Detecting Adversarial Attacks using linguistic features

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

Adversarial Attack Detection and Security Analytics

- Help to make machine learning based system robust.

Stylometry

- We consider adversarial attack on **Stylometry**: an authorship recognition system based solely on writing style.

- Supervised Stylometry:
  - Given a set of documents of known authorship, classify a document of unknown authorship.

- Unsupervised Stylometry:
  - Given a set of documents of unknown authorship, cluster them into author groups.

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.

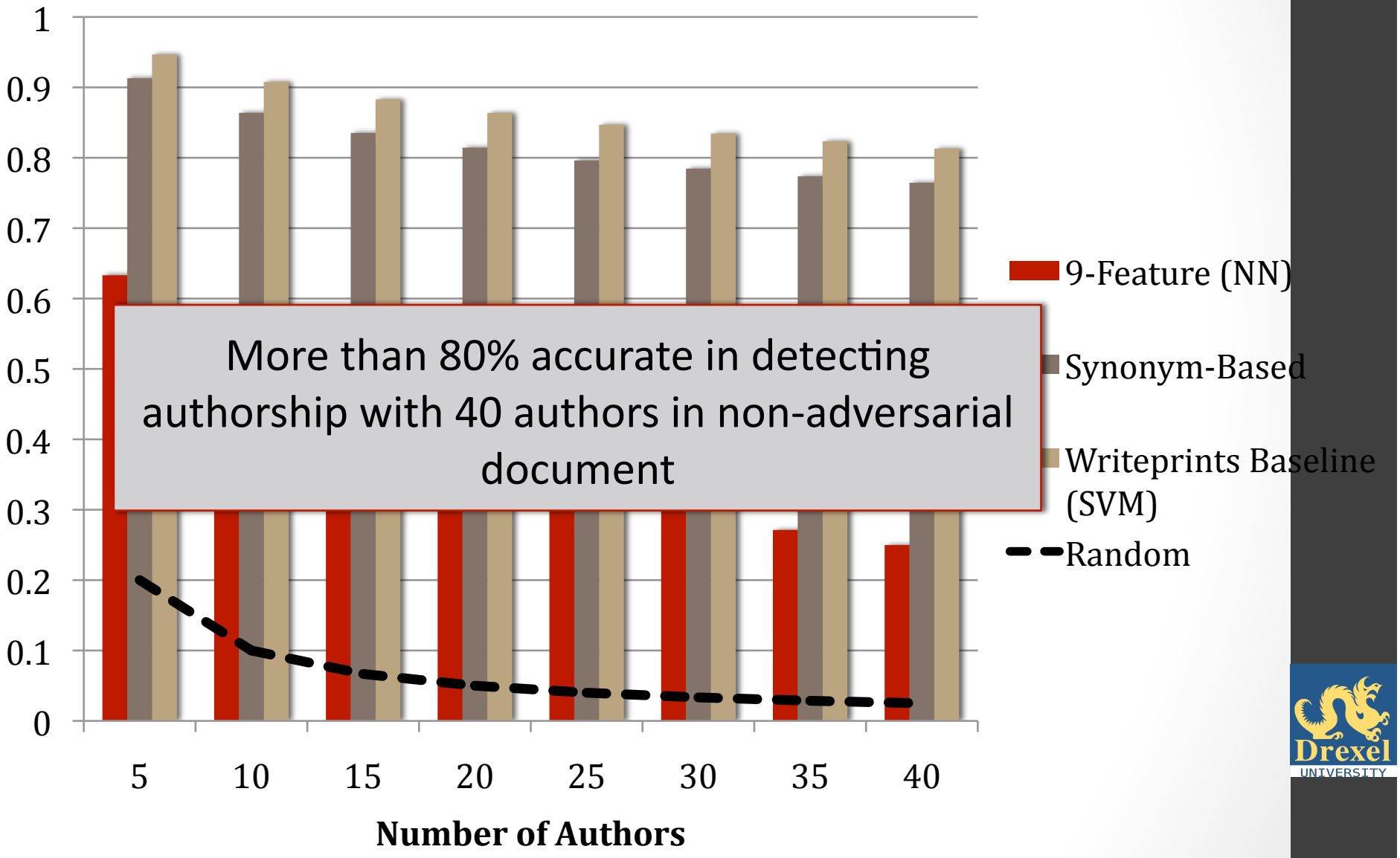
Adversarial Attacks

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

M. Brennan and R. Greenstadt. Practical attacks against authorship recognition techniques. In Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence (IAAI), Pasadena, CA, 2009.

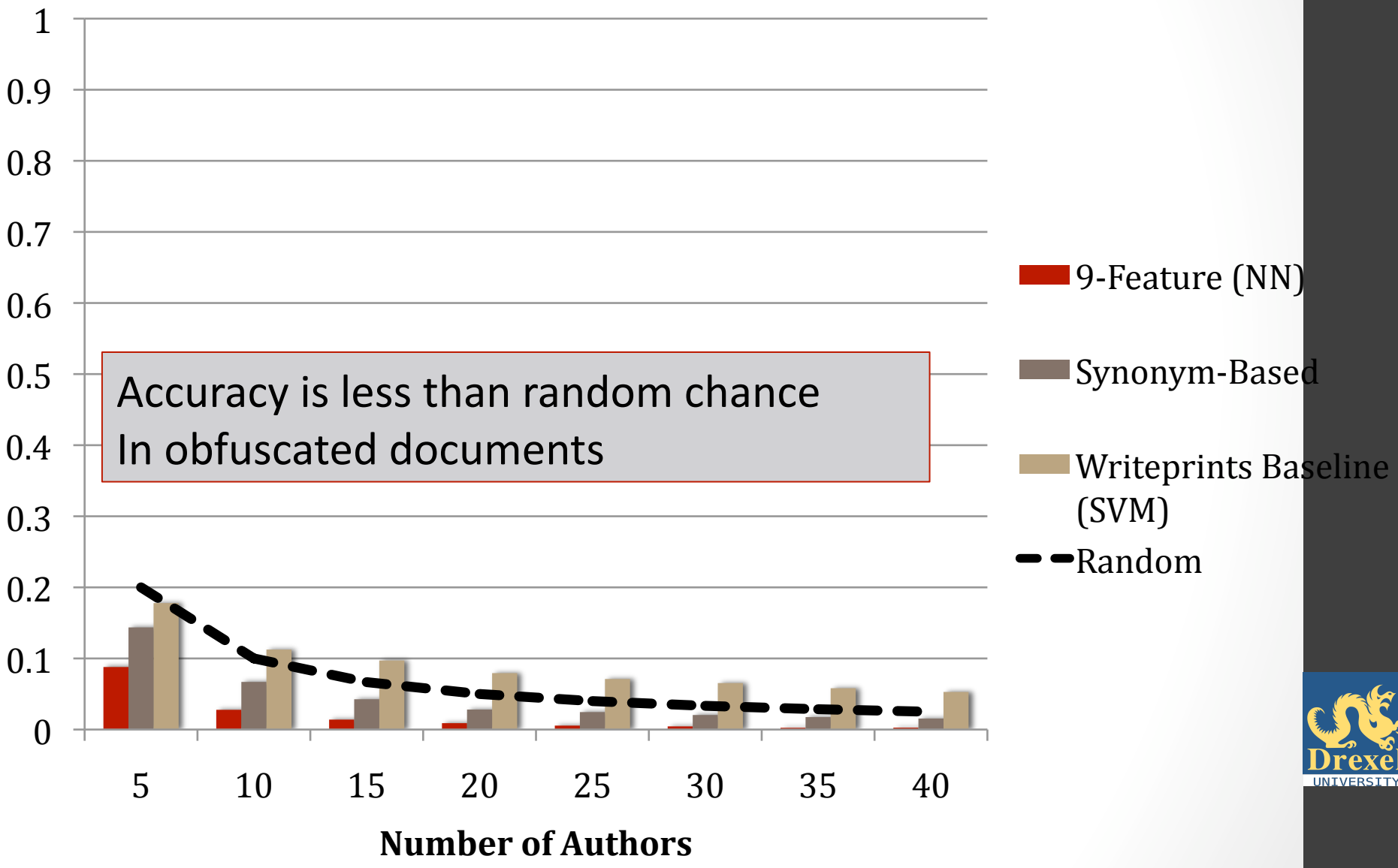
Drexel UNIVERSITY

Accuracy in detecting authorship of regular documents



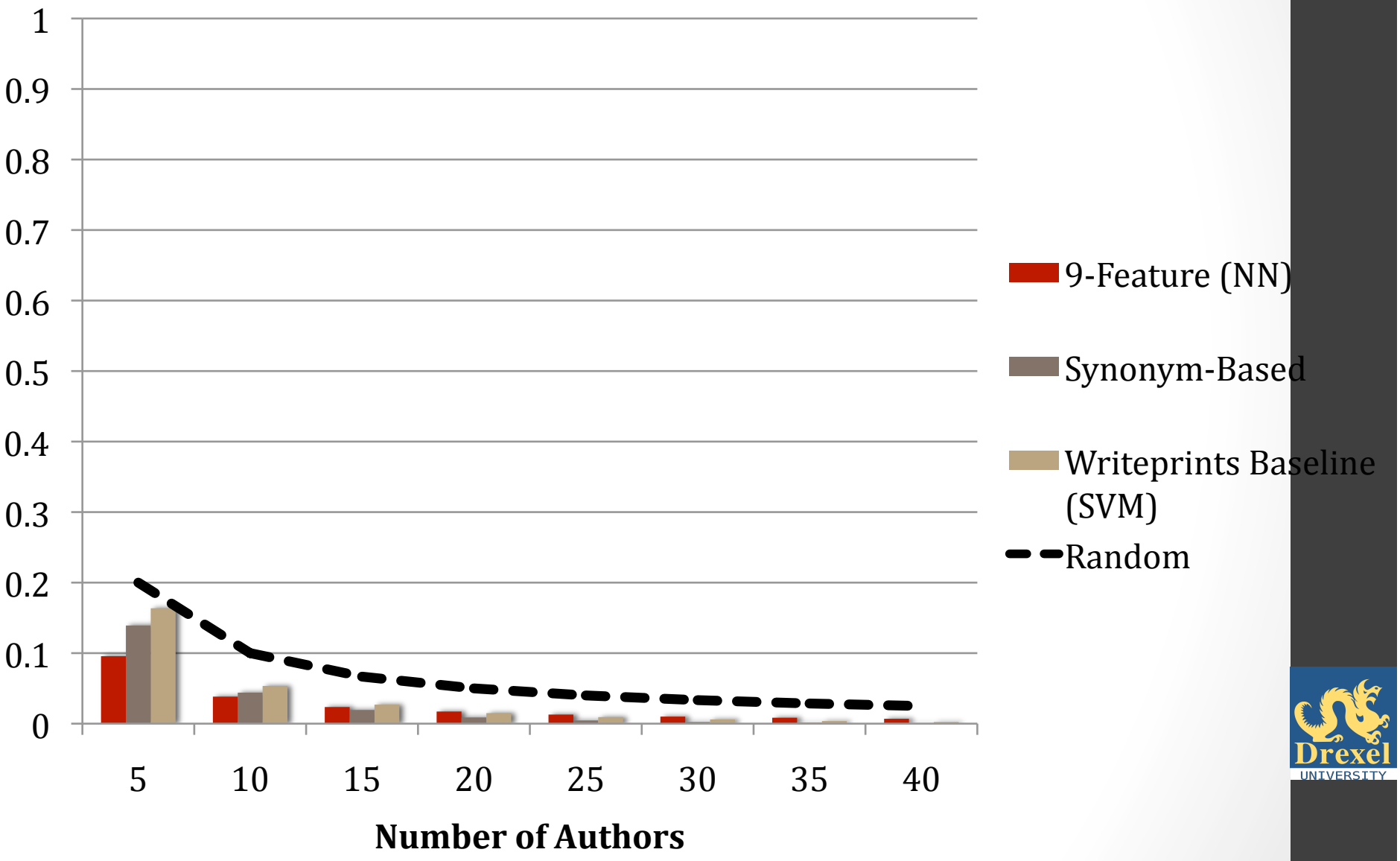
Drexel UNIVERSITY

Accuracy in detecting authorship of Obfuscated documents



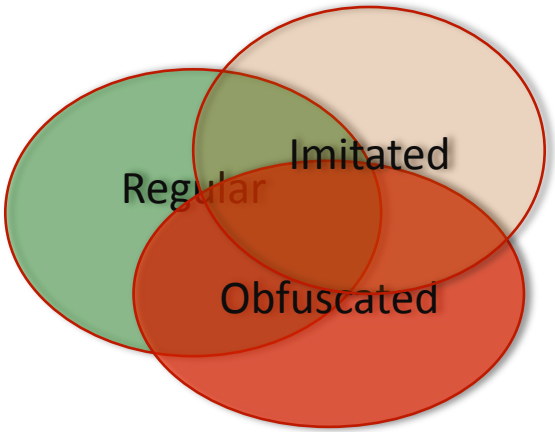
Drexel UNIVERSITY

Accuracy in detecting authorship of Imitated documents



Drexel UNIVERSITY

Can we detect Adversarial Attacks?



Drexel UNIVERSITY

Datasets

- Extended-Brennan-Greenstadt Corpus
- Hemingway-Faulkner Imitation corpus
- Long Term Deception: Blog posts from ‘A Gay Girl in Damascus’

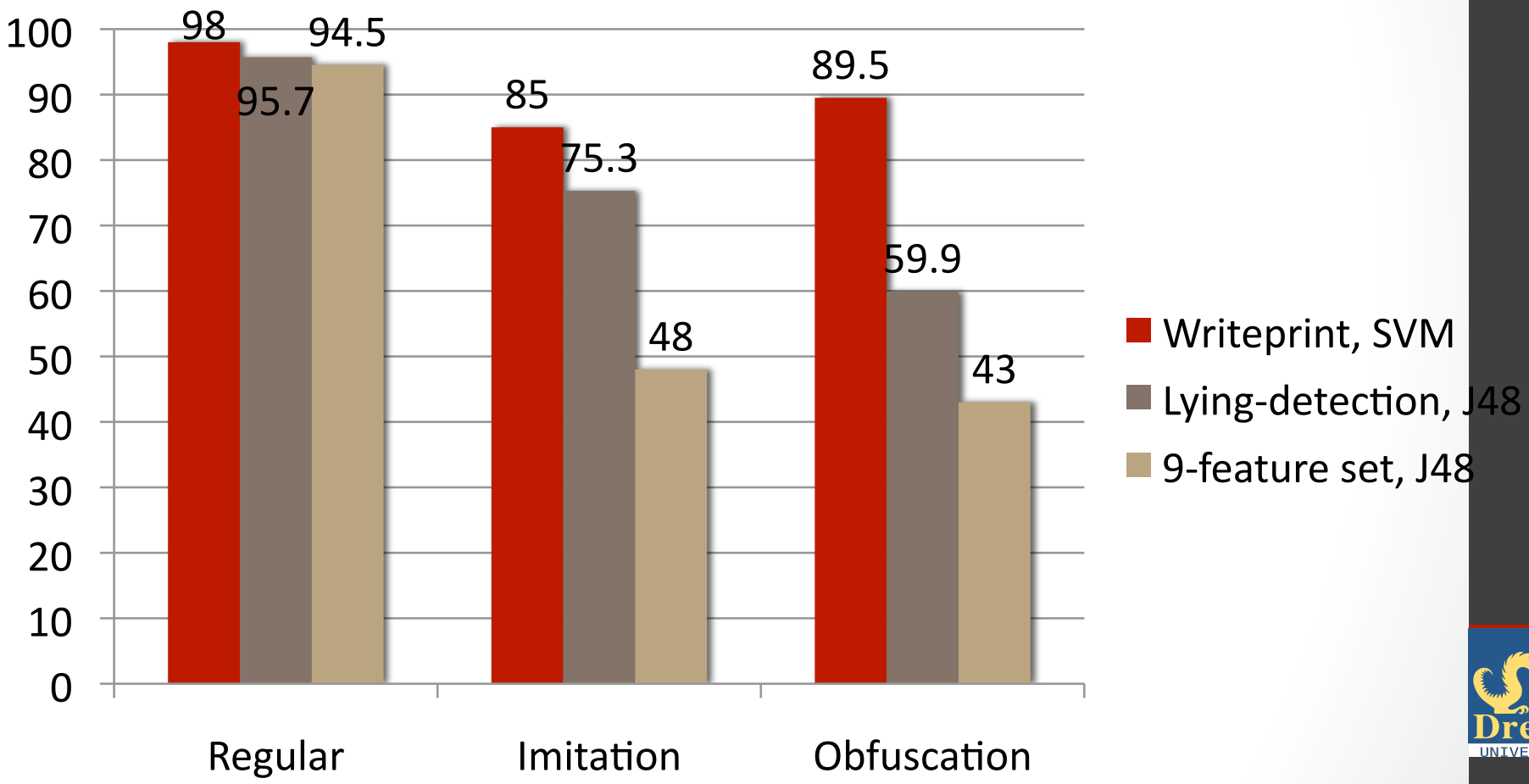
Drexel UNIVERSITY

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

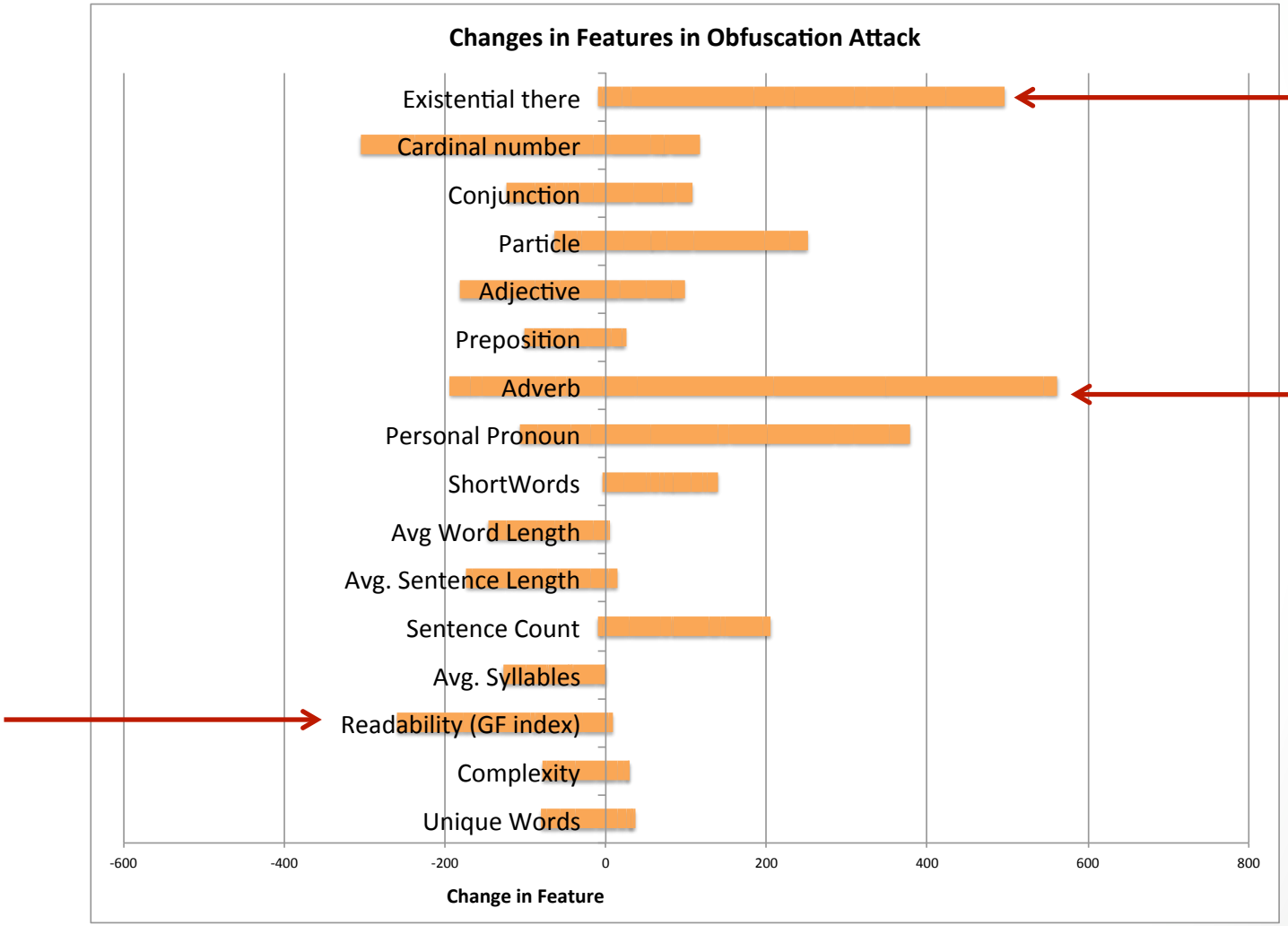
Drexel UNIVERSITY

Detecting Adversarial attack is possible



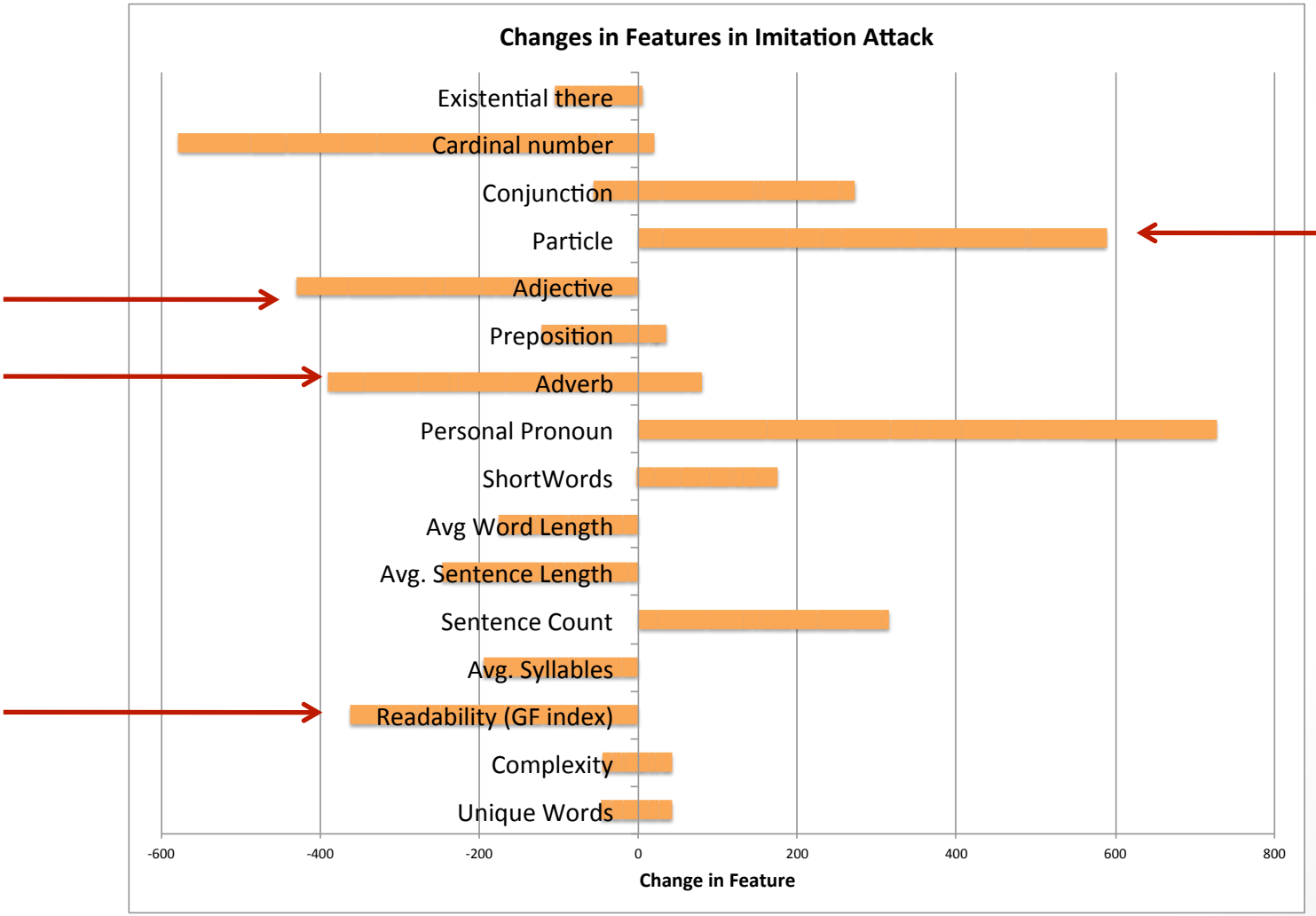
Drexel UNIVERSITY

Feature Changes in Obfuscated Passages



Drexel UNIVERSITY

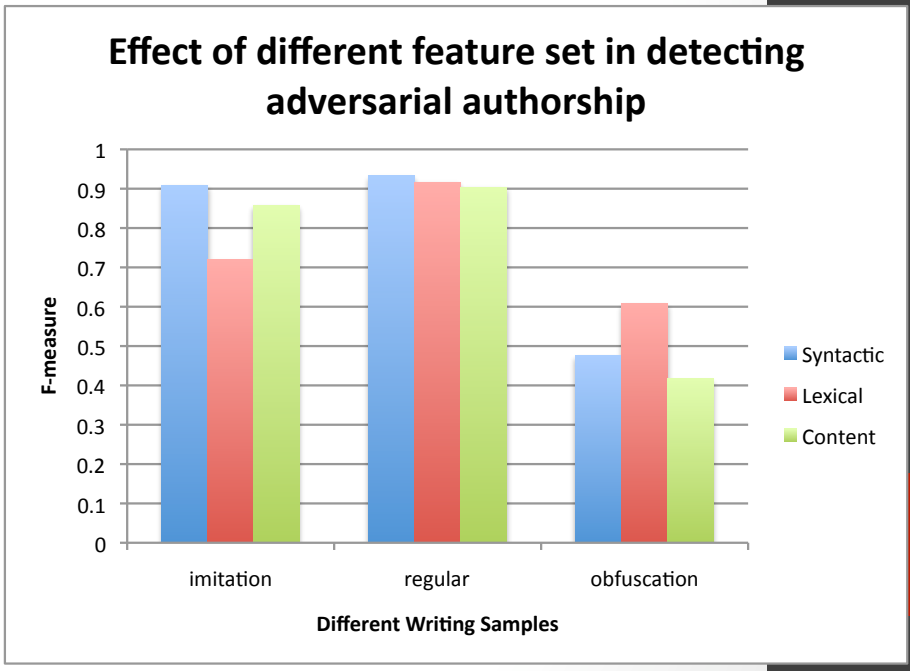
Feature Changes in Imitated Passages



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

- 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
- 36 imitation samples.

Drexel UNIVERSITY

Adversarial attack can be detected even when the topic is not similar

- 81.2% accurate in detecting imitated documents.

Long term deception: A Gay Girl In Damascus

Thomas MacMaster.

- 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.

Drexel UNIVERSITY

Long term deception is hard to detect

- DAMASCUS Last week, the Syrian government unveiled the first of the promised reforms. So far, they've been well, slight might be a kind way to refer to them. Citizenship was finally given to a large group of Kurds, a casino was closed and a ban on teachers wearing the niqab, the face veil, was rescinded. That last probably strikes many readers as hardly the sort of thing that should be greeted as a liberalizing reform. But, in my opinion, it is symbolic

Drexel UNIVERSITY

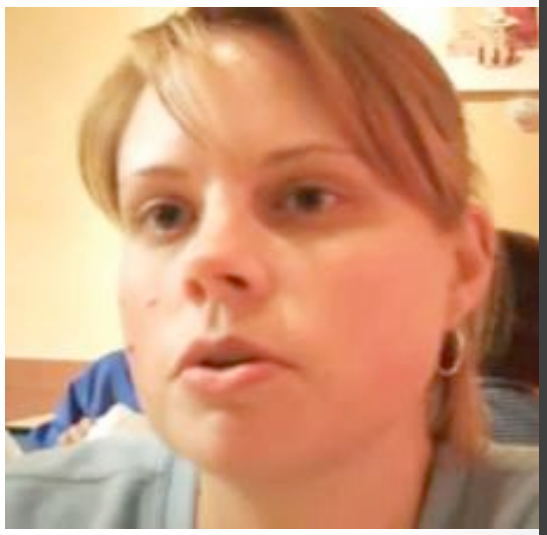
Long term deception
Authorship recognition of the blog posts

Thomas MacMaster.

54%

Amina Arraf

43%



Britta (Thomas's wife)

3%

Drexel UNIVERSITY

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.

Drexel UNIVERSITY

Thanks

- Sadia Afroz (sadia.afroz@drexel.edu)
- Michael Brennan (mb553@drexel.edu)
- Rachel Greenstadt (greenie@cs.drexel.edu)