The web is full of anonymous communication that was never meant to be analyzed for authorship attribution.

Stylistics is a form of authorship attribution that relies on the linguistic information found in a document.

Stylistics 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.

**Motivation**

- The Closed-World Assumption
- Authorship Attribution
- Security & Privacy Applications

**Contribution and Application to Security & Privacy**

- The *Classify-Verify* method
- An abstaining classification approach that augments authorship classification with a verification step.
- 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**

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

**Problems:**
- Authorship Attribution: Which $A \in A$ is the author of $D$?
- The Classify-Verify Problem: given $D, A$ and optionally $p$:
  - Determine the author $A \in A$ of $D$,
  - Determine that $D$'s author is not in $A$ (w.r.t. acceptance threshold $t$)

**Classify-Verify**

**The Classify-Verify Algorithm**

**Input:** Document of suspect author set $A = \{A_1, ..., A_n\}$, target measure to maximize $\mu$.

**Optional:** in-set prob. $p$, manual threshold $t$.

**Output:** $A_i \in A$, or $\perp$. otherwise.

- $C_i = \text{classifier trained on } A_i$,
- $V_i = \{V_i(A_1), ..., V_i(A_n)\}$ – verifiers trained on $A_i$.

1. If $t$ is not set then
   - $t$ – threshold maximizing $\mu$ of Classify-Verify cross-validation on $A$.

2. If $V_i(D) \geq t$ then
   - Return $A_i$.

**Synopsis:**

- Train one closed-world classifier $C_A$ over $A$ and $n$ verifiers $V_i$.
- Classifier $D$ using $C_A$ and let the result be $A_i$.
- 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 $i$th order statistic of the probability outputs of $C_A(D)$, then:

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

**Standalone Verifiers**

- distractorest or Sigma verification

**Verification Acceptance Threshold $t$**

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

---

**Evaluation & Results**

- Corpora:
  - EBG: The Extended-Brennan-Greenstadt Adversarial corpus [Brennan et al., ACM Trans. Inf. Secur. Secur. 16:3, 45 authors]
  - Blog: The ICWSM 2009 Spinrider Blog dataset [Jure et al., ICWSM 2010, 50 authors]

- Closed-world classifier: SVM SMO

- Feature set: 500 most common character bigrams

**Results:**

**Distractorest & Sigma Verification**

- **Distractorest – $V$** [Noecker & Ryan, LREC’12]: Verification based on vector distance between $A$’s centroid $D$ and $D$, using cosine distance

$$i_r(A_i) = \frac{A_i \cdot D}{||A_i|| \cdot ||D||} = \frac{\sum_{i=1}^{n} A_i \cdot D}{\sum_{i=1}^{n} ||A_i|| \cdot ||D||}$$

**Sigma – $V^p$**:

- $V^p$: enhances distractorest verification with per-feature SMD ($V^p$) and per-author threshold ($V^p$) normalization

Distance $|D|$ Test

$$\delta_r(D) = \max_{A_i} \frac{A_i \cdot D}{||A_i|| \cdot ||D||}$$

$$i_r(A_i) = \frac{A_i \cdot D}{||A_i|| \cdot ||D||}$$

Differences in distance calculation and threshold test for $V$, $V^p$, and $V^{p'}$ evaluation on the EBG corpus.

**ROC curves for $V$, $V^p$, and $V^{p'}$ evaluation on the EBG corpus:**

---

Dept. of Computer Science - Drexel University - 3175 JFK Blvd., Philadelphia, PA 19104

Mail: {stoleram, rj443, sa499, greenie}@cs.drexel.edu

WWW: http://psal.cs.drexel.edu