Adversarial Stylometry: Attacking Authorship Recognition to Preserve Privacy and Anonymity

Michael Brennan
Drexel University
Privacy, Security and Automation Lab
Overview

• Stylometry
  • Overview and brief history.
  • Some current methods.
• Privacy and Stylometry
  • Stylometry’s potential threat to privacy and anonymity.
  • Attacking Authorship Recognition to Protect Privacy
    • We see how it is possible to anonymize writing styles.
    • The potential for deception analysis.
• Postscript: Geeks in Washington
  • A Technologist at the Federal Trade Commission
What is Authorship Recognition?

• The basic question: “who wrote this document?”

• **Stylometry**: The study of attributing authorship to documents based only on the linguistic style they exhibit.
  
  • “Linguistic Style” Features: sentence length, word choices, syntactic structure, etc.

  • Handwriting, content-based features, and contextual features are not considered.

• Individuals have unique writing styles because language is learned on an individual basis.

• In this presentation, stylometry and authorship recognition are used interchangeably.
What is Adversarial Stylometry?

• Adversarial Stylometry: Applying deception to writing style in order to affect the outcome of stylometric analysis.
  • But, is writing style modifiable?
  • Is it possible to “attack” stylometry through altered writing style?
  • What are the implications of looking at stylometry in an adversarial context?
Stylometry: A Brief History

• The classic stylometry problem: The Federalist Papers.
  • 85 anonymous papers to persuade ratification of the Constitution. 12 of these have disputed authorship.
  • Stylometry has been used to show Madison authored the disputed documents.
• Used as a data set for countless stylometry studies.
• Modern Stylometry Based in Machine Learning
  • SVMs, Genetic Algorithms, Neural Networks, Bayesian Classifiers... used extensively.
Why is Stylometry Important?

• Great, you can figure out who wrote a 200 year old document, so what?

• From the Institute for Linguistic Evidence:
  • “In some criminal, civil, and security matters, language can be evidence... When you are faced with a suspicious document, whether you need to know who wrote it, or if it is a real threat or real suicide note, or if it is too close for comfort to some other document, you need reliable, validated methods.”

• Plagiarism, Forensics, Anonymity...
The Value of Anonymity

• The value of privacy and anonymous speech is widely recognized.

• “And we must also grapple with the issue of anonymous speech. Those who use the internet to recruit terrorists or distribute stolen intellectual property cannot divorce their online actions from their real world identities. But these challenges must not become an excuse for governments to systematically violate the rights and privacy of those who use the internet for peaceful political purposes.” – Hilary Clinton, 1/21/2010
Stylometry: the Threat to Privacy and Anonymity

• Good techniques for location privacy (Tor, Mixes, etc).
  • But it may be insufficient!
• Stylometry can identify authors based on their writing.
  • Can anonymous authors defend against this?
  • ~6500 words to leak identity – Rao, Rohatgi. 2000.
Supervised Stylometry

• Given a set of documents of known authorship, classify a document of unknown authorship.
  • Classifier trained on undisputed text.
• Scenario: Alice the Anonymous Blogger vs. Bob the Abusive Employer.
  • Alice blogs about abuses at Bob’s company.
    • Blog posted anonymously (Tor, pseudonym, etc).
  • Bob obtains 5000-10000 words of each employee’s writing.
    • Bob uses stylometry to identify Alice as the blogger.
Unsupervised Stylometry

• Given a set of documents of unknown authorship, cluster them into author groups.
  • No pre-existing author information.
  • “Similarity Detection”

• Scenario: Anonymous Forum vs. Oppressive Government.
  • Participants organize protests.
    • Posts are completely unlabeled (no pseudonyms)
    • Unknown organizational structure, number of authors, etc.
  • The government applies unsupervised stylometric techniques.
    • Number of authors may be discovered, author profiles created.
    • Results fed into supervised stylometry system to identify individuals.
Purely Hypothetical?

• Previous examples are purely hypothetical. What about a real example?
• From “Inside WikiLeaks” by Daniel Domscheit-Berg:
  • “I nudged Julian with my foot. We exchanged glances and started giggling. If someone had run WikiLeaks documents through such a program, he would have discovered that the same two people were behind all the various press releases, document summaries, and correspondence issued by the project. The official number of volunteers we had was also, to put it mildly, grotesquely exaggerated.”
Preserving Privacy: Attacking Stylometry

• Can stylometry be attacked? How so? How Easily?
• In depth study on attacking multiple methods of Stylometry.

• Results:
  • Stylometry is very vulnerable to attack by inexperienced human adversaries.
  • Attacks can be used to protect privacy.
We’re Under Attack!

• **Obfuscation Attack**
  • An author attempts to write a document in such a way that their personal writing style will not be recognized.

• **Imitation Attack**
  • An author attempts to write a document such that the writing style will be recognized as that of another specific author.

• **Translation Attack**
  • Machine translation is used to translate a document to one or more languages and then back to the original language.
Building a Corpus

• 12 Individual Authors. Participation had three parts:
  • Submit 5000 words of pre-existing writing from a formal source.
  • Write a new 500 word passage as an obfuscation attack.
    • Task: Describe your neighborhood.
  • Write a new 500 word passage as an imitation attack.
    • Task: Imitate Cormac McCarthy, describe your day.
• Authors had no formal training or knowledge in linguistics or stylometry.
• Corpus is publicly available at psal.cs.drexel.edu
Imitating Cormac McCarthy

“On the far side of the river valley the road passed through a stark black burn. Charred and limbless trunks of trees stretching away on every side. Ash moving over the road and the sagging hands of blind wire strung from the blackened lightpoles whining thinly in the wind.”
Imitation Attack Examples

• “Light sliced through the blinds, and construction began in the adjacent apartment. The harsh cacophony crashed through the wall.”
• “Hot water in the mug. Brush in the mug. The blade read ‘Wilkinson Sword’ on the layered wax paper packaging.”
• “He fills the coffee pot with water, after cleaning out the putrid remains of yesterday's brew. The beans are in the freezer, he remembers.”
Methodology

• Tested 4 methods of Authorship Recognition
  • Neural Network w/ 9 Features
  • Support Vector Machine w/ 9 Features
  • Synonym-Based Method
  • Writeprints
• Cross-validated across randomly selected data from a set of 12 unique authors.
Linguistic Features

• Function Words
  • Word with little or no meaning, used to define relationships between content words (“of”, “and” vs. “bed”, “dancing”)

• Vocabulary
  • Spelling (“color” vs. “colour”).
  • Vocabulary richness, lexical density.
  • Syllable count.

• Syntax
  • Part-of-Speech-tagging, punctuation, n-gram (word and characters).

• Gunning-Fog Readability Index:
  • 0.4 * ( Average_Sentence_Length + 100 * Complex_Word_Ratio )
  • Result: years of formal education required to read the text.

• Other Features
  • Structural features (sentence length)
  • Idiosyncratic features (misspellings)
Synonym-Based Approach

• Clark & Hannon, 2007.
• Good demonstration of single feature type stylometry. Vocabulary features based on word choice.
• Hypothesis: Repeated choice of a specific word given a selection of potential synonyms is tightly related to writing style.
Synonym-Based Approach

Correct classification of “Unknown” is author “X.”
Synonym-Based Approach

• Match value:
  • Apply stop-word list (319 most common words)
  • N: shared text frequency of word W.
  • G: global frequency of word W.
  • U: global frequencies of all synonyms of W.
  • \( M = M + \frac{f[N] \times f[U]}{f[G]} \)
• Update the match value for each word in the passage.
• Do this for every known author K given an unknown passage P.
Writeprints
Features

- “Baseline” and “Extended” feature sets (327 vs. thousands of features).
- Individual-Author-Level Feature Sets
  - Different feature set for each author.
  - Only contains features relevant to that author.

<table>
<thead>
<tr>
<th>Group</th>
<th>Category</th>
<th>Quantity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Baseline (BF)</td>
<td>Extended (EF)</td>
</tr>
<tr>
<td>Lexical</td>
<td>Word-Level</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Character-Level</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Letters</td>
<td>26</td>
<td>26</td>
<td>count of letters (e.g., a, b, c)</td>
</tr>
<tr>
<td>Character Bigrams</td>
<td></td>
<td>&lt;676</td>
<td>letter bigrams (e.g., aa, ab, ac)</td>
</tr>
<tr>
<td>Character Trigrams</td>
<td></td>
<td>&lt;17,576</td>
<td>letter trigrams (e.g., aaa, aab, aac)</td>
</tr>
<tr>
<td>Digits</td>
<td></td>
<td>10</td>
<td>digits (e.g., 1, 2, 3)</td>
</tr>
<tr>
<td>Digit Bigrams</td>
<td></td>
<td>&lt;100</td>
<td>2 digit number frequencies</td>
</tr>
</tbody>
</table>
The Writeprints Algorithm

- Writeprint Creation – Positive correlation between text and author.
  - Feature vector created for all non-zero features.
  - N-Dimensional representation of author’s feature set.
- Pattern Disruption – Negative correlation between text and author.
  - Disruptor feature – word.
  - Disruptor coefficient $d$ applied to author’s writeprint.

$$d_p = IG(c, p)K(syn_{total} + 1)(syn_{used} + 1)$$
The Writeprints Algorithm

• Comparison process:
  • Construct pattern for B using B’s text with A’s feature set.
  • Construct pattern for A using A’s text with B’s feature set.
  • Similarity between A and B is sum of the average n-dimensional euclidean distance between these two comparisons.
Study Results
Training Set Accuracy

Baseline Accuracy

Accuracy

No. of Authors

NN
SVM
Synonym
Writeprints
Random
Obfuscation Attack Detection

Obfuscation Accuracy

Accuracy

No. of Authors

- NN
- SVM
- Synonym
- Writeprints
- Random
Imitation Attack Detection

Imitation Accuracy

Accuracy

No. of Authors

- NN
- SVM
- Synonym
- Writeprints
- Random
Imitation Attack Success

Imitation Success

Accuracy

No. of Authors

NN
SVM
Synonym
Writeprints
Random
Translation Attack

Google Translate (13 Authors)

Accuracy vs. No. of Authors

- Baseline (EN)
- EN-DE-EN
- EN-JA-EN
- EN-DE-JA-EN

Accuracy percentages for different methods:
- NN
- SVM
- Synonym
- Random
Ineffective Translations

• (Original)
  • Just remember that the things you put into your head are there forever, he said.

• (English ⇒ German ⇒ English)
  • Remember that the things that you are dead set on always there, he said.

• (English ⇒ Japanese ⇒ English)
  • But things are there forever remember what you put in your head, he said.

• (English ⇒ German ⇒ Japanese ⇒ English)
  • You are dead, that there always is set, please do not forget what he said.
Interpreting the Results

• Stylometry is very susceptible to manual adversarial attacks.
  • Automated translation attacks are not useful.
• Some methods more susceptible to attacks.
  • Writeprints showed resistance to imitation attacks.
• Participant Skill Level
  • The lack of linguistic expertise of the participants strengthens our conclusion.
• Imitation attacks were more effective.
  • Guidance from a target author’s style?
• Participants showed distinct patterns in writing style changes.
Changes in Features in Obfuscation Attack

Existential there
Cardinal number
Conjunction
Particle
Adjective
Preposition
Adverb
Personal Pronoun
ShortWords
Avg Word Length
Avg. Sentence Length
Sentence Count
Avg. Syllables
Readability (GF index)
Complexity
Unique Words

Total number of changes per feature across all obfuscation documents.
Changes in Features in Imitation Attack

- Existential there
- Cardinal number
- Conjunction
- Particle
- Adjective
- Preposition
- Adverb
- Personal Pronoun
- ShortWords
- Avg Word Length
- Avg. Sentence Length
- Sentence Count
- Avg. Syllables
- Readability (GF index)
- Complexity
- Unique Words

Total number of changes per feature across all imitation documents.
Detecting Attacks

• Can attacks be detected? Kacmarcik & Gamon (2006) say yes.
• First, modified the feature vectors of Madison Federalist Papers to look like Hamilton’s writing.
• The “unmasking” approach.
  • Use a series of SVM classifiers, on each iteration remove the most salient feature.
  • Hypothesis: accuracy in distinguishing between two authors will slowly decline if they are truly different, but quickly drop off if one was an obfuscation of the other.
• They found that it was possible to detect their attacks. Though, deeper attacks (more modifications) did make it harder.
Detecting Attacks

Distinguishing Between 2 Authors

Accuracy

SVM Iteration (one per feature removal)
Detecting Deception

M

O-M
Detecting Deception

M

O-M
Detecting Deception

M

O-M
Detecting Deception
Detecting Deception
Detecting Adversaries

• We have not yet duplicated these results. Why?
  • Primarily due to an inadequate corpus – Kacmarcik & Gamon did a automated obfuscation (and only of feature frequencies)
  • We need (even more) text!
• We were able to distinguish between regular passages, obfuscation attacks, and imitation attacks with high accuracy.
  • But this is not satisfying – it may just be selecting features based on the domain (e.g. first person pronouns).
Review

• Stylometry is a deep field that has provided a great deal of opportunity for important research in linguistics and computer science.
  • No end in sight.
• Privacy and anonymity concerns are real and must continue to be looked at.
  • Download our corpus!
• Early studies in adversarial stylometry demonstrate the relative ease of fooling current methods.
• Deception detection is promising but not yet realized.
What comes next?

• 25 authors (and counting) in our new corpus built by Amazon Mechanical Turkers.
  • Finding volunteers is hard!
• Two research groups working on analysis and circumvention tools for stylometry (Stylo and Anonamouth)
• Deception detection: reliable methods?
• Potential for arms race:
  • Developing attack-resistant methods of stylometry vs. creating new attacks to preserve privacy.
Postscript: Geeks in Washington
A Technologist at the FTC

• My job: Technologist, Federal Trade Commission’s Division of Privacy and Identity Protection (DPIP)
• DPIP operates under Section 5 of the FTC Act:
  • “Unfair methods of competition in or affecting commerce, and unfair or deceptive acts or practices in or affecting commerce, are hereby declared unlawful.”
• Investigations are non-public until closed or settled.
A Technologist’s Role in Policy

• I am a technical adviser to the legal staff of DPIP.
  • Interact with parties under investigation to foster a technical understanding of the issues.
  • Research potential violations and suggest investigations.
  • Educate staff on current trends and technologies that may be of interest.
  • Help construct legal documents that are technically sound.
Recent FTC Work

- Privacy Report:
  - Industry self-regulation efforts “have failed to provide adequate and meaningful protection.”
  - Regulation and legislation are necessary.
  - Do Not Track
  - Reasonable access to data.
- Google Buzz Settlement
- Twitter Settlement
- Chitika Settlement
Thanks.

- PSAL: psal.cs.drexel.edu
- Research Team.
  - Sadia Afroz: www.cs.drexel.edu/~sa499
  - Rachel Greenstadt: www.cs.drexel.edu/~greenie
- Contact Me.
  - mb553@drexel.edu
  - @brennan_mike
  - mbrennan.net
Privacy, Security, Automation.

• Other work at our lab: