Classify, but Verify

Breaking the Closed-World Assumption in Stylometric Authorship Attribution

Ariel Stolerman Rebekah Overdorf Sadia Afroz Rachel Greenstadt
The Privacy, Security and Automation Lab, Drexel University

USENIX Security, August 14-16, 2013

Motivation

- The web is full of anonymous communication that was never meant to be analyzed for authorship attribution
- Stylometry is a form of authorship attribution that relies on the linguistic information found in a document
- Stylometry research has thus far focused on closed-world models, limited to a set of known suspect authors
- Often the closed-world assumption is broken, requiring a solution for forensic analysts and Internet activists who wish to remain anonymous

Contribution and Application to Security & Privacy

- The *Classify-Verify* method
  An abstaining classification approach that augments authorship classification with a verification step
  - Performs well in open-world problems with similar accuracy as traditional methods in closed-world problems
  - Improves closed-world solutions by replacing misclassifications with “unknown”
  - Performs well in adversarial settings where traditional methods fail without the need to train on adversarial data
- The Sigma Verification method
  An extension of the distractorless verification method [Noecker & Ryan, *LREC'12*] for author-document distance measurement
  - Incorporates pairwise distances within the author’s documents
  - Normalizes over the standard deviations of the author’s features
- Security & Privacy Applications
  Useful when the target class may absent from the suspect set:
  - Authorship Attribution/Verification (this work)
  - Website fingerprinting
  - Malware family identification

Problem Statement

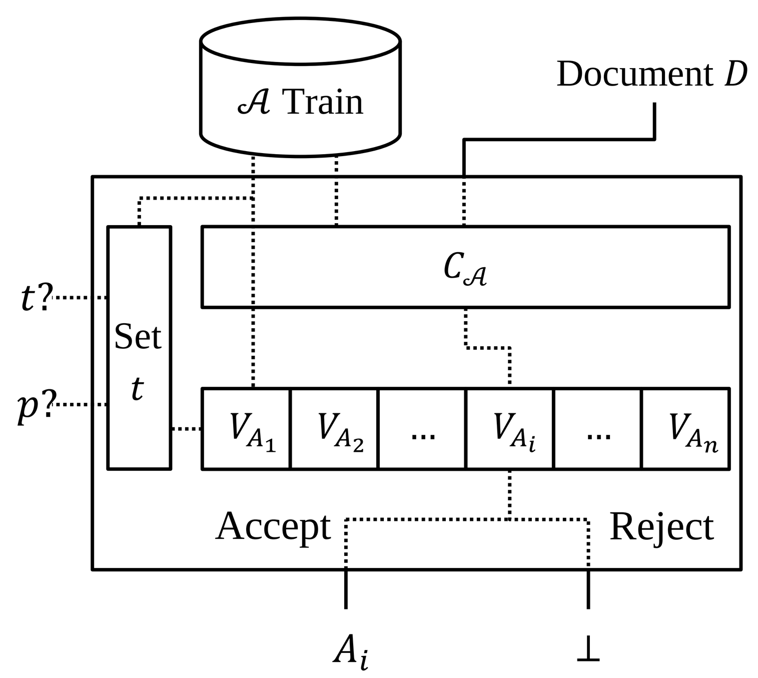
Definitions:

- D – document of unknown authorship
- A – candidate author
- $\mathcal{A} = \{A_1, ..., A_n\}$ – set of candidate authors
- $p = Pr[A_D \in \mathcal{A}]$ – the probability that D’s author is in the set of candidates $\mathcal{A}$, denoted the *in-set* prob. ($1 - p$ is the *not-in-set* prob.)
- t – verification acceptance threshold

Problems:

- *Authorship Attribution*: Which $A \in \mathcal{A}$ is the author of D?
- *Authorship Verification*: is D written by A?
- The *Classify-Verify* Problem: given D, $\mathcal{A}$ and *optionally* p:
  - Determine the author $A \in \mathcal{A}$ of D, or
  - Determine that D’s author is not in $\mathcal{A}$ (w.r.t. acceptance threshold t)

Classify-Verify



Flow of the *Classify-Verify* algorithm on a test document D and a suspect set $\mathcal{A}$, with optional acceptance threshold t and *in-set* prob. p.

The *Classify-Verify* Algorithm

Input: Document D, suspect author set $\mathcal{A} = \{A_1, ..., A_n\}$, target measure to maximize μ
Optional: in-set prob. p, manual threshold t
Output: A_D if $A_D \in \mathcal{A}$, and $\perp$ otherwise
$C_{\mathcal{A}} \leftarrow$ classifier trained on $\mathcal{A}$
$\mathcal{V}_{\mathcal{A}} = \{V_{A_1}, ..., V_{A_n}\} \leftarrow$ verifiers trained on $\mathcal{A}$
if t, p not set **then**
  $t \leftarrow$ threshold maximizing p-μ_R of *Classify-Verify* cross-validation on $\mathcal{A}$
else if t not set **then**
  $t \leftarrow$ threshold maximizing p-μ of *Classify-Verify* cross-validation on $\mathcal{A}$
end if
$A \leftarrow C_{\mathcal{A}}(D)$
if $V_A(D, t) = True$ **then**
  return A
else
  return $\perp$
end if

Synopsis:

- Train one closed-world classifier $C_{\mathcal{A}}$ over $\mathcal{A}$ and n verifiers†‡ $V_1, ..., V_n$ for each $A_i \in \mathcal{A}$
- Classify D using $C_{\mathcal{A}}$, and let the result be author A_i
- Verify D using V_i
  - If it accepts, return the author A_i
  - Otherwise, return $\perp$, which stands for “none”

†Verification Methods

Classifier-Induced Verifiers

Let P_i denote the ith order statistic of the probability outputs of $C_{\mathcal{A}}(D)$, then:

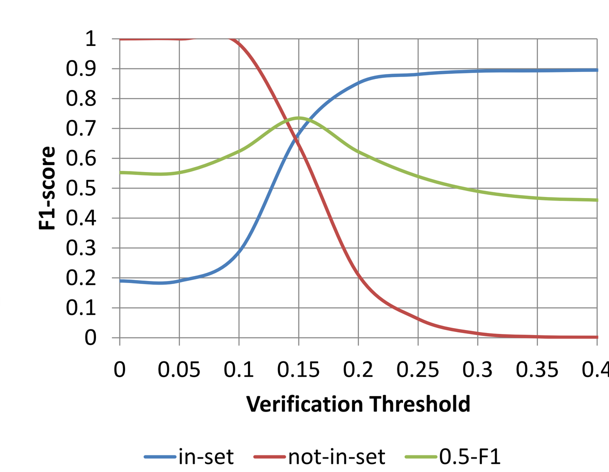
- P_1 : probability of the chosen class
- P_1-P_2-*Diff* : difference between chosen and second-to-chosen class probabilities
- *Gap-Conf* [Paskov, MIT 2010] : P_1-P_2-*Diff* based on n 1-vs-all classifiers

Standalone Verifiers *

distractorless or *Sigma* verification

‡Verification Acceptance Threshold t

- Manual: acceptance threshold t set manually
- p-induced threshold: t is set empirically using cross-validation over the training set, to maximize the target evaluation measure μ (e.g. F1-score) for given *in-set* prob. p
- p-Robust: t is set like in p-induced, but to maximize the *average* μ across any p



F1-score for *Classify-Verify* on EBG using $p = 0.5$, SMO SVM and V_σ with varying manually-set thresholds.

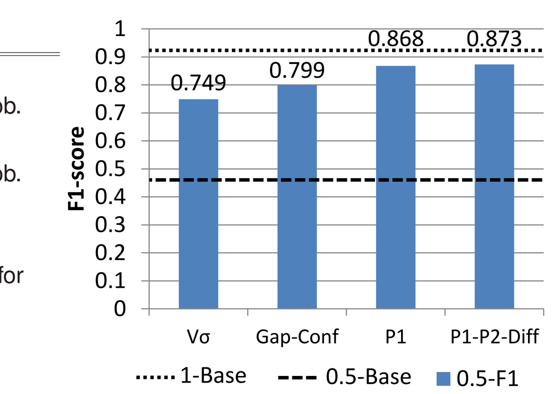
Evaluation & Results

- Corpora:
  - EBG: The Extended-Brennan-Greenstadt Adversarial corpus [Brennan et al., *ACM Trans. Inf. Syst. Secur.'12*] , 45 authors
  - Blog: The ICWSM 2009 Spinn3r Blog dataset [Burton et al., *ICWSM'09*], 50 authors
- Closed-world classifier: SVM SMO
- Feature set: 500 most common character bigrams

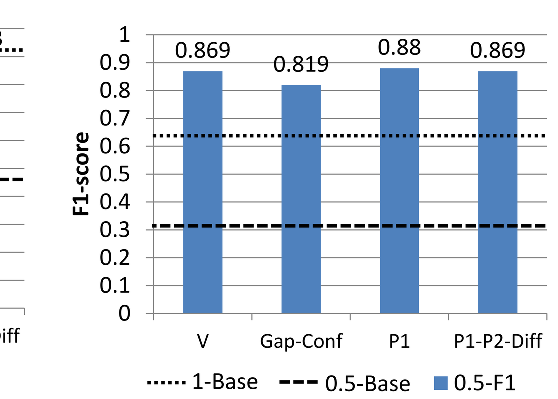
Results:

Term	Meaning
p	Prob. of *in-set* documents
p-$F1$	F1-score of *Classify-Verify* for p prob. of *in-set* documents
p-$F1_R$	F1-score of *Classify-Verify* for p prob. of *in-set* documents, using robust thresholds
p-*Base*	F1-score of closed-world classifier for p prob. of *in-set* documents 1-*Base* means pure closed-world settings For each $p \in [0, 1]$, p-*Base* $= p \cdot 1$-*Base*

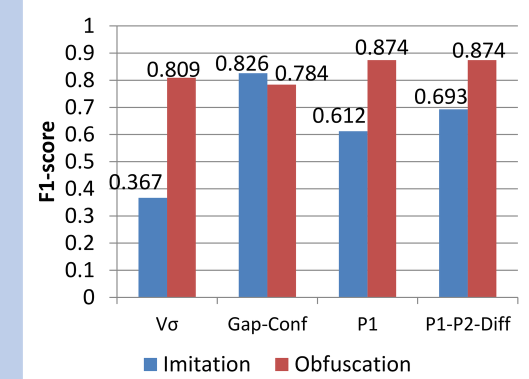
F1-score references terminology



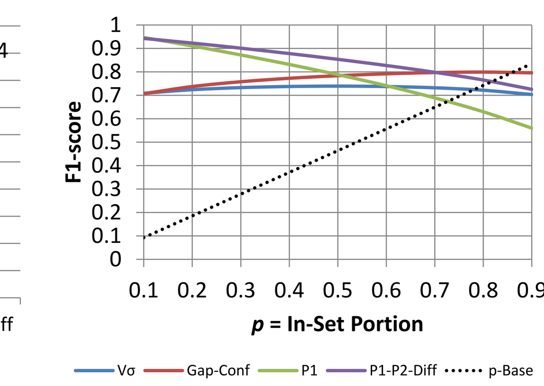
0.5-$F1$ results of the *Classify-Verify* method on the EBG corpus, where the expected prob. of *in-set* and *not-in-set* documents is equal (50%). *Classify-Verify* attains 0.5-$F1$ that outperforms 0.5-*Base* and even close to 1-*Base*.



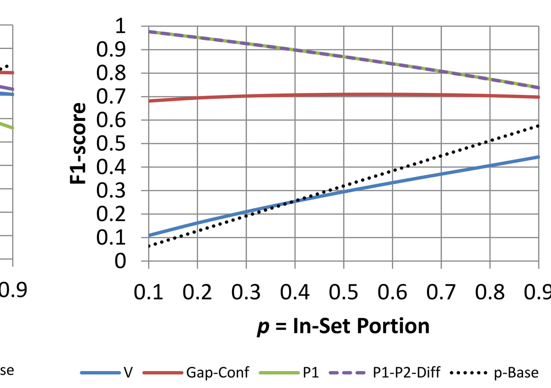
0.5-$F1$ results of the *Classify-Verify* method on the blog corpus, where the expected prob. of *in-set* and *not-in-set* documents is equal (50%). *Classify-Verify* attains 0.5-$F1$ that outperforms both 0.5-*Base and* 1-*Base*.



F1-score results of the *Classify-Verify* method on the EBG corpus under imitation and obfuscation attacks, where the expected prob. of *in-set* and *not-in-set* documents is equal (50%). *Classify-Verify* successfully thwarts most of the attacks.



p-$F1_R$ results of the *Classify-Verify* method on the EBG corpus, where the expected prob. of *in-set* documents p varies from 10% to 90% and is assumed to be unknown. Robust p-independent thresholds are used for the underlying verifiers. *Classify-Verify* attains p-$F1_R$ that outperforms the respective p-*Base*.



p-$F1_R$ results of the *Classify-Verify* method on the blog corpus, where the expected prob. of *in-set* documents p varies from 10% to 90% and is assumed to be unknown. Robust p-independent thresholds are used for the underlying verifiers. *Classify-Verify* attains p-$F1_R$ that outperforms the respective p-*Base*.

∗*Distractorless* & *Sigma* Verification

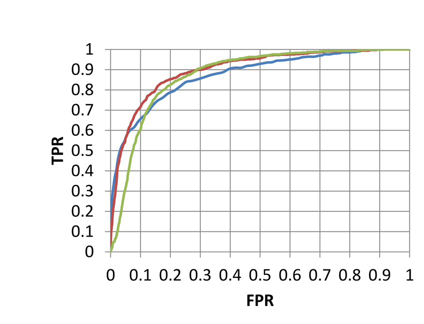
Distractorless – V [Noecker & Ryan, *LREC'12*] : verification based on vector distance between A’s centroid $\mathbb{A}$ and D, using cosine distance:

$$\delta(A, D) = \frac{\mathbb{A} \cdot D}{\|\mathbb{A}\|\|D\|} = \frac{\sum_{i=1}^{n} \mathbb{A}_i D_i}{\sum_{i=1}^{n} \mathbb{A}_i^2 \sum_{i=1}^{n} D_i^2}$$

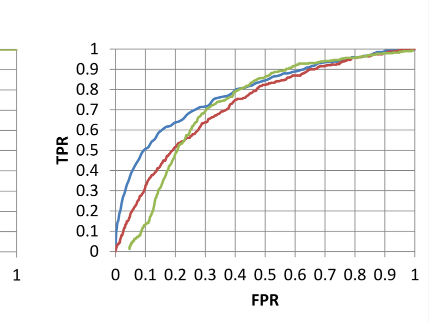
Sigma – V_σ^a : enhances distractorless verification with per-feature SD (V_σ) and per-author threshold (V^a) normalization

Distance \ Test	$\delta < t$	$\delta - \delta_A < t$
$\delta_{D,A} = \Delta(D_i, \mathbb{A}_i)_{i=1}^n$	V	V^a
$\delta_{D,A}^\sigma = \Delta(\frac{D_i}{\sigma(A)_i}, \frac{\mathbb{A}_i}{\sigma(A)_i})_{i=1}^n$	V_σ	V_σ^a

Differences in distance calculation and t-threshold test for V, V_σ and V^a.



ROC curves for V, V_σ and V_σ^a evaluation on the EBG corpus.



ROC curves for V, V_σ and V_σ^a evaluation on the blog corpus.

Dept. of Computer Science - Drexel University - 3175 JFK Blvd., Philadelphia, PA 19104
Mail: {stolerman,rjo43,sa499,greenie}@cs.drexel.edu
WWW: http://psal.cs.drexel.edu/