Classify, but Verify

Breaking the Closed-World Assumption in Stylometric Authorship Attribution

Ariel Stolerman   Rebekah Overdorf   Sadia Afroz
Rachel Greenstadt

Drexel University
Philadelphia, PA

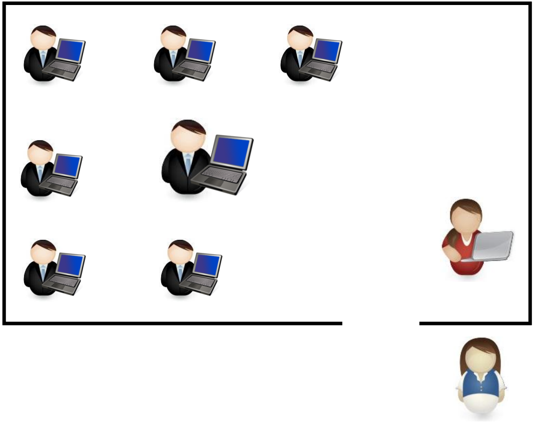
10^{th} IFIP WG 11.9, January 2014

Outline

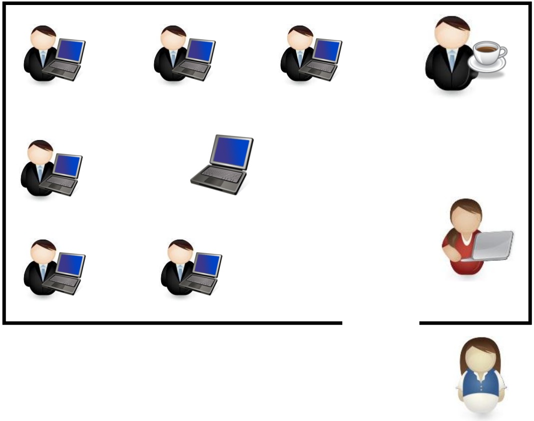
Outline

1. Motivation
2. Background
3. Corpora
4. Methodology
5. Evaluation
6. Conclusion

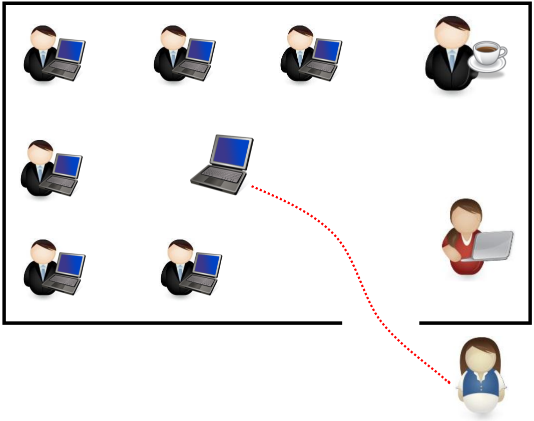
Motivation



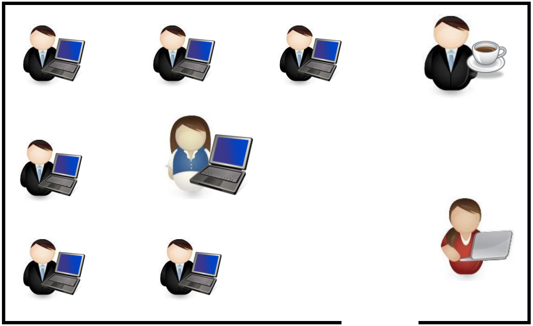
Motivation



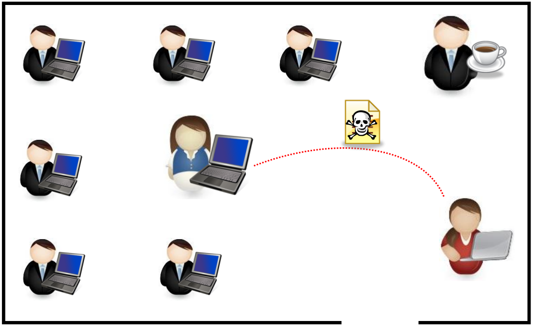
Motivation



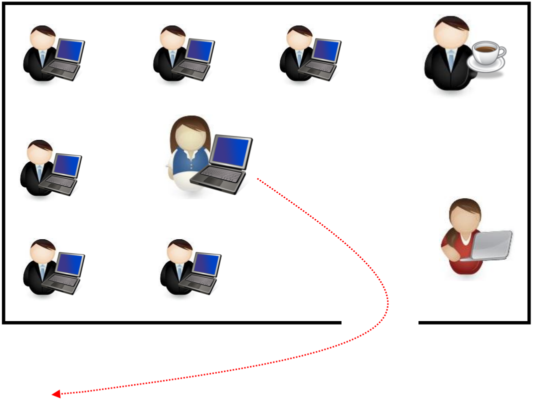
Motivation



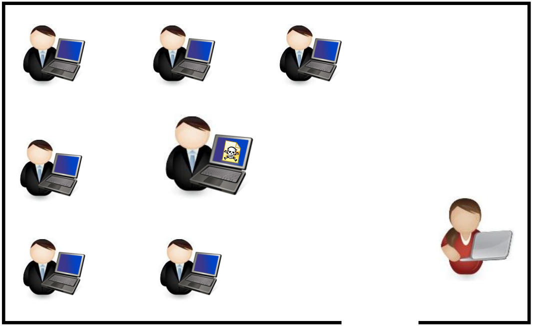
Motivation



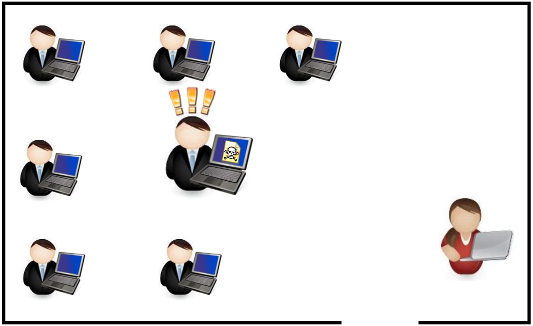
Motivation



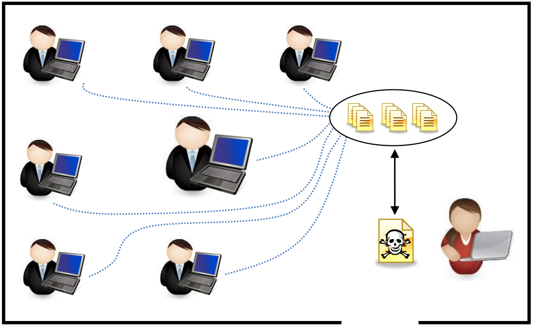
Motivation



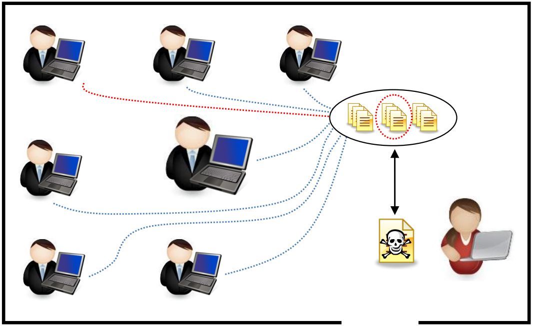
Motivation



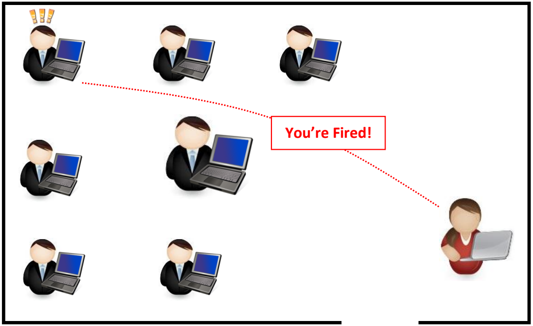
Motivation



Motivation



Motivation



Motivation – Contd.

- The web is full of anonymous communication
- Stylometry –
  The study of linguistic style applied to authorship attribution
  - "De-anonymizing" internet users
  - Accuracy & scale in constant increase
  - Used by law practitioners for forensic evidence
  - Requires closed set of candidate authors

Motivation – Contd.

- The web is full of anonymous communication
- Stylometry –
  The study of linguistic style applied to authorship attribution
  - "De-anonymizing" internet users
  - Accuracy & scale in constant increase
  - Used by law practitioners for forensic evidence
  - Requires closed set of candidate authors

Motivation – Contd.

- The web is full of anonymous communication
- Stylometry –
  The study of linguistic style applied to authorship attribution
  - "De-anonymizing" internet users
  - Accuracy & scale in constant increase
  - Used by law practitioners for forensic evidence
  - Requires closed set of candidate authors

Motivation – Contd.

- The web is full of anonymous communication
- Stylometry –
  The study of linguistic style applied to authorship attribution
  - "De-anonymizing" internet users
  - Accuracy & scale in constant increase
  - Used by law practitioners for forensic evidence
  - Requires closed set of candidate authors

Motivation – Contd.

- The web is full of anonymous communication
- Stylometry –
  The study of linguistic style applied to authorship attribution
  - “De-anonymizing” internet users
  - Accuracy & scale in constant increase
  - Used by law practitioners for forensic evidence
  - Requires closed set of candidate authors

Motivation – Contd.

- The web is full of anonymous communication
- Stylometry –
  The study of linguistic style applied to authorship attribution
  - “De-anonymizing” internet users
  - Accuracy & scale in constant increase
  - Used by law practitioners for forensic evidence
  - Requires closed set of candidate authors

Motivation – Contd.

- Pseudonymous documents published on the web:
  - Virtually ∞ suspects
  - Or lack of training data
- $\Rightarrow$ problem for:
  - Analysts: confidence in suspect pool
  - Users: may be falsely accused of authorship

Motivation – Contd.

- Pseudonymous documents published on the web:
  - Virtually ∞ suspects
  - Or lack of training data
- $\Rightarrow$ problem for:
  - Analysts: confidence in suspect pool
  - Users: may be falsely accused of authorship

Motivation – Contd.

- Pseudonymous documents published on the web:
  - Virtually ∞ suspects
  - Or lack of training data
- $\Rightarrow$ problem for:
  - Analysts: confidence in suspect pool
  - Users: may be falsely accused of authorship

Motivation – Contd.

- Pseudonymous documents published on the web:
  - Virtually ∞ suspects
  - Or lack of training data
- $\Rightarrow$ problem for:
  - Analysts: confidence in suspect pool
  - Users: may be falsely accused of authorship

Motivation – Contd.

- This work: explore mixed open/closed-world stylometry
  - Closed set of candidate authors
  - Take into account author may not be in the set
- ⇒ *Classify-Verify* algorithm: classification + binary verification
  - Intercepts misclassifications
  - Tunable rigidity – FAR/FRR
  - Performs better in adversarial settings than traditional stylometry

Motivation – Contd.

- This work: explore mixed open/closed-world stylometry
  - Closed set of candidate authors
  - Take into account author may not be in the set
- ⇒ *Classify-Verify* algorithm: classification + binary verification
  - Intercepts misclassifications
  - Tunable rigidity – FAR/FRR
  - Performs better in adversarial settings than traditional stylometry

Motivation – Contd.

- This work: explore mixed open/closed-world stylometry
  - Closed set of candidate authors
  - Take into account author may not be in the set
- ⇒ *Classify-Verify* algorithm: classification + binary verification
  - Intercepts misclassifications
  - Tunable rigidity – FAR/FRR
  - Performs better in adversarial settings than traditional stylometry

Motivation – Contd.

- This work: explore mixed open/closed-world stylometry
  - Closed set of candidate authors
  - Take into account author may not be in the set
- ⇒ *Classify-Verify* algorithm: classification + binary verification
  - Intercepts misclassifications
  - Tunable rigidity – FAR/FRR
  - Performs better in adversarial settings than traditional stylometry

Motivation – Contd.

- This work: explore mixed open/closed-world stylometry
  - Closed set of candidate authors
  - Take into account author may not be in the set
- ⇒ *Classify-Verify* algorithm: classification + binary verification
  - Intercepts misclassifications
  - Tunable rigidity – FAR/FRR
  - Performs better in adversarial settings than traditional stylometry

Motivation – Contd.

- This work: explore mixed open/closed-world stylometry
  - Closed set of candidate authors
  - Take into account author may not be in the set
- $\Rightarrow$ *Classify-Verify* algorithm: classification + binary verification
  - Intercepts misclassifications
  - Tunable rigidity – FAR/FRR
  - Performs better in adversarial settings than traditional stylometry

Motivation – Contd.

- This work: explore mixed open/closed-world stylometry
  - Closed set of candidate authors
  - Take into account author may not be in the set
- $\Rightarrow$ *Classify-Verify* algorithm: classification + binary verification
  - Intercepts misclassifications
  - Tunable rigidity – FAR/FRR
  - Performs better in adversarial settings than traditional stylometry

Outline

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problem Statement

- Problem building blocks:
  - D: document of unknown authorship
  - $\mathcal{A} = \{A_1, ..., A_n\}$: set of known authors
  - $p = Pr[A_D \in \mathcal{A}]$: probability D's author is a candidate
- Problems:
  - Closed-world – attribution: find $A_D \in \mathcal{A}$
  - Open-world: find D's author
    - Verification: is A the author of D?
  - Classify/Verify: find D's author in $\mathcal{A}$ or determine $A_D \notin \mathcal{A}$
    - Optional: given p
- Notations:
  - *in-set* : documents whose author is a candidate
  - *not-in-set* : documents whose author is not a candidate

Problems in Closed-World Models

- Closed-world models applied in open-world settings: Classifier always outputs an author
  - Chosen author is merely least-worst choice
  - Absence of true author from pool is unknown
- Extremely relevant for stylometry in online domains

Problems in Closed-World Models

- Closed-world models applied in open-world settings: Classifier always outputs an author
  - Chosen author is merely least-worst choice
  - Absence of true author from pool is unknown
- Extremely relevant for stylometry in online domains

Problems in Closed-World Models

- Closed-world models applied in open-world settings: Classifier always outputs an author
  - Chosen author is merely least-worst choice
  - Absence of true author from pool is unknown
- Extremely relevant for stylometry in online domains

Problems in Closed-World Models

- Closed-world models applied in open-world settings: Classifier always outputs an author
  - Chosen author is merely least-worst choice
  - Absence of true author from pool is unknown
- Extremely relevant for stylometry in online domains

Outline

Corpora

- Brennan-Greenstadt Adversarial Corpus (EBG) [BAG12]
  - 45 authors, > 6500 words each
  - Adversarial documents: deliberate style change
- ICWSM 2009 Spinn3r Blog dataset (blog) [BJS09]
  - 44M blogs, previously used for web-scale stylometry
  - Here using subcorpus of 50 authors, > 7500 words each
  - Used as control to avoid overfitting on EBG

Corpora

- Brennan-Greenstadt Adversarial Corpus (EBG) [BAG12]
  - 45 authors, > 6500 words each
  - Adversarial documents: deliberate style change
- ICWSM 2009 Spinn3r Blog dataset (blog) [BJS09]
  - 44M blogs, previously used for web-scale stylometry
  - Here using subcorpus of 50 authors, > 7500 words each
  - Used as control to avoid overfitting on EBG

Corpora

- Brennan-Greenstadt Adversarial Corpus (EBG) [BAG12]
  - 45 authors, > 6500 words each
  - Adversarial documents: deliberate style change
- ICWSM 2009 Spinn3r Blog dataset (blog) [BJS09]
  - 44M blogs, previously used for web-scale stylometry
  - Here using subcorpus of 50 authors, > 7500 words each
  - Used as control to avoid overfitting on EBG

Corpora

- Brennan-Greenstadt Adversarial Corpus (EBG) [BAG12]
  - 45 authors, > 6500 words each
  - Adversarial documents: deliberate style change
- ICWSM 2009 Spinn3r Blog dataset (blog) [BJS09]
  - 44M blogs, previously used for web-scale stylometry
  - Here using subcorpus of 50 authors, > 7500 words each
  - Used as control to avoid overfitting on EBG

Corpora

- Brennan-Greenstadt Adversarial Corpus (EBG) [BAG12]
  - 45 authors, > 6500 words each
  - Adversarial documents: deliberate style change
- ICWSM 2009 Spinn3r Blog dataset (blog) [BJS09]
  - 44M blogs, previously used for web-scale stylometry
  - Here using subcorpus of 50 authors, > 7500 words each
  - Used as control to avoid overfitting on EBG

Outline

1. Motivation
2. Background
3. Corpora
4. Methodology
5. Evaluation
6. Conclusion

Closed-World Setup

- SMO SVM as underlying classifier for the “Classify” phase
- Feature set
  - *Writeprints* – extensive feature set
    Lexical, syntactic, content, grammar, idiosyncrasies...
  - $k \in \{50, ..., 1000\}$ most common $n \in \{1, ..., 5\}$-grams
    $\langle k, n\rangle$*-chars*, $\langle k, n\rangle$*-words*
- Processing w/ the JStylo authorship attribution framework
- Evaluation: F1-Score of 10-fold cross-validation

Closed-World Setup

- SMO SVM as underlying classifier for the "Classify" phase
- Feature set
  - *Writeprints* – extensive feature set
    Lexical, syntactic, content, grammar, idiosyncrasies...
  - $k \in \{50, ..., 1000\}$ most common $n \in \{1, ..., 5\}$-grams
    $\langle k, n \rangle$*-chars*, $\langle k, n \rangle$*-words*
- Processing w/ the JStylo authorship attribution framework
- Evaluation: F1-Score of 10-fold cross-validation

Closed-World Setup

- SMO SVM as underlying classifier for the "Classify" phase
- Feature set
  - *Writeprints* – extensive feature set
    Lexical, syntactic, content, grammar, idiosyncrasies...
  - $k \in \{50, ..., 1000\}$ most common $n \in \{1, ..., 5\}$-grams
    $\langle k, n \rangle$*-chars*, $\langle k, n \rangle$*-words*
- Processing w/ the JStylo authorship attribution framework
- Evaluation: F1-Score of 10-fold cross-validation

Closed-World Setup

- SMO SVM as underlying classifier for the "Classify" phase
- Feature set
  - *Writeprints* – extensive feature set
    Lexical, syntactic, content, grammar, idiosyncrasies...
  - $k \in \{50, ..., 1000\}$ most common $n \in \{1, ..., 5\}$-grams
    $\langle k, n \rangle$*-chars*, $\langle k, n \rangle$*-words*
- Processing w/ the JStylo authorship attribution framework
- Evaluation: F1-Score of 10-fold cross-validation

Closed-World Setup

- SMO SVM as underlying classifier for the "Classify" phase
- Feature set
  - *Writeprints* – extensive feature set
    Lexical, syntactic, content, grammar, idiosyncrasies...
  - $k \in \{50, ..., 1000\}$ most common $n \in \{1, ..., 5\}$-grams
    $\langle k, n\rangle$*-chars*, $\langle k, n\rangle$*-words*
- Processing w/ the JStylo authorship attribution framework
- Evaluation: F1-Score of 10-fold cross-validation

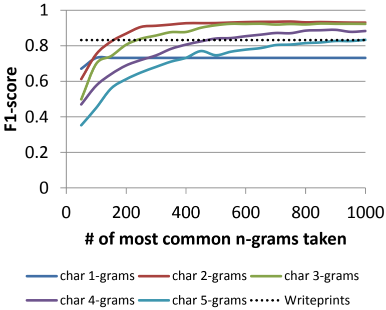
Closed-World Setup

- SMO SVM as underlying classifier for the “Classify” phase
- Feature set
  - *Writeprints* – extensive feature set
    Lexical, syntactic, content, grammar, idiosyncrasies...
  - $k \in \{50, ..., 1000\}$ most common $n \in \{1, ..., 5\}$-grams
    $\langle k, n \rangle$*-chars*, $\langle k, n \rangle$*-words*
- Processing w/ the JStylo authorship attribution framework
- Evaluation: F1-Score of 10-fold cross-validation

Closed-World Setup – Contd.

- $\langle 500, 2\rangle$-*chars* wins w/ $F1 = 0.928$ (EBG)
- Also wins on blogs: $F1 = 0.64$

Performance by Feature Sets



Verification

- Authorship Verification: is D written by A?
  - Naïve #1: reduce to 1-vs-all modeling *not*-A
  - Naïve #2: cross validate A vs D & test distinguishability
- Verification methods:
  - Classifier-induced: based on closed-world classifier outputs
  - Standalone: models built using A's training data *only*

Verification

- Authorship Verification: is *D* written by *A*?
  - Naïve #1: reduce to 1-vs-all modeling *not*-*A*
  - Naïve #2: cross validate *A* vs *D* & test distinguishability
- Verification methods:
  - Classifier-induced: based on closed-world classifier outputs
  - Standalone: models built using *A*'s training data *only*

Verification

- Authorship Verification: is D written by A?
  - Naïve #1: reduce to 1-vs-all modeling *not*-A
  - Naïve #2: cross validate A vs D & test distinguishability
- Verification methods:
  - Classifier-induced: based on closed-world classifier outputs
  - Standalone: models built using A's training data *only*

Verification

- Authorship Verification: is D written by A?
  - Naïve #1: reduce to 1-vs-all modeling *not*-A
  - Naïve #2: cross validate A vs D & test distinguishability
- Verification methods:
  - Classifier-induced: based on closed-world classifier outputs
  - Standalone: models built using A's training data *only*

Verification

- Authorship Verification: is D written by A?
  - Naïve #1: reduce to 1-vs-all modeling *not*-A
  - Naïve #2: cross validate A vs D & test distinguishability
- Verification methods:
  - Classifier-induced: based on closed-world classifier outputs
  - Standalone: models built using A's training data *only*

Verification

- Authorship Verification: is D written by A?
  - Naïve #1: reduce to 1-vs-all modeling *not*-A
  - Naïve #2: cross validate A vs D & test distinguishability
- Verification methods:
  - Classifier-induced: based on closed-world classifier outputs
  - Standalone: models built using A's training data *only*

Standalone Verification

- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char n-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for "looser" ones
  - V_σ : per-feature SD normalization
  - V^a : account for A's avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification

- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char n-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for "looser" ones
  - V_σ : per-feature SD normalization
  - V^a : account for A's avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification

- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char n-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for “looser” ones
  - V_σ : per-feature SD normalization
  - V^a : account for A’s avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification

- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char n-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for “looser” ones
  - V_σ : per-feature SD normalization
  - V^a : account for A's avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification

- *V*: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char *n*-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for "looser" ones
  - V_σ : per-feature SD normalization
  - V^a : account for *A*'s avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification

- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char *n*-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for “looser” ones
  - V_σ : per-feature SD normalization
  - V^a : account for A's avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification

- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char n-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for “looser” ones
  - V_σ : per-feature SD normalization
  - V^a : account for A's avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification

- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char n-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for "looser" ones
  - V_σ : per-feature SD normalization
  - V^a : account for A's avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification

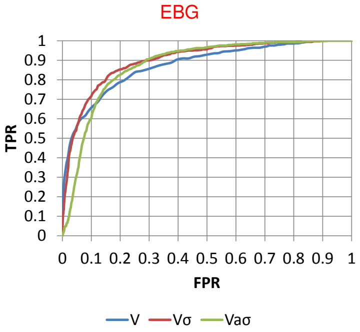
- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char n-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for “looser” ones
  - V_σ : per-feature SD normalization
  - V^a : account for A's avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

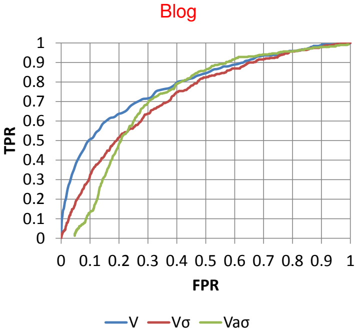
Standalone Verification

- V: Distractorless Verification [NR12]
  - Standardize char-case & whitespaces, extract word/char n-grams
  - Author model $M = \langle m_1, m_2, ..., m_n \rangle$
  - Document model $F = \langle f_1, f_2, ..., f_n \rangle$
  - Test: $\delta(M, F) < t$?
- Variants:
  - Tighten bound for less varied authors, widen for “looser” ones
  - V_σ : per-feature SD normalization
  - V^a : account for A's avg. pairwise document distances
  - Evaluation w/ 10-fold CV + $\langle 500, 2 \rangle$-*chars*

Standalone Verification – Contd.

- ROC curves: no method is strictly preferred over the other
  - EBG (left): V_σ wins, Blog (right): V wins





Classify-Verify

- Abstaining classifier: refrain when not sure
- Closed-world classifier + verifier $\rightarrow$ open-world
- Output range: $\mathcal{A} \rightarrow \mathcal{A} \cup \{\perp\}$
  - $\perp$ = "unknown"/"I don't know"
- Manual/automatically set verification threshold t
- Aim to maximize *p-F1* : weighted avg. F1-scores over
  - *p in-set* documents
  - $1 - p$ *not-in-set* documents

Classify-Verify

- Abstaining classifier: refrain when not sure
- Closed-world classifier + verifier $\rightarrow$ open-world
- Output range: $\mathcal{A} \rightarrow \mathcal{A} \cup \{\perp\}$
  - $\perp =$ "unknown"/"I don't know"
- Manual/automatically set verification threshold t
- Aim to maximize *p-F1* : weighted avg. F1-scores over
  - *p in-set* documents
  - $1 - p$ *not-in-set* documents

Classify-Verify

- Abstaining classifier: refrain when not sure
- Closed-world classifier + verifier $\rightarrow$ open-world
- Output range: $\mathcal{A} \rightarrow \mathcal{A} \cup \{\perp\}$
  - $\perp =$ "unknown"/"I don't know"
- Manual/automatically set verification threshold t
- Aim to maximize p-$F1$: weighted avg. F1-scores over
  - p *in-set* documents
  - $1 - p$ *not-in-set* documents

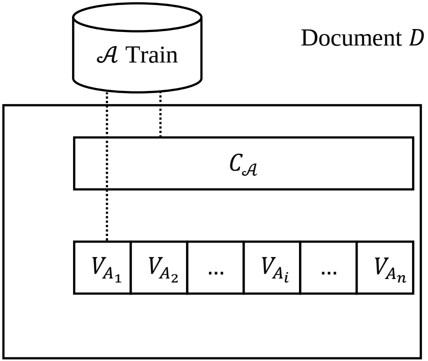
Classify-Verify

- Abstaining classifier: refrain when not sure
- Closed-world classifier + verifier $\rightarrow$ open-world
- Output range: $\mathcal{A} \rightarrow \mathcal{A} \cup \{\perp\}$
  - $\perp =$ "unknown"/"I don't know"
- Manual/automatically set verification threshold t
- Aim to maximize *p-F1* : weighted avg. F1-scores over
  - *p in-set* documents
  - $1 - p$ *not-in-set* documents

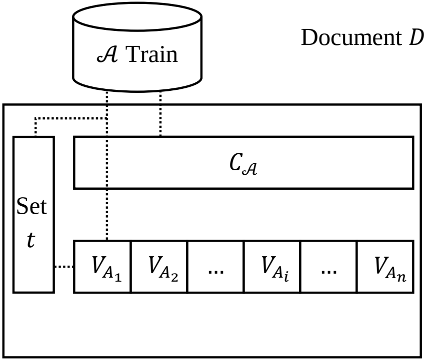
Classify-Verify

- Abstaining classifier: refrain when not sure
- Closed-world classifier + verifier $\rightarrow$ open-world
- Output range: $\mathcal{A} \rightarrow \mathcal{A} \cup \{\perp\}$
  - $\perp =$ "unknown"/"I don't know"
- Manual/automatically set verification threshold t
- Aim to maximize p-$F1$: weighted avg. F1-scores over
  - p *in-set* documents
  - $1 - p$ *not-in-set* documents

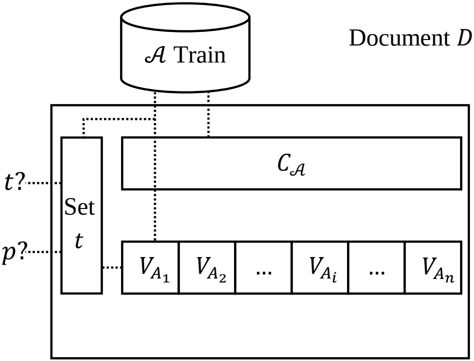
Classify-Verify – Flow



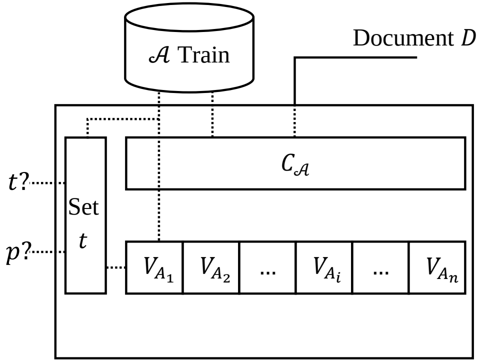
Classify-Verify – Flow



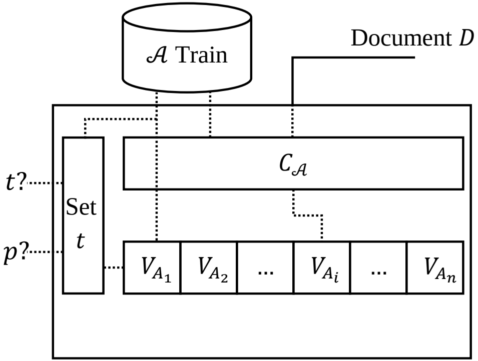
Classify-Verify – Flow



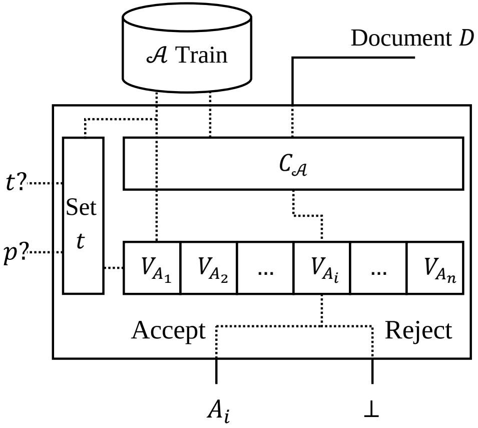
Classify-Verify – Flow



Classify-Verify – Flow

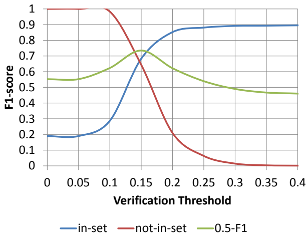


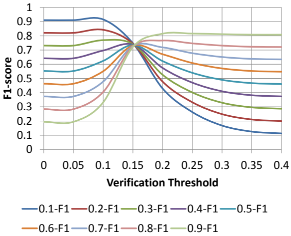
Classify-Verify – Flow



Classify-Verify – Threshold Selection

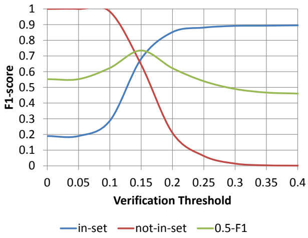
- Manual: [accept] relaxed $\rightarrow$ strict [reject]
- p-Induced: t set empirically over training set to maximize p-$F1$
- p-Robust: set t at intersection of all p-induced curves (p-$F1_R$)

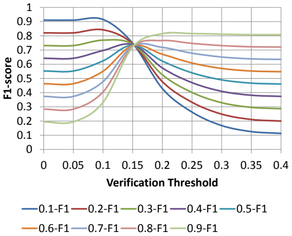




Classify-Verify – Threshold Selection

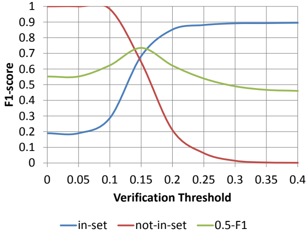
- Manual: [accept] relaxed $\rightarrow$ strict [reject]
- *p*-Induced: *t* set empirically over training set to maximize $p\text{-}F1$
- *p*-Robust: set *t* at intersection of all *p*-induced curves ($p\text{-}F1_R$)

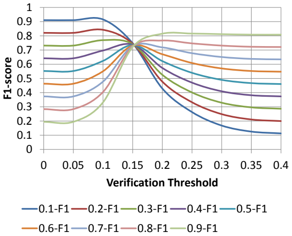




Classify-Verify – Threshold Selection

- Manual: [accept] relaxed $\rightarrow$ strict [reject]
- p-Induced: t set empirically over training set to maximize p-$F1$
- p-Robust: set t at intersection of all p-induced curves (p-$F1_R$)





Outline

Evaluation Methodology

- 10-fold cross-validation
- Credit thwarted misclassifications as true (even if $A_D \in \mathcal{A}$)
- Each D is evaluated twice: as *in-set* and *not-in-set*

Evaluation Methodology

- 10-fold cross-validation
- Credit thwarted misclassifications as true (even if $A_D \in \mathcal{A}$)
- Each D is evaluated twice: as *in-set* and *not-in-set*

Evaluation Methodology

- 10-fold cross-validation
- Credit thwarted misclassifications as true (even if $A_D \in \mathcal{A}$)
- Each D is evaluated twice: as *in-set* and *not-in-set*

Evaluation Methodology – Adversarial Settings

- Train same models, test adversarial documents
  - Where authors try to hide their style
- Here $\perp$ can be considered both “unknown” or “possible attack”
- Measure 0.5-$F1$

Evaluation Methodology – Adversarial Settings

- Train same models, test adversarial documents
  - Where authors try to hide their style
- Here $\perp$ can be considered both "unknown" or "possible attack"
- Measure 0.5-$F1$

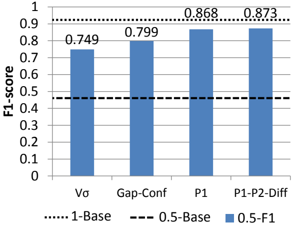
Evaluation Methodology – Adversarial Settings

- Train same models, test adversarial documents
  - Where authors try to hide their style
- Here $\perp$ can be considered both “unknown” or “possible attack”
- Measure 0.5-$F1$

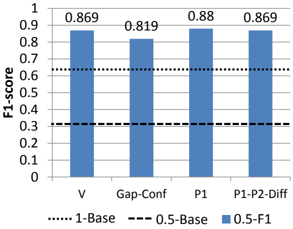
Results

- Known $p = 0.5$: *Classify-Verify* significantly outperforms closed-world classifier
- 0.5-$F1$ > 0.5-*Base*
- 0.5-$F1$ even close to/better than 1-*Base*

EBG

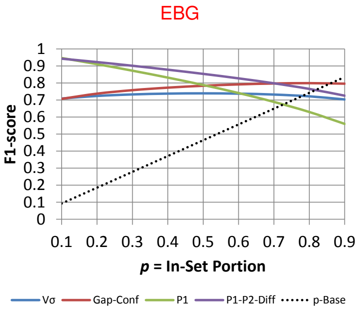


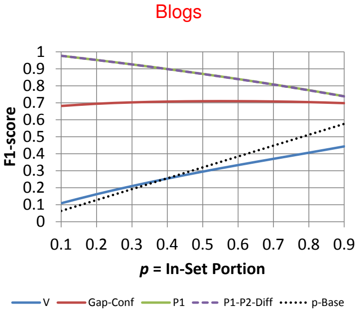
Blogs



Results – Contd.

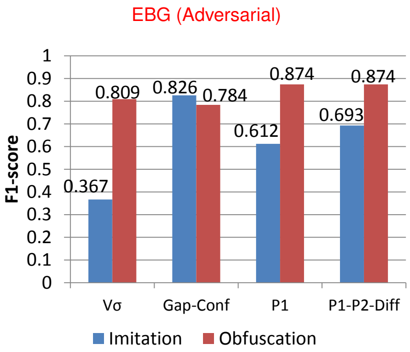
- Unknown p: *Classify-Verify* still significantly outperforms closed-world classifier
- $p\text{-}F1_R > p\text{-}Base$ almost always





Results – Adversarial Settings

- *Classify-Verify* successfully thwarts most attacks
- Similar to results under standard settings
- Baseline: F1 = 0–0.04 for imitation/obfuscation attacks



Outline

Conclusion

- *Classify-Verify* is effective in open-world settings
  - Also more effective in closed-world settings than closed-world classifiers
- Effective in thwarting attacks
  - Without special “defensive” configuration
- $\Rightarrow$ *Classify-Verify* is preferable over closed-world classifiers almost always
  - Essential tool for forensic analysis of open-world problems
- Future work:
  - Other classification-based applications
  - Fusion of verification methods in the “verify” phase
  - Utilization for scalability: divide-and-conquer

Conclusion

- *Classify-Verify* is effective in open-world settings
  - Also more effective in closed-world settings than closed-world classifiers
- Effective in thwarting attacks
  - Without special "defensive" configuration
- $\Rightarrow$ *Classify-Verify* is preferable over closed-world classifiers almost always
  - Essential tool for forensic analysis of open-world problems
- Future work:
  - Other classification-based applications
  - Fusion of verification methods in the "verify" phase
  - Utilization for scalability: divide-and-conquer

Conclusion

- *Classify-Verify* is effective in open-world settings
  - Also more effective in closed-world settings than closed-world classifiers
- Effective in thwarting attacks
  - Without special “defensive” configuration
- $\Rightarrow$ *Classify-Verify* is preferable over closed-world classifiers almost always
  - Essential tool for forensic analysis of open-world problems
- Future work:
  - Other classification-based applications
  - Fusion of verification methods in the “verify” phase
  - Utilization for scalability: divide-and-conquer

Conclusion

- *Classify-Verify* is effective in open-world settings
  - Also more effective in closed-world settings than closed-world classifiers
- Effective in thwarting attacks
  - Without special “defensive” configuration
- $\Rightarrow$ *Classify-Verify* is preferable over closed-world classifiers almost always
  - Essential tool for forensic analysis of open-world problems
- Future work:
  - Other classification-based applications
  - Fusion of verification methods in the “verify” phase
  - Utilization for scalability: divide-and-conquer

Conclusion

- *Classify-Verify* is effective in open-world settings
  - Also more effective in closed-world settings than closed-world classifiers
- Effective in thwarting attacks
  - Without special “defensive” configuration
- $\Rightarrow$ *Classify-Verify* is preferable over closed-world classifiers almost always
  - Essential tool for forensic analysis of open-world problems
- Future work:
  - Other classification-based applications
  - Fusion of verification methods in the “verify” phase
  - Utilization for scalability: divide-and-conquer

Conclusion

- *Classify-Verify* is effective in open-world settings
  - Also more effective in closed-world settings than closed-world classifiers
- Effective in thwarting attacks
  - Without special "defensive" configuration
- $\Rightarrow$ *Classify-Verify* is preferable over closed-world classifiers almost always
  - Essential tool for forensic analysis of open-world problems
- Future work:
  - Other classification-based applications
  - Fusion of verification methods in the "verify" phase
  - Utilization for scalability: divide-and-conquer

Conclusion

- *Classify-Verify* is effective in open-world settings
  - Also more effective in closed-world settings than closed-world classifiers
- Effective in thwarting attacks
  - Without special "defensive" configuration
- $\Rightarrow$ *Classify-Verify* is preferable over closed-world classifiers almost always
  - Essential tool for forensic analysis of open-world problems
- Future work:
  - Other classification-based applications
  - Fusion of verification methods in the "verify" phase
  - Utilization for scalability: divide-and-conquer

Thank You

Thank You!

Questions?

- Contact: stolerman@cs.drexel.edu
- Drexel Privacy Security & Automation Lab: http://psal.cs.drexel.edu/

For Further Reading I

Michael Brennan, Sadia Afroz, and Rachel Greenstadt.
Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity.
ACM Trans. Inf. Syst. Secur., 15(3):12:1–12:22, November 2012.

Kevin Burton, Akshay Java, and Ian Soboroff.
The icwsm 2009 spinn3r dataset.
In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA*, 2009.

John Noecker, Jr. and Michael Ryan.
Distractorless authorship verification.
In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).