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

A Gay Girl in Damascus

An out Syrian lesbian's thoughts on life, the universe and so on ...

19 February 2011

"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 inapporpriate 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

A Gay Girl in Damascus

An out Syrian lesbian's thoughts on life, the universe and so on ...

5 June 2011

ANOTHER DAY IN DAMASCUS

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

GAY RIGHTS

Will gays be 'sacrificial lambs' in Arab Spring?

May 27, 2011 | By Catriona Davies, for CNN

Share | Twitter | Email

👍 Recommend   f 880 people recommend this. Be the first of your friends.

(Page 2 of 2)

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 source 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.

Search

Free Amina Abdalla | Syrian Blogger 👍 Like

Writer

Wall

Free Amina Abdalla | Syrian Blogger
Questions about Amina's identity are surfacing. However, we think it is possible that the writer of the blog is indeed in custody, in which case, it is important to continue to support her. Many people in Syria are forced to use alternative identities to protect themselves. However, administrators of this site cannot verify

borders mean **NOTHING** when you have **WINGS**™

FREE AMINA ARRAF
حرية لامينة عارف

Wall
Info
Photos
Notes

About
We demand the release of Amina Abdallah Arraf al Omari

14,866
people like this

Create a Page
Add to my page's favourites
Subscribe via RSS

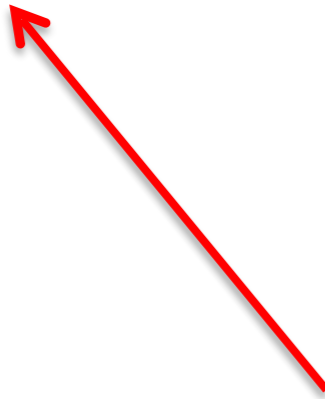Working on my social media skills to help #FreeAmina and spread the word out...Keep the momentum. #Syria

less than a minute ago via web ☆ Favorite ⟲ Retweet ↩ Reply

Sade B.
sade_la_bag

A Gay Girl In Damascus

Fake picture
(copied from Facebook)

The real "Amina"

Thomas MacMaster
A 40-year old American male

Why we are interested?

Thomas developed a new writing style for Amina

- 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

Why not Authorship Recognition?

- Many algorithms are available for authorship recognition using writing style.

- Why cannot we use that?

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

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.
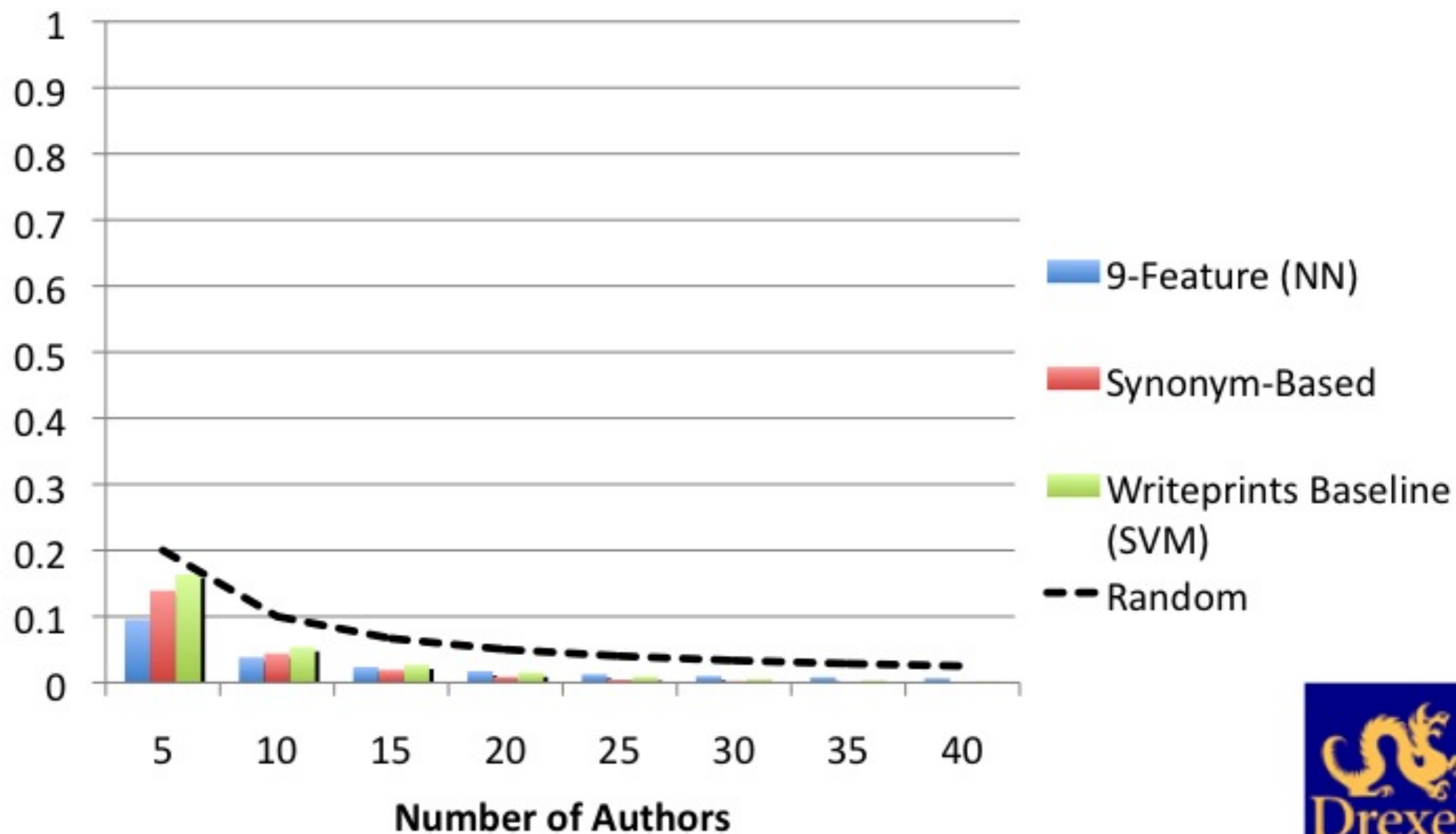
Accuracy in detecting authorship of regular documents
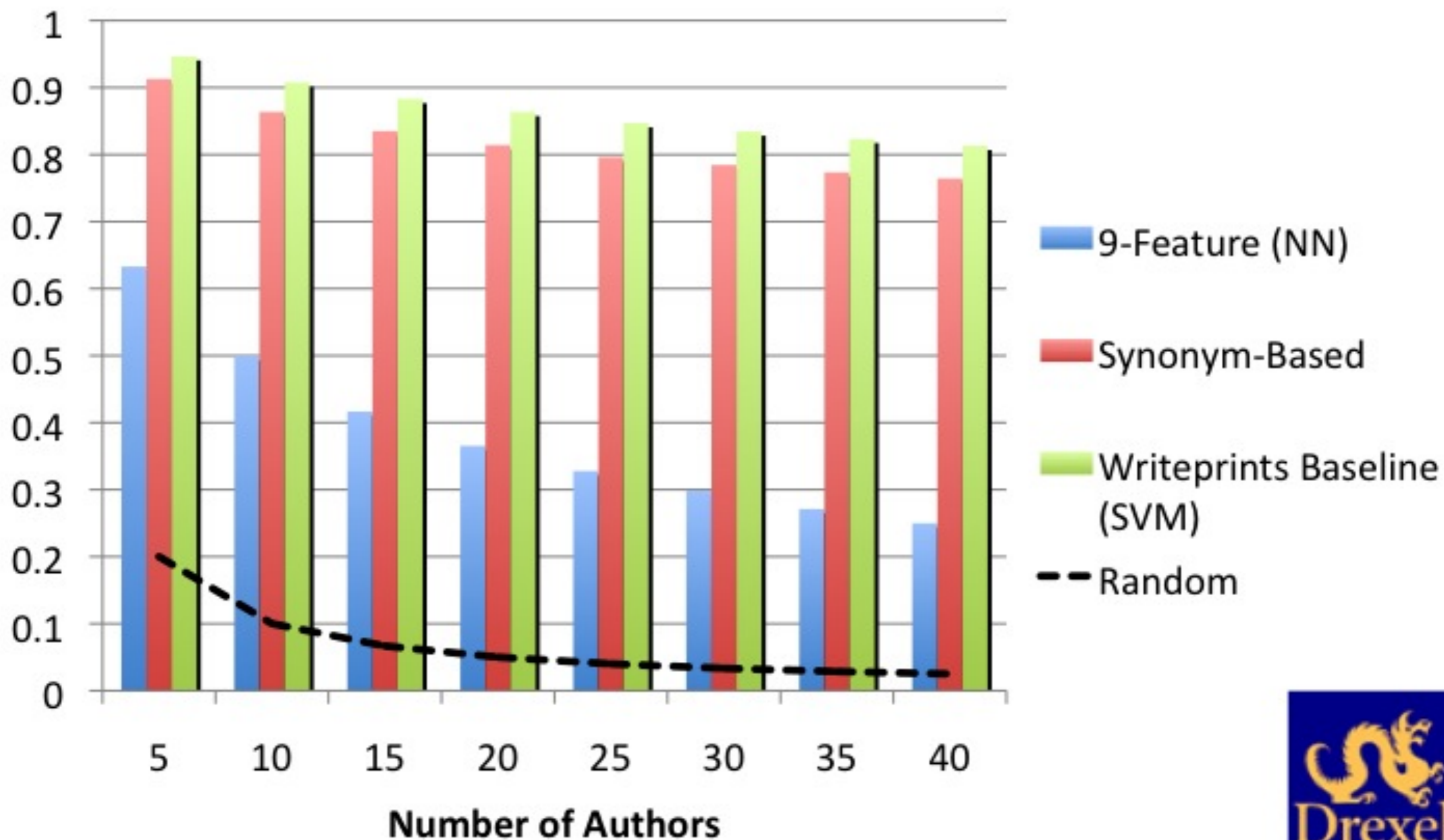
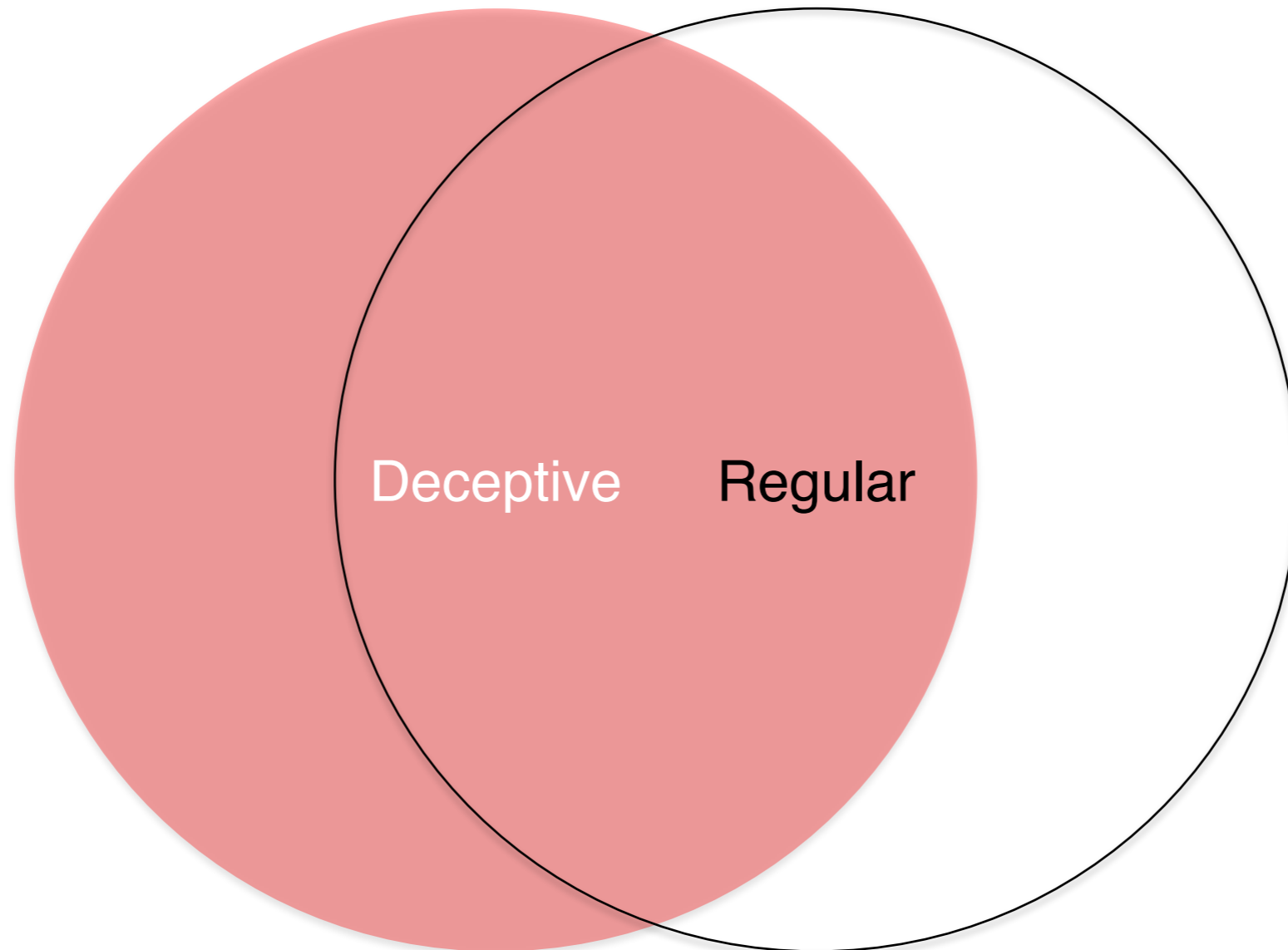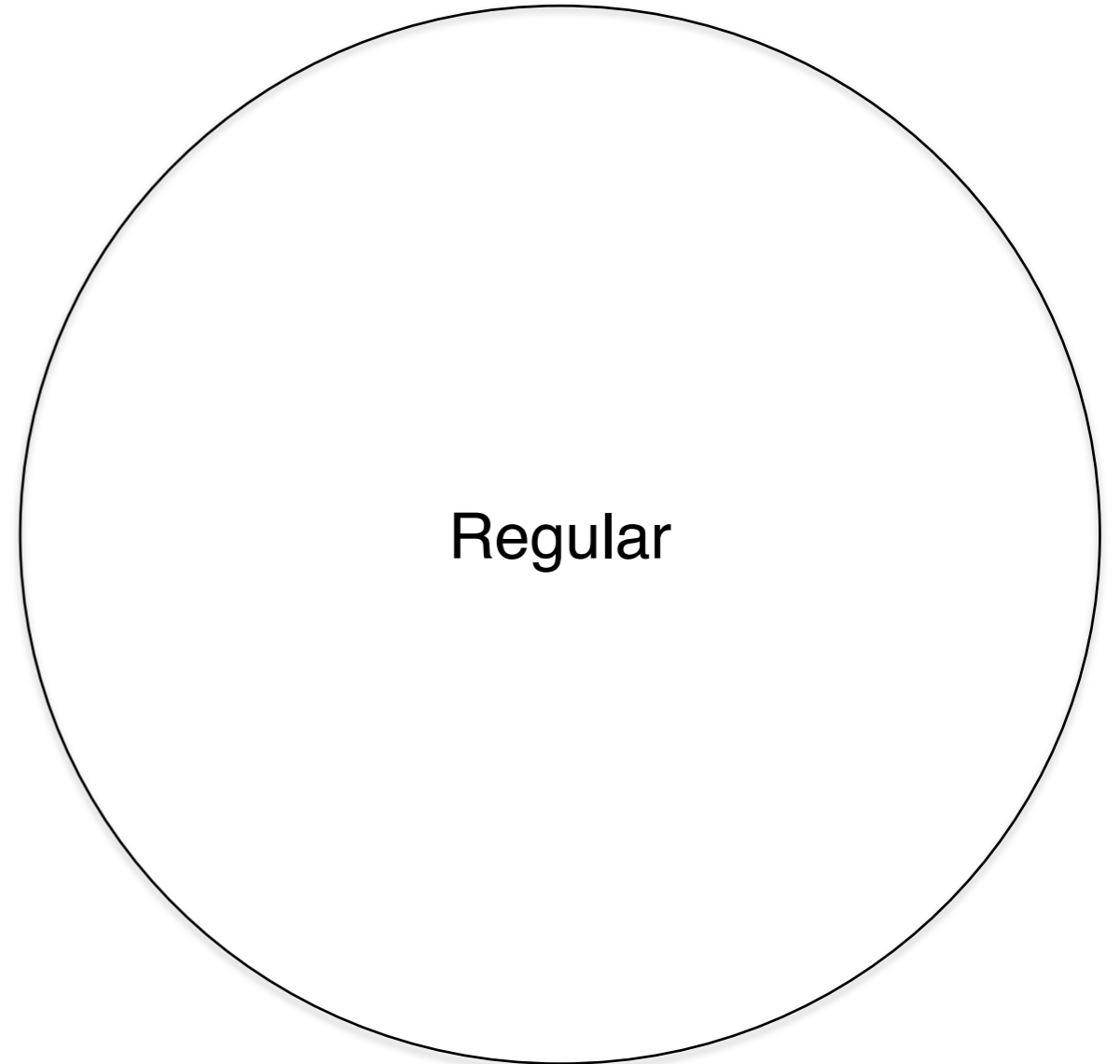# Accuracy in detecting authorship of obfuscated documents
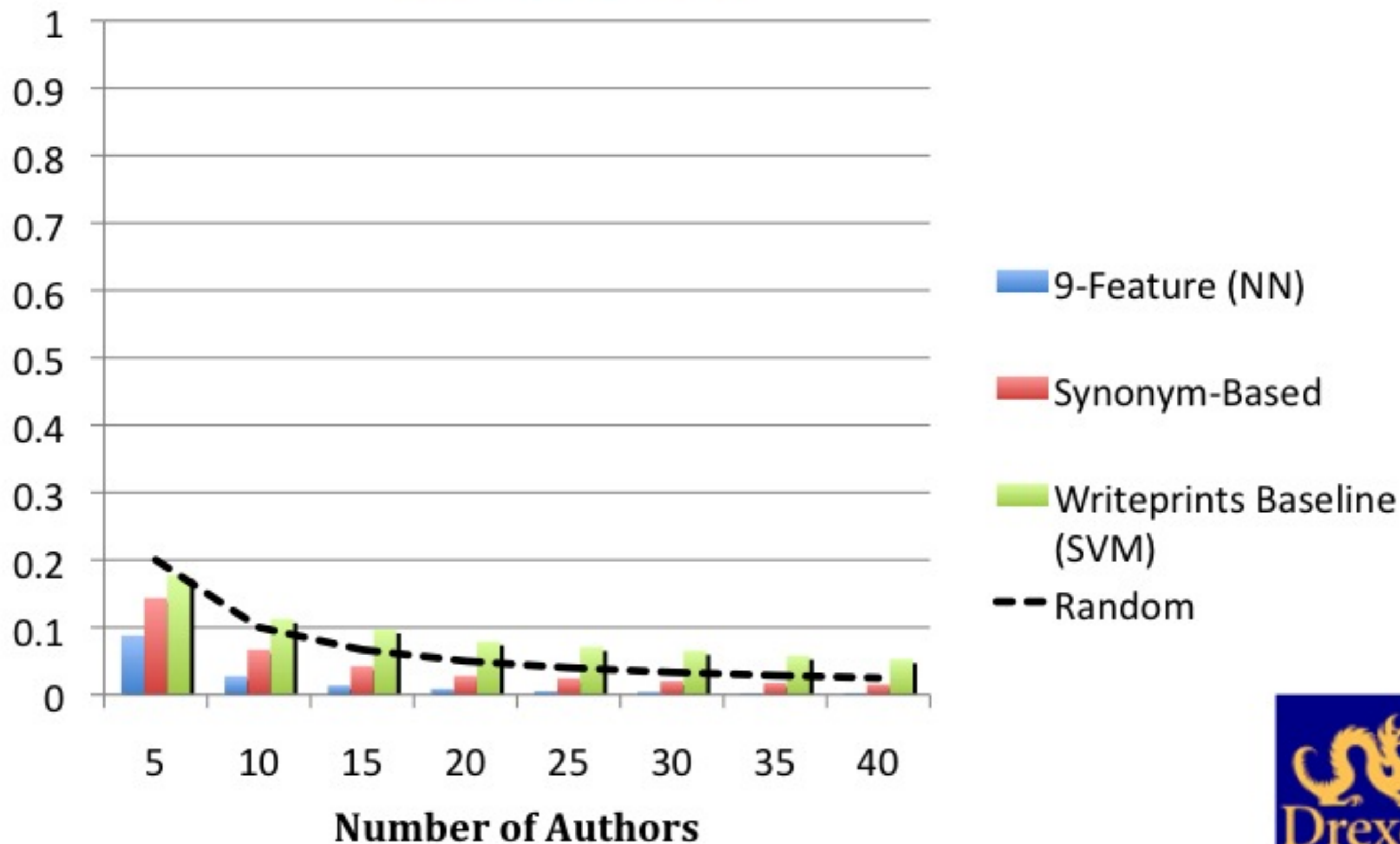
Accuracy in detecting authorship of imitated documents
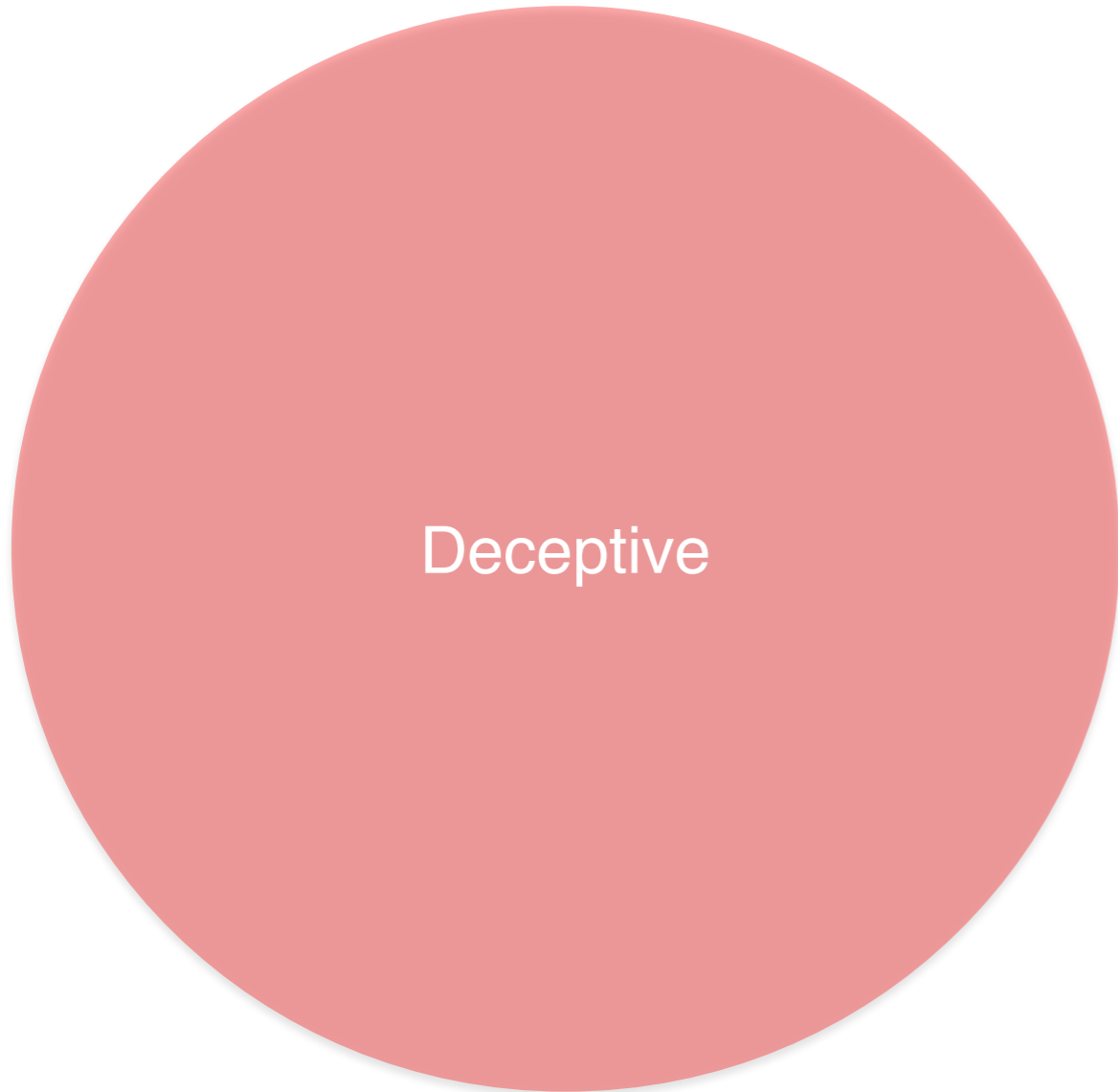
Can we detect stylistic deception?
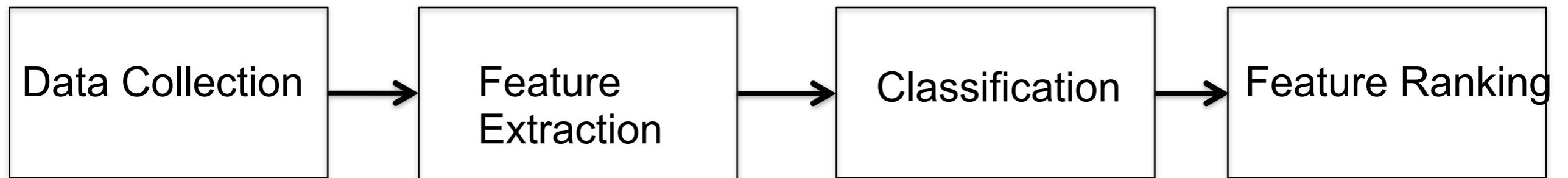
Deceptive   Regular

Can we detect stylistic deception?

Deceptive

Regular

Analytic Approach

```
┌──────────────────┐     ┌──────────────────┐     ┌──────────────────┐     ┌──────────────────┐
│                  │     │                  │     │                  │     │                  │
│ Data Collection  │────▶│ Feature          │────▶│ Classification   │────▶│ Feature Ranking  │
│                  │     │ Extraction       │     │                  │     │                  │
│                  │     │                  │     │                  │     │                  │
└──────────────────┘     └──────────────────┘     └──────────────────┘     └──────────────────┘
```

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:**
  - Extended-Brennan-Greenstadt Corpus
    - Fixed topic
    - Controlled style
  - Hemingway-Faulkner Imitation corpus
    - No fixed topic
    - Controlled style

- **Long-term deception:**

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

Feature sets

- 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.
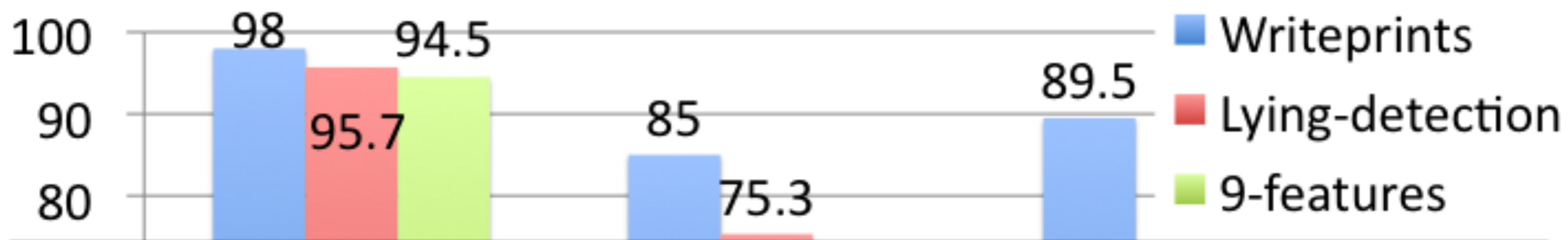
Detecting stylistic deception is possible

Detecting stylistic deception is possible
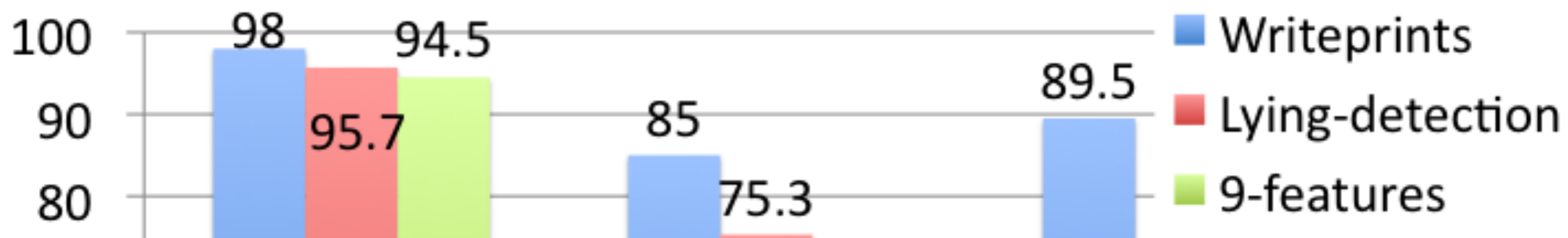
Legend: Writeprints, Lying-detection, 9-features

Regular: 98, 95.7, 94.5
Imitation: 85, 75.3, 48
Obfuscation: 89.5, 43

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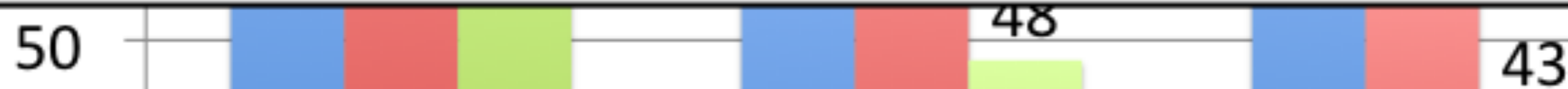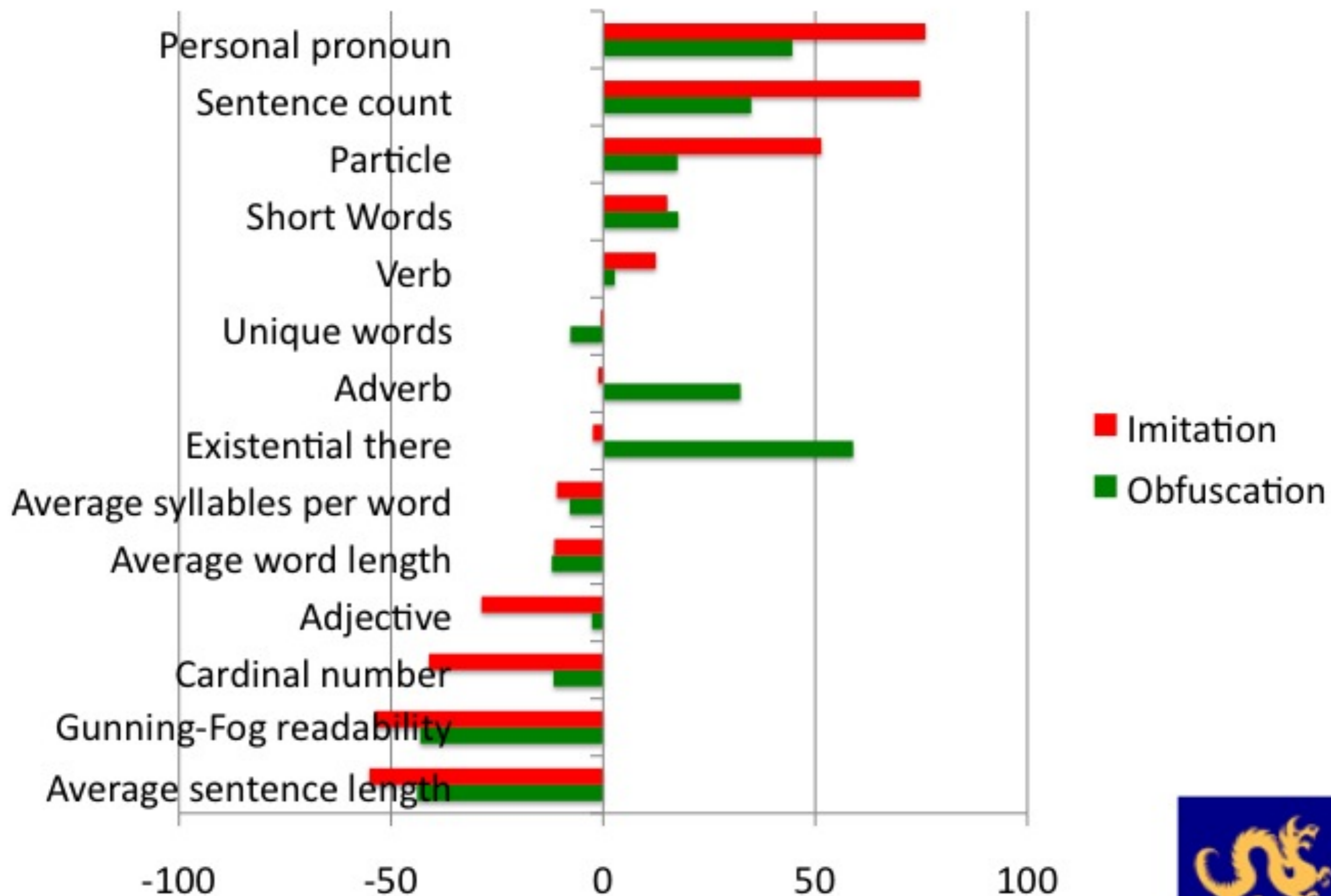
Writing style change has similarity with lying as lying-detection features can detect imitation and obfuscation.
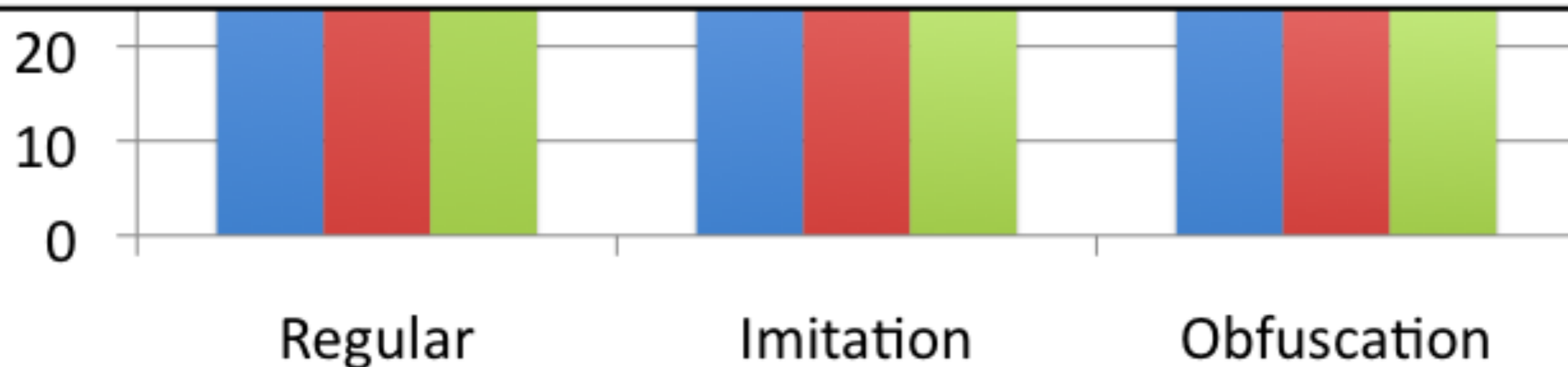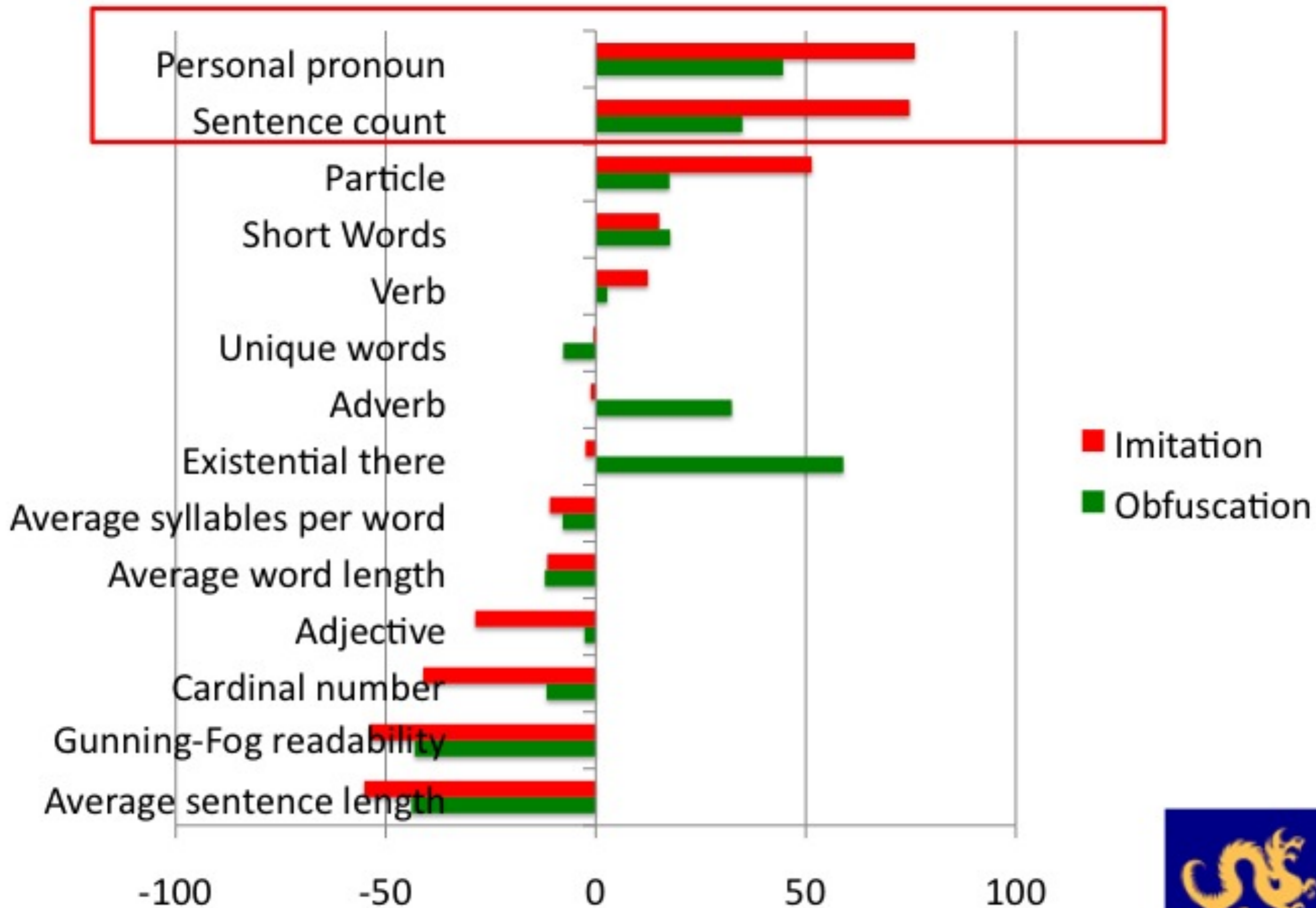
Which features are changed

Change in feature = $\dfrac{Feature_{deceptive} - Feature_{regular}}{Feature_{regular}}$
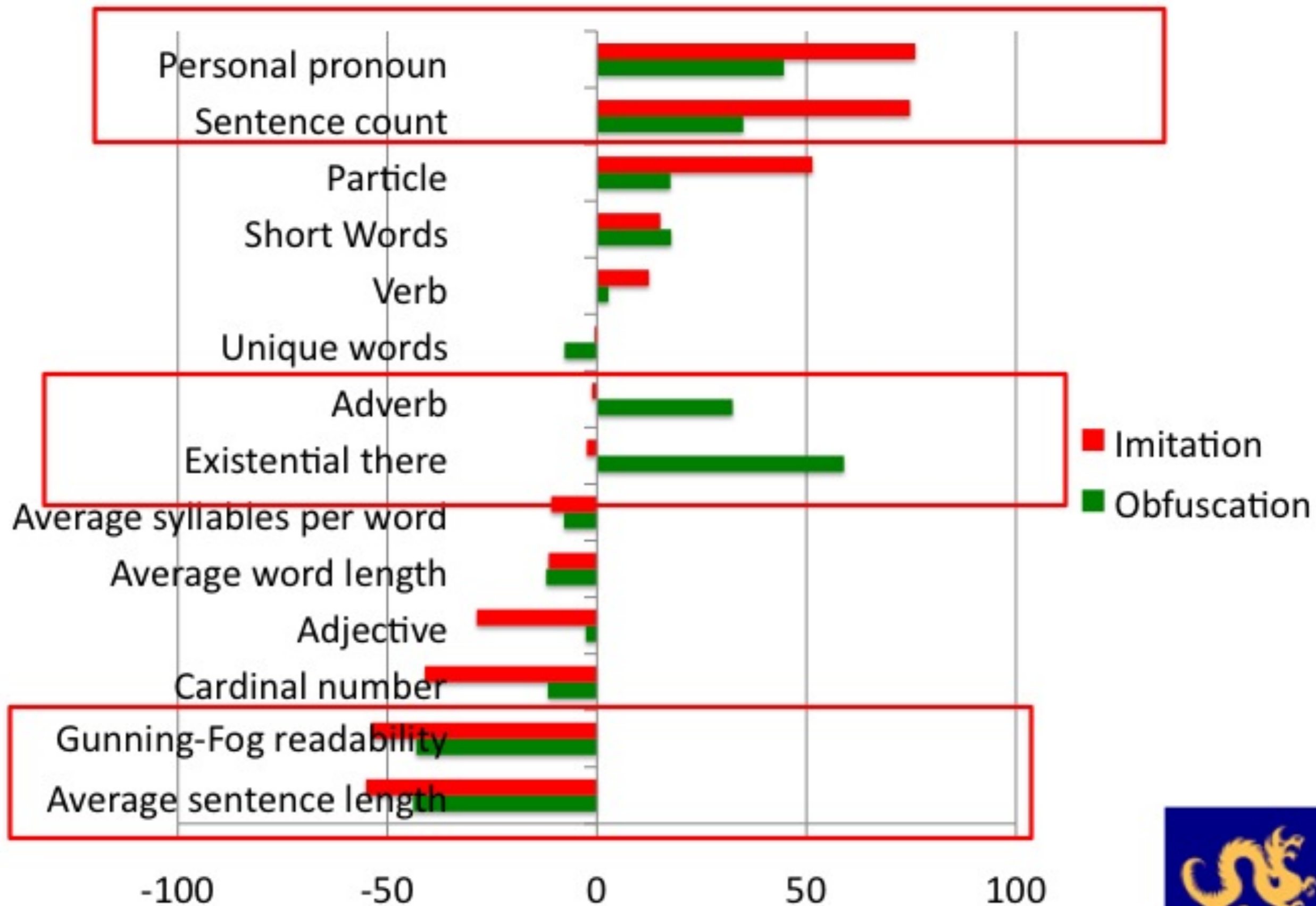
Feature Changes in Imitation and Obfuscation

Feature Changes in Imitation and Obfuscation

Bar chart showing feature changes for Imitation (red) and Obfuscation (green). Features listed top to bottom: 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. X-axis ranges from -100 to 100. Legend: Imitation (red), Obfuscation (green).

Feature Changes in Imitation and Obfuscation

Bar chart comparing Imitation (red) and Obfuscation (green) feature changes, with features listed from top to bottom: 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. X-axis ranges from -100 to 100. Legend: Imitation (red), Obfuscation (green).

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
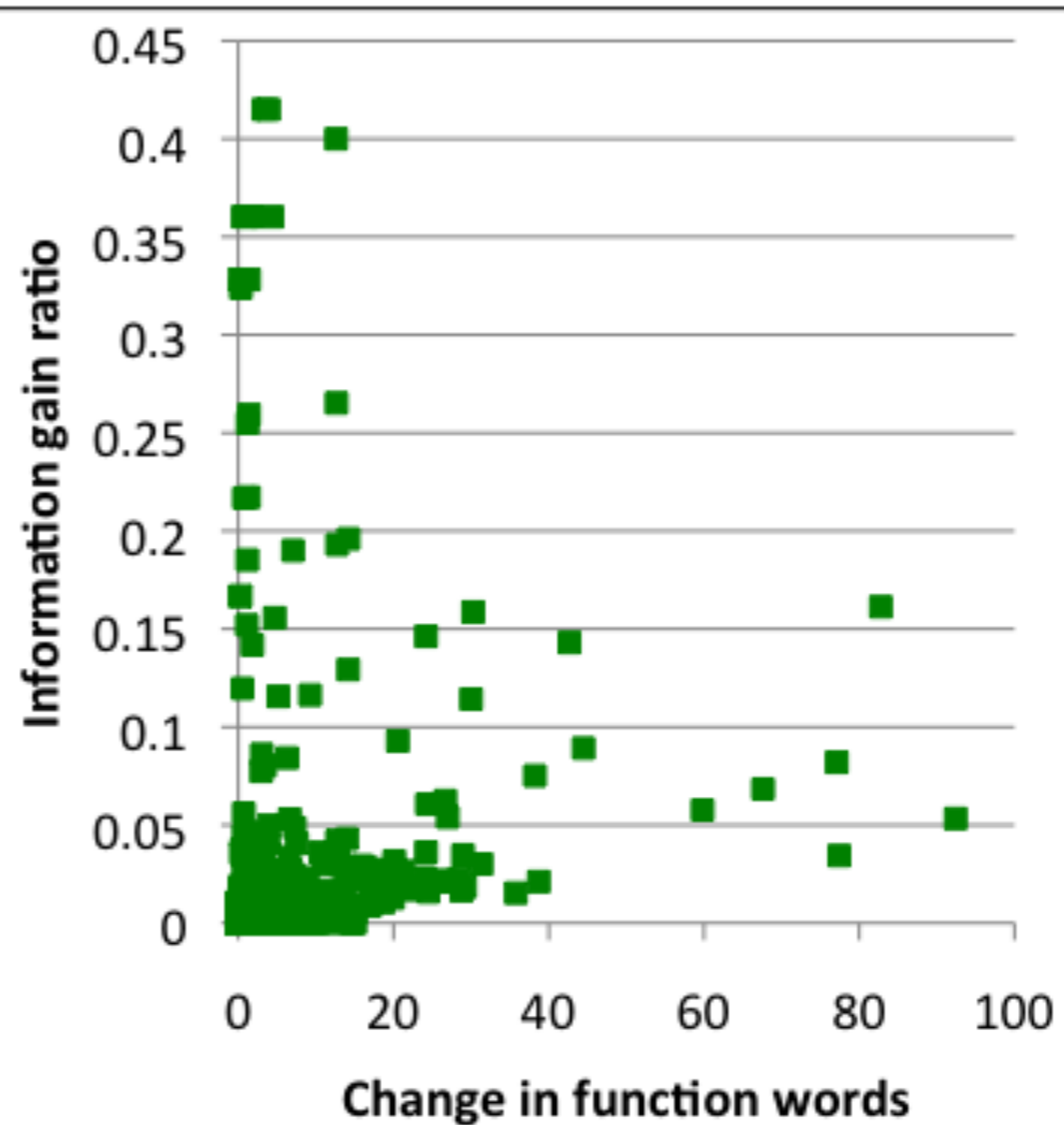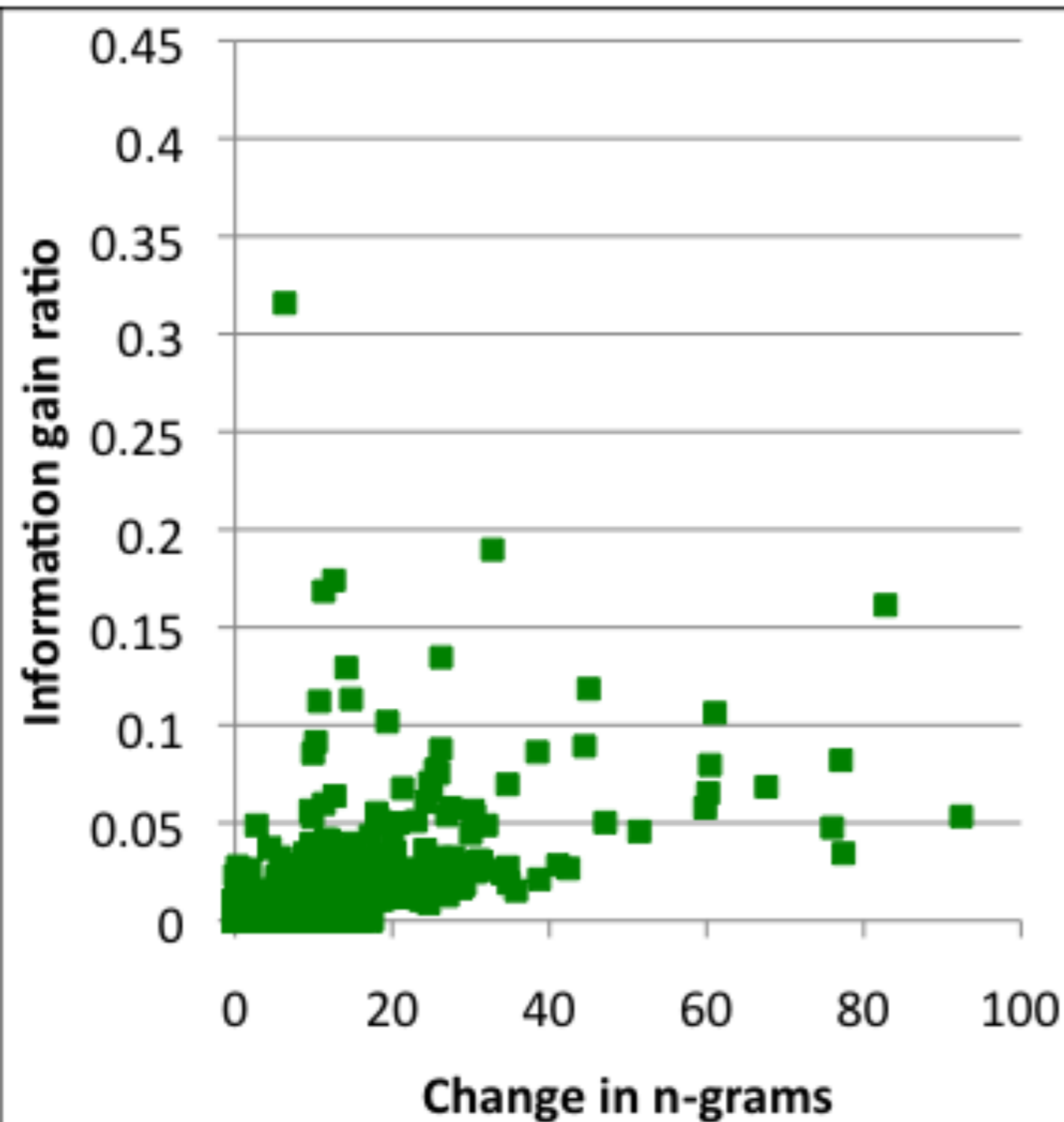
Information Gain vs Feature Change
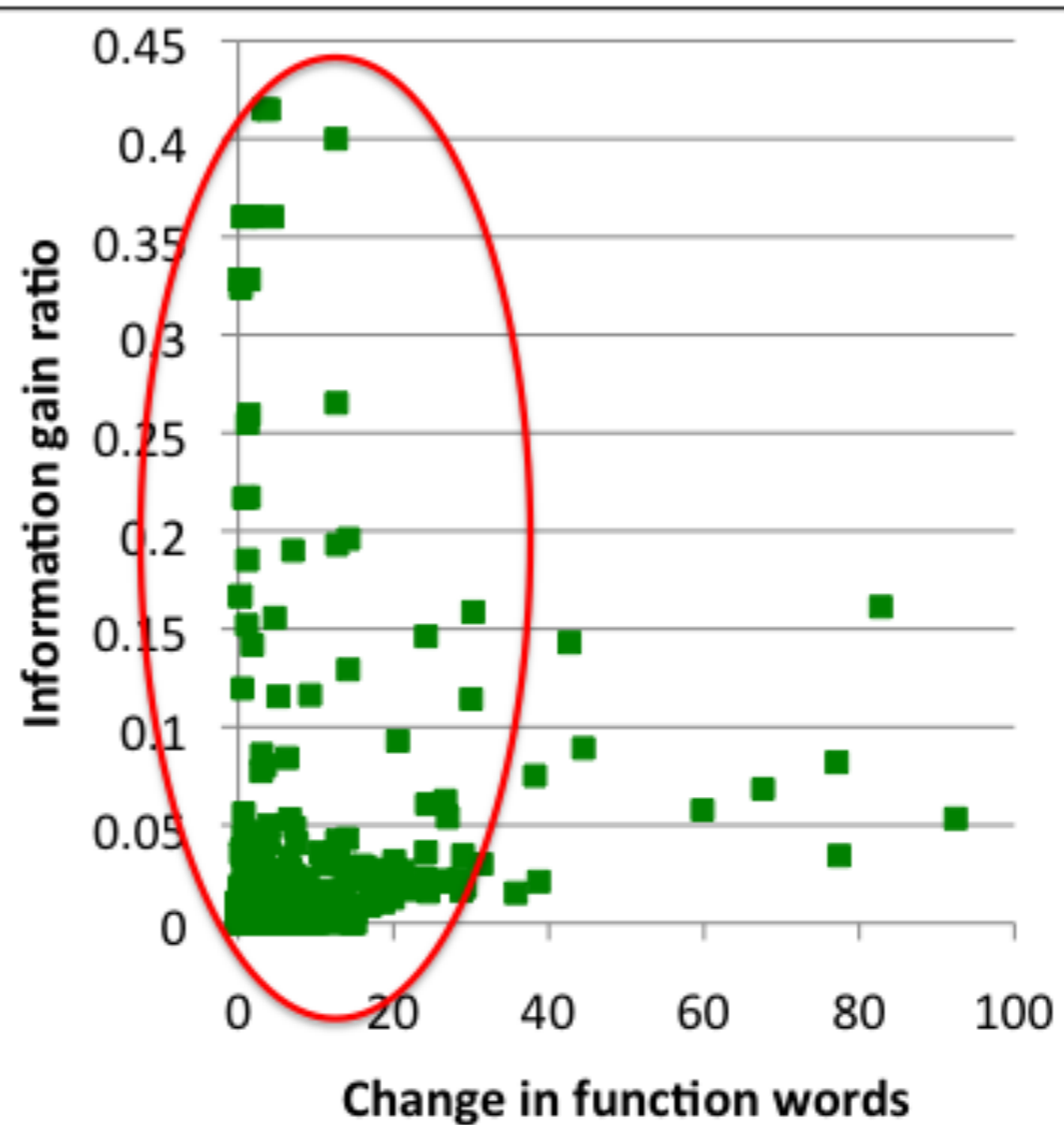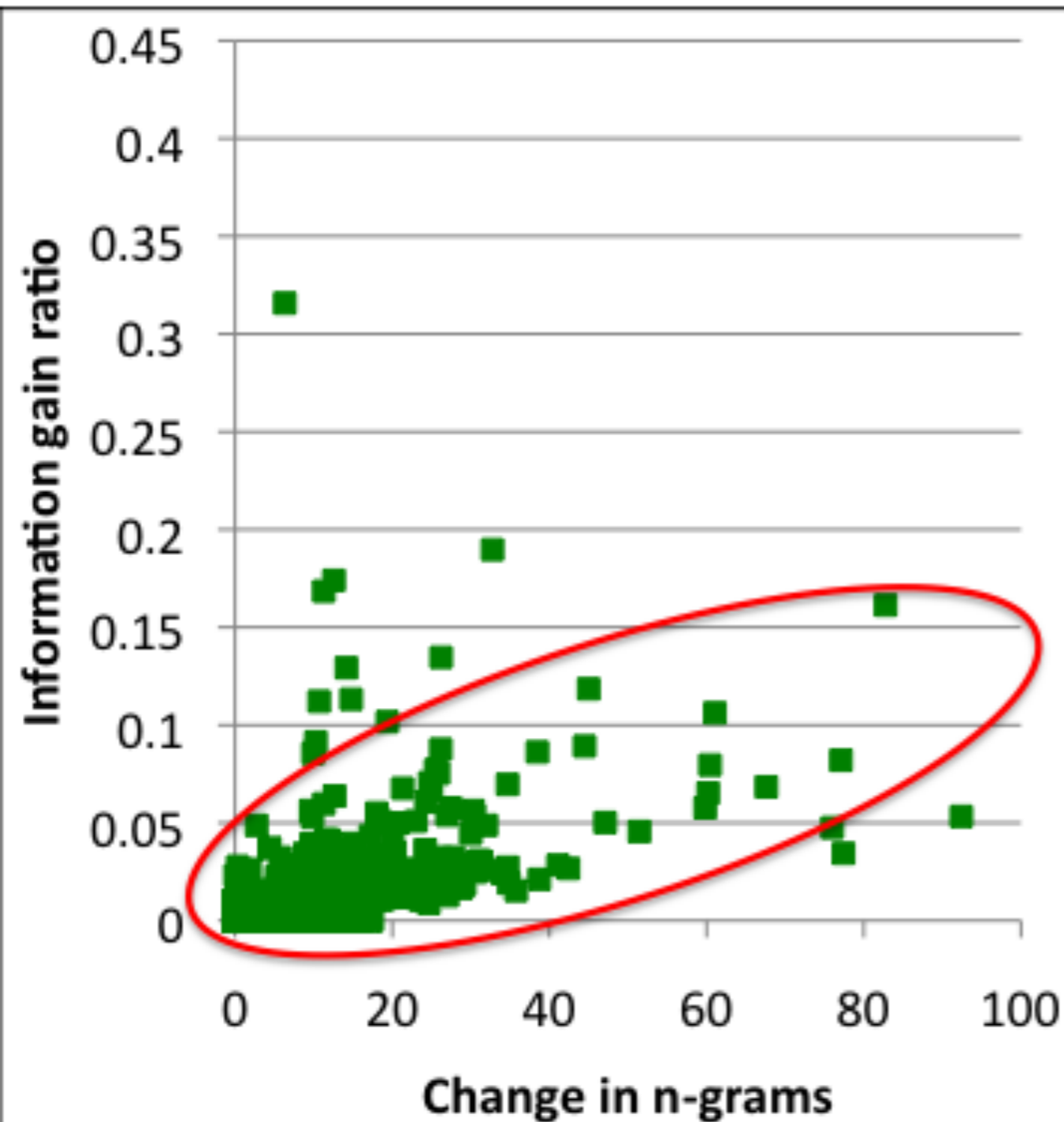
How the classifier uses changed and unchanged features

- 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.
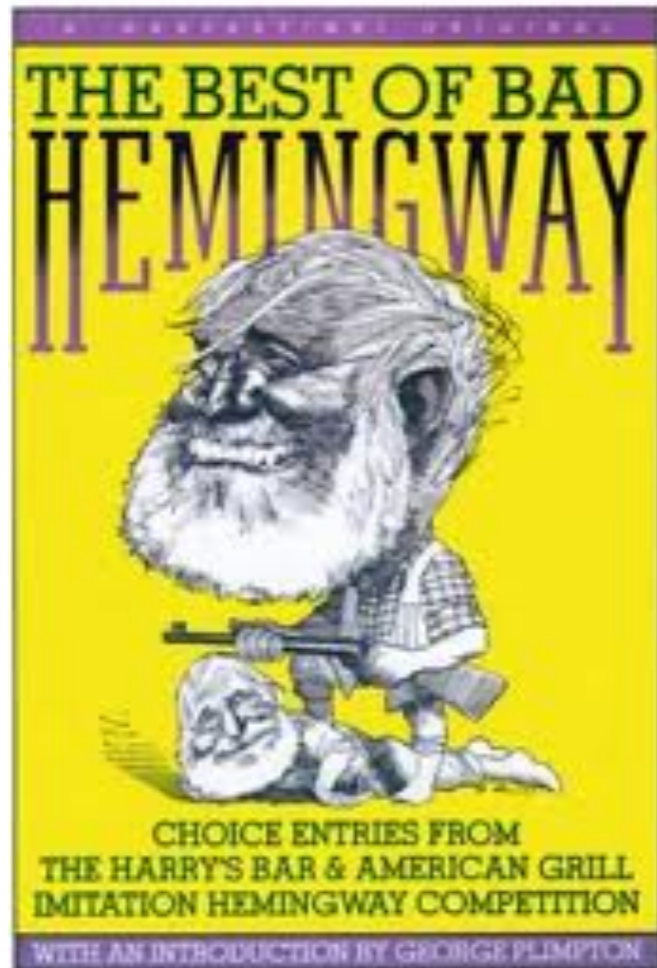
Problem with the dataset:
Topic Similarity

- All the adversarial documents were of same topic.
- Non-content-specific features have same effect as content-specific features.

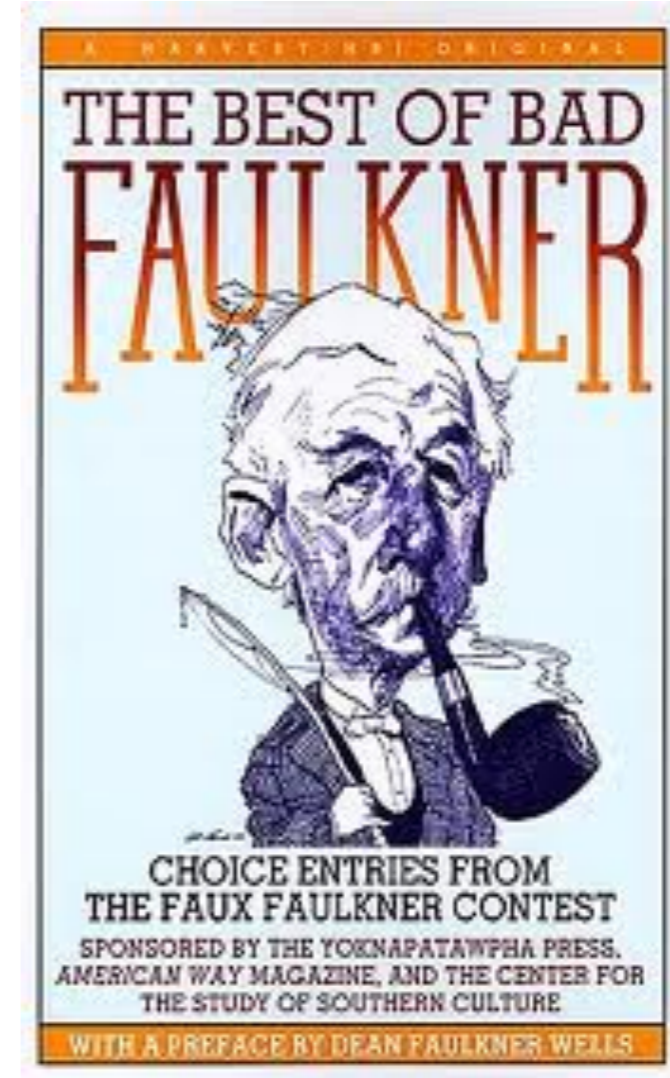Effect of different feature set in detecting adversarial authorship
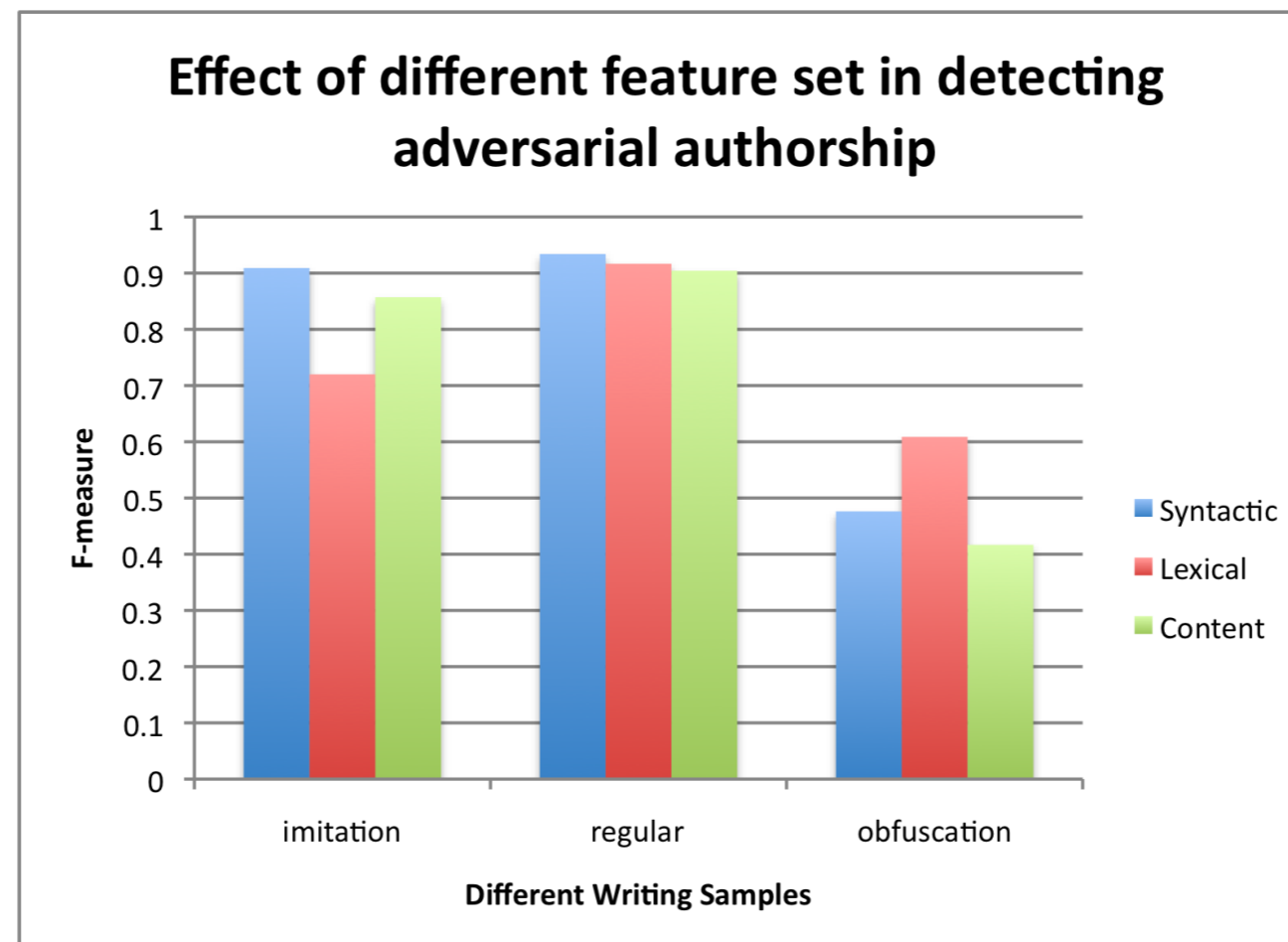
Hemingway-Faulkner Imitation Corpus

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

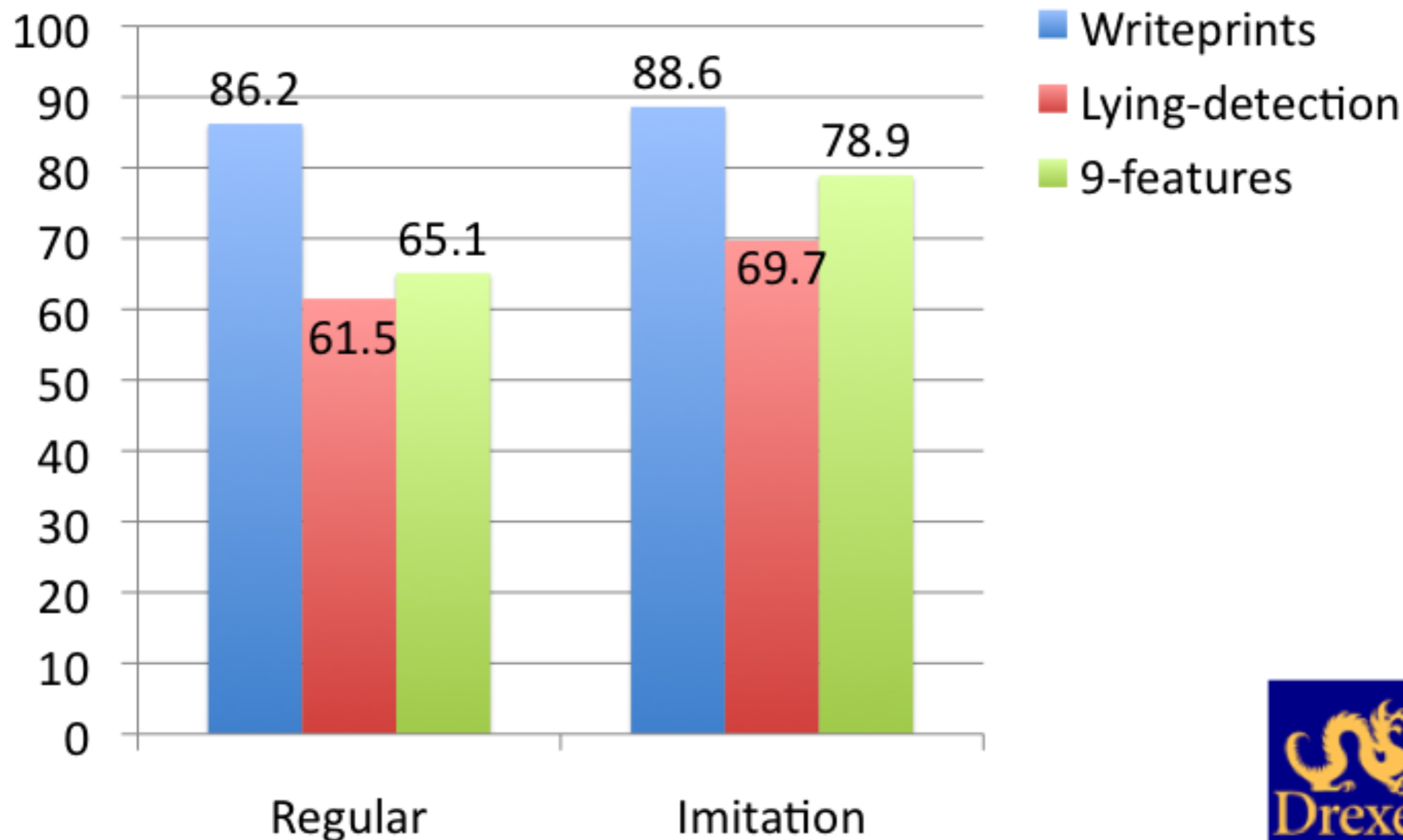# Hemingway-Faulkner Imitation Corpus

- Classification task:
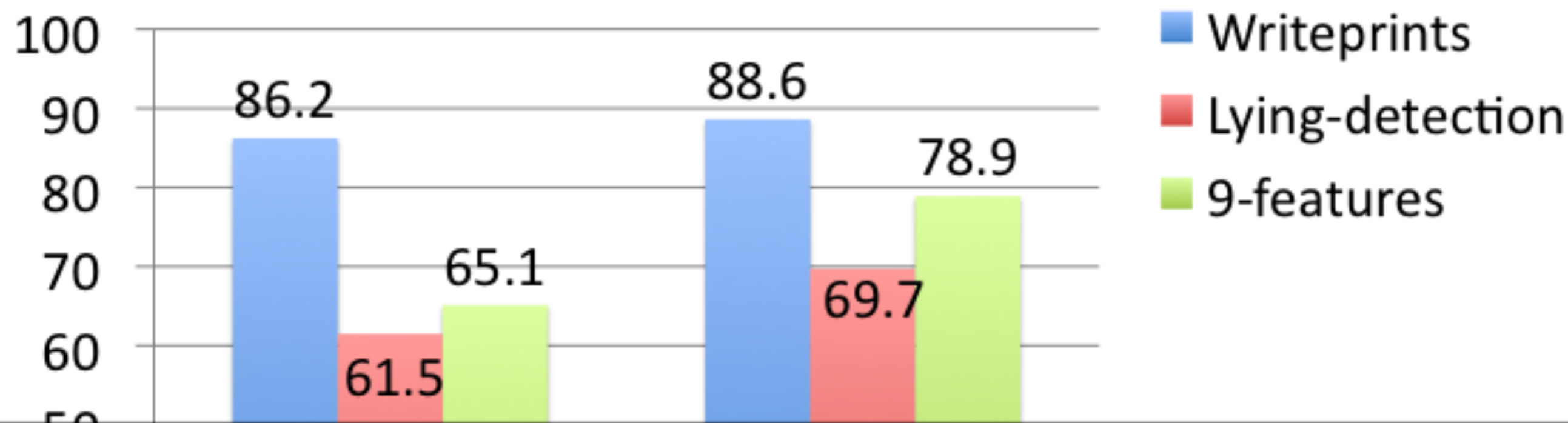 - Distinguish Regular and Imitation

Imitation success

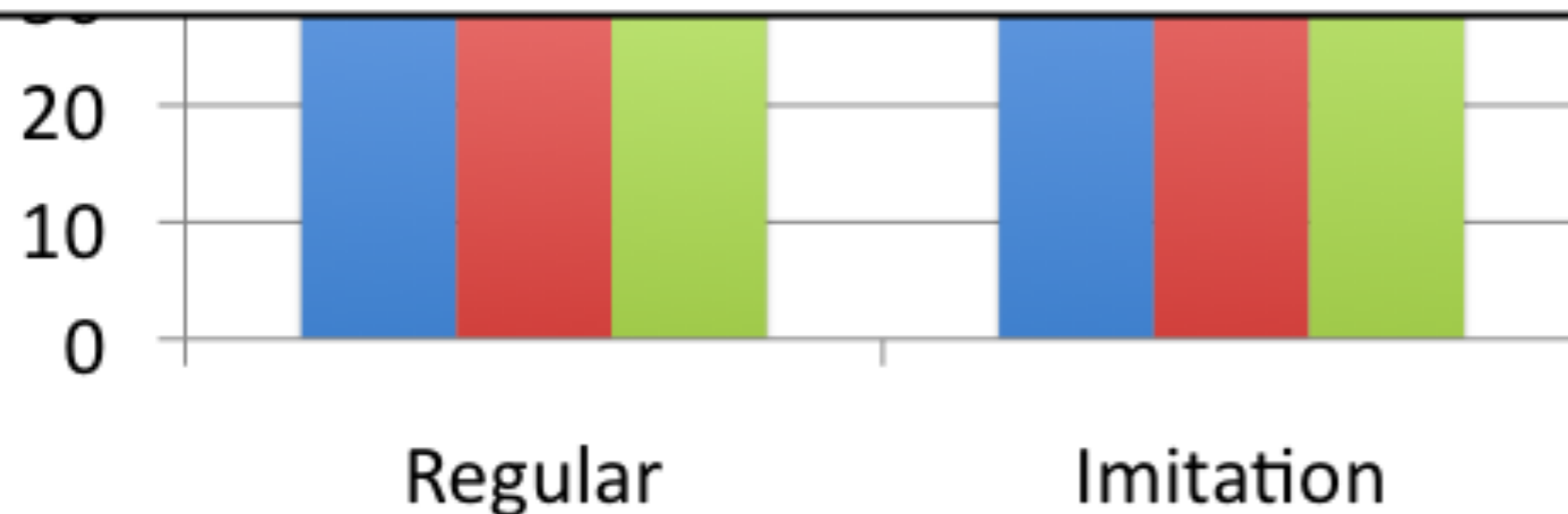| Author to imitate	Imitation success	Writer's Skill
Cormac McCarthy	*47.05%*	*Not professional*
Ernest Hemingway	*84.21%*	*Professional*
William Faulkner	*66.67%*	*Professional*

Result: Hemingway-Faulkner Imitation corpus

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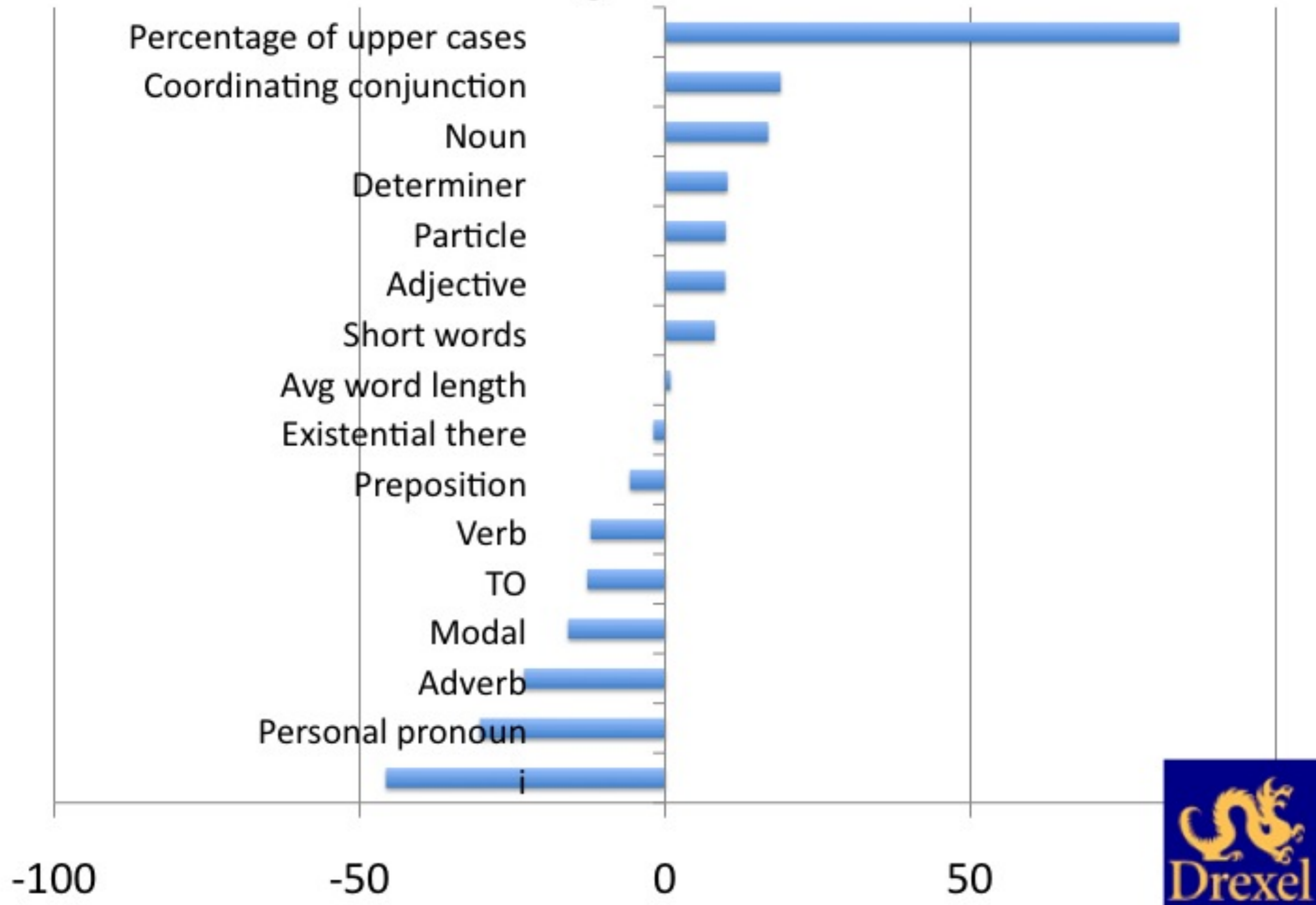Thomas MacMaster (as Amina Arraf): 43%
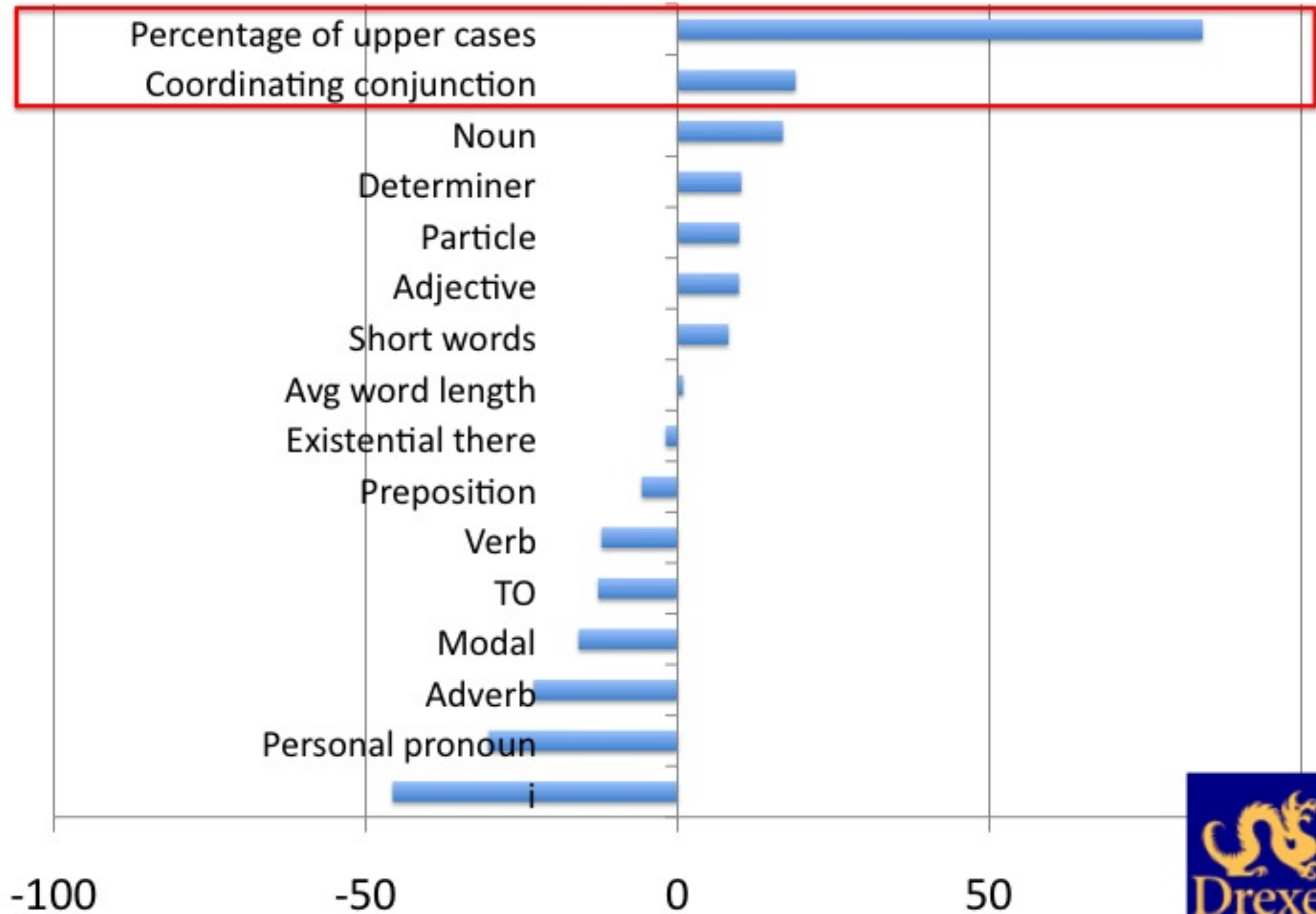Britta: 3%

Maintaining separate writing styles is hard!

Which features are changed

Change in feature = $\dfrac{Feature_{AMINA} - Feature_{THOMAS}}{Feature_{THOMAS}}$
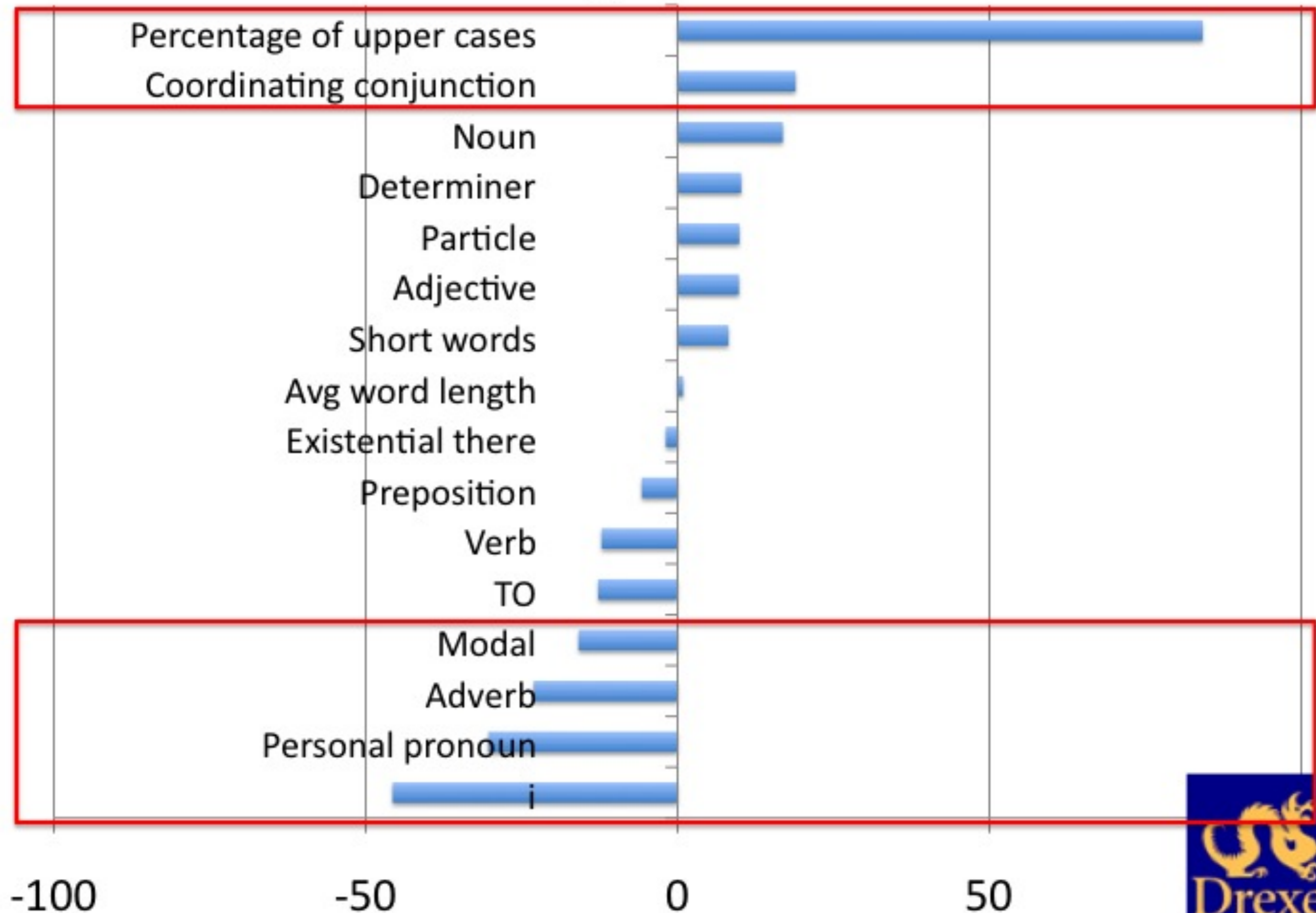
Feature Change: Thomas-Amina

Feature Change: Thomas-Amina

Feature Change: Thomas-Amina

Summary

- ## The problem:
  - Detecting stylistic deception
  - How to detect if someone is hiding his regular writing style?

- ## Why do we care:
  - For detecting hoaxes and frauds.
  - For automating writing style anonymization.

- ## Why not authorship recognition:
  - Because authorship recognition algorithms are not effective in detecting authorship when style is changed.

Summary

- Results:
  - Extended-Brennan-Greenstadt corpus:
    - We can detect imitation and obfuscation with high accuracy.
  - Hemingway-Faulkner Imitation corpus:
    - We can detect imitation with high accuracy.
  - Thomas-Amina Hoax corpus:
    - We can detect authorship of the blog posts as maintaining different writing styles is hard.
- Which linguistic features are changed when people hide their writing style:
  - Adjectives, adverbs, sentence length, average syllables per word

- Which linguistic features are not changed
  - Function words (and, or, of, for, on)

Future work

- JStylo: Authorship Recognition Analysis Tool.
- Anonymouth: Authorship Recognition Circumvention Tool.

- Free, Open Source. (GNU GPL)
- Alpha releases available at: https://psal.cs.drexel.edu

Thank you!

- Sadia Afroz: sadia.afroz@drexel.edu
- Michael Brennan: mb553@drexel.edu
- Rachel Greenstadt: greenie@cs.drexel.edu
- Privacy, Security And Automation Lab (https://psal.cs.drexel.edu)