IN SEARCH OF STYLES IN LANGUAGE

Identifying ✓ Deceptive Product Reviews
✓ Wikipedia Vandalism
✓ The Gender of Authors

via Statistical Stylometric Analysis

Yejin Choi
Stony Brook University
“So how can you spot a fake review? Unfortunately, it’s difficult, but with some technology, there are a few warning signs:”
### Styles in Language

**Research Papers?**  **New York Times?**  **Blogs?**

<table>
<thead>
<tr>
<th>“So how can you spot a fake review? Unfortunately, it’s difficult, but with some technology, there are a few warning signs:”</th>
</tr>
</thead>
</table>

| “To obtain a deeper understanding of the nature of deceptive reviews, we examine the relative utility of three potentially complementary framings of our problem.” |
“So how can you spot a fake review? Unfortunately, it’s difficult, but with some technology, there are a few warning signs:”

“To obtain a deeper understanding of the nature of deceptive reviews, we examine the relative utility of three potentially complementary framings of our problem.”

“As online retailers increasingly depend on reviews as a sales tool, an industry of fibbers and promoters has sprung up to buy and sell raves for a pittance.”
Why different Styles in Language?

Influencing factors:

• Convention / customary style of certain genres
• Expected audience
• Intent of the author
• Personal traits of the author
The Stuff of Thought: Language as a Window into Human Nature

-- Steven Pinker

The Secret Life of Pronouns: What our Words Say about Us

-- James W. Pennebaker
What constitute Styles in language?

• Lexical Choice
• Grammar / Syntactic Choice
• Cohesion / Discourse Structure
• Narrator / Point of View
• Tone (formal, informal, intimate, playful, serious, ironic, condescending)
• Imagery, Allegory, Punctuation, and more
Computational analysis of styles

Mostly limited to

lexical choices

shallow syntactic choices (part of speech)

--- notable exception: Raghavan et al. (2010)
# Previous Research in NLP

## Genre Detection
- Petrenz and Webber, 2011
- Sharoff et al., 2010
- Wu et al., 2010
- Feldman et al., 2009
- Finn et al., 2006
- Argamon et al., 2003
- Dewdney et al., 2001
- Stamatatos et al., 2000
- Kessler et al., 1997

## Authorship Attribution
- Escalante et al., 2011
- Seroussi et al., 2011
- Raghavan et al., 2010
- Luyckx and Daelemans, 2008
- Koppel and Shler, 2004
- Gamon, 2004
- van Halteren, 2004
- Spracklin et al., 2008
- Stamatatos et al., 1999
In this talk: three case studies of *stylometric analysis*

- Deceptive Product Reviews
- Wikipedia Vandalism
- The Gender of Authors
In this talk: three case studies of **stylometric analysis**

- Deceptive Product Reviews
- Wikipedia Vandalism
- The Gender of Authors

**Underlying themes:**

A. Discovering “language styles” in a broader range of real-world NLP tasks

B. Learning (statistical) stylistic cues beyond shallow lexico-syntactic patterns.
In this talk: three case studies of stylometric analysis

- Deceptive Product Reviews
- Wikipedia Vandalism
- The Gender of Authors
Motivation

• Consumers increasingly rate, review and research products online

• Potential for opinion spam
  – Disruptive opinion spam
  – Deceptive opinion spam
Motivation

• Consumers increasingly rate, review and research products online

• Potential for opinion spam
  – Disruptive opinion spam
  – Deceptive opinion spam

★☆☆☆☆ Great Customer Service!!, April 7, 2011
By akaempf - See all my reviews
Amazon Verified Purchase (What's this?)
This review is from: Apple iPad 2 MC984LL/A Tablet (64GB, Wifi + AT&T 3G, White) NEWEST MODEL (Personal Computers)
"WE SHIP TECH" is a great reliable company. I ordered the iPad2 late 3/30 @ 10:50pm and received the iPad2 4/1. When I wrote an email to them on the 3/31 they responded in about 20 min max. It's so hard to find great customer service and not get scammed these days that "We Ship Tech" is a breath of fresh air!! I would surely use them again and highly recommend them to anyone who expects great products & service. Thank you We Ship Tech!!!!!
Motivation

- Consumers increasingly rate, review and research products online
- Potential for opinion spam
  - Disruptive opinion spam
  - Deceptive opinion spam
“My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful! The area of the hotel is great, since I love to shop I couldn’t ask for more! We will definitely be back to Chicago and we will for sure be back to the James Chicago.”

Deceptive or Truthful?
“My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful! The area of the hotel is great, since I love to shop I couldn’t ask for more! We will definitely be back to Chicago and we will for sure be back to the James Chicago.”

“I have stayed at many hotels traveling for both business and pleasure and I can honestly say that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travellers and couples.”
“My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful! The area of the hotel is great, since I love to shop I couldn’t ask for more! We will definitely be back to Chicago and we will for sure be back to the James Chicago.”
Gathering Data

• Label existing reviews?
  – Can’t manually do this
Gathering Data

• Label existing reviews?
  – Can’t manually do this

• Create new reviews
  – By hiring people to write fake POSITIVE reviews
  – Using Amazon Mechanical Turk
Gathering Data

• Mechanical Turk
  – 20 hotels
  – 20 reviews / hotel
  – Offer $1 / review
  – 400 reviews
Gathering Data

• Mechanical Turk
  – 20 hotels
  – 20 reviews / hotel
  – Offer $1 / review
  – 400 reviews
Gathering Data

• Mechanical Turk
  – 20 hotels
  – 20 reviews / hotel
  – Offer $1 / review
  – 400 reviews
Gathering Data

- **Mechanical Turk**
  - 20 hotels
  - 20 reviews / hotel
  - Offer $1 / review
  - 400 reviews

- **Average time spent:**
  > 8 minutes

- **Average length:**
  > 115 words
Human Performance

• Why bother?
  – Validates deceptive opinions
  – Baseline to compare other approaches

• 80 truthful and 80 deceptive reviews
• 3 undergraduate judges
Human Performance

Accuracy

Judge 1: 61.9
Judge 2: 56.9
Judge 3: 53.1
Human Performance

Accuracy

Judge 1: 61.9
Judge 2: 56.9
Judge 3: 53.1

Performed at chance (p-value = 0.1)
Performed at chance (p-value = 0.5)
Human Performance

Judge 1: Precision 74.4, Recall 36.3
Judge 2: Precision 78.9, Recall 18.8
Judge 3: Precision 54.7, Recall 36.3
Human Performance

<table>
<thead>
<tr>
<th>Judge</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>74.4</td>
<td>36.3</td>
</tr>
<tr>
<td>Judge 2</td>
<td>78.9</td>
<td>18.8</td>
</tr>
<tr>
<td>Judge 3</td>
<td>54.7</td>
<td>36.3</td>
</tr>
</tbody>
</table>

Low Recall ➔ Truth Bias
Human Performance

Meta Judges

1. Majority
2. Skeptic
Human Performance
being skeptical helps with recall...

<table>
<thead>
<tr>
<th></th>
<th>Majority</th>
<th>Skeptic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>76</td>
<td>60.5</td>
</tr>
<tr>
<td>Recall</td>
<td>23.8</td>
<td>61.3</td>
</tr>
<tr>
<td>F-score</td>
<td>36.2</td>
<td>60.9</td>
</tr>
</tbody>
</table>
Human Performance
but not the accuracy

Accuracy

Best Single Judge: 61.9
Meta Judge - Majority: 54.8
Meta Judge - Skeptic: 60.8
Classifier Performance

• Feature sets
  – POS (Part-of-Speech Tags)
  – Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007)
  – Unigram, Bigram, Trigram

• Classifiers: SVM & Naïve Bayes
Classifier Performance

• Feature sets
  - POS (Part-of-Speech Tags)
    – Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007)
    – Unigram, Bigram, Trigram

• Classifiers: SVM & Naïve Bayes
Classifier Performance

• Viewed as *genre* identification
  – 48 part-of-speech (POS) features
  – Baseline automated approach

• Expectations
  – Truth similar to *informative* writing
  – Deception similar to *imaginative* writing
Classifier Performance

<table>
<thead>
<tr>
<th></th>
<th>Best Human Variant</th>
<th>Classifier - POS Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>61.9</td>
<td>73</td>
</tr>
<tr>
<td>F-score</td>
<td>60.9</td>
<td>74.2</td>
</tr>
</tbody>
</table>

Legend:
- **Accuracy**
- **F-score**
<table>
<thead>
<tr>
<th>Category</th>
<th>Variant</th>
<th>Weight</th>
<th>Category</th>
<th>Variant</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUNS</td>
<td>Singular</td>
<td>0.008</td>
<td>DECEPTIVE/IMAGINATIVE</td>
<td>Base</td>
<td>-0.057</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>0.002</td>
<td></td>
<td>Past tense</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>Proper, singular</td>
<td>-0.041</td>
<td></td>
<td>Present participle</td>
<td>-0.089</td>
</tr>
<tr>
<td></td>
<td>Proper, plural</td>
<td>0.091</td>
<td></td>
<td>Singular, present</td>
<td>-0.031</td>
</tr>
<tr>
<td>ADJECTIVES</td>
<td>General</td>
<td>0.002</td>
<td></td>
<td>Third person</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>Comparative</td>
<td>0.058</td>
<td></td>
<td>singular, present</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Superlative</td>
<td>-0.164</td>
<td></td>
<td>Modal</td>
<td>-0.063</td>
</tr>
<tr>
<td>PREPOSITIONS</td>
<td>General</td>
<td>0.064</td>
<td></td>
<td>General</td>
<td>0.001</td>
</tr>
<tr>
<td>DETERMINERS</td>
<td>General</td>
<td>0.009</td>
<td></td>
<td>Comparative</td>
<td>-0.035</td>
</tr>
<tr>
<td>COORD. CONJ.</td>
<td>General</td>
<td>0.094</td>
<td></td>
<td>Personal</td>
<td>-0.098</td>
</tr>
<tr>
<td>VERBS</td>
<td>Past participle</td>
<td>0.053</td>
<td></td>
<td>Possessive</td>
<td>-0.303</td>
</tr>
<tr>
<td>ADVERBS</td>
<td>Superlative</td>
<td>-0.094</td>
<td></td>
<td>Pre-determiners</td>
<td>0.017</td>
</tr>
</tbody>
</table>

*Informative* writing (left) --- nouns, adjectives, prepositions

*Imaginative* writing (right) --- verbs, adverbs, pronouns

Rayson et. al. (2001)
<table>
<thead>
<tr>
<th>TRUTHFUL/INFORMATIVE</th>
<th>WEIGHT</th>
<th>DECEPTIVE/IMAGINATIVE</th>
<th>WEIGHT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td><strong>Variant</strong></td>
<td><strong>Weight</strong></td>
<td><strong>Category</strong></td>
</tr>
<tr>
<td>NOUNS</td>
<td>Singular</td>
<td>0.008</td>
<td>Base</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>0.002</td>
<td>Past tense</td>
</tr>
<tr>
<td></td>
<td>Proper, singular</td>
<td>-0.041</td>
<td>Present participle</td>
</tr>
<tr>
<td></td>
<td>Proper, plural</td>
<td>0.017</td>
<td>Singular, present</td>
</tr>
<tr>
<td>ADJECTIVES</td>
<td>General</td>
<td>0.084</td>
<td>Third person</td>
</tr>
<tr>
<td></td>
<td>Comparative</td>
<td>0.058</td>
<td>singular, present</td>
</tr>
<tr>
<td></td>
<td>Superlative</td>
<td>-0.164</td>
<td>Modal</td>
</tr>
<tr>
<td>PREPOSITIONS</td>
<td>General</td>
<td>0.064</td>
<td></td>
</tr>
<tr>
<td>DETERMINERS</td>
<td>General</td>
<td>0.000</td>
<td>Comparative, present</td>
</tr>
<tr>
<td>COORD. CONJ.</td>
<td>General</td>
<td>0.093</td>
<td>Personal</td>
</tr>
<tr>
<td>VERBS</td>
<td>Past participle</td>
<td>0.053</td>
<td>Possessive</td>
</tr>
<tr>
<td>ADVERBS</td>
<td>Superlative</td>
<td>-0.094</td>
<td></td>
</tr>
</tbody>
</table>

*deceptive reviews -- superlatives, exaggerations*
<table>
<thead>
<tr>
<th>Category</th>
<th>Variant</th>
<th>Weight</th>
<th>Category</th>
<th>Variant</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUNS</td>
<td>Singular</td>
<td>0.008</td>
<td></td>
<td>Base</td>
<td>-0.057</td>
</tr>
<tr>
<td></td>
<td>Plural</td>
<td>0.002</td>
<td></td>
<td>Past tense</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>Proper, singular</td>
<td>-0.041</td>
<td></td>
<td>Present participle</td>
<td>-0.089</td>
</tr>
<tr>
<td></td>
<td>Proper, plural</td>
<td>0.091</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJECTIVES</td>
<td>General</td>
<td>0.002</td>
<td></td>
<td>Third person</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>Comparative</td>
<td>0.058</td>
<td></td>
<td>singular, present</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Superlative</td>
<td>-0.164</td>
<td></td>
<td>Modifying</td>
<td>0.008</td>
</tr>
<tr>
<td>PREPOSITIONS</td>
<td>General</td>
<td>0.064</td>
<td></td>
<td>Comparative</td>
<td>-0.035</td>
</tr>
<tr>
<td>DETERMINERS</td>
<td>General</td>
<td>0.009</td>
<td></td>
<td>Personal</td>
<td>-0.098</td>
</tr>
<tr>
<td>COORD. CONJ.</td>
<td>General</td>
<td>0.094</td>
<td></td>
<td>Possessive</td>
<td>-0.303</td>
</tr>
<tr>
<td>VERBS</td>
<td>Past participle</td>
<td>0.053</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADVERBS</td>
<td>Superlative</td>
<td>-0.094</td>
<td></td>
<td>Pre-determiners</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Deceptive reviews -- first person singular pronouns

- In contrast to “self-distancing” reported by previous psycholinguistics studies of deception (Newman et al., 2003)
- Deception cues are domain dependent
Classifier Performance

• Feature sets
  – POS (Part-of-Speech Tags)
    – Unigram, Bigram, Trigram

• Classifiers: SVM & Naïve Bayes
Classifier Performance

- **Linguistic Inquire and Word Count (LIWC)** (Pennebaker et al., 2001, 2007)
  - Widely popular tool for research in social science, psychology, etc
  - Counts instances of ~4,500 keywords
    - Regular expressions, actually
  - Keywords are divided into 80 dimensions across 4 broad groups
    - Linguistic processes, Psychological processes, Personal concerns, Spoken categories
Classifier Performance

- **Best Human Variant**
  - Accuracy: 61.9
  - F-score: 60.9

- **Classifier - POS**
  - Accuracy: 73
  - F-score: 74.2

- **Classifier - LIWC**
  - Accuracy: 76.8
  - F-score: 76.9
Classifier Performance

• Feature sets
  – POS (Part-of-Speech Tags)
  – Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007)

  Unigram, Bigram, Trigram

• Classifiers: SVM & Naïve Bayes
Classifier Performance

- **Best Human Variant**: 61.9 Accuracy, 60.9 F-score
- **Classifier - POS**: 73 Accuracy, 74.2 F-score
- **Classifier - LIWC**: 76.8 Accuracy, 76.9 F-score
- **Classifier - LIWC+Bigram**: 89.8 Accuracy, 89.8 F-score
## Classifier Performance

- Spatial difficulties (Vrij et al., 2009)
- Psychological distancing (Newman et al., 2003)

<table>
<thead>
<tr>
<th>LIWC+BIGRAMS</th>
<th>TRUTHFUL</th>
<th>DECEPTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>allpunctLIWC floor</td>
<td>experience</td>
<td>hilton</td>
</tr>
<tr>
<td>the_hotel</td>
<td>business</td>
<td>vacation</td>
</tr>
<tr>
<td>bathroom</td>
<td>i</td>
<td>spa</td>
</tr>
<tr>
<td>small</td>
<td>looking</td>
<td>while</td>
</tr>
<tr>
<td>helpful</td>
<td>hotel_.</td>
<td>husband</td>
</tr>
<tr>
<td>other</td>
<td>my_husband</td>
<td></td>
</tr>
</tbody>
</table>
Classifier Performance

- Spatial difficulties (Vrij et al., 2009)
- Psychological distancing (Newman et al., 2003)
### Classifier Performance

- **Spatial difficulties** (Vrij et al., 2009)
- **Psychological distancing** (Newman et al., 2003)

<table>
<thead>
<tr>
<th>LIWC+BIGRAMS</th>
<th>TRUTHFUL</th>
<th>DECEPTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>chicago</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>my</td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>hotel</td>
<td></td>
</tr>
<tr>
<td>location (</td>
<td>, _ and</td>
<td>luxury</td>
</tr>
<tr>
<td>)</td>
<td>experience</td>
<td></td>
</tr>
<tr>
<td>allpunctLIWC floor</td>
<td>experience</td>
<td>hilton</td>
</tr>
<tr>
<td>(</td>
<td>⭐business</td>
<td>⭐vacation</td>
</tr>
<tr>
<td>the_hotel bathroom</td>
<td>i</td>
<td>spa</td>
</tr>
<tr>
<td>small helpful $</td>
<td>looking</td>
<td>while</td>
</tr>
<tr>
<td>hotel other</td>
<td>⭐husband</td>
<td>⭐my_husband</td>
</tr>
</tbody>
</table>
## Classifier Performance

- **Spatial difficulties** (Vrij et al., 2009)
- **Psychological distancing** (Newman et al., 2003)

<table>
<thead>
<tr>
<th>LIWC+BIGRAMS</th>
<th>(\ast)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TRUTHFUL</strong></td>
<td><strong>DECEPTIVE</strong></td>
</tr>
<tr>
<td>-</td>
<td>chicago</td>
</tr>
<tr>
<td>...</td>
<td>my</td>
</tr>
<tr>
<td>on</td>
<td>hotel</td>
</tr>
<tr>
<td>location</td>
<td>, _ and</td>
</tr>
<tr>
<td>)</td>
<td>luxury</td>
</tr>
<tr>
<td>allpunctLIWC</td>
<td>experience</td>
</tr>
<tr>
<td>floor</td>
<td>hilton</td>
</tr>
<tr>
<td>(</td>
<td>business</td>
</tr>
<tr>
<td>the_hotel</td>
<td>vacation</td>
</tr>
<tr>
<td>bathroom</td>
<td>i</td>
</tr>
<tr>
<td>small</td>
<td>spa</td>
</tr>
<tr>
<td>helpful</td>
<td>looking</td>
</tr>
<tr>
<td>$</td>
<td>while</td>
</tr>
<tr>
<td>hotel_</td>
<td>husband</td>
</tr>
<tr>
<td>other</td>
<td>my__husband</td>
</tr>
</tbody>
</table>


Media Coverage

- ABC News
- New York Times
- Seattle Times
- Bloomberg / BusinessWeek
- NPR (National Public Radio)
- NHPR (New Hampshire Public Radio)
Conclusion (Case Study I)

- First large-scale gold-standard deception dataset
- Evaluated human deception detection performance
- Developed automated classifiers capable of nearly 90% accuracy
  - Relationship between deceptive and imaginative text
  - Importance of moving beyond universal deception cues
In this talk: three case studies of stylometric analysis

- Deceptive Product Reviews
- Wikipedia Vandalism
- The Gender of Authors
Wikipedia

• Community-based knowledge forums (collective intelligence)
• anybody can edit
• susceptible to vandalism --- 7% are vandal edits

• Vandalism – ill-intentioned edits to compromise the integrity of Wikipedia.
  – E.g., irrelevant obscenities, humor, or obvious nonsense.
Example of Vandalism
Example of Textual Vandalism

<Edit Title: *Harry Potter*>  
- Harry Potter is a teenage boy who likes to smoke crack with his buds. They also run an illegal smuggling business to their headmaster dumbledore. He is dumb!
Example of Textual Vandalism

<Edit Title: Harry Potter>

• Harry Potter is a teenage boy who likes to smoke crack with his buds. They also run an illegal smuggling business to their headmaster dumbledore. He is dumb!

<Edit Title: Global Warming>

• Another popular theory involving global warming is the concept that global warming is not caused by greenhouse gases. The theory is that Carlos Boozer is the one preventing the infrared heat from escaping the atmosphere. Therefore, the Golden State Warriors will win next season.
Vandalism Detection

• Challenge:
  – Wikipedia covers a wide range of topics (and so does vandalism)
    • vandalism detection based on topic categorization does not work.
  – Some vandalism edits are very tricky to detect
Previous Work I

Most work outside NLP

– Rule-based Robots:
  – e.g., Cluebot (Carter 2007)

– Machine-learning based:
  • features based on hand-picked rules, meta-data, and lexical cues
  • capitalization, misspellings, repetition, compressibility, vulgarism, sentiment, revision size etc

→ works for easier/obvious vandalism edits, but...
Some recent work started exploring NLP, but most based on shallow lexico-syntactic patterns

- Wang and McKeown (2010), Chin et al. (2010), Adler et al. (2011)
Vandalism Detection

• Our Hypothesis: textual vandalism constitutes a unique genre where **a group of people share a similar linguistic behavior**
Wikipedia Manual of Style

Extremely detailed prescription of style:

- **Formatting / Grammar Standards**
  - layout, lists, possessives, acronyms, plurals, punctuations, etc

- **Content Standards**
  - *Neutral point of view, No original research* (always include citation), *Verifiability*
  - “What Wikipedia is Not”: propaganda, opinion, scandal, promotion, advertising, hoaxes
Example of Textual Vandalism

<Edit Title: Harry Potter>

• Harry Potter is a teenage boy who likes to smoke crack with his buds. They also run an illegal smuggling business to their headmaster dumbledore. He is dumb!

<Edit Title: Global Warming>

• Another popular theory involving global warming is the concept that global warming is not caused by greenhouse gases. The theory is that Carlos Boozer is the one preventing the infrared heat from escaping the atmosphere. Therefore, the Golden State Warriors will win next season.
Language Model Classifier

• Wikipedia Language Model \((P_w)\)
  – trained on normal Wikipedia edits

• Vandalism Language Model \((P_v)\)
  – trained on vandalism edits

• Given a new edit \((x)\)
  – compute \(P_w(x)\) and \(P_v(x)\)
  – if \(P_w(x) < P_v(x)\), then edit ‘x’ is vandalism
Language Model Classifier

1. N-gram Language Models
   -- most popular choice

2. PCFG Language Models
   -- Chelba (1997), Raghavan et al. (2010),

\[
P(w_1^n) = \prod_{k=1}^{n} P(w_k | w_{k-1})
\]

\[
P(w_1^n) = \prod P(A \rightarrow \beta)
\]
Classifier Performance

Baseline

Baseline + ngram LM

Baseline + PCFG LM

Baseline + ngram LM + PCFG LM

F-Score

52.6
Classifier Performance

Baseline: 52.6
Baseline + ngram LM: 53.5

F-Score
Classifier Performance

Baseline
Baseline + ngram LM
Baseline + PCFG LM

F-Score

52.6
53.5
57.9
Classifier Performance

- Baseline: 52.6
- Baseline + ngram LM: 53.5
- Baseline + PCFG LM: 57.9
- Baseline + ngram LM + PCFG LM: 57.5
Classifier Performance

- Baseline: 91.6
- Baseline + ngram LM: 91.7
- Baseline + PCFG LM: 92.9
- Baseline + ngram LM + PCFG LM: 93

AUC
Vandalism Detected by PCFG LM

One day rodrigo was in the school and he saw a girl and she love her now and they are happy together.
# Ranking of features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of author contributions</td>
<td>0.106</td>
</tr>
<tr>
<td>How long the author has been registered</td>
<td>0.098</td>
</tr>
<tr>
<td>How frequently the author contributed in the training set</td>
<td>0.097</td>
</tr>
<tr>
<td>If the author is registered</td>
<td>0.0885</td>
</tr>
<tr>
<td>Difference in the maximum PCFG scores</td>
<td>0.0437</td>
</tr>
<tr>
<td>Difference in the mean PCFG scores</td>
<td>0.0377</td>
</tr>
<tr>
<td>How many times the article has been reverted</td>
<td>0.0372</td>
</tr>
<tr>
<td>Total contributions of author to Wikipedia</td>
<td>0.0343</td>
</tr>
<tr>
<td>Previous vandalism count of the article</td>
<td>0.0325</td>
</tr>
<tr>
<td>Difference in the sum of PCFG scores</td>
<td>0.0320</td>
</tr>
</tbody>
</table>
Conclusion (Case Study II)

- There are unique language styles in vandalism, and stylometric analysis can improve automatic vandalism detection.
- Deep syntactic patterns based on PCFGs can identify vandalism more effectively than shallow lexico-syntactic patterns based on n-gram language models.
In this talk: three case studies of 
stylometric analysis

- Deceptive Product Reviews
- Wikipedia Vandalism
- The Gender of Authors
“STEVE JOBS was an enemy of nostalgia. (......) One of the keys to Apple’s success under his leadership was his ability to see technology with an unsentimental eye and keen scalpel, ready to cut loose whatever might not be essential. This editorial mien was Mr. Jobs’s greatest gift — he created a sense of style in computing because he could edit.”
“More important, you worked with that little blinking cursor before you. No one in the world particularly cared if you wrote and, of course, you knew the computer didn’t care, either. But it was waiting for you to type something. It was not inert and passive, like the page. It was listening. It was your ally. It was your audience.”
“More important, you worked with that little blinking cursor before you. No one in the world of you wrote and, of course, computer didn’t care, either. But it it you to type something. It was like the page. It was

Gish Jen
a novelist
“STEVE JOBS was an enemy of nostalgia. (......) One of the keys to Apple’s success under his leadership was his ability to see technology with an unsentimental eye and keen scalpel, ready to cut loose whatever might not be essential. This was Mr. Jobs’s greatest gift — he created a sense of style in computing because he could edit.”

Mike Daisey
an author and performer
Motivations

Demographic characteristics of user-created web text

– New insight on social media analysis
– Tracking gender-specific styles in language over different domain and time
– Gender-specific opinion mining
– Gender-specific intelligence marketing
Women’s Language

Robin Lakoff (1973)

1. Hedges: “kind of”, “it seems to be”, etc.
3. Hyper-polite: “would you mind ...”, “I’d much appreciate if ...”
4. Apologetic: “I am very sorry, but I think...”
5. Tag questions: “you don’t mind, do you?”

...
Related Work

Sociolinguistic and Psychology
- Crosby and Nyquist (1977)
- Tannen (1991)
- Coates, Jennifer (1993)
- Holmes (1998)
- Argamon et al. (2003, 2007)
- McHugh and Hambaugh (2010)
Related Work

Machine Learning

– Koppel et al. (2002)
– Mukherjee and Liu (2010)
Concerns: Gender Bias in Topics

“Considerable gender bias in topics and genres”

– Herring and Paolillo (2006)
– Argamon et al. (2007)
We want to ask...

• Are there indeed gender-specific styles in language?

• If so, what kind of statistical patterns discriminate the gender of the author?
  – morphological patterns
  – shallow-syntactic patterns
  – deep-syntactic patterns
We want to ask...

• Can we trace gender-specific styles beyond topics and genres?
  – train in one domain and test in another
We want to ask...

• Can we trace gender-specific styles beyond topics and genres?
  – train in one domain and test in another
  – what about scientific papers?

*Gender specific language styles are not conspicuous in formal writing.*
Dataset

*Balanced topics to avoid gender bias in topics*

- **Blog Dataset**
  -- informal language

- **Scientific Dataset**
  -- formal language
Dataset

Balanced topics to avoid gender bias in topics

- Blog Dataset
  - informal language
  - 7 topics – education, entertainment, history, politics, etc.
  - 20 documents per topic and per gender
  - first 450 (+/- 20) words from each blog
Dataset

*Balanced topics to avoid gender bias in topics*

- Scientific Dataset
  - formal language
  - 5 female authors, 5 male authors
  - include multiple subtopics in NLP
  - 20 papers per author
  - first 450 (+/- 20) words from each paper
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
Balanced-Topic / Cross-Topic

I. balanced-topic

II. cross-topic
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic

- Scientific dataset
  3. balanced-topic
  4. cross-topic

- Both datasets
  5. cross-topic & cross-genre
Language Model Classifier

- Wikipedia Language Model ($P_w$)
  - trained on normal Wikipedia edits
- Vandalism Language Model ($P_v$)
  - trained on vandalism edits
- Given a new edit (x)
  - compute $P_w(x)$ and $P_v(x)$
  - if $P_w(x) < P_v(x)$, then edit ‘x’ is vandalism
Language Model Classifier

1. N-gram Language Models
   -- most popular choice

2. PCFG Language Models
   -- Chelba (1997), Raghavan et al. (2010),

\[
P(w_1^n) = \prod_{k=1}^{n} P(w_k | w_{k-1})
\]

\[
P(w_1^n) = \prod P(A \rightarrow \beta)
\]
Statistical Stylometric Analysis

1. Shallow Morphological Patterns
   ➔ Character-level Language Models (Char-LM)

2. Shallow Lexico-Syntactic Patterns
   ➔ Token-level Language Models (Token-LM)

3. Deep Syntactic Patterns
   ➔ Probabilistic Context Free Grammar (PCFG)
   – Chelba (1997), Raghavan et al. (2010),
Baseline

1. Gender Genie:
   http://bookblog.net/gender/genie.php

2. Gender Guesser
   http://www.genderguesser.com/
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre
Experiment I: balanced-topic, blog

Accuracy of Gender Attribution (%) -- overall

- Baseline: 50
- Char-LM: 71.3 (N = 2)
- Token-LM: 66.1
- PCFG: 64.1 (N = 2)
Experiment I: balanced-topic, blog

can detect gender even after removing bias in topics!
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre
Experiment II: cross-topic, blog

Accuracy of Gender Attribution (%) -- overall

- Baseline: 50
- Char-LM: 68.3 (N = 2)
- Token-LM: 61.5 (N = 2)
- PCFG: 59

Avg
Experiment II: cross-topic, blog

Accuracy of Gender Attribution (%) -- overall

- **Baseline**: 50
- **Char-LM**: 68.3 (N = 2)
- **Token-LM**: 61.5 (N = 2)
- **PCFG**: 59

---

can trace gender-specific styles even across topics!
Plan for the Experiments

• Blog dataset (7 different topics)
  I. balanced-topic
  II. cross-topic

• Scientific paper dataset (10 different authors)
  III. balanced-topic (balanced-author)
  IV. cross-topic (cross-author)

• Both datasets
  V. cross-topic & cross-genre
Experiment I & II: balanced-topic v.s. crossed-topic

Accuracy of Gender Attribution (%) -- overall

<table>
<thead>
<tr>
<th>Model</th>
<th>Baseline</th>
<th>Char-LM</th>
<th>Token-LM</th>
<th>PCFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>50</td>
<td>50</td>
<td>71.3</td>
<td></td>
</tr>
<tr>
<td>Char-LM</td>
<td></td>
<td>68.3</td>
<td>66.1</td>
<td>61.5</td>
</tr>
<tr>
<td>Token-LM</td>
<td></td>
<td></td>
<td></td>
<td>64.1</td>
</tr>
<tr>
<td>PCFG</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiment I & II: balanced-topic v.s. crossed-topic

Accuracy of Gender Attribution (%) -- overall

- Baseline: 50, 50
- Char-LM: 71.3, 68.3
- Token-LM: 66.1, 61.5
- PCFG: 64.1, 59

char-LM the most robust against topic change
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic

- Scientific dataset
  3. balanced-topic
  4. cross-topic

- Both datasets
  5. cross-topic & cross-genre
Experiment III: balanced-topic, scientific

Accuracy of Gender Attribution (%) -- overall

- Baseline: 47
- Char-LM: 91.5 (N = 3)
- Token-LM: 87 (N = 3)
- PCFG: 85
Experiment III: balanced-topic, scientific

Accuracy of Gender Attribution (%) -- overall

- **Baseline**: 47
- **Char-LM**: 91.5 (N = 3)
- **Token-LM**: 87 (N = 3)
- **PCFG**: 85

could be authorship attribution – upper bound
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic

- Scientific dataset
  3. balanced-topic
  4. cross-topic

- Both datasets
  5. cross-topic & cross-genre
Experiment IV: cross-topic, scientific

Accuracy of Gender Attribution (%): overall

- Baseline: 47% (N = 3)
- Char-LM: 76% (N = 3)
- Token-LM: 63.5% (N = 3)
- PCFG: 76%

[Bar chart showing accuracy levels for different models]
Experiment IV: cross-topic, scientific

can detect the gender of previously unseen authors!
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre
Experiment II & IV: cross-topic, scientific v.s. blog

Accuracy of Gender Attribution (%) -- overall

- Baseline: 47
- Char-LM: 91.5
- Token-LM: 87
- PCFG: 85

Baseline: 47
Char-LM: 91.5
Token-LM: 87
PCFG: 85

- Cross-topic: 47
- Cross-topic: 76
- Cross-topic: 63.5
- Cross-topic: 76
Experiment II & IV: cross-topic, scientific v.s. blog

1. PCFG most robust against topic change
2. token-level least robust against topic change
Plan for the Experiments

- Blog dataset
  1. balanced-topic
  2. cross-topic
- Scientific dataset
  3. balanced-topic
  4. cross-topic
- Both datasets
  5. cross-topic & cross-genre
Experiment V: cross-topic/genre, blog/scientific

Accuracy of Gender Attribution (%) -- overall

- Baseline: 47%
- Char-LM: 58.5%
- Token-LM: 51%
- PCFG: 47.5%
- BOW: 61.5%

Avg
Experiment V: cross-topic/genre, blog/scientific

Accuracy of Gender Attribution (%) -- overall

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>51</td>
</tr>
<tr>
<td>Char-LM</td>
<td>58.5</td>
</tr>
<tr>
<td>Token-LM</td>
<td></td>
</tr>
<tr>
<td>PCFG</td>
<td></td>
</tr>
<tr>
<td>BOW</td>
<td>61.5</td>
</tr>
</tbody>
</table>

weak signal of gender specific styles beyond topic & genre
Conclusions (Case Study III)

• comparative study of machine learning techniques for gender attribution consciously removing gender bias in topics.

• statistical evidence of gender-specific language styles beyond topics and genres.
Collaborators

• @ Stony Brook University:
  Kailash Gajulapalli, Manoj Harpalani, Rob Johnson, Michael Hart, Ruchita Sarawgi, Sandesh Singh

• @ Cornell University:
  Claire Cardie, Jeffrey Hancock, Myle Ott

• Based on
  – Ott et al., 2011 (ACL)
  – Harpalani et al., 2011 (ACL)
  – Sarawgi et al., 2011 (CoNLL)
THANK YOU!