Detection of Stereotypes, Bias, and Prejudice in Text

Anjalie Field
Online data is riddled with SOCIAL STEREOTYPES
Social media platforms contain hate speech, offensive language, and microaggressions

Newspaper articles, online encyclopedias, movie scripts, etc. perpetuate stereotypes

Field et al. (2019) Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories. ICWSM

Sap et al. (2017) Connotation Frames of Power and Agency in Modern Films. EMNLP
Pitfalls and Opportunities for NLP Research

- NLP models can absorb and amplify stereotypes and prejudice from data
“Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.”


Birhane, Abeba and Vinay Uday Prabhu. (2021) **Large Image Datasets: A Pyrrhic Win for Computer Vision?**. WACV

Bender, Emily M., et al. (2021) **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?**. FAccT
Pitfalls and Opportunities for NLP Research

- NLP models have the potential to identify and mitigate bias, stereotypes, and prejudice in text
- Reactive approaches:
  - Content moderation (hate speech, offensive language)
- Proactive approaches:
  - Detection of implicit biases, microaggressions
  - Training data audits, selection, and modifications (contrast post-hoc model de-biasing)
- Technical Challenges
  - Difficult to collect annotated data
  - Models learn correlations which may not be indicative of bias
This talk

- Dialog (Social Media): Detecting subtle gender bias in social media using methodology from causal inference
  - Field and Tsvetkov (2020) Unsupervised Discovery of Implicit Gender Bias. *EMNLP*

- Narratives (Wikipedia and Newspaper articles): Identifying disparities in how people of different social groups are portrayed
  - Field et al. (2019) Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories. *ICWSM*
  - Field and Tsvetkov. (2019) Entity-Centric Contextual Affective Analysis. *ACL*
  - Field, Park, and Tsvetkov. An Algorithm for Controlled Text Analysis on Wikipedia. *Ongoing*
Biased comments are prevalent online

Our goal: detect subtle gender biases like microaggressions, objectifications, and condescension in 2\textsuperscript{nd}-person text

- “Oh, you work at an office? I bet you’re a secretary”
- “Total tangent I know, but you’re gorgeous”

Current classifiers that detect hate speech, offensive language, or negative sentiment cannot detect these comments
Automated detection can improve civility

Our goal: detect subtle gender biases like microaggressions, objectifications, and condescension in 2nd-person text

- “Oh, you work at an office? I bet you’re a secretary”
- “Total tangent I know, but you’re gorgeous”

Why does automated detection help?
- Posters are often unaware that their comments contain bias -- if they were, they may choose not to post them (proactive response)
- Users can choose not to read flagged comments (reactive response)
- Remove comments from training data (Han and Tsvetkov, EMNLP 2020)
Naive Approach 1: Supervised Classification

- Bro, golf is better
- Isn't it too cold to play in Canada?
- Me too <3
- UR hot!

Supervised Classifier
Naive Approach 1: Supervised Classification

Problem: Biases are *subtle* and *implicit* even experts are bad at identifying them
Naive Approach 2: Comments contain bias if they are highly predictive of gender

- Train a classifier that predicts the gender of the person the text is addressed to
- If the classifier makes a prediction with high confidence, the text likely contains bias

<table>
<thead>
<tr>
<th>W_GENDER: M</th>
<th>W_GENDER: F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bro, golf is better</td>
<td>Me too &lt;3</td>
</tr>
<tr>
<td>Isn’t it too cold to play in Canada?</td>
<td>UR hot!</td>
</tr>
</tbody>
</table>
Naive Approach 2: Comments contain bias if they are highly predictive of gender

- Problem: Text may contain *confounds* that are predictive of gender but not indicative of bias
Proposed Model: Comments contain bias if they are highly predictive of gender despite confound control

- Balance observed confounders through propensity matching
- Demote latent confounders through adversarial training
- Substitute overt indicators

OW: Person 1
W_GENDER: M
W_TRAIT: Canadian

Tennis is great!

OW: Person 2
W_GENDER: F
W_TRAIT: ?

I love tennis!

Me too <3

Do I look ok?

UR hot!

Isn’t it too cold to play in Canada?

Bre <title>, golf is better

UR hot!
Propensity matching for *observed* confounding variables

- Language in comments may be caused by preceding context and not gender of the addressee
Propensity matching for *observed* confounding variables

- Balance the data set so that comments addressed to men have a similar distribution of confounding variables as comments addressed to women
  - Match posts with similar indicators of confounding variables
  - Discard posts that are unable to be matched
Propensity matching for *observed* confounding variables
Adversarial training for *latent* confounding variables

- Comments may reference traits of the addressee (such as occupation, nationality, nicknames, etc.) other than gender
- Difficult to enumerate all of them
- Often unique to individuals (difficult to make matches)

---

OW: Person 1
\[W\_GENDER: M\]
\[W\_TRAIT: Canadian\]

Tennis is great!

Bro, golf is better

Isn't it too cold to play in **Canada**?

OW: Person 2
\[W\_GENDER: F\]
\[W\_TRAIT: ?\]

I love tennis!

Do I look ok?

Me too <3

**UR hot!**

**UR hot!**

OW:

COM.TXT

O.TXT
Represent latent confounding variables as a vector

- For each *addressee* in the training corpus, for each word type *w* in training comments, compute \( \text{association}(\text{addressee}, w) \rightarrow p(w | k) \)
- For each *comment* in the training corpus, for each training *addressee* (*k*), define:

\[
p(\text{addressee} = k | \text{comment}) \propto p(\text{addressee} = k) \prod_{w \in \text{comment}} p(w | k)
\]

\( k = 1 \ldots k = n \)
Adversarial training demotes latent vector

- Traits are inferred from comments using log-odds scores and represented in a vector. GAN-like training procedure discourages the model from learning them

Evaluation: Performance improvement on held-out data

<table>
<thead>
<tr>
<th></th>
<th>Public Figures</th>
<th></th>
<th></th>
<th>Politicians</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>Acc.</td>
<td>F1</td>
<td>Acc.</td>
<td></td>
</tr>
<tr>
<td>base</td>
<td>74.9</td>
<td>63.8</td>
<td>23.2</td>
<td>73.2</td>
<td></td>
</tr>
<tr>
<td>+demotion</td>
<td><strong>76.1</strong></td>
<td><strong>65.1</strong></td>
<td>17.4</td>
<td><strong>77.1</strong></td>
<td></td>
</tr>
<tr>
<td>+match</td>
<td>65.4</td>
<td>56.0</td>
<td>28.5</td>
<td>46.7</td>
<td></td>
</tr>
<tr>
<td>+match+demotion</td>
<td>68.2</td>
<td>59.7</td>
<td><strong>28.8</strong></td>
<td>51.4</td>
<td></td>
</tr>
</tbody>
</table>
Word substitutions for overt signals

- Remove overtly gendered terms (Mrs. Ms. Mr., etc.) using keyword substitution

```
I love tennis!
Tennis is great!
Do I look ok?
Bro, golf is better
Isn’t it too cold to play in Canada?

Person 1
W_GENDER: M
W_TRAIT: Canadian

Person 2
W_GENDER: F
W_TRAIT: ?

Me too <3
UR hot!
```
### Evaluation: detection of gender-based microaggressions

<table>
<thead>
<tr>
<th>Model Description</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model (Trained on Public Figures)</td>
<td><strong>51.0</strong></td>
<td>50.7</td>
<td>50.9</td>
<td><strong>57.0</strong></td>
</tr>
<tr>
<td>Our model (Trained on Politicians)</td>
<td>45.7</td>
<td><strong>75.3</strong></td>
<td>56.9</td>
<td>49.9</td>
</tr>
<tr>
<td>Random baseline</td>
<td>43.5</td>
<td>48.7</td>
<td>46.0</td>
<td>49.8</td>
</tr>
</tbody>
</table>

*“You're pretty for a black girl.”*
Findings: characteristics of bias against women Politicians

Influential words:
- Competence and domesticity

Examples:
- “DINO I hope another real Democrat challenges you next election”
- “I did not vote for you and have no clue why anyone should have. You do not belong in politics”
Findings: characteristics of bias against women

Influential words:
- Appearance and sexualization

Examples:
- “Total tangent I know but, you’re gorgeous.”
- “I like Bob, but you’re hot, so kick his butt.”
Conclusions and future work

Unsupervised approach to detecting implicit biases that are difficult to annotate
- Bias differs in different domains (Politicians vs. Public Figures)
- Scope for technical improvement

Extension to other domains:
- Comments on newspaper articles, reviews of essays/job applications

Extension to other attributes:
- “Persona vectors”, reduce dependency on gender labels
This talk

- **Dialog (Social Media):** Detecting subtle gender bias in social media using methodology from causal inference
  - Field and Tsvetkov (2020) *Unsupervised Discovery of Implicit Gender Bias*. *EMNLP*

- **Narratives (Wikipedia and Newspaper articles):** Identifying disparities in how people of different social groups are portrayed
  - Field et al. (2019) *Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories*. *ICWSM*
  - Field and Tsvetkov. (2019) *Entity-Centric Contextual Affective Analysis*. *ACL*
  - Field, Park, and Tsvetkov. *An Algorithm for Controlled Text Analysis on Wikipedia*. *Ongoing*
How do we measure people portrayals?

Sentiment: is this person portrayed positively or negatively?

“As first lady, [Michelle] Obama served as a role model for women and worked as an advocate for poverty awareness, education, nutrition, physical activity, and healthy eating.”

“She has said she was overwhelmed during her first year, attributing this to the fact that neither of her parents had graduated from college, and that she had never spent time on a college campus.”
Measure of affective meaning: valence, arousal, dominance

Three most important, largely independent, dimensions of affective meaning are:

- **Sentiment** Valence (positiveness–negativeness/pleasure–displeasure)
- **Agency** Arousal (active–passive)
- **Power** Dominance (dominant–submissive)

Form the basis of *affective control theory*

Mohammad, Saif. (2018) *Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words*. ACL

Connotation Frames (Rashkin 2016)

\[ X \text{ pushed } Y \]

How the writer feels about \( X \):
Positive  Either Positive or Neutral  Neutral  Either Negative or Neutral  Negative

Who has higher authority and is more powerful?
The subject \( X \) has more authority and is more powerful
\( Y / X \) have similar authority and power (the difference is unclear)
The object \( Y \) has more authority and is more powerful

Sap, Maarten, et al. (2017) *Connotation frames of power and agency in modern films*. EMNLP
How can we improve off-the-shelf lexicons?

- Each verb has a single annotation for each dimension, but verbs have different connotations in different contexts

  The hero deserves appellation
  The boy deserves punishment
Generating Contextualized Lexicons

Corpus A

Extract ELMo embeddings

De-contextualized embeddings

Supervised Classifier

Corpus B

Extract ELMo embeddings

Contextualized Connotation Frames

Connotation Frames

Language Technologies Institute
Evaluation of Contextualization

<table>
<thead>
<tr>
<th></th>
<th>Verb-level</th>
<th>Sent.-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment(theme)</td>
<td>41.05</td>
<td>44.35</td>
</tr>
<tr>
<td>Sentiment(agent)</td>
<td>51.37</td>
<td>52.80</td>
</tr>
</tbody>
</table>

_F1 scores over sentence-level annotations_
# Evaluation of Entity Scoring

<table>
<thead>
<tr>
<th>Off-the-shelf</th>
<th>Frequency</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.1</td>
<td>59.1</td>
<td>71.4</td>
</tr>
</tbody>
</table>

**Accuracy, compared to hand-annotations**

Task: Given a pair of entities mentioned in the same set of newspaper articles, choose which entity is more powerful
Analysis of articles about #MeToo

- The #MeToo movement has largely been viewed as “empowering” but journalists have a choice in how they portray people (victim vs. survivor)

- Tarana Burke described her goal in founding the #MeToo Movement as: “empowerment through empathy”
Analysis of articles about #MeToo

- 27,602 articles across 1,576 outlets containing the keyword #metoo
  - November 2, 2017 - January 31, 2018
  - February 28, 2018 - May 29, 2018

- Caveats:
  - Sample of articles may not be representative
  - Model performance is imperfect
Corpus Level: Who are the most powerful, sympathetic, and high agency people?

**Most Positive:** Kara Swisher, Tarana Burke, Meghan Markle, Frances McDormand, Oprah Winfrey

**Most Negative:** Bill Cosby, Harvey Weinstein, Eric Schneiderman, Kevin Spacey, Ryan Seacrest

**Highest Power:** the #MeToo movement, Judge Steven O’Neill, The New York Times, Congress, Facebook Twitter, Eric Schneiderman, Donald Trump

**Lowest Power:** Kevin Spacey, Andrea Constand, Uma Thurman, Dylan Farrow, Leeann Tweeden

**Highest Agency:** Judge Steven O’Neill, Eric Schneiderman, Russell Simmons, The New York Times, Frances McDormand, CNN, Donald Trump, Hillary Clinton

**Lowest Agency:** Kara Swisher, the United States, Hollywood, Meryl Streep
Cross-outlet comparison: Sentiment scores for Brett Kavanaugh Nomination

![Sentiment score comparison chart](image-url)
Limitations

- Do contextualized embeddings really add enough context?

<table>
<thead>
<tr>
<th></th>
<th>Verb-level</th>
<th>Sent.-level</th>
<th>Contextualized Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment(theme)</td>
<td>41.05</td>
<td>44.35</td>
<td>50.16</td>
</tr>
<tr>
<td>Sentiment(agent)</td>
<td>51.37</td>
<td>52.80</td>
<td>54.11</td>
</tr>
</tbody>
</table>

F1 scores over sentence-level annotations
Limitations

- Do contextualized embeddings really add enough context?

<table>
<thead>
<tr>
<th></th>
<th>Verb-level</th>
<th>Sent.-level</th>
<th>Contextualized Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment(theme)</td>
<td>41.05</td>
<td>44.35</td>
<td>50.16</td>
</tr>
<tr>
<td>Sentiment(agent)</td>
<td>51.37</td>
<td>52.80</td>
<td>54.11</td>
</tr>
</tbody>
</table>

F1 scores over sentence-level annotations

- Original connotation frames are only in English
- Lots of work in NLP on binary gender and very little on other social dimensions
Multilingual Connotation Frames

- **English, Russian, and Spanish**
  - Facilitates within-language analysis and across language analysis
  - Challenges: translating instructions, identifying native language speakers (Chinese, French)

- **Collect annotations in complete contexts drawn from newspaper articles**
  - \( X \) rescues \( Y \)
  - The firefighter rescues the boy

- **Analyze connotations in Wikipedia articles about LGBT people**

---

Generating Contextualized Lexicons

Corpus A

Extract ELMo embeddings

De-contextualized embeddings

Supervised Classifier

Corpus B

Extract ELMo embeddings

Contextualized Connotation Frames

Connotation Frames

[AGENT] X

[THEME] Y

pushed

PERSPECTIVE (writer → agent) [neutral]

PERSPECTIVE (writer → theme) [neutral]

Agency

Power

Power
Generating Contextualized Lexicons

Corpus B

XLM
Extract ELMo embeddings

Supervised Classifier

XLM
Extract ELMo embeddings

Multilingual Contextualized Connotation Frames

Contextualized Connotation Frames
Wikipedia page for Alan Turing in English, Russian, and Spanish

English Wikipedia:
He *accepted* the option of injections of what was then called stilboestrol.

Spanish Wikipedia:
Finalmente escogió las inyecciones de estrógenos.
*Finally he chose estrogen injections.*

Russian Wikipedia:
Учёный предпочёл инъекции стильтэстрола
*The scientist preferred stilbestrol injections.*
LGBTBio Corpus

- 1,340 Wikipedia articles about LGBT people
- 1,340 articles about non-LGBT people with similar characteristics
- Do Wikipedia language editions reflect cultural biases about LGBT people?

WIKIPEDIA
The Free Encyclopedia

WIKIPEDIA
La enciclopedia libre

ВИКИПЕДИЯ
Свободная энциклопедия
Difference in verb connotations for LGBT people and non-LGBT people

![Bar chart showing difference scores for sentiment, power, and agency across different languages.](chart.png)
Conclusions and Ongoing Work

- Online text contains microaggressions, stereotypes, and content disparities
- Detecting these phenomena can improve data quality, support pro-active mitigation approaches, and lead to more equitable NLP models

Facebook

The New York Times

FOX News

Wikipedia
Conclusions and Ongoing Work

- Online text contains microaggressions, stereotypes, and content disparities
- Detecting these phenomena can improve data quality, support pro-active mitigation approaches, and lead to more equitable NLP models
- Significant technical challenges: What NLP methodology do we need to accomplish this?
  - Reducing the influence of confounding variables on classification and analyses
  - Exploring dimensions of meaning other than sentiment or toxicity
  - Moving beyond binary gender
This talk

- Dialog (Social Media): Detecting subtle gender bias in social media using methodology from causal inference
  - Field and Tsvetkov. (2020) Unsupervised Discovery of Implicit Gender Bias. *EMNLP*
- Narratives (Wikipedia and Newspaper articles): Identifying disparities in how people of different social groups are portrayed
  - Field et al. (2019) Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories. *ICWSM*
  - Field and Tsvetkov. (2019) Entity-Centric Contextual Affective Analysis. *ACL*
  - Field, Park, and Tsvetkov. An Algorithm for Controlled Text Analysis on Wikipedia. *Ongoing*