Detecting Deception in Writing Style

Sadia Afroz, Michael Brennan and Rachel Greenstadt.
Privacy, Security and Automation Lab
Drexel University
Overview

• Authorship recognition
• Authorship recognition in adversarial environment
• Deception detection
• Experiments on different datasets
Authorship recognition

Who wrote the document?
Authorship recognition

**Stylometry:**

- An authorship recognition system based solely on writing style.
- Not handwriting
- Only linguistic style: word choice, sentence length, parts-of-speech usage, ...
Why it works?

• Everybody has learned language differently
How regular authorship recognition works

Extract features

Machine Learning System
Document of unknown authorship

Extract features

Machine Learning System

Determine authorship

Drexel University
Assumptions

• Writing style is invariant.

  – It’s like a fingerprint, you can’t really change it.

  – Authorship recognition can identify you if there are sufficient writing samples and a set of suspects.
Adversarial Attacks

• Imitation or framing attack
  – Where one author imitates another author
  – Participants were asked to imitate Cormac McCarthy in writing about their day

• Obfuscation attack
  – Where an author hides his regular style
  – Participants were asked to write about their neighborhood in a different style

Accuracy in detecting authorship of regular documents

More than 80% accurate in detecting authorship with 40 authors in regular document

Number of Authors

9-Feature (NN)
Synonym-Based
Writeprints Baseline (SVM)
Random
Accuracy in detecting authorship of Obfuscated documents

- 9-Feature (NN)
- Synonym-Based
- Writeprints Baseline (SVM)

Accuracy is less than random chance in obfuscated documents
Accuracy in detecting authorship of Imitated documents

Accuracy is less than random chance in imitated documents
Can we detect Stylistic Deception?
Extended-Brennan-Greenstadt Corpus

• 56 authors
  – 12 of the participants are from Drexel University
  – The rest are paid workers from Amazon mechanical turk

• Three kinds of writing samples
  – Regular writings (5000 words)
  – Imitated writing
    • A 500-word article describing a day
    • Imitate Cormac McCarthy from ‘The Road’
  – Obfuscated writing
    • A 500-word article describing neighborhood
    • Hide own writing style
Detecting stylistic deception is possible

<table>
<thead>
<tr>
<th></th>
<th>Writeprint, SVM</th>
<th>Lying-detection, J48</th>
<th>9-feature set, J48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular</td>
<td>98</td>
<td>95.7</td>
<td></td>
</tr>
<tr>
<td>Imitation</td>
<td>85</td>
<td>75.3</td>
<td>48</td>
</tr>
<tr>
<td>Obfuscation</td>
<td>89.5</td>
<td>59.9</td>
<td>43</td>
</tr>
</tbody>
</table>
Feature Changes in Imitation and Obfuscation

- Personal pronoun
- Sentence count
- Particle
- Short Words
- Verb
- Unique words
- Adverb
- Existential there
- Average syllables per word
- Average word length
- Adjective
- Cardinal number
- Gunning-Fog readability index
- Average sentence length

Graph showing comparisons between Imitation and Obfuscation with specific features.
Problem with the dataset: Topic Similarity

• All the deceptive documents were of same topic.

• Non-content-specific features have same effect as content-specific features.
Hemingway-Faulkner Imitation Corpus

• Articles from the International Imitation Hemingway Contest (2000-2005)
• Articles from the Faux Faulkner Contest (2001-2005)
• Original excerpts of Ernest Hemingway and William Faulkner
Deception detection is possible even when the topic is not similar

• 81.2% accurate in detecting imitated documents.
Long term deception: A Gay Girl In Damascus

– Original author was a 40-year old American citizen, Thomas MacMaster.
– Pretended to be a Syrian gay woman, Amina Arraf.
– The author worked for at least 5 years to create a new style.
Long term deception is hard to detect

• None of the blog posts were found to be deceptive.
• But regular authorship recognition can help.
• We tried to attribute authorship of the blog posts using Thomas (as himself), Thomas (as Amina), Britta (Thomas’s wife).
Long term deception
Authorship recognition of the blog posts

Thomas MacMaster.  54%
Amina Arraf  43%
Britta (Thomas’s wife)  3%
Future works

• Intrusion detection
• Social spam detection
• Identifying quality discourse
Two Tools

• JStylo: Authorship Recognition Analysis Tool.
• Anonymouth: Authorship Recognition Evasion Tool.

• Free, Open Source. (GNU GPL)
• Alpha releases available today at https://psal.cs.drexel.edu
  – Migrating to GitHub soon.
Privacy, Security and Automation Lab
(https://psal.cs.drexel.edu)

• Faculty
  – Dr. Rachel Greenstadt

• Graduate Students
  – Sadia Afroz (Deception Detection Lead)
  – Diamond Bishop
  – Michael Brennan
  – Aylin Caliskan
  – Ariel Stolerman (JStylo Lead Developer)

• Undergraduate Students
  – Pavan Kantharaju
  – Andrew McDonald (Anonymouth Lead Developer)