Detecting Hoaxes, Frauds and Deception in Writing Style Online

Sadia Afroz, Michael Brennan and Rachel Greenstadt
Privacy, Security and Automation Lab
Drexel University
What do we mean by “deception?”

Let me give an example...
A Gay Girl In Damascus

A blog by
Amina Arraf

Facts about Amina:
A Syrian-American activist
Lives in Damascus
“HALFWAY OUT OF THE DARK”: ON BEING A GAY GIRL IN DAMASCUS

Almost every time I speak or write to other LGBT people outside the Middle East, they always seem to wonder what it’s like to be a lesbian here in Damascus. Well, I always find myself answering, it’s not as easy as I’d like it to be but it’s probably easier than you might think. And that, of course, opens up a whole endless stream of questions. To answer fully, I suppose, I almost have to give a little autobiographical detail.

I’m a dual-national and I grew up between Damascus, Syria and the American South, neither of which was exactly the easiest place to be struggling with what I considered inappropriate desires. When I was fifteen, I realized I was gay and the thought absolutely terrified me. I was suicidal and self-destructive until, I thought, I found a way out of sinful desires; I became what might be described by some people as an ‘Islamic extremist’, by others...
ANOTHER DAY IN DAMASCUS

5 June 2011

Well, we had a scare here but it looks like we're back; the internet was down for virtually the whole country for a day and came back on yesterday. Before posting again, I needed to be sure of safety (as well as giving highest priority to those who had greater need than me of internet use!) and here I am ...

In my ever humble opinion, the regime shut down the internet out of desperation; they are beginning to really feel how far they’ve fallen. I’m not the only one who thinks that they will not be able to get back up from this. However, the days and weeks and months ahead are not going to be simple ones. We know that they will be pushing back as much as they can and, among them, there are elements who’d rather pull the whol edifice of our society down than hand over power to anyone else.
A Gay Girl in Damascus becomes a heroine of the Syrian revolt
Blog by half-American 'ultimate outsider' describes dangers of political and sexual dissent

Katherine Marsh in Damascus

guardian.co.uk, Friday 6 May 2011 11.24 BST
Article history
The blog's author, Amina Abdallah, is a 35-year-old English teacher who says she returned to Syria last year after many years in the United States. In an email interview Abdallah said she believed that political change could improve gay rights.

She said: "A whole lot of long time changes are coming suddenly bubbling to the surface and views towards women, gay people and minorities are rapidly changing."

Abdallah said the reaction to her blog had been "almost entirely positive."

"What has really startled me has been the fact that I have received no criticism from Islamic sources she said. "Instead, they've been entirely positive."

Syrian blogger Amina Abdallah kidnapped by armed men

Author of A Gay Girl in Damascus had shot to prominence for her frank views on Syrian uprising, politics and being a lesbian

Nidaa Hassan in Damascus
The Guardian, Monday 6 June 2011
Article history

A blogger whose frank and witty thoughts on Syria’s uprising, politics and being a lesbian in the country shot her to prominence was last night seized by armed men in Damascus.

Amina Arraf, who blogged under the name Amina Abdallah, holds dual Syrian and American citizenship and is the author of the blog A Gay Girl in Damascus, which has drawn fans from Syria and across the world.

She was kidnapped last night as she and a friend were on their way to a meeting in Damascus. The kidnapping was reported on her blog by a cousin.

"Amina was seized by three men in their early 20s. According to the witness (who does not want her identity known), the men were armed," wrote Rania Ismail.
Working on my social media skills to help #FreeAmina and spread the word out...Keep the momentum. #Syria

less than a minute ago via web  Favorite  Retweet  Reply

Sade B.
sade_la_bag
A Gay Girl In Damascus
A Gay Girl In Damascus

Fake picture
copied from Facebook

Sunday, October 21, 12
A Gay Girl In Damascus

Fake picture
(copied from Facebook)

The real “Amina”

Thomas MacMaster
A 40-year old American male
Why we are interested?
Why we are interested?

Thomas developed a new writing style for Amina
Why we are interested?

Thomas developed a new writing style for Amina

One member of alternate-history Yahoo! group wrote:

“If you read through her blog entries, its pretty clear its our Amina. Same background, same interests, same style of writing. I can confirm she's the same.”
• Deception in Writing Style:
  – Someone is hiding his regular writing style

• Research question:
  – If someone is hiding his regular style, can we detect it?
Why do we care?

• Security:
  – To detect fake internet identities, astroturfing, and hoaxes

• Privacy and anonymity:
  – To understand how to anonymize writing style
Overview

- How to detect authorship of a document?
- Can we circumvent authorship recognition?
- Can we detect if someone is trying to circumvent authorship recognition?
- How to anonymize writing style?
Overview

• How to detect authorship of a document?
  • Can we circumvent authorship recognition?
  • Can we detect if someone is trying to circumvent authorship recognition?
  • How to anonymize writing style?
Authorship recognition

Who wrote the document?
Can be determined using writing style
Does everybody have unique writing style?

• Most people do!
• Because everybody learns language differently
WHAT IS THIS OBJECT?

Thanks to Patrick Juola for this example
WHAT IS THIS OBJECT?

Is this a couch?

Thanks to Patrick Juola for this example
WHAT IS THIS OBJECT?

Is this a couch?
... a sofa?

Thanks to Patrick Juola for this example
WHAT IS THIS OBJECT?

Is this a couch?
... a sofa?
... a davenport?

Thanks to Patrick Juola for this example
WHAT IS THIS OBJECT?

Is this a couch?
... a sofa?
... a davenport?
... a chesterfield?

Thanks to Patrick Juola for this example
WHAT IS THIS OBJECT?

Is this a couch?
... a sofa?
... a davenport?
... a chesterfield?
... a divan?

Thanks to Patrick Juola for this example
WHAT IS THIS OBJECT?

Is this a couch?
... a sofa?
... a davenport?
... a chesterfield?
... a divan?
... a settee?

Thanks to Patrick Juola for this example
WHAT IS THIS OBJECT?

Is this a couch?
... a sofa?
... a davenport?
... a chesterfield?
... a divan?
... a settee?

Regional differences

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

- “next to” the plate?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

• “next to” the plate?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

- “next to” the plate?
- “to the left of”?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

• “next to” the plate?

• “to the left of”?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

- “next to” the plate?
- “to the left of”?
- “on the left of”?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

- “next to” the plate?
- “to the left of”?
- “on the left of”?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

- “next to” the plate?
- “to the left of”?
- “on the left of”?
- “at the plate’s left”?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

- “next to” the plate?
- “to the left of”?
- “on the left of”?
- “at the plate’s left”?

Thanks to Patrick Juola for this example
WHERE IS THE DINNER FORK?

- “next to” the plate?
- “to the left of”?
- “on the left of”?
- “at the plate’s left”?
- “left of” the plate?

Thanks to Patrick Juola for this example
FUNCTION WORDS

Thanks to Patrick Juola for this example
FUNCTION WORDS

FINISHED FILES ARE NOT THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF MANY YEARS.

Thanks to Patrick Juola for this example
FUNCTION WORDS

FINISHED FILES ARE NOT THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF MANY YEARS.

• How many times does the letter ‘F’ appear in this passage?
FUNCTION WORDS

• How many times does the letter ‘F’ appear in this passage?

Thanks to Patrick Juola for this example
FUNCTION WORDS

• How many times does the letter ‘F’ appear in this passage?
• Many people (most?) only count three
FUNCTION WORDS

• How many times does the letter ‘F’ appear in this passage?
• Many people (most?) only count three
• They miss the word ‘OF.’

Thanks to Patrick Juola for this example
Authorship Recognition

- Modern authorship recognition systems are machine learning based.
- Supervised
- Unsupervised
How good are current authorship recognition algorithms?

• **100 authors** *(Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. Abbasi et al.)*

• **10,000 authors** *(content-based approach)* *(“Authorship attribution in the wild,” Koppel et al.)*

• **100,000 authors** *(“On the Feasibility of Internet-Scale Author Identification,” Narayanan et al.)*
Threat

- 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 authorship recognition to identify Alice as the blogger.
Overview

- How to detect authorship of a document?
- Can we circumvent authorship recognition?
- Can we detect if someone is trying to circumvent authorship recognition?
- How to anonymize writing style?
Assumption of Authorship recognition

• Writing style is invariant.
  – It’s like a fingerprint, you can’t really change it.
Wrong Assumption!

- Imitation or framing attack
  - Where one author imitates another author
- Obfuscation attack
  - Where an author hides his regular style

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.”
Obfuscating writing style

Your goal is to fool the computer into thinking that your passage was NOT written by you.

You may use whatever means you wish so long as the writing would not raise any eyebrows when a human reads over it (no scrambled words, mixed up semantics, etc) and the point is still clearly conveyed.
Accuracy in detecting authorship of regular documents

Number of Authors

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

![Graph showing accuracy of detecting authorship with different number of authors. The graph compares 9-Feature (NN), Synonym-Based, Writeprints Baseline (SVM), and Random methods.]
Accuracy in detecting authorship of imitated documents

Number of Authors

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

• How to detect authorship of a document?

• Can we circumvent authorship recognition?

• Can we detect if someone is trying to circumvent authorship recognition?

• How to anonymize writing style?
Can we detect stylistic deception?
Can we detect stylistic deception?

- Deceptive
- Regular
Analytic Approach
Analytic Approach

Data Collection
Analytic Approach

Data Collection → Feature Extraction
Analytic Approach

Data Collection → Feature Extraction → Classification
Analytic Approach

Data Collection → Feature Extraction → Classification → Feature Ranking
Data collection

• Short-term deception:  • Long-term deception:
Data collection

- Short-term deception:
  - Extended-Brennan-Greenstadt Corpus
    - Fixed topic
    - Controlled style

- Long-term deception:
Data collection

• Short-term deception:  • Long-term deception:
  – Extended-Brennan-Greenstadt Corpus
    • Fixed topic
    • Controlled style
  – Hemingway-Faulkner Imitation corpus
    • No fixed topic
    • Controlled style
Data collection

• Short-term deception:
  – Extended-Brennan-Greenstadt Corpus
    • Fixed topic
    • Controlled style
  – Hemingway-Faulkner Imitation corpus
    • No fixed topic
    • Controlled style

• Long-term deception:
  - Thomas-Amina Hoax corpus
    • No fixed topic
    • No control in style
Extended-Brennan-Greenstadt Corpus

• Writing samples
  – Regular (5000-word)
  – Imitation (500-word)
    – Imitate Cormac McCarthy
    – Topic: A day
  – Obfuscation (500-word)
    – Write in a way they don’t usually write
    – Topic: Neighborhood

• Participants
  – 12 Drexel students
  – 56 AMT authors
Extended-Brennan-Greenstadt Corpus

- Classification task:
  - Distinguish Regular, Imitation and Obfuscation
Classification

• We used WEKA for machine learning.

• Classifier:
  – Experimented with several classifiers
  – Choose the best classifier for a feature set

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

• We experimented with 3 feature sets:
  – Writeprints
  – Lying-detection features
  – 9-features
Feature sets

• We experimented with 3 feature sets:
  – Writeprints
    • 700+ features, SVM
    • Includes features like frequencies of word/character n-grams, parts-of-speech n-grams.
  – Lying-detection features
  – 9-features
We experimented with 3 feature sets:

- **Writeprints**
  - 700+ features, SVM
- **Lying-detection features**
  - 20 features, J48 decision tree
  - Previously used for detecting lying.
  - Includes features like rate of Adjectives and Adverbs, sentence complexity, frequency of self-reference.
- **9-features**
Feature sets

• We experimented with 3 feature sets:
  – Writeprints
    • 700+ features, SVM
  – Lying-detection features
    • 20 features, J48 decision tree
  – 9-features
    • 9 features, J48 decision tree
    • Used for authorship recognition
    • Includes features like readability index, number of characters, average syllables.
Detecting stylistic deception is possible

- **Writeprints**: 98, 85, 89.5
- **Lying-detection**: 95.7, 75.3, 59.9
- **9-features**: 94.5, 48, 43

---

Sunday, October 21, 12

Drexel UNIVERSITY
Detecting stylistic deception is possible

Writeprints features can distinguish Regular, Imitation and Obfuscation documents with over 80% accuracy.
Detecting stylistic deception is possible

Writeprints features can distinguish Regular, Imitation and Obfuscation documents with over 80% accuracy.

Writing style change has similarity with lying as lying-detection features can detect imitation and obfuscation.
Which features are changed

Change in feature = \frac{Feature_{deceptive} - Feature_{regular}}{Feature_{regular}}
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

Red: Imitation
Green: Obfuscation
How the classifier uses changed and unchanged features

• We measured
  – How important a feature is to the classifier (using information gain ratio)
  – How much it is changed by the deceptive users
Information Gain vs Feature Change

- Diagram showing the relationship between information gain ratio and change in n-grams.
- Diagram showing the relationship between information gain ratio and change in function words.
How the classifier uses changed and unchanged features

• We measured
  – How important a feature is to the classifier (using information gain ratio)
  – How much it is changed by the deceptive users

• We found
  – For words, characters and parts-of-speech n-grams
    information gain increased as features were changed more.
  – The opposite is true for function words (of, for, the)

• Deception detection works because deceptive users changed n-grams but not function words.
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

International Imitation Hemingway Competition

Faux Faulkner Contest
Hemingway-Faulkner Imitation Corpus

- Writing samples
  - Regular
    - Excerpts of Hemingway
    - Excerpts of Faulkner
  - Imitation
    - Imitation of Hemingway
    - Imitation of Faulkner

- Participants
  - 33 contest winners
Hemingway-Faulkner Imitation Corpus

- Classification task:
  - Distinguish Regular and Imitation
# Imitation success

<table>
<thead>
<tr>
<th>Author to imitate</th>
<th>Imitation success</th>
<th>Writer’s Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cormac McCarthy</td>
<td>47.05%</td>
<td>Not professional</td>
</tr>
<tr>
<td>Ernest Hemingway</td>
<td>84.21%</td>
<td>Professional</td>
</tr>
<tr>
<td>William Faulkner</td>
<td>66.67%</td>
<td>Professional</td>
</tr>
</tbody>
</table>
Result: Hemingway-Faulkner
Imitation corpus

<table>
<thead>
<tr>
<th></th>
<th>Writeprints</th>
<th>Lying-detection</th>
<th>9-features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular</td>
<td>86.2</td>
<td>65.1</td>
<td>61.5</td>
</tr>
<tr>
<td>Imitation</td>
<td>88.6</td>
<td>69.7</td>
<td>78.9</td>
</tr>
</tbody>
</table>
Result: Hemingway-Faulkner Imitation corpus

Writeprints features can distinguish Regular and Imitation documents with over 80% accuracy.
Long term deception

• Writing samples
  – Regular
    • Thomas’s writing sample at alternate-history Yahoo! group
  – Deceptive
    • Amina’s writing sample at alternate-history Yahoo! group
    • Blog posts from “A Gay Girl in Damascus”

• Participant
  – 1 (Thomas)
Long term deception

• Classification:
  • Train on short-term deception corpus
  • Test blog posts to find deception

• Result:
  • 14% of the blog posts were deceptive (less than random chance).
Long term deception: Authorship Recognition

• We performed authorship recognition of the Yahoo! group posts.

• None of the Yahoo! group posts written as Amina were attributed to Thomas.
Long term deception: Authorship Recognition

• We tested authorship recognition on the blog posts.
• Training:
  – writing samples of Thomas (as himself),
  – writing samples of Thomas (as Amina),
  – writing samples of Britta (Another suspect of this hoax).
Long term deception: Authorship Recognition

Thomas MacMaster (as himself): 54%
Thomas MacMaster (as Amina Arraf): 43%
Britta: 3%
Long term deception: Authorship Recognition

Thomas MacMaster (as himself): 54%
Thomas MacMaster (as Amina Arraf): 43%
Britta: 3%

Maintaining separate writing styles is hard!
Overview

• How to detect authorship of a document?

• Can we circumvent authorship recognition?

• Can we detect if someone is trying to circumvent authorship recognition?

• How to anonymize writing style?
Why not machine translation?

They passed through the city at noon of the day following.

(German)

(Japanese)
Why not machine translation?

They passed through the city at noon of the day following.

(German)

(Japanese)

They passed the city at noon the following day.
Why not machine translation?

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

(German)

(Japanese)
Why not machine translation?

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

(German)

(Japanese)

You are dead, that there always is set, please do not forget what he said.
Why not machine translation?

Machine translation does not anonymize writing style because:

- A good translator does not change the style that much
- A bad translator completely changes the meaning
Hackers Meeting in Berlin 28C3

Whoever imitates Hemingway writes, anonymous

Philip Elsbrock

Logo of the 28C3: The Chaos Communication Congress in Berlin ends this Friday
How about imitation?

- Task: Change a pre-existing document by imitating Cormac McCarthy
I can't pinpoint the exact moment I started to break.

After Imitation

The girl sitting in the pristine and serene and sterile psychiatrist office couldn't pinpoint the moment she started breaking.
Authorship attribution with writeprints features

![Bar chart showing accuracy for authors 1 to 10, comparing original and imitated works.](chart.png)
How to anonymize writing style?

JStylo

Authorship Recognition Tool
(Lead developer: Ariel Stolerman)

Anonymouth

Authorship Recognition Circumvention Tool
(Lead developer: Andrew McDonald)

Alpha release available: https://psal.cs.drexel.edu
How Anonymouth works

1. Background corpus

2. Document
How Anonymouth works

1. Background corpus
2. Document

Is the document anonymized? (Jstylo)
How Anonymouth works

1. Background corpus
2. Document
3. Is the document anonymized? (Jstypo)
   - YES!
   - Document
   - No
   - Which feature to change (Rank features)
How Anonymouth works

1. Background corpus
2. Document
3. Is the document anonymized? (Jstylo)
   - YES!
   - Document
   - No
     - Which feature to change (Rank features)
     - How to change feature (Feature clustering)
How Anonymouth works

1. Background corpus
2. Document
3. Is the document anonymized? (Jstylo)
   - Yes!
   - No
4. Which feature to change (Rank features)
5. Shows suggestion

How to change feature (Feature clustering)
How Anonymouth works

1. Background corpus
2. Is the document anonymized? (Jstylo)
3. No
4. How to change feature (Feature clustering)
5. Shows suggestion
6. Document

YES!

Document
Anonymouth user study

- 10 participants
  - 6500+ pre-existing documents
  - 500-word document to modify
- Background corpus: 6 authors documents
- Classifier: 9-features and SVM
Limitations

• On an extensive feature set, Anonymouth gives suggestions like:
  – Use fewer instances of the letter “l”
  Hard for users to follow
Summary

• How to detect authorship of a document?
  – Using writing style

• Can we circumvent authorship recognition?
  – Yes! By imitating or obfuscating.

• Can we detect if someone is trying to circumvent authorship recognition?

• How to anonymize writing style?
  – Anonymouth (https://psal.cs.drexel.edu)
Thank you!

- Sadia Afroz: sadia.afroz@drexel.edu
- Michael Brennan: mb553@drexel.edu
- Ariel Stolerman: ams573@drexel.edu
- Andrew McDonald: awm32@drexel.edu
- Aylin Caliskan: ac993@drexel.edu
- Rachel Greenstadt: greenie@cs.drexel.edu

- Privacy, Security And Automation Lab (https://psal.cs.drexel.edu)