning.

Keywords Plagiarism detection · Authorship verification · Stylometry · One-class classification

1 Problem Statement

In the following, the term plagiarism refers to *text* plagiarism, i.e., the use of another author's information, language, or writing, when done without proper acknowledgment of the original source. Plagiarism detection refers to the unveiling of text plagiarism. Existing approaches to *computer-based* plagiarism detection break down this task into manageable parts:

Faculty of Media, Media Systems
Bauhaus-Universität Weimar
99421 Weimar, Germany
E-mail: <first.last>@uni-weimar.de

"Given a text d and a reference collection D, does d contain a section s for which one can find a document $d_i \in D$ that contains a section s_i such that under some retrieval model $\mathcal{R}$ the similarity $\varphi_{\mathcal{R}}$ between s and s_i is above a threshold θ?"

Observe that research on automated plagiarism detection presumes a closed world where a reference collection D is given. Since D can be extremely large—possibly the entire indexed part of the World Wide Web—the main research focus is on efficient search technology: near-similarity search and near-duplicate detection [3, 19, 2, 16, 18, 65], tailored indexes for near-duplicate detection [10, 2, 4], or similarity hashing techniques [28, 23, 12, 54, 55]. This article, however, deals with technology to identify plagiarized sections in a text if no reference collection is given. We distinguish the two analysis challenges as external and intrinsic analysis respectively. Note that human readers are able to identify plagiarism without having a reference collection at their disposal: changes between brilliant and baffling passages, or the change of person narrative give hints to multiple authorship.

1.1 Intrinsic Plagiarism Analysis and Authorship Verification

Intrinsic plagiarism analysis is closely related to authorship verification: goal of the former is to identify potential plagiarism by analyzing a document with respect to undeclared changes in writing style. Similarly, in an authorship verification problem one is given writing examples of an author A, and one is asked to determine whether or not a text with doubtful authorship is also from A. Intrinsic plagiarism analysis can be understood as a more general form of the authorship verification problem:

1. one is given a single document only, and
2. one is faced with the problem of finding the suspicious sections.

Intrinsic plagiarism analysis and authorship verification are one-class classification problems. A one-class classification problem defines a target class for which a certain number of examples exist. Objects outside the target class are called outliers, and the classification task is to tell apart outliers from target class members. Actually, the set of "outliers" can be much bigger than the target class, and an arbitrary number of outlier examples could be collected. Hence a one-class classification problem may look like a two-class discrimination problem, but there is an important difference: members of the target class can be considered as representatives for their class, whereas one will not be able to compile a set of outliers that is representative for some kind of "non-target class". This fact is rooted in the huge number and the diversity of possible non-target objects. Put another way, solving a one-class classification problem means to learn a concept (the concept of the target class) in the absence of discriminating features. However, in rare cases, knowledge about outliers can be used to construct representative counter examples related to the target class. Then a standard discrimination strategy can be followed.

1.2 Decision Problems

Within the classical authorship verification problem the target class is comprised of writing examples of a known author A, and each piece of text written by an author B, $B \neq A$, is considered as a (style) outlier. Intrinsic plagiarism analysis is an intricate variant of authorship verification, imposing particular constraints and assumptions on the availability of writing style examples. To organize existing research we introduce the following authorship verification problems, formulated as decision problems.

1. **Problem.** AVEXTERN
   Given. A text d, written by author A, and a set of texts, $D = \{d_1, \ldots, d_n\}$, written by authors $\mathbf{B}$, $A \notin \mathbf{B}$.
   Question. Does d contain a section whose similarity to a section in d_i, $d_i \in D$, is above a threshold θ?

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

3. **Problem.** AVOUTLIER
   Given. A set of texts $D = \{d_1, \ldots, d_n\}$, written by author A, and a text d, allegedly written by author A.
   Question. Is d written by an author B, $B \neq A$?

The problem class AVEXTERN corresponds to the external plagiarism analysis problem mentioned at the outset; the problem class AVFIND corresponds to the general intrinsic plagiarism analysis problem, and the problem class AVOUTLIER corresponds to the classical authorship verification problem. An instance π of AVFIND can be reduced to m instances of AVOUTLIER, AVFIND $\leq^p_{tt}$ AVOUTLIER, by applying a canonical chunking strategy that splits a document into m sections while asking for each section whether it forms an outlier or not. If at least one instance of AVOUTLIER is answered with yes, the answer to π is yes.[1] Likewise, an instance π of AVOUTLIER can be reduced to an instance of AVFIND, AVOUTLIER $\leq$ AVFIND, by simply merging d and all documents in D into a single document. The different complexity of the problem classes is reflected by the reductions $\leq^p_{tt}$ and $\leq$.

If the answer to an instance π of AVFIND is given via a reduction of π to m AVOUTLIER problems, one can try to raise the evidence of this answer by a post-processing step: from the m potential outlier sections two sets D_1 and D_2 are formed, comprising those sections that have been classified as targets into one set, and those that have been classified as outliers into the other. Again, we ask whether the documents in these two sets are written by a single author, this time applying an analysis method which takes advantage of the two sample sets, D_1, D_2, and which hence is more reliable than the outlier analysis. Since this decision problem is important from an algorithmic viewpoint we introduce a respective problem class:

[1] The reduction $\leq^p_{tt}$ is in $O(|d|^2)$; within this time all possible outliers can be constructed for a document d. The reduction $\leq^p_{tt}$ computes the answer to AVFIND from the m answers to AVOUTLIER by means of a truth table tt, which is a disjunction here.

3.′ **Problem.** AVBATCH
Given. A set of texts $D_1 = \{d_{1_1}, \ldots, d_{1_k}\}$ written by author A, and a second set of texts, $D_2 = \{d_{2_1}, \ldots, d_{2_l}\}$, allegedly written by author A.
Q. Does D_2 contain a text written by an author B, $B \neq A$?

Obviously AVOUTLIER and AVBATCH can be reduced to each other in polynomial time, hence AVOUTLIER $\equiv$ AVBATCH. However, it is important to note that both reductions, AVFIND $\leq^p_{tt}$ AVOUTLIER and AVOUTLIER $\leq$ AVBATCH, are constrained by a minimum text length that is necessary to perform a sensible style analysis. Experience shows that a style analysis becomes statistically unreliable for text lengths below 250 words [56].

1.3 Existing Research

Authorship analysis divides into authorship *verification* problems and authorship *attribution* problems. The by far larger part of the research addresses the attribution problem: given a document d of unknown authorship and a set D of candidate authors with writing examples, and one is asked to attribute d to one author. In a verification problem (see above) one is given writing examples of an author A, and one is asked to verify whether or not a document d of unknown authorship in fact is written by A. Recent contributions to the authorship attribution problem include [46, 50, 6, 24, 35, 49, 52, 51]; the authorship verification problem is addressed in [31, 63, 39, 33, 15, 40, 56, 58, 43].

Several research areas are related to authorship verification, in particular: (*i*) stylometry, i.e., the construction of models for the quantification of writing style, text complexity, and grading level assessment, (*ii*) outlier analysis and meta learning [60, 61, 36, 44, 30, 31, 32], and (*iii*) symbolic knowledge processing, i.e., knowledge representation, deduction, and heuristic inference [47, 53].

In their excellent paper from 2004 Koppel and Schler give an illustrative discussion of authorship verification as a one-class classification problem [31]. At the same place they introduce the unmasking approach to determine whether a set of writing examples is a subset of the target class. Observe the term "set" in this connection: unmasking does not solve the one-class classification problem for a single object but requires a batch of objects all of which must stem either from the target class or not.

2 Building Blocks to Operationalize Authorship Verification

Plagiarism detection can be operationalized by decomposing a document into natural sections, such as sentences, chapters, or topically related blocks, and analyzing the variance of stylometric features for these sections. In this regard the decision problems in Subsection 1.2 are of decreasing complexity: instances of AVFIND are comprised of both a selection problem (finding suspicious sections) and an AVOUTLIER problem; instances of AVBATCH are a restricted variant of AVOUTLIER since one has the additional knowledge that all elements of a batch are (or are not) outliers at the same time.

Table 1 Building blocks to operationalize authorship verification. The first column lists pre-analysis methods, the second to the fourth column list the modeling and classifier methods which form the heart of a verification process, and the last column lists post-processing methods to improve the analysis quality. The highlighted building blocks indicate the employed technology of the analysis chain in this article.

Impurity assessment	Decomposition strategy	Style model construction	Outlier identification	Outlier post-processing
Document length analysis	Uniform length	Lexical character features	One-class density estimation	Heuristic voting
Genre Analysis	Structural boundaries	Lexical word features	One-class boundary estimation	Citation analysis
Analysis of issuing institution	Text element boundaries	Syntactical features	One-class reconstruction	Human inspection
	Topical boundaries	Structural features	Two-class discrimination	Unmasking
	Stylistic boundaries	Language modeling		Qsum
				Batch means

Solving instances of AVFIND involves various subtasks; Table 1 organizes them as building blocks—from left to right—following the logical text processing chain. Among others the building blocks denote alternative decomposition strategies, alternative style models, alternative classification technology, as well as post-processing options whose objective is to improve the analysis' overall precision and recall. The table highlights those building blocks that are combined in our analysis chain; the following subsections discuss them in greater detail. Note that even with a skillful combination and adaptation of these building blocks it is pretty difficult to end up with an analysis process comparable to the power of a human reader.

2.1 Impurity Assessment

How likely is the fact that a document d contains a section of another author? We expect that the lengths, the places, and the entire fraction θ of such sections depend on particular document characteristics. Hence it makes sense to analyze the document type (paper, dissertation), its genre (novel, factual report, research, dictionary entry), but also the issuing institution (university, company, public service). Algorithmic means to reveal such information interpret document lengths, genres, and occurring named entities.

2.2 Decomposition Strategy

The simplest strategy is to decompose a text d into sections $s_1, \ldots, s_n$ of uniform length; in [39] the authors integrate an additional sentence detection. However, a more sensible interpretation of structural boundaries (chapters, paragraphs) is possible, which may consider special text elements like tables, formulas, footnotes, or quotations as well [45]. Though quite difficult, the detection of topical boundaries

Table 2 Stylometric features. Compilation of important and well-known features used within a stylometric analysis. Features that are implemented within our style model are marked with an asterisk.

Stylometric feature			**Reference**
Lexical features (character-based)		Character frequency	[67]
	*	Character n-gram frequency/ratio	[27, 48, 24, 34]
		Frequency of special characters ('(', '&', '/', etc.)	[67]
		Compression rate	[51]
Lexical features (word-based)	*	Average word length	[20, 67]
		Average sentence length	[20, 67]
	*	Average number of syllables per word	[20]
		Word frequency	[42, 20, 34]
		Word n-grams frequency/ratio	[48]
		Number of hapax legomena	[62, 67]
		Number of hapax dislegomena	[62, 67]
		Dale-Chall index	[9, 5]
	*	Flesch Kincaid grade level	[11, 26]
	*	Gunning Fog index	[14]
	*	Honore's R measure	[22, 62, 67]
		Sichel's S measure	[62, 67]
	*	Yule's K measure	[66, 20, 62, 67]
		Type-token ratio	[66, 20, 67]
	*	Average word frequency class	[38]
Syntactic features		Part-of-speech	[51, 34]
	*	Part-of-speech n-gram frequency/ratio	[30, 34]
	*	Frequency of function words	[42, 20, 1, 30, 67, 34]
		Frequency of punctuations	[67]
Structural features		Average paragraph length	[67]
		Indentation	[67]
		Use of greetings & farewells	[67, 51]
		Use of signatures	[67, 51]

has a significant impact on the usefulness of a decomposition [8]. In [13] the authors even try to identify stylistic boundaries.

2.3 Style Model Construction

The statistical analysis of literary style is called stylometry, and the first ideas date back to 1851 [21]. The automation of this task requires a quantifiable style model, and efforts in this direction became a more active research field in the 1930s [68, 66, 11]. In the meantime various stylometric features, also termed style markers, have been proposed. They measure writer-specific aspects like vocabulary richness [22, 66], text complexity and understandability [11], or reader-specific grading levels that are necessary to understand a text [9, 26, 5]. Note that the mentioned style features have been developed to judge longer texts, ranging from a few pages up to book size.

Style model construction must consider the decomposition strategy: different stylometric features have different strengths and also pose different constraints on text length, text genre, or topic variation. Since text plagiarism typically relates to sec-

tions that are shorter than a single page [37], the decomposition of a document into sections $s_1, \ldots, s_n$ must not be too coarse, and, it is questionable which of the stylometric features will work for short sections. It should be clear that style features that employ measures like average paragraph length are not reliable in general. The authors in [40] investigate the robustness of the vocabulary richness measures Yule's K, Honore's R, and the average word frequency class. They observe that 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. In [39] connections of this type have been analyzed for the Flesch Kincaid Grade Level [11, 26], the Dale-Chall formula [9, 5], Yule's K [66], Honore's R [22], the Gunning Fog index [14], and the averaged word frequency class [38].

Table 2 compiles an overview of important stylometric features that have been proposed so far; we distinguish between lexical features (character-based and word-based), syntactic features, and structural features. Our overview is restricted to the well-known style features and omits esoteric variants. Those features marked with an asterisk have been reported to be particularly discriminative for authorship analysis and are used within our stylometric analysis.

2.4 Outlier Identification

The decomposition of a document d gives a sequence $s_1, \ldots, s_n$ of sections, for which the computation of a style model gives a sequence $\mathbf{s}_1, \ldots, \mathbf{s}_n$ of feature vectors, which in turn are analyzed with respect to outliers. The identification of outliers among the $\mathbf{s}_i$ has to be solved on the basis of positive examples only and hence poses a one-class classification problem. Following Tax, one-class classification approaches fall into one of the following three classes [60]:

(a) Density methods, which directly estimate the probability distributions of features for the target class. Outliers are assumed to be uniformly distributed, and, for example, Bayes' rule can be applied to separate outliers from target class members.

(b) Boundary methods, which avoid the estimation of the multi-dimensional density function but try to define a boundary around the set of target objects. The boundary computation is based on the distances between the objects in the target set.

(c) Reconstruction methods come into play if prior knowledge for the generation process of target objects is available. Outliers can be be distinguished from targets because of the higher reconstruction error they incur during the model fit.

The main advantage of boundary methods, namely to get by without assessing the multi-dimensional density function, can also be achieved with a density-based approach under Naive Bayes. Moreover, for our domain it is not clear how a boundary around the target set should be defined. We have also developed and analyzed reconstruction methods that rely on factor analysis and principal component analysis, but experienced difficulties due to unsatisfactory generalization behavior. Here, within our analysis chain, we resort to a one-class classifier of Type (a), which is outlined in the following.

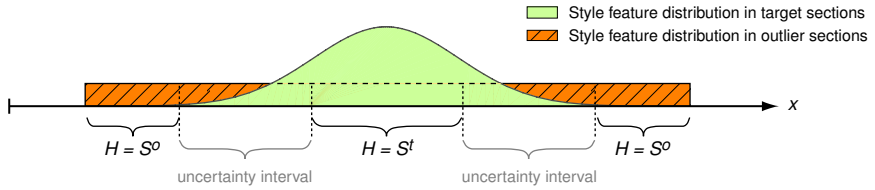

Figure 1 Targets and outliers can be separated if they are differently distributed.

Let S^t denote the event that a section $s \in \{s_1, \ldots, s_n\}$ belongs to the target group (= not plagiarized); likewise, let S^o denote the event that s belongs to the outlier group (= plagiarized). Given a document d and a single style features x, the maximum a-posteriori hypothesis $H \in \{S^t, S^o\}$ can be determined with Bayes' rule:

$$H = \underset{S \in \{S^t, S^o\}}{\operatorname{argmax}} \frac{P(x(s) \mid S) \cdot P(S)}{P(x(s))} \tag{1}$$

where $x(s)$ denotes the style features value for section s, and $P(x(s) \mid S^t)$ and $P(x(s) \mid S^o)$ denote the respective conditional probabilities that $x(s)$ is observed in the target group or the outlier group. Since the fraction of outliers is small compared to all sections it is sensible to estimate the $P(x(s) \mid S^t)$ with a Gaussian distribution; the expectation and the variance for x are estimated from $x(s_1), \ldots, x(s_n)$, omitting those sections s_i that maximize or minimize $x(s_i)$. The outliers can stem from different authors, and hence the $P(x(s) \mid S^o)$ are estimated with a uniform distribution, following a least commitment consideration [60]. See Figure 1 for an illustration of the assumed style feature distributions in target and outlier sections. The priors $P(S^t)$ and $P(S^o)$ correspond to $1-\theta$ and θ respectively and require an impurity assessment (see Subsection 2.1). If no information about θ is available a uniform distribution is assumed for the priors, i.e., we resort to the maximum likelihood estimator.

Multiple style features $x_1, \ldots, x_m$ 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 follows:

$$H = \underset{S \in \{S^o, S^t\}}{\operatorname{argmax}} P(S) \cdot \prod_{i=1}^{m} P(x_i(s) \mid S) \tag{2}$$

For the maximum a-posteriori decision (2) only those style features x are considered whose values fall outside the uncertainty intervals (cf. Figure 1), which are defined by 1.0 and 2.0 times the estimated standard deviation.

2.5 Outlier Post-Processing

The post-processing methods in Table 1 can be distinguished in knowledge-based methods and meta learning approaches. To the former count heuristic voting, citation analysis, and human inspection. Heuristic voting, which is applied here, is the estimation and use of acceptance and rejection thresholds based on the number of classified outlier sections. Meta learning is brought into play if from the solution of

several AVOUTLIER problems two sets D_1 (sections labeled as targets) and D_2 (sections labeled as outliers) are formed, obtaining this way an instance of the AVBATCH problem. Possible meta learning approaches are:

(a) Unmasking [31], which is a representative of what Tax terms "reconstruction method" [60]; it measures the increase of a sequence of reconstruction errors, starting with a good reconstruction which then is successively impaired.
(b) The Qsum heuristic [41, 17], which compares the growth rates of two cumulative sums over a sequence of sentences. Basis for the sums are the deviations from the mean sentence length and the deviations of function word frequencies.
(c) Batch means, which is applied within the analysis of simulation data in order to detect the end of a transient phase. For a series of values the variance development of the sample mean is measured while the sample size is successively increased.

Unmasking has been successfully applied to solve instances of AVBATCH [49, 29, 33, 59]. The robustness of the approach is also reported by Kacmarcik, and Gamon,, who develop methods for obfuscating document stylometry in order to preserve author anonymity [25]. Since unmasking is a building block in our analysis chain it is explained in greater detail now. The use of unmasking for intrinsic plagiarism analysis was proposed in [57], who consider a style outlier analysis as a heuristic to compile a potentially plagiarized and sufficiently large auxiliary document.

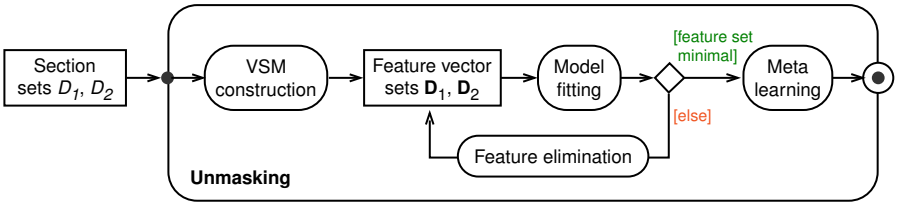

Figure 2 Given are two sets of sections D_1 and D_2, allegedly written by a single author. Unmasking measures the separability of $\mathbf{D}_1$ versus $\mathbf{D}_2$ when the style model is successively impaired.

Recall that the set D_1 (targets) is attributed to author A, while the authorships of the sections in D_2 (outliers) is considered as unsettled. With unmasking we seek further evidence for the hypothesis whether a text in D_2 is written by an author B, $B \neq A$. At first, 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. Unmasking then happens in the following steps (see Figure 2):

1. *Model Fitting.* Training of a classifier that separates $\mathbf{D}_1$ from $\mathbf{D}_2$. In [31] the authors implement a ten-fold cross-validation experiment with a linear kernel SVM to determine the achievable accuracy.
2. *Impairing.* Elimination of the most discriminative features with respect to the model obtained in Step 1; construction of new collections $\mathbf{D}_1$, $\mathbf{D}_2$, which now contain impaired representations. [31] reports on convincing results by eliminat-

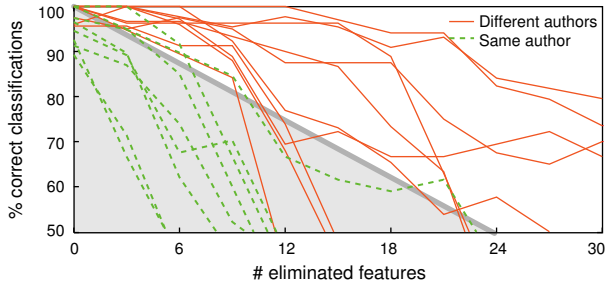

Figure 3 Unmasking at work: each line corresponds to a comparison of two papers. A solid red line belongs to papers of two different authors; a dashed green line belongs to papers of the same author.

ing the six most discriminating features. This heuristic depends on the section length which in turn depends on the length of d.

3. Go to Step 1 until the feature set is sufficiently reduced. About 5-10 iterations are typical.
4. *Meta Learning*. Analyze the degradation in the quality of the model fitting process: if after the last impairing step the sets $\mathbf{D}_1$ and $\mathbf{D}_2$ can still be separated with a small error, assume that d_1 and d_2 stem from different authors. Figure 3 shows a characteristic plot where unmasking is applied to short papers of 4-8 pages.

The rationale of unmasking: Two sets of sections, 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) retrieval model is chosen. The VSM 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.[2] Among these 250 most-frequent words a small number does the major part of the discrimination job; these words capture topical differences, differences that result from genre, purpose, or the like. By eliminating them, one approaches 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. But, if d_1 and d_2 stem from two different authors, the remaining features will still quantify significant differences between $\mathbf{D}_1$ and $\mathbf{D}_2$.

3 Analysis

This section reports on the performance of the operationalized analysis chain. Figure 4 gives an illustration: the top row shows documents with original sections (green), plagiarized sections (red), and sections spotted by the classifier (hashed); the middle row shows the micro- and macro-averaged outlier classification performance; the bottom row shows three alternative post-processing strategies. These strategies

[2] Function words and stop words are not disjunct sets: most function words in fact are stop words; however, the converse does not hold.

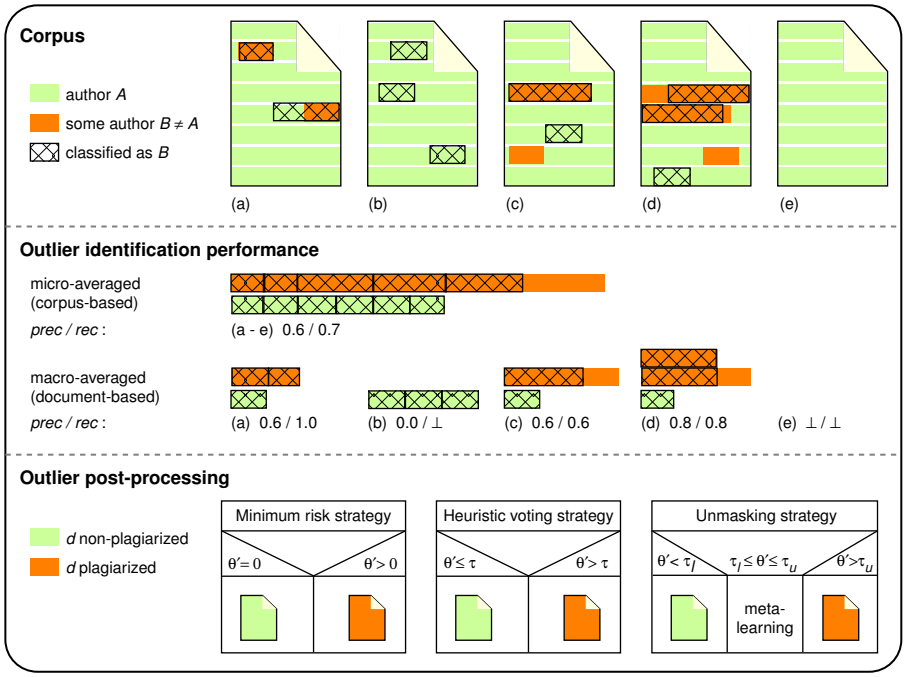

Figure 4 Illustration of the analysis chain. Top: corpus with five documents of author A, containing sections of some author $B \neq A$. Middle: micro- and macro-averaged analysis of the outlier identification performance. Bottom: outlier post-processing according to three alternative strategies; θ' denotes the fraction of sections per document that are classified as outliers.

differ with respect to the interpretation of the fraction θ' of sections per document that are classified as outliers: under the minimum risk strategy a document d is considered as plagiarized if at least one outlier section is spotted, under the heuristic voting strategy θ' is compared to a threshold τ, and under the unmasking strategy meta learning is applied if θ' falls into an uncertainty interval. The remainder of this section gives particulars.

3.1 Corpus

To run analyses on a large scale one has to resort to artificially plagiarized documents. Here, we use a subset of the corpus that has been constructed for the intrinsic plagiarism analysis task of the PAN'09 competition [64]. The PAN'09 corpus comprises about 3000 generated cases of intrinsic plagiarism—more precisely: cases of style contamination—exhibiting varying degrees of obfuscation. The corpus is based on books from the English part of the Project Gutenberg and contains mainly narrative text. Sections of varying length, ranging from a few sentences up to many pages, are inserted into other documents according to heuristic placement rules. In addition,

Table 3 Selected summary statistics of the four test collections. The statistics of the columns 2-5 are per collection and consider both the plagiarized and the non-plagiarized documents; the statistics of the columns 6-8 are per document; the statistics of the columns 6-7 consider both the plagiarized and the non-plagiarized documents, whereas column 8 considers only the plagiarized documents of a collection.

Collection	**# Documents**		**# Sections** (total)		**# Sections** (avg.)		**Impurity** θ (avg.)
	plag.	non-plag.	plag.	non-plag.	plag.	non-plag.	
1	231	231	2067	44316	4.5	96	0.09
2	178	178	451	9560	1.3	27	0.09
3	178	178	4744	21896	13.3	62	0.30
4	188	188	1871	7814	5.0	21	0.33

obfuscation of the inserted sections is performed by replacing, shuffling, deleting, or adding words.[3]

For our experiments the documents of the PAN'09 corpus are uniformly decomposed into candidate sections of 5 000 characters; each candidate section s in turn is categorized as being either non-plagiarized, if s contains no word from an inserted section, or plagiarized, if s consists to more than 50% of an inserted section. Otherwise s is discarded and excluded from further investigations. Documents with less than seven sections are removed from the corpus because they are considered to be too short for a reliable stylometric analysis.

In order to study the effect of document length and impurity on the performance of our analysis chain, four disjoint collections are compiled. For this purpose two levels of document lengths are introduced (short versus long) and combined with two levels of impurity (light versus strong). Short documents consist of less than 250 000 characters, which corresponds to approximately 40 000 words. The impurity θ of a document is defined as the portion of plagiarized characters, i.e., characters that belong to an inserted section. A document is considered to have a light impurity if $\theta \leq 0.15$; it has a strong impurity if $\theta > 0.15$. Finally, the number of plagiarized documents per collection is set to 50%. The resulting test collections exhibit varying degrees of difficulty, both in terms of training data scarcity (document length) and class imbalance (impurity). We number the collections according to their level of difficulty and show selected summary statistics in Table 3.

3.2 Performance of Outlier Identification

Outlier identification is addressed with the density estimation method as described in Section 2.4. To capture a broad range of writing styles a diverse set of stylometric features is employed, belonging to three of the four categories introduced in Section 2.3: lexical character features, lexical word features, and syntactical features. Among the employed stylometric features are the classical measures for vocabulary richness, text complexity, as well as stylometric features that have been reported to be particularly discriminative for authorship analysis, such as character n-grams and the frequency

[3] The corpus can be downloaded at `http://www.webis.de/research/corpora`.

Table 4 Feature ranking. Stylometric features ranked by their F-Measure performance within in a style outlier detection task. The classification decision is given by the maximum a-posterior hypothesis from Equation (1).

Stylometric feature	F**-Measure**
Flesch Reading Ease Score	0.208
Average number of syllables per word	0.205
Frequency of term: of	0.192
Noun-Verb-Noun tri-gram	0.189
Noun-Noun-Verb tri-gram	0.182
Verb-Noun-Noun tri-gram	0.179
Gunning Fog index	0.179
Yule's K measure	0.176
Flesch Kincaid grade level	0.175
Average word length	0.173
Noun-Preposition-ProperNoun tri-gram	0.173
Honore's R measure	0.165
Average word length	0.165
Average word frequency class	0.162
Consonant-Vowel-Consonant tri-gram	0.154
Frequency of term: is	0.151
Noun-Noun-CoordinatingConjunction tri-gram	0.150
NounPlural-Preposition-Determiner tri-gram	0.149
Determiner-NounPlural-Preposition tri-gram	0.148
Consonant-Vowel-Vowel tri-gram	0.146
Verb-Noun-Verb tri-gram	0.146
Vowel-Vowel-Consonant tri-gram	0.146
Frequency of term: the	0.141
Determiner-Noun-Preposition tri-gram	0.139
Frequency of term: been	0.136
Noun-Noun-Noun tri-gram	0.134
Noun-Preposition-Determiner tri-gram	0.133
Vowel-Vowel-Vowel tri-gram	0.129
Noun-Preposition-Noun	0.128
Verb-Preposition-Determiner tri-gram	0.127

of function words (see Table 2). To capture syntactic variations in writing style, part-of-speech information in the form of part-of-speech trigrams is exploited; the tagging is done with the probabilistic part-of-speech tagger QTAG.

Table 4 shows the top 30 stylometric features with respect to their discriminative power; the F-Measure-value pertains to the outlier class and is computed as micro-averaged mean over the four collections. The decision whether or not a section is classified as an outlier is given by the maximum a-posteriori hypothesis of the univariate model in Equation (1). Note that this ranking serves merely for illustration purposes and is not used for feature selection: the outlier analysis in the analysis chain is based on the multivariate use of all stylometric features. For each document in a collection an individual style classifier according to Equation (2) is constructed and applied to each section of that document. The correctness of each classification decision is pooled over all documents. Table 5 summarizes the achieved classification results in terms of micro-averaged F-Measure for both the outlier class and the target class.

Table 5 Performance of the one-class classifier. The target class relates to sections of author A; the outlier class relates to sections of foreign authors $B \neq A$.

Collection	Target class			Outlier class		
	$prec$	rec	F	$prec$	rec	F
1	0.98	0.91	0.94	0.20	0.52	0.29
2	0.89	0.90	0.89	0.34	0.32	0.33
3	0.98	0.64	0.77	0.10	0.78	0.18
4	0.89	0.64	0.74	0.27	0.64	0.38

Recall that the four collections are compiled in a way that sections with less than 50% plagiarism are discarded. If all sections with less than 90% plagiarism are discarded, the precision of the outlier class is unchanged, but its recall increases by 9% on average over all collections. On the other hand, if sections with less than 50% plagiarism are kept, the precision and the recall of the outlier class decrease by 4% on average.

3.3 Performance of Meta Learning

To illustrate the performance of the unmasking approach we evaluate the meta learner that is used in Step 4 of the unmasking procedure. Unmasking is parameterized as follows: documents are represented under the term frequency vector space model, defined by the 500 most frequent words of the input document sets, without applying stemming or stop wording. In each iteration i of 30 unmasking iterations the best 10 features according to the information gain heuristic are removed and the classification accuracy, acc_i, of a linear kernel SVM is computed, based on 5-fold cross validation.

In practice the distribution of the outlier and target class is extremely unbalanced. In order to correct this class imbalance, the outlier class is over-sampled. Here, the SMOTE approach is used to create new, synthetic instances of the outlier class by interpolating between the original instances [7]. A meta learner is trained with vectors each of which comprising the following elements: the acc-values of iteration i, the Δ-acc-values to iteration $i-1$, the Δ-acc-values to iteration $i-2$, and a class label "plagiarized" or "non-plagiarized". This meta learner is also realized as a linear kernel SVM; Table 6 reports on its performance.

Table 6 Evaluation of the unmasking meta learner. Setting: 10-fold cross validated with 100 plagiarized documents and 100 non-plagiarized documents drawn randomly from the corresponding collection.

Collection	Non-plagiarized documents			Plagiarized documents		
	$prec$	rec	F	$prec$	rec	F
1	0.78	0.86	0.82	0.82	0.73	0.77
2	0.77	0.88	0.82	0.48	0.30	0.37
3	0.95	0.94	0.95	0.94	0.95	0.95
4	0.70	0.69	0.70	0.68	0.70	0.69

Table 7 Overall performance of the solution of the AVFIND problem under different strategies: minimum risk (columns 2-4), heuristic voting (columns 5-8), and unmasking (columns 9-12).

Collection	**Minimum risk**			**Heuristic voting**				**Unmasking**			
	$prec$	rec	F	τ	$prec$	rec	F	$[\tau_l; \tau_u]$	$prec$	rec	F
1	0.50	1.00	0.66	0.1	0.55	0.57	0.63	[0.1; 0.5]	**0.83**	0.50	0.62
2	0.50	1.00	0.66	0.1	0.50	1.00	0.66	[0.1; 0.5]	**0.66**	0.57	0.67
3	0.50	1.00	0.66	0.2	0.69	0.30	0.42	[0.2; 0.8]	**0.72**	0.30	0.43
4	0.50	1.00	0.66	0.2	0.52	0.97	0.68	[0.2; 0.8]	**0.98**	0.60	0.74

The unmasking approach of Koppel and Schler decides for two sets of documents whether or not all documents stem from a single author. If both sets belong to the same author the associated unmasking curve drops away (cf. the dashed green lines in Figure 3). This fact is exploited within our analysis chain in order to reduce the number of misclassified non-plagiarized documents, which are caused by the insufficient precision of the one-class classifier.

3.4 Performance of the Analysis Chain

We evaluate three strategies, from naive to sophisticated, to solve AVFIND for a document d. Under the minimum risk strategy d is classified as plagiarized if at least one style outlier has been announced for d. Under the heuristic voting strategy d is classified as plagiarized if the detected fraction of outlier text is above a threshold τ. Under the unmasking strategy d is classified as plagiarized if the detected fraction of outlier text is above an upper threshold τ_u; d is classified as non-plagiarized if the detected fraction of outlier text is below a lower threshold τ_l; for all other cases unmasking is applied. Note that the values for τ, τ_u, and τ_l are collection-dependent. In our experiments τ and τ_l are fitted to the averaged impurities of the collections, while τ_u is chosen overly optimistic. Table 7 summarizes the results: the minimum risk strategy classifies all documents as plagiarized because of the imprecision of the outlier detection, which claims at least one section in each document as outlier. Heuristic voting and unmasking consider the outlier detection characteristic. A main observation is that especially unmasking can be used to substantially increase the precision when solving instances of AVFIND.

4 Summary

Intrinsic plagiarism detection is the spotting of sections with undeclared writing style changes in a text document. Intrinsic plagiarism detection is a one-class classification problem that cannot be tackled with a single technique but requires the combination of algorithmic and statistical building blocks. Our article provides an overview of these building blocks and presents ideas to operationalize analysis chains that cope with the intrinsic plagiarism challenge.

Intrinsic plagiarism detection and authorship verification are two sides of the same coin. This fact is explained in this article, and, in order to organize existing

research and to work out the intricate difficulties between problem variants, we introduce four problem classes for authorship verification problems. We propose and implement an analysis chain that integrates document chunking, style model computation, style outlier identification, and outlier post-processing. Style outlier identification is unreliable, among others because it is difficult to quantify style and to spot style changes in short sections. Since we feel that plagiarism detection technology should avoid the announcement of wrongly claimed plagiarism at all costs, we propose to post-process the results of the outlier identification step. We employ the unmasking technology for this purpose, which has been developed to settle the authorship for a text in question—if sufficient sample text is at one's disposal. The combination of outlier identification with unmasking entails a significant improvement of the precision (see Table 7 for details). However, we see different places and room to improve certain building blocks in the overall picture, among others: knowledge-based chunking, better style models, multivariate one-class classification, and bootstrapping for outlier identification.

References

1. Shlomo Argamon, Marin šarić, and Sterling S. Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480, New York, NY, USA, 2003. ACM. ISBN 1-58113-737-0. .
2. Y. Bernstein and J. Zobel. A scalable system for identifying co-derivative documents. In A. Apostolico and M. Melucci, editors, *Proceedings of the String Processing and Information Retrieval Symposium (SPIRE)*, pages 55–67, Padova, Italy, September 2004. Springer. Published as LNCS 3246.
3. Sergey Brin, James Davis, and Hector Garcia-Molina. Copy detection mechanisms for digital documents. In *SIGMOD '95*, pages 398–409, New York, NY, USA, 1995. ACM Press. ISBN 0-89791-731-6.
4. Andrei Z. Broder, Nadav Eiron, Marcus Fontoura, Michael Herscovici, Ronny Lempel, John McPherson, Runping Qi, and Eugene J. Shekita. Indexing Shared Content in Information Retrieval Systems. In *EDBT '06*, pages 313–330, 2006.
5. J.S. Chall and E. Dale. *Readability Revisited: The new Dale-Chall Readability Formula*. Brookline Books, 1995.
6. Carole E. Chaski. Who's at the keyboard? authorship attribution in digital evidence investigations. *IJDE*, 4(1), 2005.
7. Nitesh V. Chawla, Kevin W. Bowyer, and Philip W. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
8. Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 26–33, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
9. E. Dale and J.S. Chall. A formula for predicting readability. *Educational Research Bulletin*, 27, 1948.
10. Raphael A. Finkel, Arkady Zaslavsky, Krisztian Monostori, and Heinz Schmidt. Signature Extraction for Overlap Detection in Documents. In *Proceedings of the 25th Australian conference on Computer science*, pages 59–64. Australian Computer Society, Inc., 2002. ISBN 0-909925-82-8.
11. R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 1948.
12. Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity Search in High Dimensions via Hashing. In *Proceedings of the 25th VLDB Conference Edinburgh, Scotland*, pages 518–529, 1999.
13. Neil Graham, Graeme Hirst, and Bhaskara Marthi. Segmenting a document by stylistic character. *Natural Language Engineering*, 11(4):397–415, December 2005. Supersedes August 2003 workshop version.

14. R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.
15. Hans Van Halteren,. Author verification by linguistic profiling: An exploration of the parameter space. *ACM Trans. Speech Lang. Process.*, 4(1):1, 2007. ISSN 1550-4875.
16. Monika Henzinger. Finding Near-Duplicate Web Pages: a Large-Scale Evaluation of Algorithms. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 284–291, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-369-7. .
17. Michael L. Hilton and David I. Holmes. An assessment of cumulative sum charts for authorship attribution. *Literary and Linguistic Computing*, 8(2), 1993.
18. G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313:504–507, July 2006.
19. Timothy C. Hoad and Justin Zobel. Methods for Identifying Versioned and Plagiarised Documents. *American Society for Information Science and Technology*, 54(3):203–215, 2003.
20. David I. Holmes. The evolution of stylometry in humanities scholarship. *Lit Linguist Computing*, 13 (3):111–117, September 1998. .
21. David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3), 1998.
22. A. Honore. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177, 1979.
23. Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbor—Towards Removing the Curse of Dimensionality. In *Proceedings of the 30th Symposium on Theory of Computing*, pages 604–613, 1998.
24. Patrick Juola,. Authorship attribution. *Found. Trends Inf. Retr.*, 1(3):233–334, 2006. ISSN 1554-0669.
25. Gary Kacmarcik, and Michael Gamon,. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 444–451, Morristown, NJ, USA, 2006. Association for Computational Linguistics.
26. J. Kincaid, R.P. Fishburne, R.L. Rogers, and B.S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75 Millington TN: Naval Technical Training US Naval Air Station, 1975.
27. Bradley Kjell, Addison W. Woods, and Ophir Frieder. Discrimination of authorship using visualization. *Inf. Process. Manage.*, 30(1):141–150, 1994. ISSN 0306-4573.
28. J. Kleinberg. Two Algorithms for Nearest-Neighbor Search in High Dimensions. In *STOC '97: Proceedings of the Twenty-Ninth annual ACM symposium on Theory of computing*, 1997.
29. Moshe Koppel, and Jonathan Schler,. Authorship verification as a one-class classification problem. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 62, New York, NY, USA, 2004. ACM.
30. Moshe Koppel and Jonathan Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, 2003.
31. Moshe Koppel and Jonathan Schler. Authorship Verification as a One-Class Classification Problem. In *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004. ACM Press.
32. Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7.
33. Moshe Koppel,, Jonathan Schler,, and Elisheva Bonchek-Dokow,. Measuring differentiability: Unmasking pseudonymous authors. *J. Mach. Learn. Res.*, 8:1261–1276, 2007. ISSN 1533-7928.
34. Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009.
35. M. B. Malyutov. Authorship attribution of texts: A review. *Lecture Notes in Computer Science*, 2063:362–380, 2006.
36. Larry M. Manevitz and Malik Yousef. One-Class SVMs for Document Classification. *Journal of Machine Learning Research*, 2:139–154, 2001.
37. J. S. 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.
38. Sven Meyer zu Eißen and Benno Stein. Genre classification of web pages: User study and feasibility analysis. In Susanne Biundo, Thom Frühwirth, and Günther Palm, editors, *KI 2004: Advances in Artificial Intelligence*, volume 3228 LNAI of *Lecture Notes in Artificial Intelligence*, pages 256–269, Berlin Heidelberg New York, September 2004. Springer. ISBN 0302-9743.
39. Sven Meyer zu Eissen and Benno Stein. Intrinsic plagiarism detection. In Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, *Proceedings of the European Conference on Information Retrieval (ECIR 2006)*, volume 3936 of *Lecture Notes in Computer Science*, pages 565–569. Springer, 2006. ISBN 3-540-33347-9.
40. Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. Plagiarism Detection without Reference Collections. In Reinhold Decker and Hans J. Lenz, editors, *Advances in Data Analysis*, pages 359–366. Springer, 2007. ISBN 978-3-540-70980-0.
41. A. Q. Morton and S. Michaelson. The QSUM plot. Technical report, University of Edinburgh, 1990.
42. Frederick Mosteller and D. L. Wallace. *Inference and Disputed Authorship: Federalist Papers*. Addison-Wesley Educational Publishers Inc, 1964. ISBN 0-201-04865-5.
43. Daniel Pavelec, Luiz S. Oliveira, Edson J. R. Justino, and Leonardo Vidal Batista. Using conjunctions and adverbs for author verification. *J. UCS*, 14(18):2967–2981, 2008.
44. Gunnar Ratsch, Sebastian Mika, Bernhard Scholkopf, and Klaus-Robert Muller. Constructing Boosting Algorithms from SVMs: An Application to One-Class Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1184–1199, 2002. ISSN 0162-8828. .
45. Jerey C. Reynar. *Topic Segmentation: Algorithms and Applications*. PhD thesis, University of Pennsylvania, 1998.
46. J. Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31:351–365, 1997.
47. Stuart J. Russel and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, N.J., 1995.
48. C. Sanderson and S. Guenter. On authorship attribution via markov chains and sequence kernels. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 437–440, 2006.
49. Conrad Sanderson and Simon Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, July 2006.
50. E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35:193–214, 2001.
51. Efstathios Stamatatos. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556, 2009. ISSN 1532-2882.
52. Efstathios Stamatatos. Author Identification Using Imbalanced and Limited Training Texts. In A. M. Tjoa and R. R. Wagner, editors, *18th International Conference on Database and Expert Systems Applications (DEXA 07)*, pages 237–241. IEEE, September 2007. ISBN 0-7695-2932-1. .
53. Mark Stefik. *Introduction to Knowledge Systems*. Morgan Kaufmann, 1995.
54. Benno Stein. Fuzzy-Fingerprints for Text-Based Information Retrieval. In Klaus Tochtermann and Hermann Maurer, editors, *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05), Graz*, Journal of Universal Computer Science, pages 572–579. Know-Center, July 2005.
55. Benno Stein. Principles of hash-based text retrieval. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen de Vries, editors, *30th Annual International ACM SIGIR Conference*, pages 527–534. ACM, July 2007. ISBN 978-1-59593-597-7.
56. Benno Stein and Sven Meyer zu Eissen. Intrinsic Plagiarism Analysis with Meta Learning. In Benno Stein, Moshe Koppel, and Efstathios Stamatatos, editors, *SIGIR Workshop Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*, pages 45–50. CEUR-WS.org, July 2007. URL `http://ceur-ws.org/Vol-276`.
57. Benno Stein and Sven Meyer zu Eißen. Topic-Identifikation: Formalisierung, Analyse und neue Verfahren. *KI – Künstliche Intelligenz*, 3:16–22, July 2007. ISSN 0933-1875. URL `http://www.kuenstliche-intelligenz.de/index.php?id=7758`.
58. Benno Stein, Nedim Lipka, and Sven Meyer zu Eißen. Meta Analysis within Authorship Verification. In A. M. Tjoa and R. R. Wagner, editors, *19th International Conference on Database and Expert Systems Applications (DEXA 08)*, pages 34–39. IEEE, September 2008. ISBN 978-0-7695-3299-8. .

59. Razvan Surdulescu. Verifying authorship. Final Project Report CS391L, University of Texas at Austin, 2004.
60. David M. J. Tax. *One-Class Classification*. PhD thesis, Technische Universiteit Delft, 2001.
61. David M. J. Tax and Robert P. W. Duin. Combining One-Class Classifiers. In *Proceedings of the Second International Workshop on Multiple Classifier Systems*, pages 299–308. Springer, 2001. ISBN 3-540-42284-6.
62. Fiona J. Tweedie and Harald R. Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.
63. Hans van Halteren,. Linguistic profiling for author recognition and verification. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 199, Morristown, NJ, USA, 2004. Association for Computational Linguistics.
64. Webis at Bauhaus-Universität Weimar and NLEL at Universidad Polytécnica de Valencia. PAN Plagiarism Corpus 2009 (PAN-PC-09). `http://www.webis.de/research/corpora`, 2009. Martin Potthast, Andreas Eiselt, Benno Stein, Alberto Barrón Cedeño, and Paolo Rosso (editors).
65. Hui Yang and James P. Callan. Near-Duplicate Detection by Instance-Level Constrained Clustering. In Efthimis N. Efthimiadis, Susan Dumais, David Hawking, and Kalervo Järvelin, editors, *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 421–428, 2006. ISBN 1-59593-369-7.
66. G. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.
67. Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006. .
68. G. K. Zipf. Selective studies and the principle of relative frequency in language, 1932.