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:

– Study of writing style
– Not handwriting
– Only linguistic style: word choice, sentence length, parts-of-speech usage, ...
Why it works?

Everybody has learned language differently

Toward vs Towards
Though vs Although

Number of characters
Number of syllables
New bigram al, It
New trigram alt, lth
How regular authorship recognition works

Extract features

Machine Learning System
Document of unknown authorship

Extract features

Machine Learning System

Determine authorship
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

• Obfuscation attack
  – Where an author hides his regular 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

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

Accuracy is less than random chance in imitated documents
Can we detect Stylistic Deception?
Difference from Authorship Recognition

Authorship Recognition models individual authors
Why it is important?

- Security
  - Detecting fraud, imitation, masking attack
- Privacy
  - Writing style anonymization
Analytic Approach

1. Data collection
2. Feature extraction
3. Classification
4. Feature Rank
Data collection

• Extended-Brennan-Greenstadt Corpus
  – Available at https://psal.cs.drexel.edu

• Hemingway-Faulkner Imitation corpus
  – Available here

• Long Term Deception: Blog posts from “A Gay Girl in Damascus”
Feature sets

• Writeprints:
  – Around 700 features
  – Lexical, syntactic, content specific

• Lying-detection features:
  – 16 features
  – Uncertainty, Vocabulary Complexity, Grammatical Complexity, Specificity and Expressiveness, Verbal Non-immediacy.

• Basic-9 features:
  – Readability index, character count, average syllables per word, sentence count, average sentence length.
Classification

• Classifier:
  – Support Vector Machine (SVM)
  – J48 Decision Tree

• 10-fold cross-validation
  – 90% of data used for training
  – 10% of data used for testing
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

- Regular: Writeprint, SVM: 98%, Imitation: 85%, Obfuscation: 89.5%
- Imitation: Lying-detection, J48: 95.7%, 9-feature set, J48: 48%
- Obfuscation: 9-feature set, J48: 43%
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
- Average sentence length

Legend:
- Imitation
- Obfuscation
Feature Changes in Imitation and Obfuscation

- Word unigram
- Function words
- Character trigram
- Parts-of-speech
- Punctuation
- Short Words
- Total Words
- Character unigram
- Percent digits
- Average word length
- Percent uppercase
- Character bigram
- Word bigram
- Word trigram

Red bars represent Change in imitation, and blue bars represent Change in obfuscation.
Problem with the dataset: Topic Similarity

• All the adversarial documents were of same topic.

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

• Winning entries from the International Imitation Hemingway Contest (2000-2005)
• Winning entries from the Faux Faulkner Contest (2001-2005)
• Original excerpts of Ernest Hemingway and William Faulkner
• 36 imitation samples.
• Each sample contains 500 words.
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

• 14% 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 writing samples of Thomas (as himself), Thomas (as Amina), Britta (Thomas’s wife).
Long term deception:
Authorship recognition of the blog posts

Thomas MacMaster (as himself): 54%
Thomas MacMaster (as Amina Arraf): 43%
Britta (Thomas’s wife): 3%
Conclusion

• We showed stylistic deception detection is possible
• We showed people can change writing style to circumvent authorship attribution
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)
Thank you!

- "Detecting Hoaxes, Frauds, and Deception in Writing Style Online." Afroz/Brennan/Greenstadt. IEEE Symposium on Security and Privacy '12
- "Practical Attacks Against Authorship Recognition Techniques." Brennan/Greenstadt. IAAI '09.