Breaking the Closed-World Assumption in Stylometric Authorship Attribution

Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, Rachel Greenstadt

Chapter 13

BREAKING THE CLOSED-WORLD ASSUMPTION IN STYLOMETRIC AUTHORSHIP ATTRIBUTION

Ariel Stolerman, Rebekah Overdorf, Sadia Afroz and Rachel Greenstadt

Abstract Stylometry is a form of authorship attribution that relies on the linguistic information found in a document. While there has been significant work in stylometry, most research focuses on the closed-world problem where the author of the document is in a known suspect set. For open-world problems where the author may not be in the suspect set, traditional classification methods are ineffective. This paper proposes the "classify-verify" method that augments classification with a binary verification step evaluated on stylometric datasets. This method, which can be generalized to any domain, significantly outperforms traditional classifiers in open-world settings and yields an F1-score of 0.87, comparable to traditional classifiers in closed-world settings. Moreover, the method successfully detects adversarial documents where authors deliberately change their styles, a problem for which closed-world classifiers fail.

Keywords: Forensic stylometry, authorship attribution, authorship verification

1. Introduction

The web is full of anonymous communications that are often the focus of digital forensic investigations. Forensic stylometry is used to analyze anonymous communications in order to "de-anonymize" them. Classic stylometric analysis requires an exact set of suspects in order to perform reliable authorship attribution, settings that are often not met in real-world problems. This paper breaks the closed-world assumption and explores a novel method for forensic stylometry that addresses the possibility that the true author is not in the set of suspects.

Stylometry is a form of authorship recognition that relies on the linguistic information found in a document. While stylometry existed before computers and artificial intelligence, the field is currently dominated by artificial intelligence techniques such as neural networks and statistical pattern recognition. State-of-the-art stylometric approaches can identify individuals in sets of 50 authors with better than 90% accuracy [1], and even scaled to more than 100,000 authors [22]. Stylometry is currently used in intelligence analysis and forensics, with increasing applications in digital communications analysis. The increase in rigor, accuracy and scale of stylometric techniques has led legal practitioners to turn to stylometry for forensic evidence [15], albeit stylometry is considered to be controversial at times and may not be admitted in court [9].

The effectiveness of stylometry has key implications with regard to anonymous and pseudonymous speech. Recent work has exposed limits on stylometry through active circumvention [4, 21]. Stylometry has thus far focused on limited, closed-world models. In the classic stylometry problem, there are relatively few authors (usually fewer than 20 and nearly always fewer than 100), the set of possible authors is known, every author has a large training set and all the text is from the same genre. However, real-world problems often do not conform to these restrictions.

Controversial pseudonymous documents that are published on the Internet often have an unbounded suspect list. Even if the list is known with certainty, training data may not exist for all suspects. Nonetheless, classical stylometry requires a fixed list and training data for each suspect, and an author is always selected from this list. This is problematic for forensic analysts who have no way of knowing when widening their suspect pool is required, as well as for Internet activists who may appear in the suspect lists and be falsely accused of authorship.

This paper explores a mixed closed-world and open-world authorship attribution problem with a known set of suspect authors, but with some probability (known or unknown) that the author is not in the set.

The primary contribution of this paper is the novel classify-verify (CV) method, which augments authorship classification with a verification step and obtains similar accuracy on open-world problems as traditional classifiers in closed-world problems. Even in the closed-world case, CV can improve results by replacing wrongly identified authors with "unknown." The method can be tuned to different levels of rigidity to achieve the desired false positive and false negative error rates. However, it can also be automatically tuned, whether or not the expected proportion of documents by authors in the suspect list versus those who are absent is known.

CV performs better in adversarial settings than traditional classification. Previous work has shown that traditional classification performs near random chance when faced with writers who change their styles. CV filters most of the attacks in the Extended-Brennan-Greenstadt Adversarial Corpus [4], an improvement over previous work that requires training on adversarial data [2].

This paper also presents the sigma verification method, which is based on Noecker and Ryan's [24] distractorless verification method that measures the distance between an author and a document. Sigma verification incorporates pairwise distances within an author's documents and the standard deviations of the author's features. Although it is not proven to statistically outperform the distractorless method in all scenarios, it has been shown to be a better alternative for datasets with certain characteristics.

2. Problem Statement

The closed-world stylometry or authorship attribution problem is: given a document D of unknown authorship and documents by a set of known authors $\mathcal{A} = \{A_1, ..., A_n\}$, determine the author $A_i \in \mathcal{A}$ of D. This problem assumes that the author of D is in $\mathcal{A}$.

The open-world stylometry problem is: given a document D, identify the author of D.

Authorship verification is a slightly relaxed version: given a document D and an author A, determine whether or not D is written by A.

The problem explored in this paper is a mixture of the two problems above: given a document D of unknown authorship and documents by a set of known authors $\mathcal{A}$, determine the author $A_i \in \mathcal{A}$ of D or that the author of D is not in $\mathcal{A}$. This problem is similar to the attribution problem but with the addition of the class "unknown." An extended definition includes $p = Pr[A_D \in \mathcal{A}]$, the probability that the author of D is in the set of candidates.

In the remainder of this paper, test documents are examined in two settings: when the authors of the documents are in the set of suspects, denoted by "in-set," and when the documents are by an author outside the suspect set, denoted by "not-in-set."

Applying closed-world stylometry in open-world settings suffers from a fundamental flaw: a closed-world classifier always outputs some author in the suspect set. If it outputs an author, it merely means that the document in question is written in a style that is more similar to that author's style than the other suspects, and the probability estimates of the classifier reflect only the suspect who is the "least-worst" choice.

Meanwhile, the absence of the document author from the set of suspects remains unknown. This problem is significant in online domains where the number of potential suspects can be virtually unbounded. Failing to address the limitations of closed-world models can result in falsely-attributed authors with consequences for both forensic analysts and innocent Internet users.

3. Related Work

Open-world classification deals with scenarios in which the set of classes is not known in advance. Approaches include unsupervised, semi-supervised and abstaining classification. Unsupervised stylometry clusters instances based on their feature vector distances [1]. Semi-supervised methods are used to identify clusters [26] that are later used in supervised classification. Abstaining classifiers refrain from classification to improve classifier reliability in certain situations, for example, to minimize the misclassification rate by rejecting the results when the confidence of the classifier is low [8, 12]. The CV method is an abstaining classifier that rejects/accepts an underlying classifier output using a verification step based on the distance between the test author and the predicted author. The novelty of this approach is that, unlike other techniques, CV considers the open-world situation where the author may not be in the suspect set.

Another method is to create a model of the closed-world and reject everything that does not fit the model. In biometric authentication systems, such distance-based anomaly detection methods perform well [3].

In authorship classification, one of the authors in a fixed suspect set is attributed to the test document. Current stylometry methods achieve in excess of 80% accuracy with 100 authors [1], more than 30% accuracy with 10,000 authors [18] and greater than 20% precision with 100,000 authors [22]. None of the methods consider the case where the true author is missing. Although stylometric techniques work well, they are easily circumvented by imitating another person or by deliberate obfuscation [5].

The goal of authorship verification is to determine whether or not a document D is written by an author A. This problem is harder than the closed-world stylometry discussed above. Authorship verification is essentially a one-class classification problem. Research in this area primarily employs support vector machines [20, 28], but little work has focused on stylometry.

Most previous work addresses verification for plagiarism detection [10, 29]. The unmasking algorithm [19] is an example of a general approach to

verification, which relies on the "depth-of-difference" between document and author models. The algorithm yields 99% accuracy with similar false positive and false negative rates, but it is limited to problems with large training datasets.

Noecker and Ryan [24] propose the distractorless verification method, which avoids using negative samples to model the "not the author" class. They use simplified feature sets constructed only of character or word n-grams, normalized dot-product (cosine distance) and an acceptance threshold. The approach has been evaluated on two corpora [13, 27] with accuracy results up to 88% and 92%, respectively. Noecker and Ryan provide a robust verification framework across different types of writings (language, genre and length independent). However, their results also suffer from low F-scores (up to 47% and 51%), which suggest a skew in the test data (testing more non-matching document-author pairs than matching ones). Section 6.2 takes a closer look at this method along with the error rates.

4. Corpora

The experiments described in this paper focused on two corpora, the Extended-Brennan-Greenstadt (EBG) Adversarial Corpus [4] and the ICWSM 2009 Spinn3r Blog Dataset (Blog Corpus) [6].

The EBG Corpus contains writings of 45 different authors, with at least 6,500 words per author. It also contains adversarial documents, where the authors change their writing styles either by imitating another author (imitation attack) or hiding their styles (obfuscation attack). Most of the evaluations in this paper are performed using the EBG Corpus.

The Spinn3r Blog Corpus from `Spinn3r.com` is a set of 44 million blog posts made between August 1, 2008 and October 1, 2008. The posts include the text as syndicated, as well as metadata such as the blog homepage, timestamps, etc. This dataset was previously used in Internet-scale authorship attribution [22]. Our experiments used a subcorpus of 50 blogs with at least 7,500 words as the Blog Corpus. This Blog Corpus was used as control and evaluated under the same settings as the EBG Corpus in order to avoid overfitting configurations on the latter and generalizing the conclusions.

5. Closed-World Setup

A closed-world classifier was used throughout the research – for the baseline results used to evaluate different methods and as the underlying classifier for the CV method. The linear kernel sequential minimal opti-

mization (SMO) support vector machine from Weka [11] was used with complexity parameter $C = 1$. Support vector machines were selected due to their proven effectiveness for stylometry [14].

In addition to classifier selection, another important part of a stylometric algorithm is the feature set used to quantify the documents prior to learning and classification. The EBG Corpus was originally quantified using the Writeprints feature set [4] based on the Writeprints algorithm [1], which has been shown to be accurate for a large number of authors (more than 90% accuracy for 50 authors). Writeprints uses a complex feature set that quantifies the different linguistic levels of text, including lexical, syntactic and content related features; however, for simplicity, a feature set consisting of one type of feature was chosen. The EBG Corpus was evaluated using ten-fold cross validation with the k most common word n-grams or character n-grams, with k between 50 and 1,000 and n between 1 and 5. The most-common feature selection heuristic is commonly used in stylometry to improve performance and avoid over-fitting [1, 17, 24], as are the chosen ranges of k and n.

Character n-grams performed best, yielding the highest F1-score at approximately 0.93 (for $k = 400$ and higher). The word and character feature sets both outperformed the original EBG evaluation with Writeprints (F1-score of 0.832 [4]). We chose the 500 most common character bigrams as the feature set (F1-score of 0.928) for all the experiments; this is denoted by $\langle 500, 2\rangle$-Chars. The choice was made due to its simplicity, performance and effectiveness. Note that $\langle 500, 2\rangle$-Chars also outperformed Writeprints on the control corpus with an F1-score of 0.64 compared with 0.509. All the feature extractions were performed using the JStylo [21] and JGAAP authorship attribution framework APIs.

6. Verification

Authorship verification seeks to determine if a document D is written by an author A. Two naïve approaches suggest themselves. The first and most intuitive is to reduce the problem to closed-world settings by creating a model for not-A (simply from documents not written by A) and to train a binary classifier. This method suffers from a fundamental flaw: if D is attributed to A, it merely means that D's style is less distant from A than it is from not-A. The second approach is to train a binary model of D versus A, and test the model on itself using cross-validation. If D is written by A, the accuracy should be close to random due to the indistinguishability of the models. However, this method does not work well and requires D to contain a substantial amount of text for cross-validation, an uncommon privilege in real-world scenarios.

The following sections discuss and evaluate several verification methods. The first family of methods comprises classifier-induced verifiers, which require an underlying (closed-world) classifier and utilize its class probabilities output for verification. The second family of methods comprises standalone verifiers, which rely on a model built using author training data, independent of other authors and classifiers. Two verification methods are evaluated. The first is the distractorless verification method [24] denoted by V. It is presented as a baseline because it is a straightforward verification method that has been shown to be robust across different domains and does not use a distractor set (model of not-A). The second is the novel sigma verification method, which applies adjustments to V for increased accuracy: adding per-feature standard deviation normalization (denoted by V_σ) and adding per-author threshold normalization (denoted by V^a). Finally, V is evaluated and compared with its new variants.

6.1 Classifier-Induced Verification

A promising aspect of the closed-world model that can be used in open-world scenarios is the confidence in solutions provided by distance-based classifiers. A higher confidence in an author may, naturally, indicate that the author is a suspect while a lower confidence may indicate that he is not and this problem is, in fact, an open-world one.

Following classification, verification can be formulated simply by setting an acceptance threshold t, measuring the confidence of the classifier in its result, and accepting the classification if it is above t.

Next, we discuss several verification schemes based on the classification probabilities output by a closed-world classifier. For each test document D with suspect authors $\mathcal{A} = \{A_1, ..., A_n\}$, a classifier produces a list of probabilities P_{A_i} which is, according to the classifier, the probability D is written by A_i ($\sum_{i=1}^{n} P_{A_i} = 1$). We denote the probabilities $P_1, ..., P_n$ as the reverse order statistic of P_{A_i}, i.e., P_1 is the highest probability given to some author (chosen author), P_2 the second highest, and so on.

These methods are obviously limited to classify-verify scenarios because verification is dependent on classification results (thus, they are not evaluated in this section, but in Section 8 as part of the CV evaluation). For this purpose and in order to extract the probability measurements required by the following methods, SMO support vector machine classifiers were used with the $\langle 500, 2 \rangle$-Chars feature set for all the experiments described in Section 8. Logistic regression models were fitted

to the support vector machine outputs to obtain proper probability estimates.

The following classifier-induced verification methods are evaluated:

- **P_1:** This measurement is simply the classifier probability output for the chosen author, namely P_1. The hypothesis behind this measurement is that as the likelihood that the top author is the true author increases, relative to all others, so does its corresponding probability.

- **P_1-P_2-Diff:** This measurement captures the difference between the classifier probability outputs of the chosen and second-choice authors, i.e., $P_1 - P_2$; it is referred to as the P_1-P_2-*Diff* method.

- **Gap-Conf:** In the case of the gap confidence method [25], one support vector machine classifier is not trained; instead, for all n authors, the corresponding n one-versus-all support vector machines are trained. For a given document D, each classifier i produces two probabilities: the probability that D is written by A_i and the probability that it is written by an author other than A_i. For each i, if $p^i(\text{Yes}|D)$ denotes the probability that D is written by A_i, then the gap confidence is the difference between the highest and second-highest $p^i(\text{Yes}|D)$, which we denote briefly as *Gap-Conf*. The hypothesis is similar to P_1-P_2-*Diff*: the probability of the true author should be much higher than that of the second-best choice.

6.2 Standalone Verification

The following methods are evaluated:

- **Distractorless Verification (V):** As discussed above, V uses the straightforward distance combined with a threshold: set an acceptance threshold t, model document D and author A as feature vectors, measure their distance and determine that D is written by A if it is below t.

  For n denoting the size of the chosen feature set, a model $M = \langle m_1, m_2, ..., m_n \rangle$ is built from the centroid of the character or word n-gram feature vectors of A's documents. For each i, m_i is the average relative frequency of feature i across A's documents, where the relative frequency is used to eliminate document length variation effects. In addition, a feature vector $F = \langle f_1, f_2, ..., f_n \rangle$ is extracted from D, where f_i corresponds to the relative frequency of feature i in D.

Finally, a distance function δ and a threshold t are set, such that if $\delta(x, y) < \delta(x, z)$, x is considered to be closer to y than to z. A normalized dot-product (cosine distance) is used:

$$\delta(M, F) = \frac{M \cdot F}{\|M\|\|F\|} = \frac{\sum_{i=1}^{n} m_i f_i}{\sqrt{\sum_{i=1}^{n} m_i^2}\sqrt{\sum_{i=1}^{n} f_i^2}}.$$

This measure has been shown to be effective for stylometry [23] and efficient for large datasets. Note that "closer to" is defined using $>$ instead of $<$, which is consistent with the cosine distance (where a value of one is a perfect match). However, we use $<$ as the more intuitive direction (according to which a smaller distance means a better match), and adjust the cosine distance δ in the equation above to $1 - \delta$.

The threshold t is set such that D is written by A is determined when $\delta(M, F) < t$. Ideally, it is empirically determined by analysis of the average δ between the author's training documents; however, the evaluation in [24] uses a hardcoded threshold that does not take the author-wise δ values into account (which V^a does, as shown below).

- **Per-Feature SD Normalization (V_σ):** The first suggested improvement to V is based on the variance of the author's writing style. If an author has a style that does not vary much, a tighter bound for verification is required, whereas for a more varied style, the model can be loosened to be more accepting. To do so, the standard deviation (SD) of an author on a per-feature basis is used. For each author, the SD of all of the author's features is determined. When computing the distance between an author and a document, each feature-distance is divided by its SD, so if the SD is smaller, then A and D move closer together, otherwise they move farther apart. This idea is applied in [3] for authentication using typing biometrics; however, its application to stylometric verification is novel.

- **Per-Author Threshold Normalization (V^a):** The second proposed improvement is to adjust the verification threshold t on a per-author basis based on the average pairwise distance between all the author's documents; this is denoted by δ_A. V does not take this into account and instead uses a hard threshold. Using δ_A to determine the threshold is, intuitively, an improvement because it accounts for the spread of the documents written by an author. This allows the model to relax if the author has a more varied

style. As in the case of V, this "varying" threshold is still applied by setting a fixed threshold t across all authors, determined empirically over the training set. However, for V^a, every author-document distance measurement δ is adjusted by subtracting δ_A before being compared with t, thus allowing per-author thresholds but still requiring the user to set only one fixed threshold value.

The three methods described above were evaluated based on their false positive (FP) and false negative (FN) rates measured on the EBG Corpus and the Blog Corpus as control. The EBG Corpus was evaluated only on the non-adversarial documents, and the Blog Corpus was evaluated in its entirety. Ten-fold cross-validation with the $\langle 500, 2\rangle$-Chars feature set was used; the author models were built using the training documents. In each fold, every test document was tested against every one of the author models, including its own (trained on other documents by the author).

When applied to the EBG Corpus, V_σ significantly outperformed V for $FP \geq 0.05$ with a confidence level of $p\text{-}val < 0.01$. Similarly, V_σ^a outperformed V for $FP \geq 0.114$. However, different results were obtained for the Blog Corpus, where V significantly outperformed both V_σ and V_σ^a. The differences could be explained by the corpora characteristics: the EBG Corpus is a cleaner and more stylistically consistent corpus consisting of all English formal writing samples (essays originally written for business or academic purposes), whereas the Blog Corpus contains less structured and formal language, which reduces the distinguishable effects of style variance normalization. This notion is supported by the better performance for the EBG Corpus compared with the Blog Corpus (larger area under the receiver operator curve). Clearly, the results suggest that one method is not preferred over the other, and selecting a verifier for a problem should rely on empirical testing over stylistically similar training data.

For both corpora, V_σ^a outperformed V_σ starting at $FP = 0.27$ and $FP = 0.22$ for the EBG and Blog Corpora, respectively. These properties allow verification approaches to be used according to need, dependent on the FP and FN error rate constraints that a specific problem may impose.

7. Classify-Verify

The CV method employs an abstaining classifier [8], i.e., a classifier that refrains from classification in certain cases to reduce misclassifications. CV combines classification with verification to expand closed-world authorship problems to the open-world essentially by adding the

"unknown" class. Another aspect of the novelty of the CV method is the utilization of abstaining classification methods to upgrade from closed-world to open-world, where the methods for thwarting misclassifications are evaluated based on how they apply to those that originate outside the assumed suspect set, instead of simply missing the true suspect.

First, closed-world classification is applied to the document D in question and the author suspect set $\mathcal{A} = \{A_1, ..., A_n\}$ (i.e., sample documents). Then, the output of the classifier $A_i \in \mathcal{A}$ is given to the verifier to determine the final output. Feeding only the classifier result to the verifier leverages the high accuracy of classifiers, which outperform verifiers in closed-world settings, thus focusing the verifier only on the top choice author in $\mathcal{A}$. The verifier determines whether to accept A_i or reject by returning $\perp$ based on a verification threshold t. CV is essentially a classifier over the suspect set $\mathcal{A} \cup \{\perp\}$.

The threshold t selection process can be automated with respect to varying expected portions of in-set and not-in-set documents. The likelihood of D's author being in $\mathcal{A}$, the expected in-set documents fraction, is denoted by $p = Pr[A_D \in \mathcal{A}]$ (making the likelihood of the expected not-in-set documents $1 - p$). In addition, $p\langle measure \rangle$ refers to the weighted average of the measure with respect to p. For instance, p-$F1$ is the weighted F1-score, weighted over F1-scores of p expected in-set documents and $1 - p$ expected not-in-set documents. Thus, the threshold t can be determined in several ways:

- **Manual:** The threshold t can be manually set by the user. The threshold determines the sensitivity of the verifier, so setting t manually adjusts it from strict to relaxed, where the stricter it is, the less likely it is to accept the classifier output. This enables the algorithm to be tuned to different settings, imposing the desired rigidity.
- **p-Induced Threshold:** The threshold can be set empirically over the training set to maximize the target measurement, e.g., F1-score, in an automated process. If p is given, then the algorithm applies cross-validation on the training data alone using the range of all relevant manually-set thresholds and chooses the threshold that yields the best target measurement. This essentially applies CV recursively on the training data one level deeper with a range of manual thresholds. The relevant threshold search range is determined automatically by the minimum and maximum distances observed in the verify phase of CV.
- **In-Set/Not-In-Set-Robust:** If the expected in-set and not-in-set documents proportion is unknown, the same idea as in the

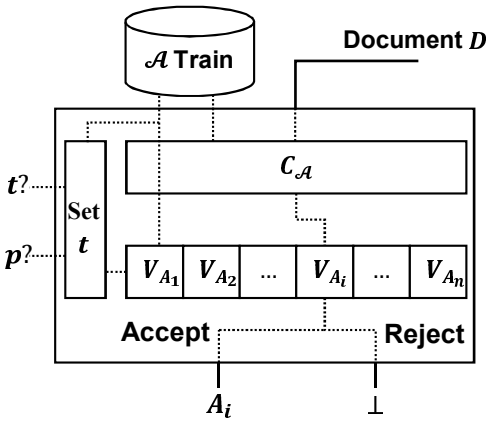

Figure 1. CV method flow.

previously described threshold can be applied. Upon examining the CV F1-score curve for some p along a range of thresholds, if p increases, then it favors smaller (more accepting) thresholds, therefore the curve behaves differently for different values of p. However, all the curves intersect at one t – at which the in-set and not-in-set curves intersect. This can be utilized to automatically obtain a robust threshold for any value of p by taking the thresholds that minimize the difference between the p-$F1$ and q-$F1$ curves for arbitrary $p, q \in [0, 1]$ (for simplicity, values of 0.3 and 0.7 are used). The robust threshold does not guarantee the highest measurement; it does, however, guarantee a relatively high expected value of the measure independent of p, and is, thus, robust for any open-world settings. This measurement is denoted by p-$\langle measure \rangle_R$ (for robust), e.g., p-$F1_R$.

Figure 1 shows the flow of the CV algorithm on a test document D and a suspect set $\mathcal{A}$ with optional threshold t and in-set portion p.

8. Evaluation and Results

This section describes the evaluation methodology and the experimental results.

8.1 Evaluation Methodology

Main Experiment. The main experiment evaluated the CV method on the EBG Corpus (excluding the adversarial documents) and the Blog Corpus as control. The corpora were evaluated in two settings: when the

authors of the documents under test were in the set of suspects (in-set) and when they were not (not-in-set).

Each classification over n authors $\mathcal{A} = \{A_1, ..., A_n\}$ produces one of $n+1$ outputs: an author $A_i \in \mathcal{A}$ or $\perp$ ("unknown"). Therefore, when the verifier accepts, the final result is the author A_i chosen by the classifier, and when it rejects, the final result is $\perp$.

In the evaluation process, the CV algorithm was credited when the verification step thwarted misclassifications in in-set settings. For instance, if D was written by A, classified as B, but the verifier replaced B with $\perp$, the result was considered to be true. This approach for abstaining classifiers [12] relies on the fact that a result of "unknown" is better than an incorrect author.

The overall performance was evaluated using ten-fold cross-validation. For each fold experiment, with nine of the folds as the training set and one fold as the test set, every test document was evaluated twice: once as in-set and once as not-in-set.

For the classification phase of CV over n authors, n $(n-1)$-class classifiers were trained, where each classifier C_i was trained on all the authors except for A_i. A test document by some author A_i was then classified as in-set using one of the $n-1$ classifiers that were trained on A_i training data; for simplicity, we chose C_{i+1} (and C_1 for A_n). For the not-in-set classification, the classifier that was not trained on A_i, i.e., C_i, was employed.

For the verification phase of CV, several methods were evaluated: one standalone method (V_σ for the EBG Corpus and V for the Blog Corpus), *Gap-Conf*, P_1 and P_1-P_2-*Diff*. V_σ was used for the EBG Corpus and V for the Blog Corpus because these methods outperformed the other standalone methods evaluated per corpus as discussed in Section 6.

The more the verifiers reject, the higher the precision (because bad classifications are thrown away), but the recall decreases (as good classifications are thrown away as well), and vice-versa – higher acceptance increases recall but decreases precision. Therefore, the overall performance is measured using the F1-score, since it provides a balanced measurement of precision and recall:

$$F1\text{-}score = 2 \times \frac{precision \times recall}{precision + recall}$$

where

$$precision = \frac{tp}{tp + fp}; \qquad recall = \frac{tp}{tp + fn}.$$

The two automatic verification threshold selection methods discussed in Section 7 were used. For the scenario in which the proportion of

in-set and not-in-set is known with the in-set proportion $p = 0.5$ (the not-in-set proportion is $1 - p$), we used the p-induced threshold that maximizes the F1-score on the training set. For the scenario in which p is unknown, we used the robust threshold configured as described in Section 7. In order to calculate the F1-score of evaluating the test set, the confusion matrices produced by the in-set and not-in-set evaluations were combined to form a p-weighted average matrix, from which the weighted F1-score was calculated. The p-induced F1-scores is denoted by p-$F1$ and robust threshold induced F1-scores evaluated at some p is denoted by p-$F1_R$.

The threshold optimization phase of the CV method discussed in Section 7 was performed using nine-fold cross-validation with the same experimental settings as in the main experiment. Since the F1-score was used to evaluate the overall performance, it was also used as the target measurement to maximize in the automatic threshold optimization phase. When p was known, the threshold that maximizes p-$F1$ was selected, and when it was unknown, the robust threshold was selected as the one for which the F1-scores of different p values intersect (arbitrarily set to 0.3-$F1$ and 0.7-$F1$).

As a baseline, the F1-scores were compared with ten-fold cross-validation results of closed-world classification using the underlying classifier, SMO support vector machine with the $\langle 500, 2 \rangle$-Chars feature set. Let p-$Base$ be the baseline F1-score of the closed-world classifier where the in-set proportion is p. It follows that 1-$Base$ is the performance in pure closed-world settings (i.e., only in-set documents) and for any $p \in [0, 1]$, p-$Base = p \cdot 1$-$Base$ (since the classifier is always wrong for not-in-set documents).

Adversarial Settings. To evaluate the CV method in adversarial settings, the models were trained on the non-adversarial documents in the EBG Corpus, and tested on the imitation and obfuscation attack documents to measure how well CV thwarted attacks (by returning $\perp$ instead of a wrong author). In this context, $\perp$ can be considered as either "unknown" or "possible-attack." The term 0.5-$F1$ was measured, i.e., how well CV performed on attack documents in an open-world scenario where the verification threshold was set independent of a possible attack, tuned only to maximize performance on expected in-set and not-in-set document portions of 50% each. As a baseline, the results with standard classification using SMO support vector machine were compared with the $\langle 500, 2 \rangle$-Chars feature set.

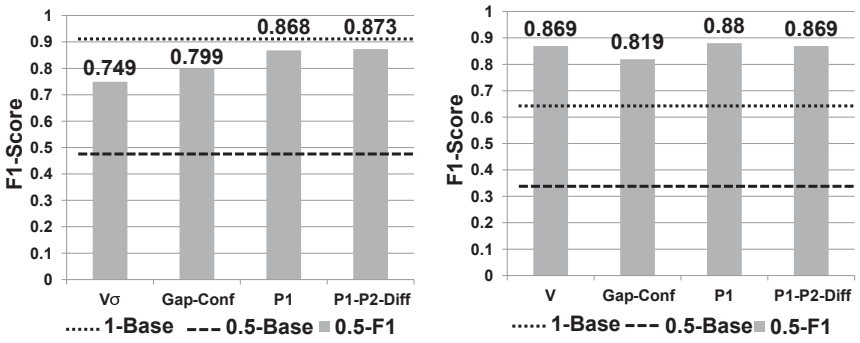

Figure 2. 0.5-$F1$ results for the EBG Corpus (left) and Blog Corpus (right).

8.2 Results

Main Experiment. For the EBG Corpus, the baseline closed-world classifier attained 1-$Base$ = 0.928 in perfect in-set settings, which implies that 0.5-$Base$ = 0.464. For the Blog Corpus, 1-$Base$ = 0.64, which implies that 0.5-$Base$ = 0.32. Figure 2 shows the 0.5-$F1$ results for the CV method on the EBG and Blog Corpora, where the authors are equally likely to be in-set or not-in-set ($p = 0.5$) and the verification thresholds were automatically selected to maximize 0.5-$F1$. For the EBG and Blog Corpora, the CV 0.5-$F1$ results significantly outperform 0.5-$Base$ (dashed lines) for all the underlying verification methods at a confidence level of $p\text{-}val < 0.01$.

Furthermore, the results are not only better than the obviously bad 0.5-$Base$, but produce similar results to 1-$Base$, giving an overall 0.5-$F1$ for open-world settings up to approximately 0.87. For the EBG Corpus, moving to open-world settings only slightly decreases the F1-score compared with the closed-world classifier performance in closed-world settings (dotted line in Figure 2), which is a reasonable penalty for upgrading to open-world settings. However, for the Blog Corpus, where the initial 1-$Base$ is low (0.64), CV manages to upgrade to open-world settings and outperform 1-$Base$. These results suggest that out of the in-set documents, many misclassifications were thwarted by the underlying verifiers, leading to an overall increase in the F1-score.

Next, the robust threshold selection scheme was evaluated. In this scenario, the portion of in-set documents p was not known in advance. Figure 3 shows the p-$F1_R$ results for the EBG and Blog Corpora, where different p scenarios were "thrown" at the CV classifier with robust verification thresholds. The expected portion of in-set documents p var-

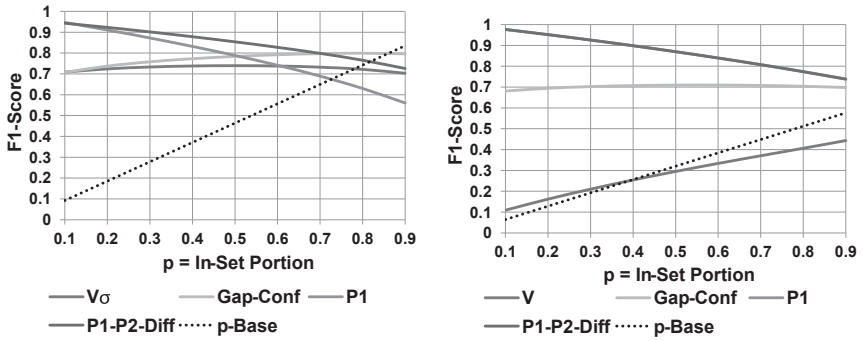

Figure 3. p-$F1_R$ results for the EBG Corpus (left) and Blog Corpus (right).

ied from 10% to 90% and was assumed to be unknown, and robust p-independent thresholds were used for the underlying verifiers.

In the robust thresholds scenario with the EBG Corpus, CV still significantly outperforms the respective closed-world classifier (p-$Base$ results) for $p < 0.7$ with any of the underlying verifiers at a confidence level of p-$val < 0.01$. For the Blog Corpus, CV significantly outperforms p-$Base$ using any of the classifier-induced verifiers for all p at a confidence level of p-$val < 0.01$.

Moreover, the robust threshold selection hypothesis holds true, and for both corpora all the methods (with the exception of V on the Blog Corpus) manage to guarantee a high F1-score at approximately 0.7 and above for almost all values of p. For the EBG Corpus, at $p \geq 0.7$ the in-set portion is large enough that the overall p-$Base$ becomes similar to p-$F1_R$. For the Blog Corpus, using V fails and has a similar performance as 0.5-$Base$.

P_1-P_2-*Diff* is the preferred verification method. It consistently outperforms the other methods across almost all values of p for both corpora, which implies that it is robust to domain variation.

Adversarial Settings. Evaluated on the EBG Corpus under imitation and obfuscation attacks, the baseline closed-world classifier yields F1-scores of 0 and 0.044 for the imitation and obfuscation attack documents, respectively. The results imply that the closed-world classifier is highly vulnerable to these types of attacks. Figure 4 shows the F1-scores for CV on the attack documents. Note that all attack documents were written by in-set authors and were, thus, handled as in-set documents.

The results suggest that CV successfully manages to thwart the majority of the attacks, with F1-scores up to 0.826 and 0.874 for the imitation and obfuscation attacks, respectively. These results are very close to the

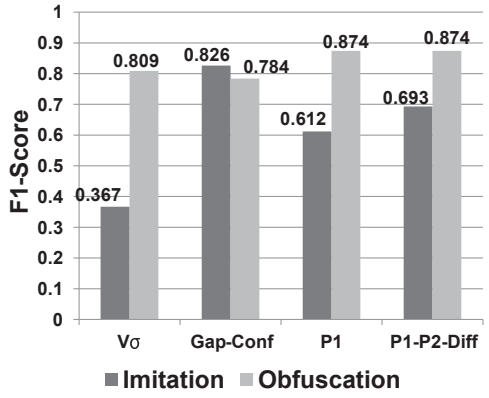

Figure 4. F1-scores for the EBG Corpus under imitation and obfuscation attacks.

deception detection results reported in [2] with F1-scores of 0.895 for imitation and 0.85 for obfuscation attacks. A major difference is that the results in this paper were obtained in open-world settings with a threshold configuration that does not consider inside-attacks. Moreover, unlike the methods applied in [2], no attack documents were used as training data.

Interestingly, the results presented above were obtained for a standard $p = 0.5$ open-world scenario, without possible attacks in mind. Still, the overall results were affected very little, if at all, depending on the underlying verifier. For example, when using *Gap-Conf*, 0.5-$F1$ is at 0.799 in non-attack scenarios and the F1-scores range from 0.784 to 0.826 under attack.

9. Discussion

The evaluation results suggest that classifier-induced verifiers consistently outperform standalone verifiers. However, this trend may be limited to large datasets with many suspect authors in the underlying classifier, like those evaluated in this paper, on which classifier-induced verifications rely. It may be the case that standalone verifiers perform better for small author sets; this direction should be explored in future research. Moreover, standalone verifiers provide reasonable accuracy that enables them to be used in pure one-class settings, where only true author data exists (a scenario in which classifier-induced methods are useless).

The CV 0.5-$F1$ results for the EBG and Blog Corpora (Figure 2) suggest that using P_1 or P_1-P_2-*Diff* as the underlying verification method yield domain-independent results for which 0.5-$F1$ is approximately 0.87.

The superiority of P_1-P_2-*Diff* is emphasized by the p-$F1_R$ results in Figure 3, where p-$F1_R$ over 0.7 is obtained for both corpora independent of p. Therefore, P_1-P_2-*Diff* is a robust, domain and in-set/not-in-set proportion independent verification method to be used with CV.

CV is effective in adversarial settings, where it outperforms the traditional closed-world classifier without any training on adversarial data (which is required in [2]). Furthermore, no special threshold tuning is needed to achieve this protection, i.e., standard threshold selection schemes can be used for non-adversarial settings while still thwarting most attacks. Finally, it appears that the results in adversarial settings can potentially be improved if p is tuned not to the likelihood of in-set documents, but to the likelihood of an attack.

10. Conclusions

From a forensic perspective, the possibility of authors outside the suspect set renders closed-world classifiers unreliable. In addition, whether linguistic authorship attribution can handle open-world scenarios has important privacy ramifications for authors of anonymous texts and individuals who are falsely implicated by erroneous results. Indeed, when the closed-world assumption is violated, traditional stylometric approaches do not fail in a graceful manner.

The CV method proposed in this paper can handle open-world settings where the author of a document may not be in the training set, and can also improve the results in closed-world settings by abstaining from low-confidence classification decisions. Furthermore, the method can filter attacks as demonstrated on the adversarial samples in the EBG Corpus. In all these settings, the CV method replaces wrong assertions with the more honest and useful result of "unknown."

The CV method is clearly preferable to the standard closed-world classifier. This is true regardless of the expected in-set/not-in-set ratio, and in adversarial settings as well. Moreover, the general nature of the CV algorithm enables it to be applied with any stylometric classifiers and verifiers.

Our future research will pursue three avenues. The first is to apply the CV method to other problems such as behavioral biometrics using, for example, the Active Linguistic Authentication dataset [16]. The second is to attempt to fuse verification methods using the Chair-Varshney optimal fusion rule [7] to reduce error rates. The third avenue is to investigate the scalability of the CV method to large problems while maintaining its accuracy.

References

[1] A. Abbasi and H. Chen, Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace, *ACM Transactions on Information Systems*, vol. 26(2), pp. 7:1–7:29, 2008.

[2] S. Afroz, M. Brennan and R. Greenstadt, Detecting hoaxes, frauds and deception in writing style online, *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 461–475, 2012.

[3] L. Araujo, L. Sucupira, M. Lizarraga, L. Ling and J. Yabu-Uti, User authentication through typing biometrics features, *IEEE Transactions on Signal Processing*, vol. 53(2), pp. 851–855, 2005.

[4] M. Brennan, S. Afroz and R. Greenstadt, Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity, *ACM Transactions on Information and System Security*, vol. 15(3), pp. 12:1–12:22, 2012.

[5] M. Brennan and R. Greenstadt, Practical attacks against authorship recognition techniques, *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence*, pp. 60–65, 2009.

[6] K. Burton, A. Java and I. Soboroff, The ICWSM 2009 Spinn3r Dataset, *Proceedings of the Third Annual Conference on Weblogs and Social Media*, 2009.

[7] Z. Chair and P. Varshney, Optimal data fusion in multiple sensor detection systems, *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-22(1), pp. 98–101, 1986.

[8] C. Chow, On optimum recognition error and reject tradeoff, *IEEE Transactions on Information Theory*, vol. 16(1), pp. 41–46, 1970.

[9] A. Clark, Forensic Stylometric Authorship Analysis Under the Daubert Standard, University of the District of Comumbia, Washington, DC (`papers.ssrn.com/sol3/papers.cfm?abstract_id=2039824`), 2011.

[10] P. Clough, Plagiarism in Natural and Programming Languages: An Overview of Current tools and Technologies, Technical Report, Department of Computer Science, University of Sheffield, Sheffield, United Kingdom, 2000.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, The Weka Data Mining Software: An update, *SIGKDD Explorations Newsletter*, vol. 11(1), pp. 10–18, 2009.

[12] R. Herbei and M. Wegkamp, Classification with reject option, *Canadian Journal of Statistics*, vol. 34(4), pp. 709–721, 2006.

[13] P. Juola, Ad hoc Authorship Attribution Competition, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2004.

[14] P. Juola, Authorship attribution, *Foundations and Trends in Information Retrieval*, vol. 1(3), pp. 233–334, 2008.

[15] P. Juola, Stylometry and immigration: A case study, *Journal of Law and Policy*, vol. 21(2), pp. 287–298, 2013.

[16] P. Juola, J. Noecker, A. Stolerman, M. Ryan, P. Brennan and R. Greenstadt, A dataset for active linguistic authentication, in *Advances in Digital Forensics IX*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 385–398, 2013.

[17] M. Koppel and J. Schler, Authorship verification as a one-class classification problem, *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

[18] M. Koppel, J. Schler and S. Argamon, Authorship attribution in the wild, *Language Resources and Evaluation*, vol. 45(1), pp. 83–94, 2011.

[19] M. Koppel, J. Schler and E. Bonchek-Dokow, Measuring differentiability: Unmasking pseudonymous authors, *Journal of Machine Learning Research*, vol. 8(2), pp. 1261–1276, 2007.

[20] L. Manevitz and M. Yousef, One-class document classification via neural networks, *Neurocomputing*, vol. 70(7-9), pp. 1466–1481, 2007.

[21] A. McDonald, S. Afroz, A. Caliskan, A. Stolerman and R. Greenstadt, Use fewer instances of the letter "i:" Toward writing style anonymization, in *Privacy Enhancing Technologies*, S. Fischer-Hubner and M. Wright (Eds.), Springer-Verlag, Berlin, Germany, pp. 299–318, 2012.

[22] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, R. Shin and D. Song, On the feasibility of Internet-scale author identification, *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 300–314, 2012.

[23] J. Noecker and P. Juola, Cosine distance nearest-neighbor classification for authorship attribution, presented at the *Digital Humanities Conference*, 2009.

[24] J. Noecker and M. Ryan, Distractorless authorship verification, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pp. 785–789, 2012.

[25] H. Paskov, A Regularization Framework for Active Learning from Imbalanced Data, M. Engg. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2010.

[26] E. Sorio, A. Bartoli, G. Davanzo and E. Medvet, Open world classification of printed invoices, *Proceedings of the Tenth ACM Symposium on Document Engineering*, pp. 187–190, 2010.

[27] B. Stein, M. Potthast, P. Rosso, A. Barron-Cedeno, E. Stamatatos and M. Koppel, Workshop report: Fourth International Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, *ACM SIGIR Forum*, vol. 45(1), pp. 45-48, 2011.

[28] D. Tax, One-Class Classification, Ph.D. Dissertation, Faculty of Applied Physics, Delft University of Technology, Delft, The Natherlands, 2001.

[29] H. van Halteren, Linguistic profiling for authorship recognition and verification, *Proceedings of the Forty-Second Annual Meeting of the Association for Computational Linguistics*, art. 199, 2004.