# Quantifying Societal Biases Towards Entities

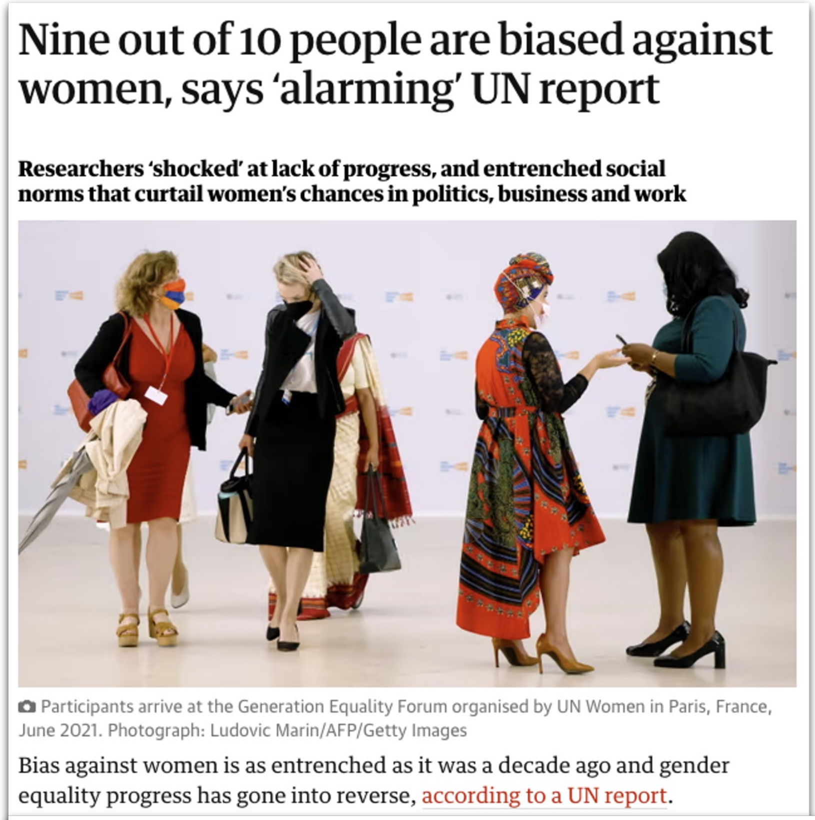
**Isabelle Augenstein**

Gender Bias in NLP Workshop @ ACL 2024
16 August 2024

COPENLU
NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP

UNIVERSITY OF COPENHAGEN

# Overview: Gender Bias Research in Natural Language Processing

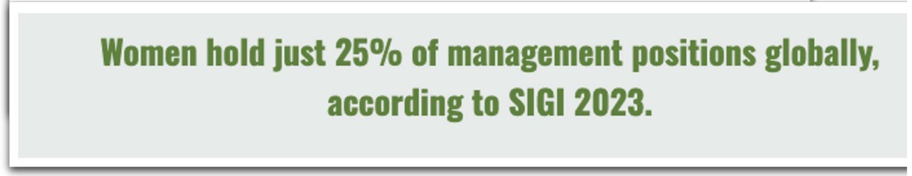


## Nine out of 10 people are biased against women, says 'alarming' UN report

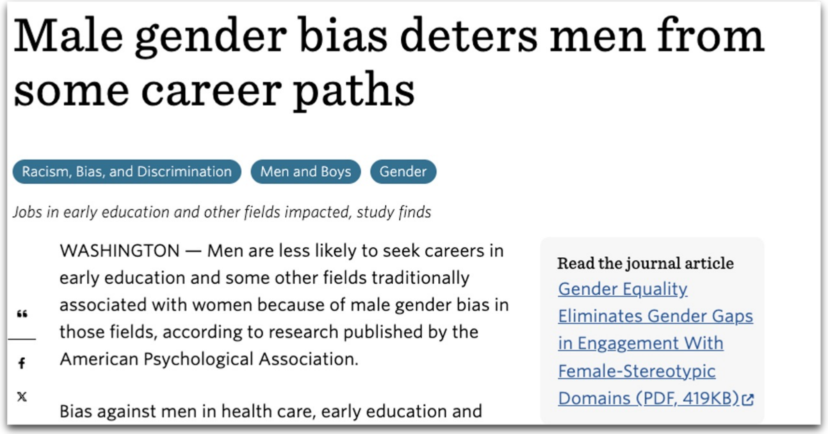
**Researchers 'shocked' at lack of progress, and entrenched social norms that curtail women's chances in politics, business and work**

Participants arrive at the Generation Equality Forum organised by UN Women in Paris, France, June 2021. Photograph: Ludovic Marin/AFP/Getty Images

Bias against women is as entrenched as it was a decade ago and gender equality progress has gone into reverse, according to a UN report.



**Women hold just 25% of management positions globally, according to SIGI 2023.**



## Male gender bias deters men from some career paths

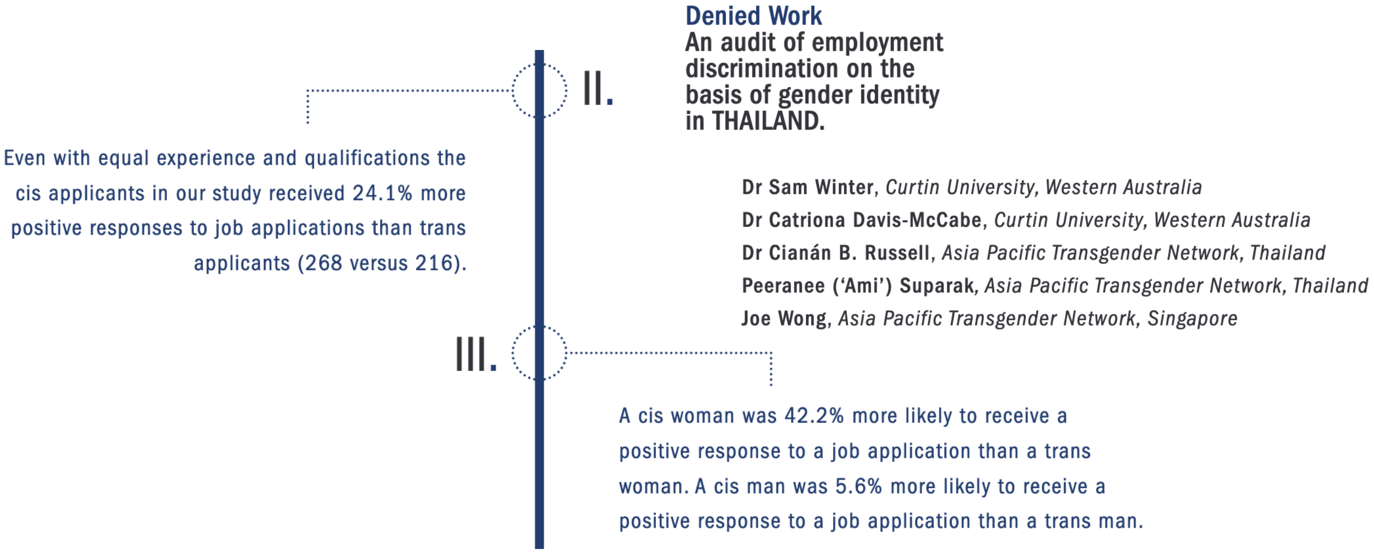
Racism, Bias, and Discrimination | Men and Boys | Gender

*Jobs in early education and other fields impacted, study finds*

WASHINGTON — Men are less likely to seek careers in early education and some other fields traditionally associated with women because of male gender bias in those fields, according to research published by the American Psychological Association.

Bias against men in health care, early education and

Read the journal article
Gender Equality Eliminates Gender Gaps in Engagement With Female-Stereotypic Domains (PDF, 419KB)



**Denied Work**
**An audit of employment discrimination on the basis of gender identity in THAILAND.**

Dr Sam Winter, *Curtin University, Western Australia*
Dr Catriona Davis-McCabe, *Curtin University, Western Australia*
Dr Cianán B. Russell, *Asia Pacific Transgender Network, Thailand*
Peeranee ('Ami') Suparak, *Asia Pacific Transgender Network, Thailand*
Joe Wong, *Asia Pacific Transgender Network, Singapore*

II.

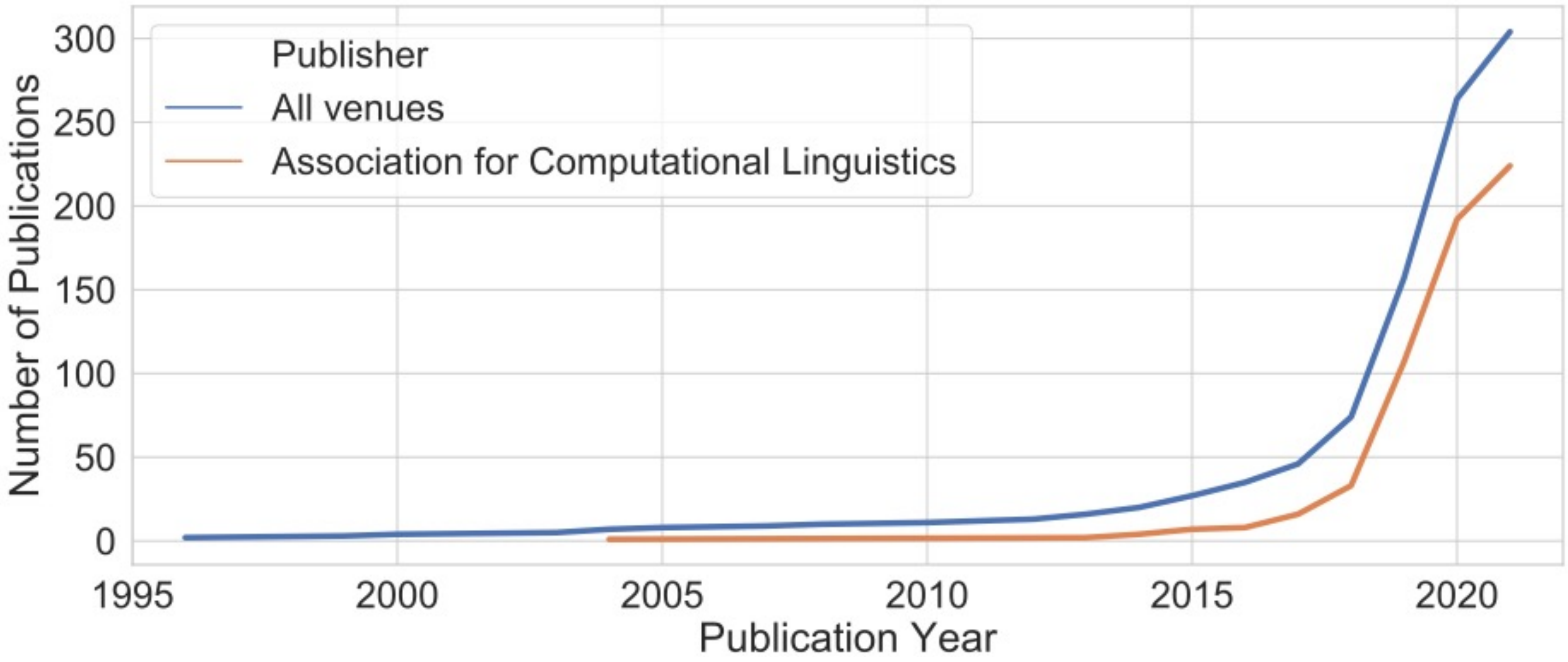
Even with equal experience and qualifications the cis applicants in our study received 24.1% more positive responses to job applications than trans applicants (268 versus 216).

III.

A cis woman was 42.2% more likely to receive a positive response to a job application than a trans woman. A cis man was 5.6% more likely to receive a positive response to a job application than a trans man.

Sources: https://www.theguardian.com/global-development/2023/jun/12/nine-out-of-10-people-are-biased-against-women-says-alarming-un-report ; https://www.oecd.org/stories/gender/social-norms-and-gender-discrimination/ ; https://www.apa.org/news/press/releases/2022/12/male-gender-bias-career-paths ; https://www.weareaptn.org/wp-content/uploads/2019/11/APTN-DeniedWork-Thailand.pdf

# Overview: Gender Bias Research in Natural Language Processing



Karolina Stańczak, **Isabelle Augenstein**. A Survey on Gender Bias in Natural Language Processing. CoRR, abs/2112.14168, December 2021.

# Overview: Gender Bias Research in Natural Language Processing

- What is gender?
  - **Grammatical gender** of nouns -- in simplest case, *masculine* vs *feminine*)
  - **Referential gender**: gender that is expressed, inferred and used by a perceiver to classify a referent – often *masculine*, *feminine*, *neuter*
  - **Lexical gender**: existence of lexical units carrying the property of gender, male- or female-specific words such as *father* and *waitress*
  - (Bio-)**social gender**: imposition of gender roles or traits based on phenotype, social and cultural norms, gender expression, and identity (such as *gender roles*)

Karolina Stańczak, **Isabelle Augenstein**. A Survey on Gender Bias in Natural Language Processing. CoRR, abs/2112.14168, December 2021.

# Overview: Gender Bias Research in Natural Language Processing

- Emergence of gender bias in text
  - **Overall**: use of *words or syntactic constructs* that connote or imply an inclination or prejudice against one gender
  - **Structural bias**: arises when the *construction of sentences* shows patterns closely tied to the presence of gender bias
    - Gender generalisation (i.e., when a gender-neutral term is assumed to refer to a specific gender based on some (stereotypical) assumptions)
      - ”A programmer must always carry his laptop with him.”
    - Explicit labelling of sex
      - ” There’s a reason the airplane stewardess asks you to put on your own oxygen mask before trying to help anyone else with theirs.”

Karolina Stańczak, **Isabelle Augenstein**. A Survey on Gender Bias in Natural Language Processing. CoRR, abs/2112.14168, December 2021.

# Overview: Gender Bias Research in Natural Language Processing

- Emergence of gender bias in text
  - **Contextual bias**: tone, words used, or context of a sentence. Cannot be observed through grammatical structure but *requires contextual background* information and human perception.
    - Societal stereotypes (which showcase traditional gender roles that reflect social norms)
      - ”The event was kid-friendly for all mothers working in the company.”
    - Behavioural stereotypes (attributes and traits used to describe a specific person or gender)
      - ”Mary must love dolls because all girls like playing with them.”

Karolina Stańczak, **Isabelle Augenstein**. A Survey on Gender Bias in Natural Language Processing. CoRR, abs/2112.14168, December 2021.

# Overview: Gender Bias Research in Natural Language Processing

- **Why** is it important to study gender bias detection?
  - **Social science** research mainly on smaller phenomena
  - New research possible using computational methods
    - Detection can be scaled up using NLP methods
    - Different types of gender bias can be quantified
    - New research questions can be answered

Karolina Stańczak, **Isabelle Augenstein**. A Survey on Gender Bias in Natural Language Processing. CoRR, abs/2112.14168, December 2021.

# Overview: Gender Bias Research in Natural Language Processing

- **Why** is it important to study gender bias detection?
  - Implications on downstream usage
    - **Gender gap**: Available texts *mainly discuss and quote men*, which leads to biased corpora for training NLP models including LLMs
    - **Representation harm**: *Associations* between gender with certain concepts are captured in representations of words/sentences and model parameters
    - **Allocation harm**: Models often perform better on data associated with the *majority gender*
  - Biases in text -> biases in LLMs -> biases in downstream applications -> reinforcement of human biases

Karolina Stańczak, **Isabelle Augenstein**. A Survey on Gender Bias in Natural Language Processing. CoRR, abs/2112.14168, December 2021.

# Quantifying Societal Biases Towards Entities

- **Overview**: Gender Bias in NLP (Preprint, 2022)
- Detecting Gender Biases in **Text**
  - Correlations between adjective/verb choice and noun gender (ACL 2019)
  - Towards politicians on Reddit (PLoS One, October 2023) and
    on Twitter / X (Preprint, 2023)
  - Measuring intersectional biases in historical documents (ACL 2023)
- Quantifying Societal Biases In **Language Models**
  - Gender bias in multilingual language models, politicians (PLoS One, June 2023)
  - Multiple social biases in language models, any entity (Preprint, 2023)

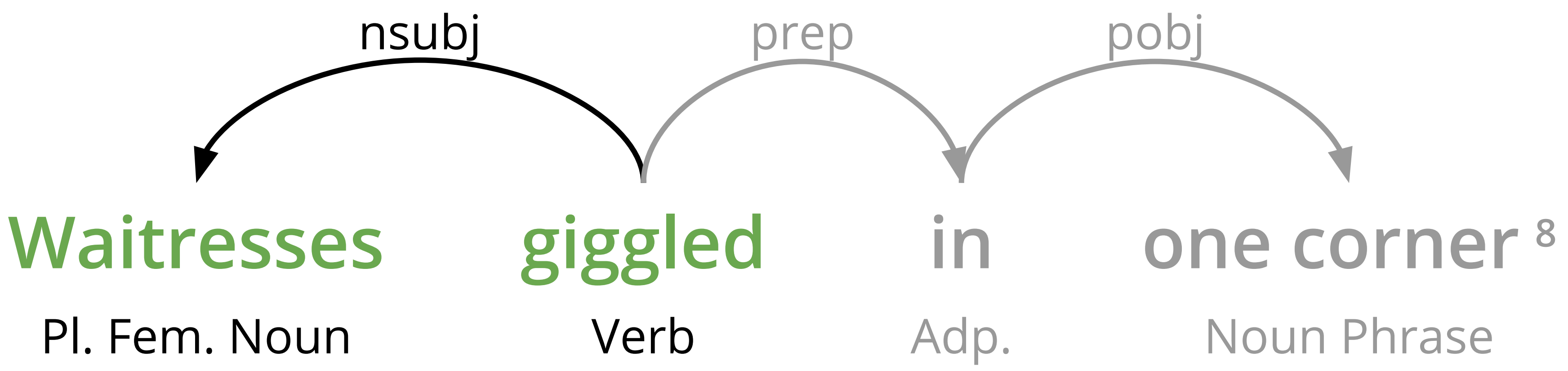
# Quantifying Societal Biases Towards Entities

| | Text or LM | Bias Targets | Domain | Languages | Bias Types |
|---|---|---|---|---|---|
| Hoyle et al. 2019 | Text | Common nouns | Books | English | Gender |
| Marjanovic et al. 2022 | Text | Politicians | Reddit | English | Gender |
| Golovchenko et al. 2023 | Text | Politicians | Twitter / X | 65 languages | Gender |
| Borenstein et al. 2023 | Text | People | Historical newspapers from colonial period | English | Gender<br>Race |
| | | | | | |
| Stańczak et al. 2023 | LM | Politicians | Multi-domain | 7 languages | Gender |
| Manerba et al. 2023 | LM | Societal groups | Multi-domain | English | Gender<br>Religion<br>Disability<br>Nationality |

# Detecting Gender Biases in Text

# Measuring Gendered Language In Text

## Measure differences in syntactic collocations



nsubj prep pobj

Waitresses giggled in one corner [8]

Pl. Fem. Noun Verb Adp. Noun Phrase

[8] *Paraphrase of* Orczy, B. 1908. The Old Man in the Corner.

Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

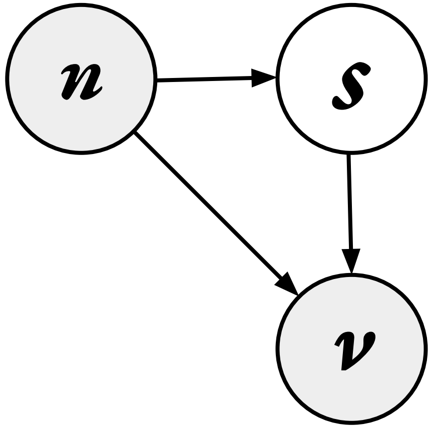
# Measuring Gendered Language In Text

## Model: a joint representation of nouns, adjectives or verbs, and sentiment

$$p(\nu, n, s) = p(\nu \mid n, s)\, p(s \mid n)\, p(n)$$

Corpus is that of Goldberg and Orwant (2013)

- ~3.5 million books
- ~11 billion words
- Years 1900-2008



Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

# Measuring Gendered Language In Text

## Components: a noun vector of lexical features

$$p(\nu, n, s) = \underline{p(\nu \mid n, s)\, p(s \mid n)}\, p(n)$$

$n \in \mathcal{G}$

$f_n \in \{0, 1\}^T$

Waitresses → [ WAITER, FEM, PL ] → [ ..., *1*, *1* ]

Waiter → [ WAITER, MASC, S ] → [ ..., *0*, *0* ]

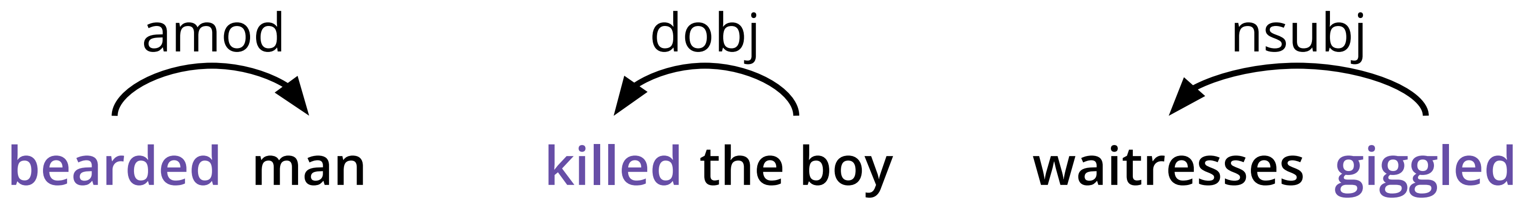
Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

# Measuring Gendered Language In Text

## Components: neighbors and categorical sentiment

$$p(v, n, s) = \underline{p(v \mid n, s)\, p(s \mid n)}\, p(n)$$

$v \in \mathcal{V}$



$s \in \mathcal{S} = \{\mathrm{POS}, \mathrm{NEG}, \mathrm{NEU}\}$

Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

# Measuring Gendered Language In Text

Overall results: stark gender differences that align with human intuition

| 👨 | | 👩 |
|---|---|---|
| Hostile<br>Violent<br>Abusive<br>Brutal | amod | Helpless<br>Disagreeable<br>Unmarried<br>Widowed |
| Flourish<br>Kill | nsubj | Giggle<br>Gossip |
| Praise<br>Kill | dobj | Eye<br>Woo |

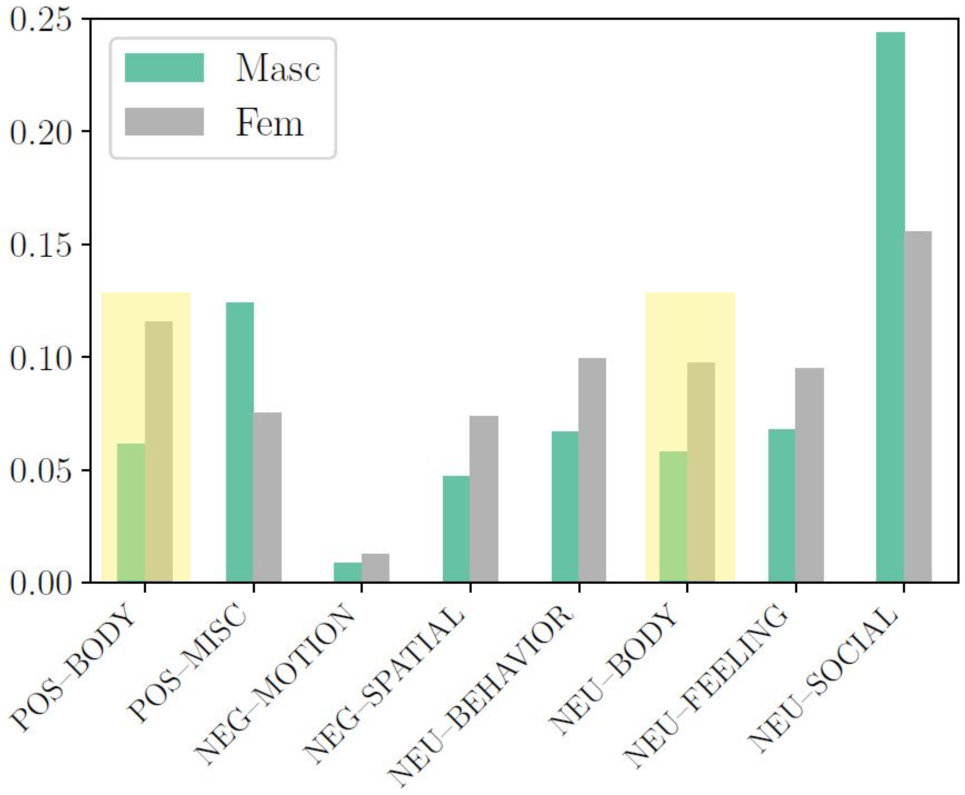
Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

# Measuring Gendered Language In Text

## Female bodies receive disproportionate attention

**"Cute"**[9]

| | |
|---|---|
| BODY | 0.78 |
| FEELING | 0.05 |
| BEHAVIOR | 0.04 |
| SUBSTANCE | 0.03 |
| SOCIAL | 0.02 |

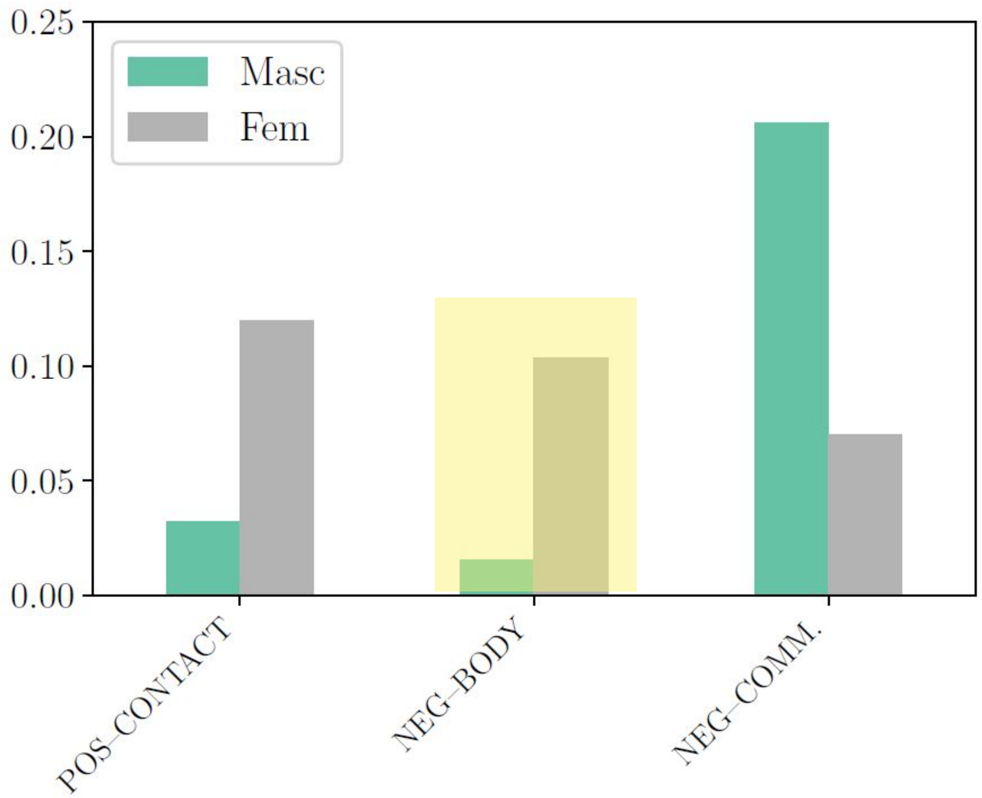


[9] Tsvetkov et al, 2014

Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

# Measuring Gendered Language In Text

## “BODY” also a more likely NSUBJ verb category



Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

# Measuring Gendered Language In Text

## Caveats

Ignore speaker & source (e.g., fiction or nonfiction)

Language changes over time, in particular that relating to gender[11]

Reporting bias (“Black sheep”[12])

Limited to binary gender

Not linked to entities

[11] Underwood et al. (2018)

[12] Meg Mitchell

Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

# Detecting Gender Biases Towards Politicians on Social Media

- Goal: measuring real-world gender biases on social media
  - Quantifying political gender gap on **Reddit**
    - Linguistic as well as extra-linguistic cues
    - Nuanced analysis of gender bias
  - Measuring scope and nature of gender bias on **Twitter** towards ambassadors
    - Systematic, global analysis
    - Large-scale massively cross-lingual study

Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

# Detecting Gender Biases Towards Politicians on Social Media

**Bias Measures:**

- There is less visibility of women in the media **(Coverage biases)**
  - To obtain a Wikipedia page, women must be more notable
  - Articles are shorter and less edited

- Women are peripheral figures in a core network of men **(Combinatorial biases)**
  - Male pages are more central
  - Women are described by their male relations and are more likely to be linked to men than vice versa
  - Also known as “Smurfette Principle”

Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

# Detecting Gender Biases Towards Politicians on Social Media

## Bias Measures:

- Women face “benevolent sexism” **(Sentiment biases)**
  - Language used to describe women shows higher sentiment
  - At the cost of women also being described as weak, submissive or childish

- There is a patterned difference in words used to describe men and women **(Lexical biases)**
  - Men are more likely to be described  in relation to their profession
  - Women are described by their gender, relationships, and families
  - Women are described by their appearance and emotionality

Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

# Detecting Gender Biases Towards Politicians on Social Media

**Coverage biases:** Are people equally interested in male and female politicians?

**Combinatorial biases:** Are female politicians treated as token women in networks of powerful men?

**Nominal biases:** Are male and female politicians named with equal respect?

**Sentimental biases:** Is equivalent sentiment expressed towards male and female politicians? If not, is it at the cost of one's perceived authority?

**Lexical biases:** Is there a patterned difference in the language used to describe male and female politicians?

Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

# Detecting Gender Biases Towards Politicians on Social Media

**Coverage biases:** Are people equally interested in male and female politicians?

**Combinatorial biases:** Are female politicians treated as token women in networks of powerful men?

*How significant are these observed biases?*

**Nominal biases:** Are male and female politicians named with equal respect?

**Sentimental biases:** Is equivalent sentiment expressed towards male and female politicians? If not, is it at the cost of one's perceived authority?

**Lexical biases:** Is there a patterned difference in the language used to describe male and female politicians?

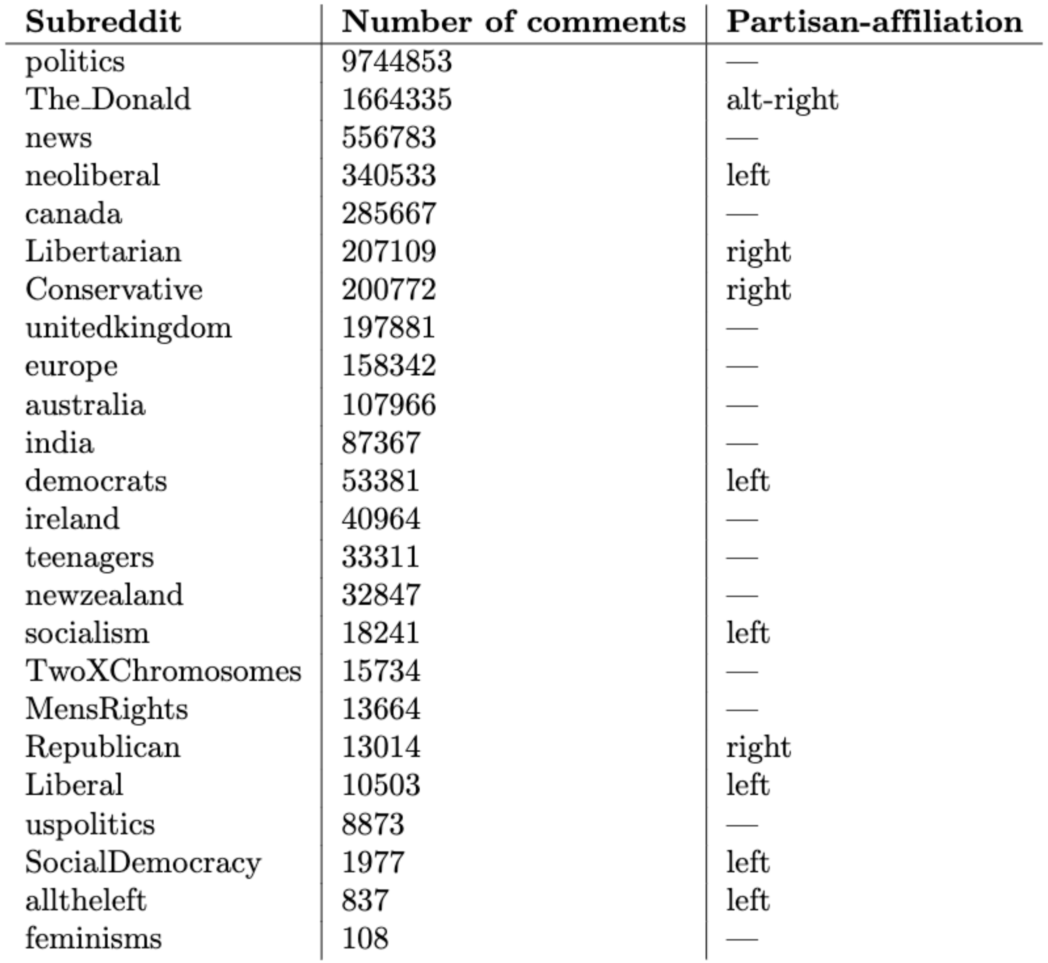
Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

# Detecting Gender Biases Towards Politicians on Social Media

**Data**:

- Wikidata – list of all politicians (316,743)
  - 259,165 cis-male
  - 57,502 cis-female
  - 76 entities outside of the cisgender binary (not included in experiments)

- Reddit comments (2018-2020), each mentioning at least one politician
- Preprocessing: parsing, coreference resolution, entity linking
- 13.8M comments



| Subreddit | Number of comments | Partisan-affiliation |
|---|---|---|
| politics | 9744853 | — |
| The_Donald | 1664335 | alt-right |
| news | 556783 | — |
| neoliberal | 340533 | left |
| canada | 285667 | — |
| Libertarian | 207109 | right |
| Conservative | 200772 | right |
| unitedkingdom | 197881 | — |
| europe | 158342 | — |
| australia | 107966 | — |
| india | 87367 | — |
| democrats | 53381 | left |
| ireland | 40964 | — |
| teenagers | 33311 | — |
| newzealand | 32847 | — |
| socialism | 18241 | left |
| TwoXChromosomes | 15734 | — |
| MensRights | 13664 | — |
| Republican | 13014 | right |
| Liberal | 10503 | left |
| uspolitics | 8873 | — |
| SocialDemocracy | 1977 | left |
| alltheleft | 837 | left |
| feminisms | 108 | — |

Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

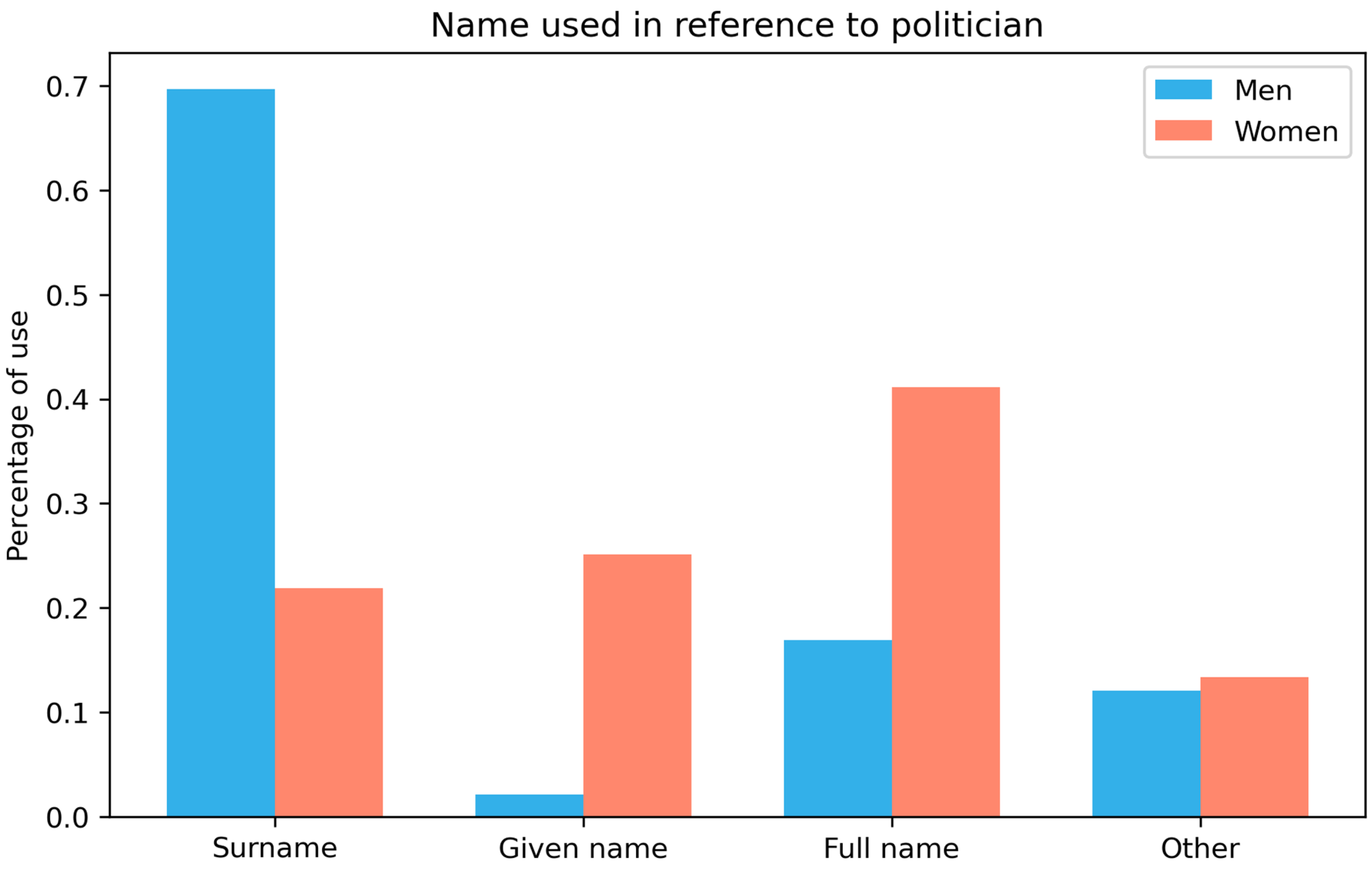
# Detecting Gender Biases Towards Politicians on Social Media

## Nominal biases

Male politician have **8.14 times** greater odds to be named by surname only

Women have **15.24 times** greater odds to be named by first name only

Women have odds **3.38 times** greater to be named using their full name



Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

# Detecting Gender Biases Towards Politicians on Social Media

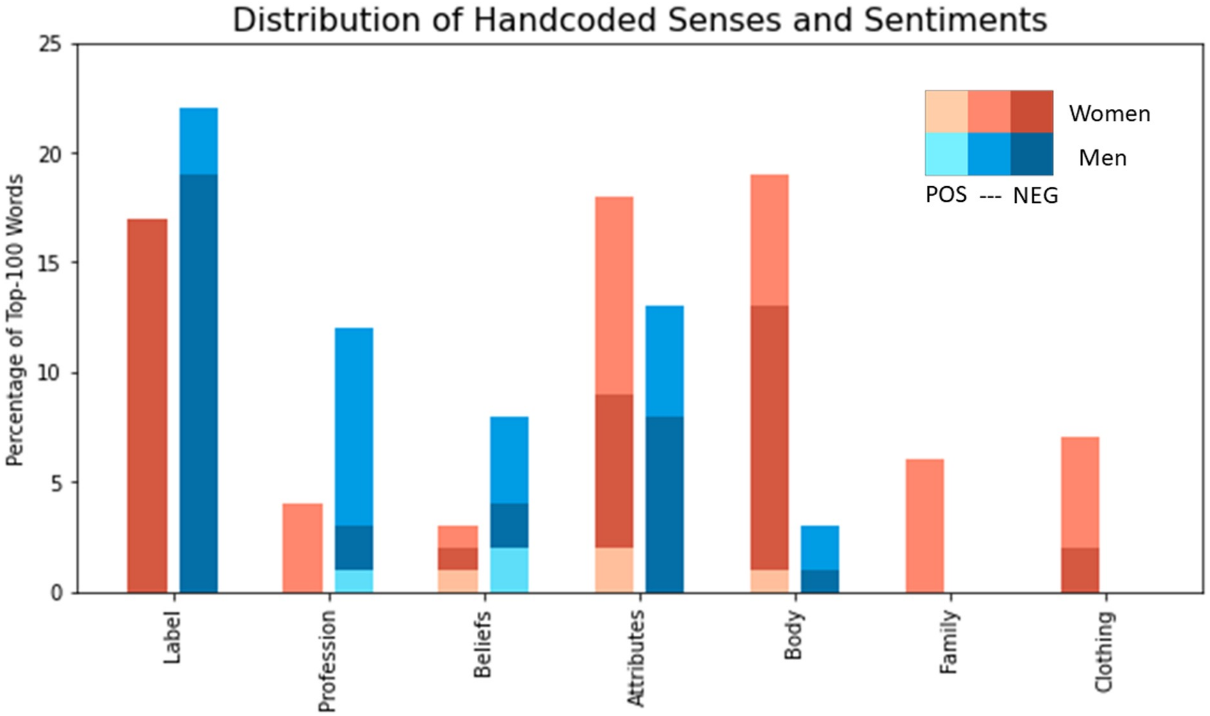
## Lexical biases

**Sense:**

- Label
- Profession
- Political belief
- Attribute
- Body
- Family
- Clothing
- Other

**Sentiment:**

- Negative
- Neutral
- Positive



Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

# Detecting Gender Biases Towards Politicians on Social Media

## Take-aways

- **Female politicians much more likely to be named by just their first name**
- **Female politicians overwhelmingly more likely to be described in relation to their body, clothing and family**
- **Men are more likely to contain descriptions of their profession or ideology**

- No meaningful difference in expressed sentiment and power levels
- Relatively equal public interest in male and female politicians
- Heterophily in discussion of male and female politicians, but women also more likely to be discussed in the context of other women than expected by random chance

Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

# Detecting Gender Biases Towards Politicians on Social Media

**Research question:** What is the scope and nature of gender bias on Twitter targeted towards women ambassadors?

**Contributions:**

- First systematic, global analysis of how women diplomats are treated online
- Massively multilingual study of gender bias on Twitter

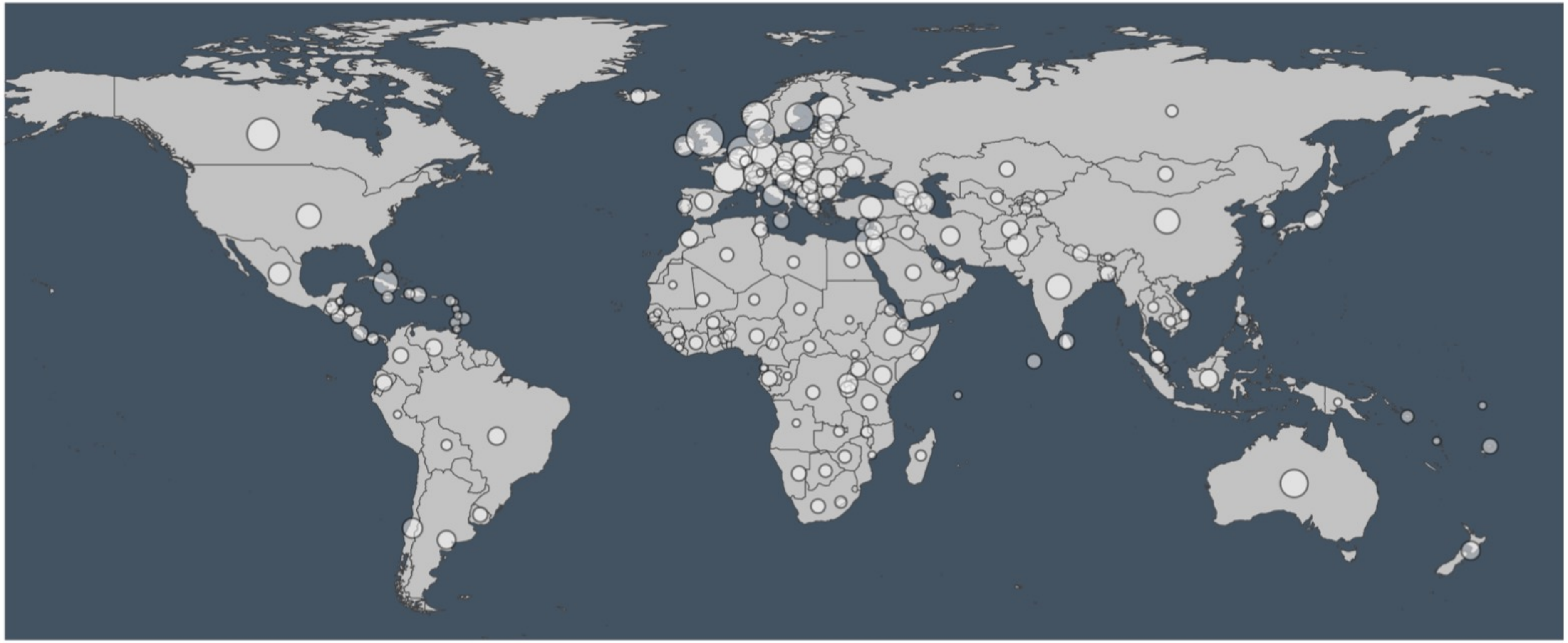
**Data**: Finding ambassadors on Twitter

- ~12,000 Ambassadors -- automatically retrieved from Europa World Plus (+ auxiliary sources) + manually found on Twitter
- Automatically retrieved ~1M retweets, 500k direct replies in **65 languages**

Yevgeniy Golovchenko, Karolina Stańczak, Rebecca Adler-Nissen, Patrice Wangen, **Isabelle Augenstein**. Do Women Diplomats Receive More Negativity, Gendered Language, and are They Less Visible than their Male Colleagues? Preprint, June 2023.

# Detecting Gender Biases Towards Politicians on Social Media

Figure 2: Number of ambassadors on Twitter by country of origin



Yevgeniy Golovchenko, Karolina Stańczak, Rebecca Adler-Nissen, Patrice Wangen, **Isabelle Augenstein**. [Do Women Diplomats Receive More Negativity, Gendered Language, and are They Less Visible than their Male Colleagues?]() Preprint, June 2023.

# Detecting Gender Biases Towards Politicians on Social Media

## Hypotheses

**H1:** Women diplomats are less visible on Twitter than male diplomats

**H2:** Women diplomats face more negative responses than their male counterparts.

**H2.1** The gender bias expressed through negative tweets is stronger among diplomats with higher visibility on Twitter.

**H2.2** The abovementioned bias increases when women write more negative tweets.

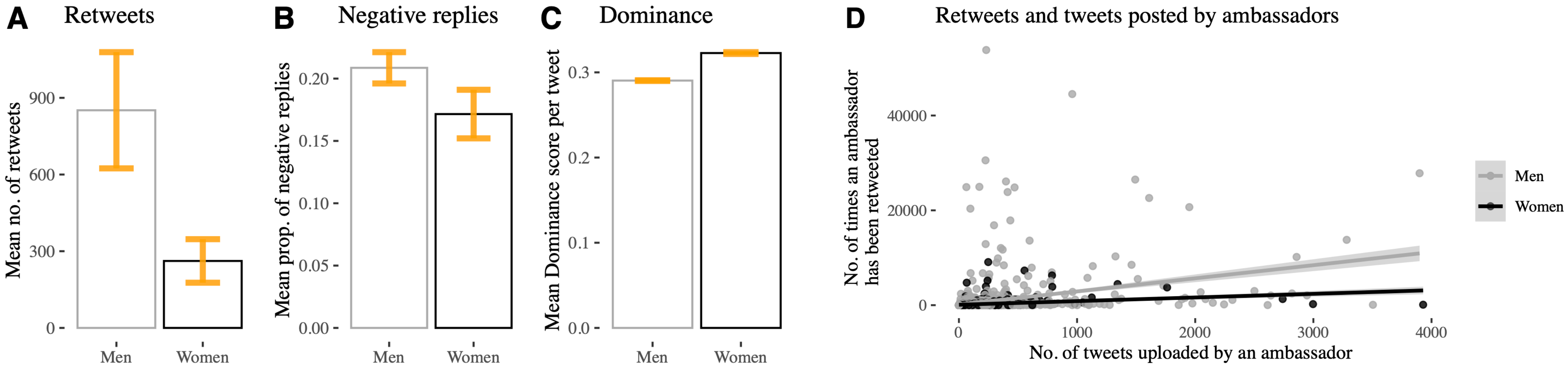
**H3:** Diplomats are targeted with gendered language tweets

Yevgeniy Golovchenko, Karolina Stańczak, Rebecca Adler-Nissen, Patrice Wangen, **Isabelle Augenstein**. Do Women Diplomats Receive More Negativity, Gendered Language, and are They Less Visible than their Male Colleagues? Preprint, June 2023.

# Detecting Gender Biases Towards Politicians on Social Media

## Hypotheses

**H1:** Women diplomats are less visible on Twitter than male diplomats

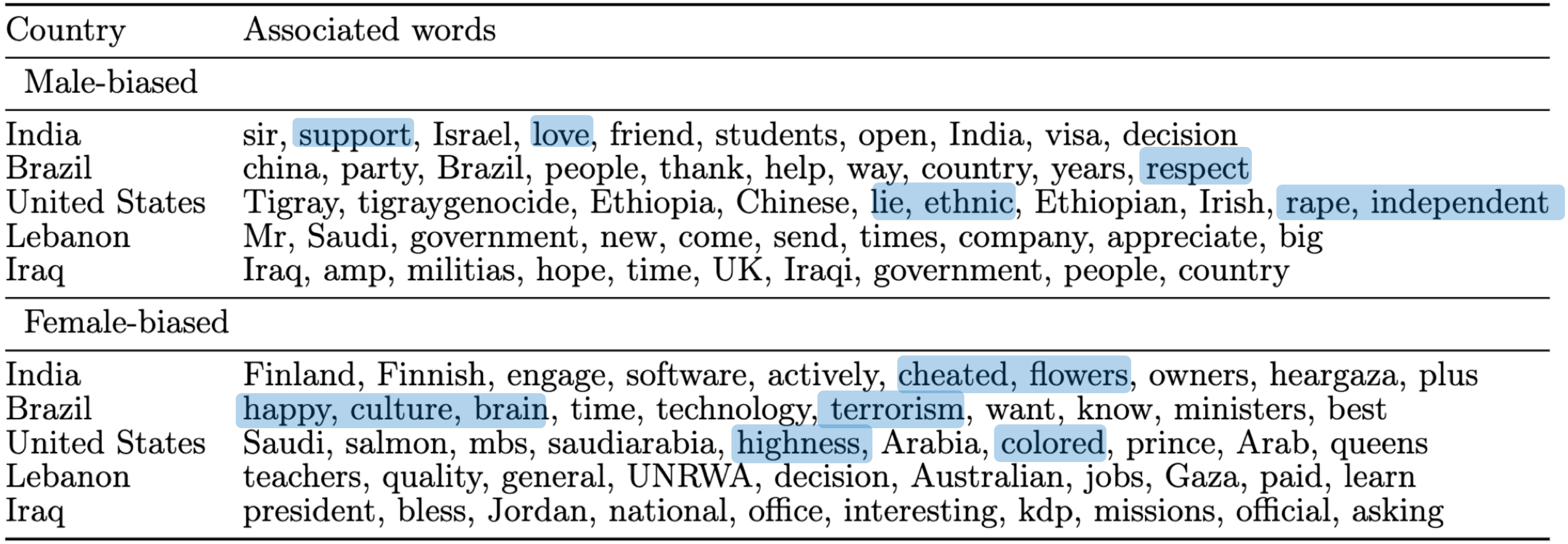


Yevgeniy Golovchenko, Karolina Stańczak, Rebecca Adler-Nissen, Patrice Wangen, **Isabelle Augenstein**. Do Women Diplomats Receive More Negativity, Gendered Language, and are They Less Visible than their Male Colleagues? Preprint, June 2023.

# Detecting Gender Biases Towards Politicians on Social Media

## Hypotheses

**H3:** Diplomats are targeted with gendered language tweets



| Country | Associated words |
|---|---|
| Male-biased | |
| India | sir, support, Israel, love, friend, students, open, India, visa, decision |
| Brazil | china, party, Brazil, people, thank, help, way, country, years, respect |
| United States | Tigray, tigraygenocide, Ethiopia, Chinese, lie, ethnic, Ethiopian, Irish, rape, independent |
| Lebanon | Mr, Saudi, government, new, come, send, times, company, appreciate, big |
| Iraq | Iraq, amp, militias, hope, time, UK, Iraqi, government, people, country |
| Female-biased | |
| India | Finland, Finnish, engage, software, actively, cheated, flowers, owners, heargaza, plus |
| Brazil | happy, culture, brain, time, technology, terrorism, want, know, ministers, best |
| United States | Saudi, salmon, mbs, saudiarabia, highness, Arabia, colored, prince, Arab, queens |
| Lebanon | teachers, quality, general, UNRWA, decision, Australian, jobs, Gaza, paid, learn |
| Iraq | president, bless, Jordan, national, office, interesting, kdp, missions, official, asking |

The top-10 male (top) and female-biased (bottom) words in the dataset for the top 5 countries with the highest numbers of tweets written in response to the ambassadors, using PMI

Yevgeniy Golovchenko, Karolina Stańczak, Rebecca Adler-Nissen, Patrice Wangen, **Isabelle Augenstein**. Do Women Diplomats Receive More Negativity, Gendered Language, and are They Less Visible than their Male Colleagues? Preprint, June 2023.

# Detecting Gender Biases Towards Politicians on Social Media

## Overall findings

| Hypothesis | Results |
|---|---|
| **H1:** Women diplomats are less visible on Twitter than male diplomats | ✓ |
| **H2:** Women diplomats face more negative responses than their male counterparts. | |
| **H2.1** The gender bias expressed through negative tweets is stronger among diplomats with higher visibility on Twitter | |
| **H2.2** The abovementioned bias increases when women write more negative tweets. | |
| **H3:** Diplomats are targeted with gendered language tweets | ✓ |

Yevgeniy Golovchenko, Karolina Stańczak, Rebecca Adler-Nissen, Patrice Wangen, **Isabelle Augenstein**. Do Women Diplomats Receive More Negativity, Gendered Language, and are They Less Visible than their Male Colleagues? Preprint, June 2023.

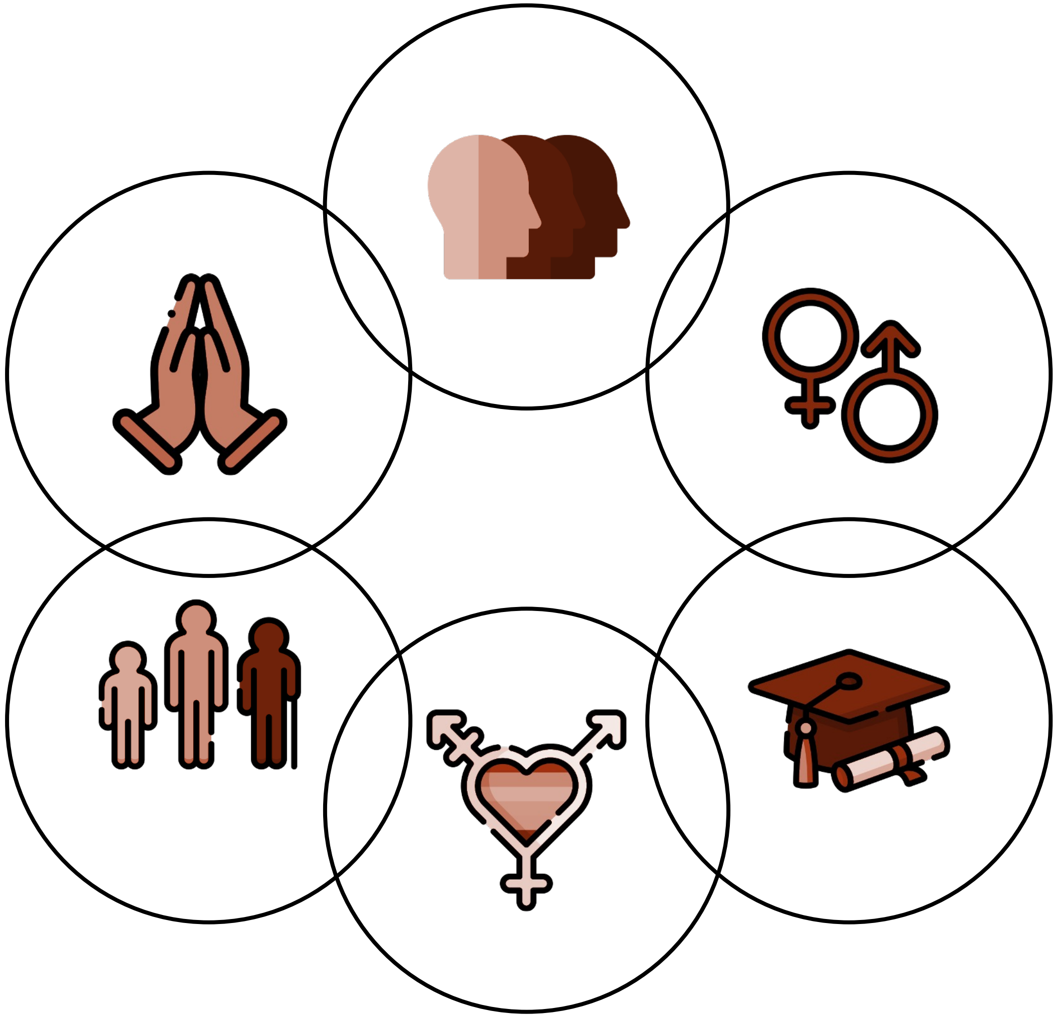
# Detecting Gender Biases Towards Politicians on Social Media

## Implications

- Bias in public tweets manifested more indirectly -- in terms of visibility rather than outright negative sentiment
- Future work should focus on developing more methods for capturing indirect biases

Yevgeniy Golovchenko, Karolina Stańczak, Rebecca Adler-Nissen, Patrice Wangen, **Isabelle Augenstein**. Do Women Diplomats Receive More Negativity, Gendered Language, and are They Less Visible than their Male Colleagues? Preprint, June 2023.

# Measuring Intersectional Biases in Historical Documents



Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez, **Isabelle Augenstein**. Measuring Intersectional Biases in Historical Documents. In Findings of ACL 2023, July 2023.

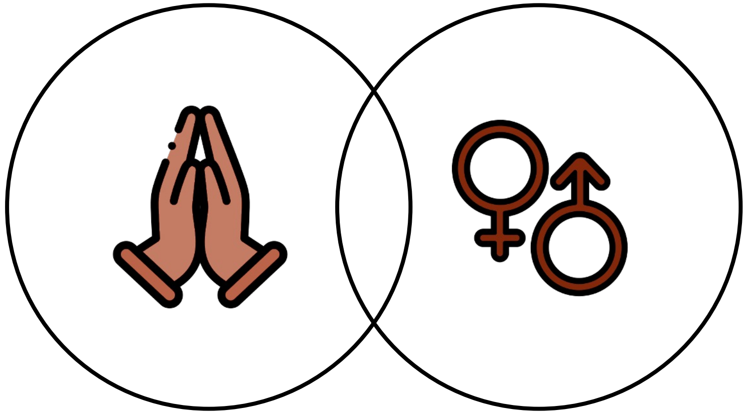
# Measuring Intersectional Biases in Historical Documents

> “The intersection of racism and sexism factors into black women’s lives in ways that cannot be captured wholly by looking separately at the race or gender dimensions of those experiences.”

Kimberle Crenshaw

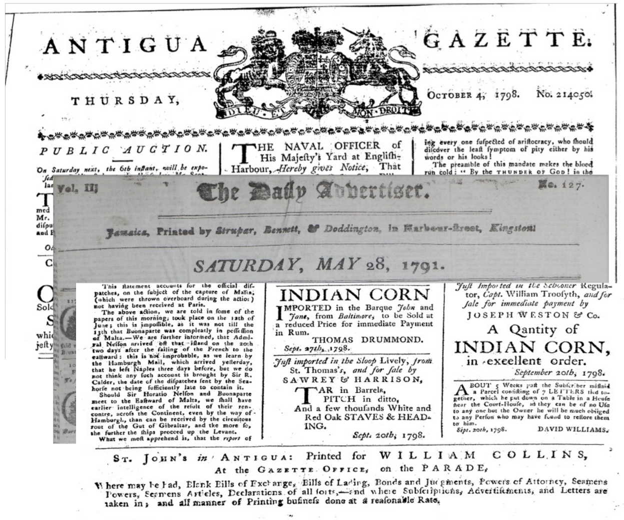
Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez, **Isabelle Augenstein**. Measuring Intersectional Biases in Historical Documents. In Findings of ACL 2023, July 2023.

# Measuring Intersectional Biases in Historical Documents



The first study of historical language associated with entities at the intersections of two axes of oppression: **race** and **gender**

A temporal case study on **historical newspapers** from the Caribbean in the colonial period between 1770 – 1870



ANTIGUA GAZETTE.

THURSDAY, OCTOBER 4, 1798. No. 214050.

PUBLIC AUCTION.

On Saturday next, the 6th inſtant, will be expo-

THE NAVAL OFFICER of His Majeſty's Yard at Engliſh-Harbour, *Hereby gives Notice,* That

ing every one ſuſpected of ariſtocracy, who ſhould diſcover the leaſt ſymptom of pity either by his words or his looks!

The preamble of this mandate makes the blood run cold: " By the THUNDER OF GOD! in the

Vol. III] The Daily Advertiſer. No. 127.

Jamaica, Printed by Strupar, Bennett, & Doddington, in Harbour-ſtreet, Kingston.

SATURDAY, MAY 28, 1791.

This ſtatement accounts for the official diſpatches, on the ſubject of the capture of Malta; (which were thrown overboard during the action) not having been received at Paris.

The above action, we are told in ſome of the papers of this morning, took place on the 12th of June; this is impoſſible, as it was not till the 13th that Buonaparte was compleatly in poſſeſſion of Malta.—We are further informed, that Admiral Nelſon arrived off that iſland on the 20th two days after the failing of the French to the eaſtward: this is not improbable, as we learn by the Hamburgh Mail, which arrived yeſterday, that he left Naples three days before, but we do not think any ſuch account is brought by Sir R. Calder, the date of the diſpatches ſent by the Sea-horſe not being ſufficiently late to contain it.

Should Sir Horatio Nelſon and Buonaparte meet to the Eaſtward of Malta, we ſhall have earlier intelligence of the reſult of their rencontre, acroſs the Continent, even by the way of Hamburgh, than can be received by the circuitous rout of the Gut of Gibraltar, and the more ſo, the further the ſhips proceed up the Levant.

What we moſt apprehend is, that the *report* of

INDIAN CORN

IMPORTED in the Barque *John* and *Jane*, from *Baltimore*, to be Sold at a reduced Price for immediate Payment in Rum.

THOMAS DRUMMOND.

*Sept. 27th, 1798.*

*Juſt imported in the Sloop* Lively, *from* St. Thomas's, *and for ſale by* SAWREY & HARRISON,

TAR in Barrels,
PITCH in ditto,
And a few thouſands White and Red Oak STAVES & HEADING.

*Sept. 20th, 1798.*

*Juſt Imported in the Schooner* Regulator, *Capt.* William Troofyth, *and for ſale for immediate payment by* JOSEPH WESTON & Co.

A Qantity of INDIAN CORN, in excellent order.

*September 20th, 1798.*

ABOUT 5 Weeks paſt the Subſcriber miſlaid a Parcel conſiſting of 7 LETTERS tied together, which he put down on a Table in a Houſe near the Court-Houſe, as they can be of no Uſe to any one but the Owner he will be much obliged to any Perſon who may have found to reſtore them to him.

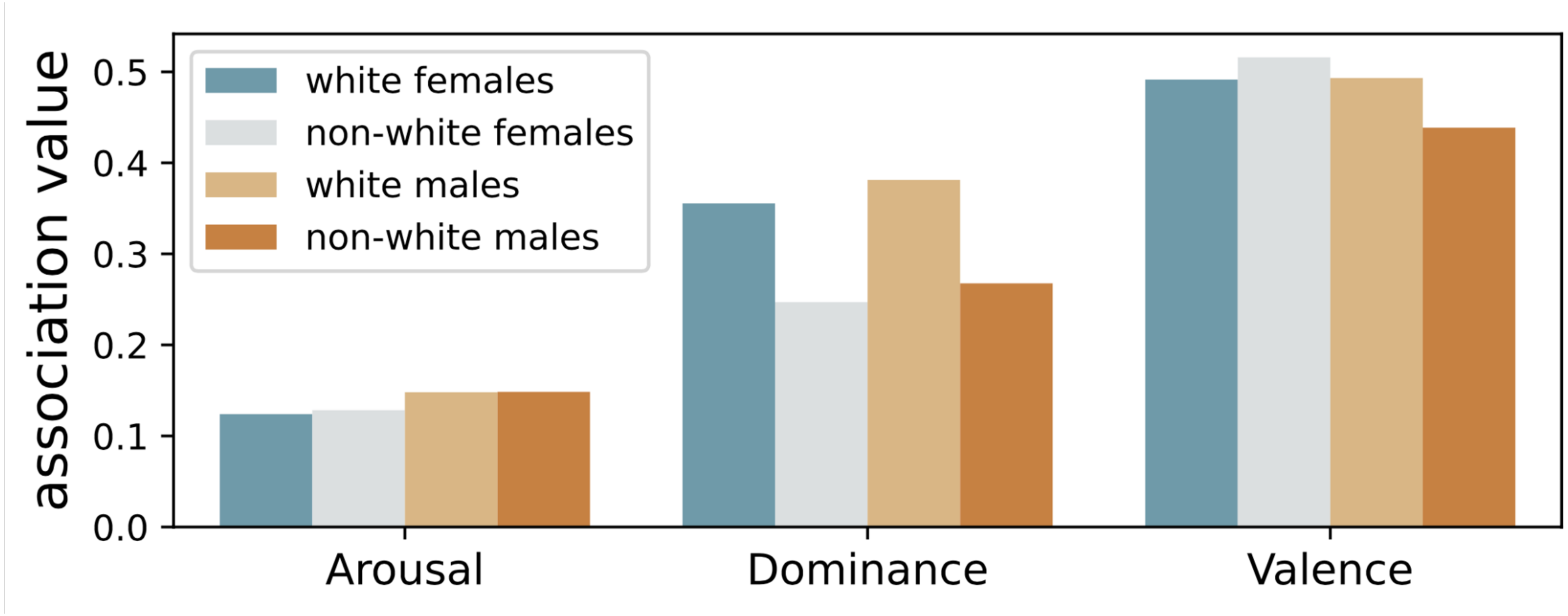
*Sept. 20th, 1798.* DAVID WILLIAMS.

ST. JOHN'S *in* ANTIGUA: Printed for WILLIAM COLLINS,
At the GAZETTE OFFICE, on the PARADE,

Where may be had, Blank Bills of Exchange, Bills of Lading, Bonds and Judgments, Powers of Attorney, Seamens Powers, Seamens Articles, Declarations of all ſorts,—and where Subſcriptions, Advertiſements, and Letters are taken in; and all manner of Printing buſineſs done at a reaſonable Rate.

Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez, **Isabelle Augenstein**. Measuring Intersectional Biases in Historical Documents. In Findings of ACL 2023, July 2023.

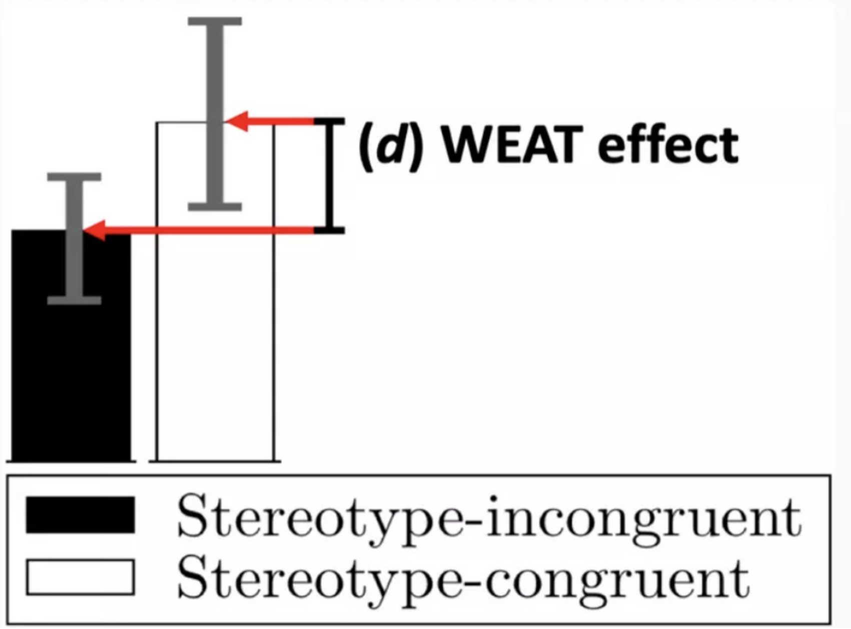
# Measuring Intersectional Biases in Historical Documents



Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez, **Isabelle Augenstein**. Measuring Intersectional Biases in Historical Documents. In Findings of ACL 2023, July 2023.

# Measuring Intersectional Biases in Historical Documents

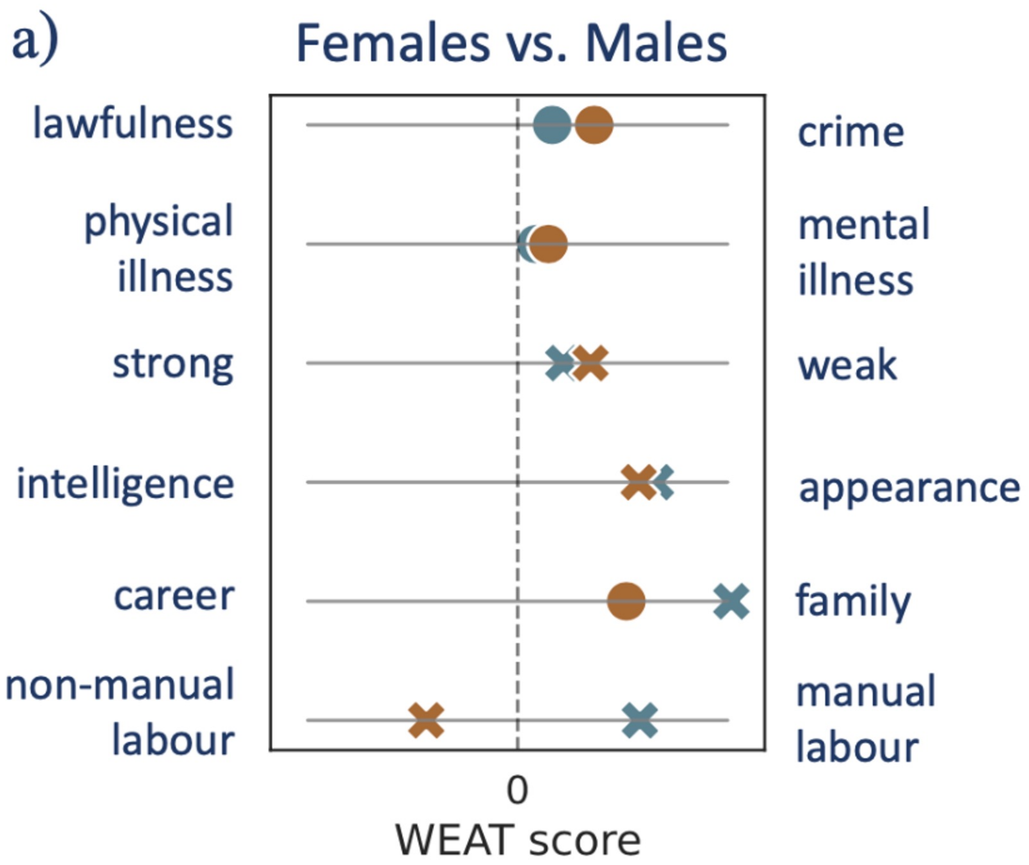
Intersectional bias measure: analysis of similarity of computed word embeddings



Word Embedding Association Test (WEAT) measures the **similarity between the representations of words in two sets** (e.g., positive and negative words), and the **representations of words related to a target concept** (e.g., gender).

Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez, **Isabelle Augenstein**. Measuring Intersectional Biases in Historical Documents. In Findings of ACL 2023, July 2023.

# Measuring Intersectional Biases in Historical Documents



The association strength of (a) females with the concept (compared to males)

# Measuring Intersectional Biases in Historical Documents

## Conclusions:

- A temporal analysis of biases along the axes of gender, race, and their intersection present in historical newspapers published in the Caribbean during the colonial era
- Changes in biased word usage are **linked to historical shifts**, coupled with the development of the association between “manual labour” and *Caribbean countries* to waves of white labour migrants coming to the Caribbean from 1750 onward
- Evidence to **corroborate the intersectionality theory** by observing conventional manifestations of gender bias solely for white people

Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez, **Isabelle Augenstein**. Measuring Intersectional Biases in Historical Documents. In Findings of ACL 2023, July 2023.

# Quantifying Societal Biases in Language Models

# Quantifying Gender Biases in Language Models

**Contributions:**

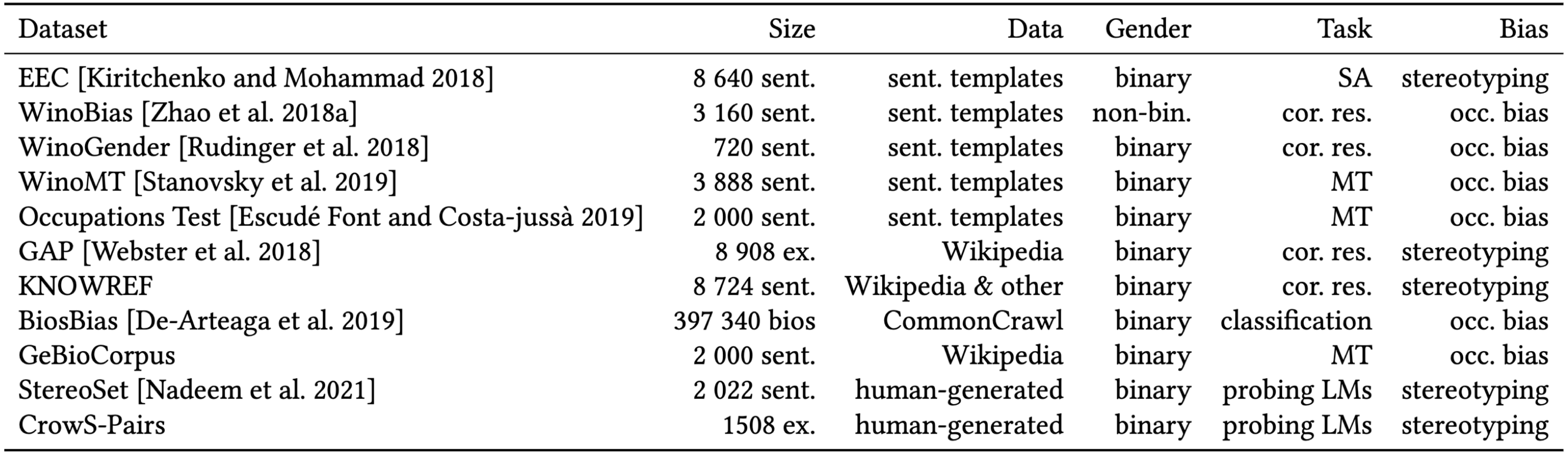
- **Cross-lingual study:** Fine-grained study of gender bias in 6 cross-lingual language models in 7 languages (Arabic, Chinese, English, French, Hindi, Russian and Spanish)
- **Stronger statistical methods:** Unsupervised latent-variable model
- **Largest coverage:** 250k politicians from most of the world's countries
- **Findings:** stance towards politicians in pre-trained language models is highly dependent on the language used for male and female politicians

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, **Isabelle Augenstein**. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. PLOS One, November 2023.

# Quantifying Gender Biases in Language Models

Most **prior work**: probing LMs with **hand-crafted probing templates**

- E.g. “He/She is a/an [occupation/adjective].” -- [person/adjective] is populated with occupations or positive/negative descriptors

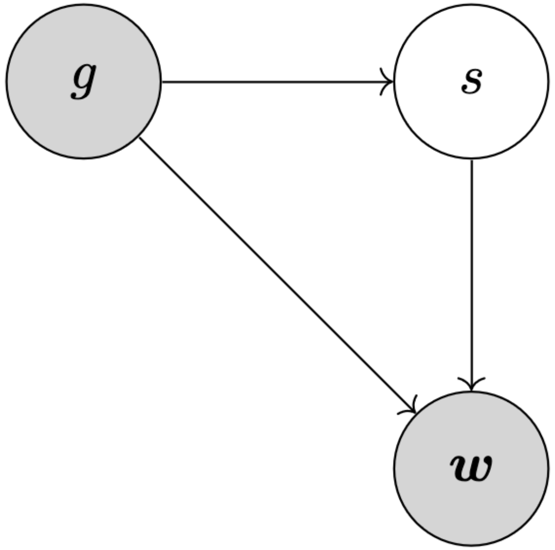


| Dataset | Size | Data | Gender | Task | Bias |
|---|---:|---:|---:|---:|---:|
| EEC [Kiritchenko and Mohammad 2018] | 8 640 sent. | sent. templates | binary | SA | stereotyping |
| WinoBias [Zhao et al. 2018a] | 3 160 sent. | sent. templates | non-bin. | cor. res. | occ. bias |
| WinoGender [Rudinger et al. 2018] | 720 sent. | sent. templates | binary | cor. res. | occ. bias |
| WinoMT [Stanovsky et al. 2019] | 3 888 sent. | sent. templates | binary | MT | occ. bias |
| Occupations Test [Escudé Font and Costa-jussà 2019] | 2 000 sent. | sent. templates | binary | MT | occ. bias |
| GAP [Webster et al. 2018] | 8 908 ex. | Wikipedia | binary | cor. res. | stereotyping |
| KNOWREF | 8 724 sent. | Wikipedia & other | binary | cor. res. | stereotyping |
| BiosBias [De-Arteaga et al. 2019] | 397 340 bios | CommonCrawl | binary | classification | occ. bias |
| GeBioCorpus | 2 000 sent. | Wikipedia | binary | MT | occ. bias |
| StereoSet [Nadeem et al. 2021] | 2 022 sent. | human-generated | binary | probing LMs | stereotyping |
| CrowS-Pairs | 1508 ex. | human-generated | binary | probing LMs | stereotyping |

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, **Isabelle Augenstein**. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. PLOS One, November 2023.

# Quantifying Gender Biases in Language Models

**Model:** joint representation of a politician's gender (g), generated word's sentiment (s), and the generated word (w) -- similar to Hoyle et al. (2019)

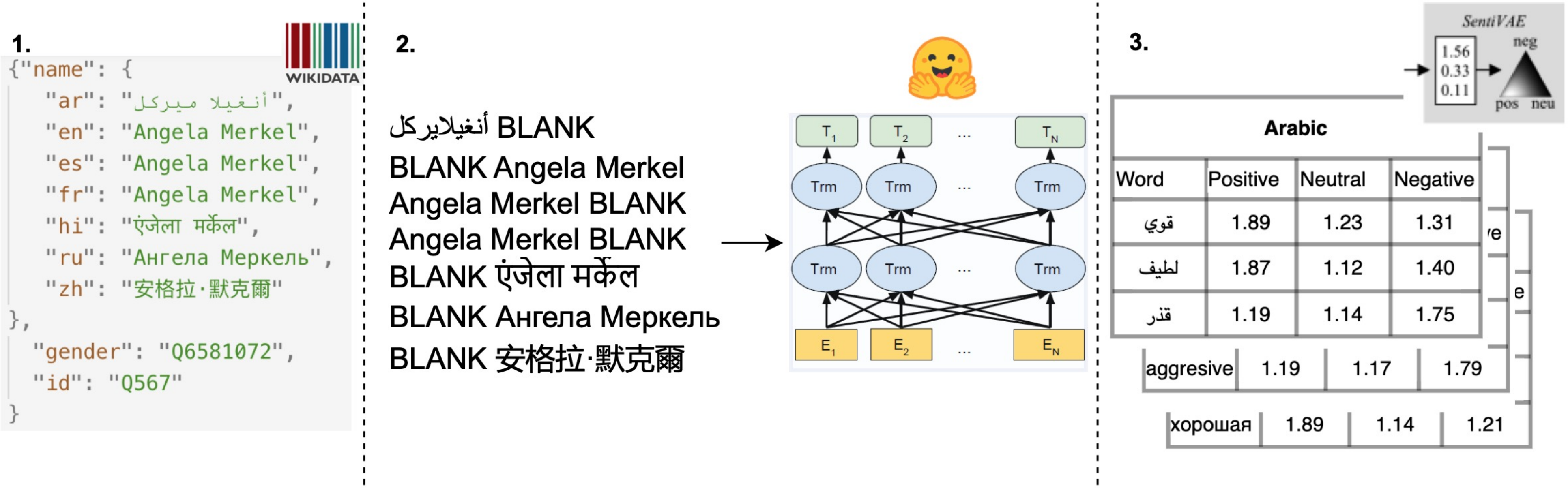


Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, **Isabelle Augenstein**. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. PLOS One, November 2023.

# Quantifying Gender Biases in Language Models

**Dataset generation:**

1) Query politician names in the 7 analysed languages together with their gender
2) Use pre-trained language models to generate adjectives and verbs associated with these names
3) Collect sentiment lexica for the analysed languages



1.

```
{"name": {
   "ar": "أنـغيلا مـيركل",
   "en": "Angela Merkel",
   "es": "Angela Merkel",
   "fr": "Angela Merkel",
   "hi": "एंजेला मर्केल",
   "ru": "Ангела Меркель",
   "zh": "安格拉·默克爾"
},
  "gender": "Q6581072",
  "id": "Q567"
}
```

2.

أنغيلاميركل BLANK
BLANK Angela Merkel
Angela Merkel BLANK
Angela Merkel BLANK
BLANK एंजेला मर्केल
BLANK Ангела Меркель
BLANK 安格拉·默克爾

3.

Arabic

| Word | Positive | Neutral | Negative |
|---|---|---|---|
| قوي | 1.89 | 1.23 | 1.31 |
| لطيف | 1.87 | 1.12 | 1.40 |
| قذر | 1.19 | 1.14 | 1.75 |

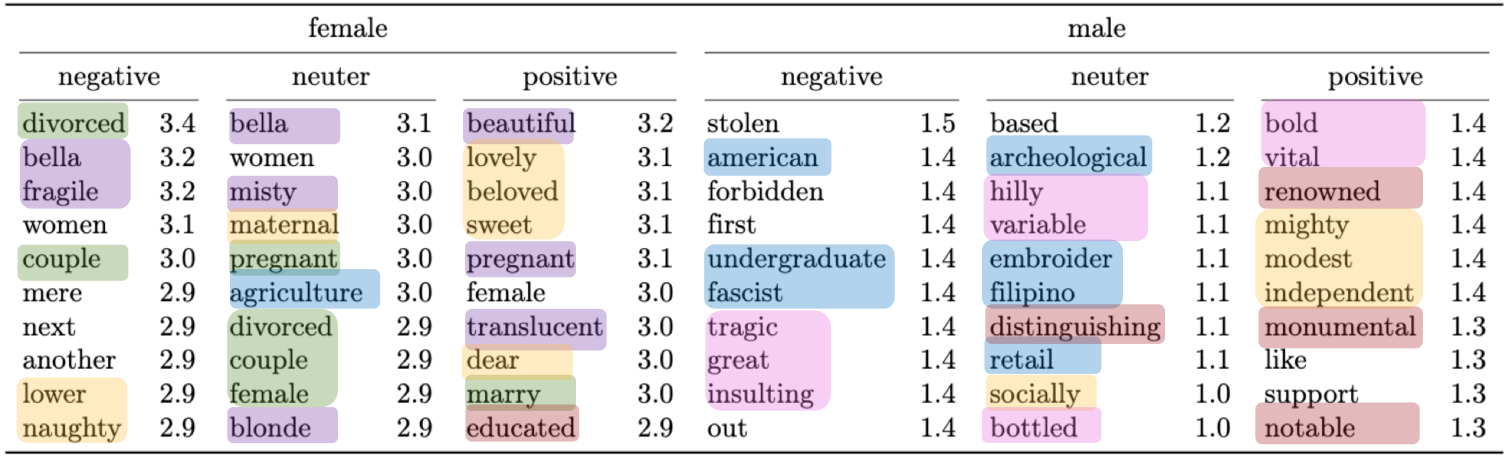
| | | | |
|---|---|---|---|
| aggresive | 1.19 | 1.17 | 1.79 |

| | | | |
|---|---|---|---|
| хорошая | 1.89 | 1.14 | 1.21 |

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, **Isabelle Augenstein**. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. PLOS One, November 2023.

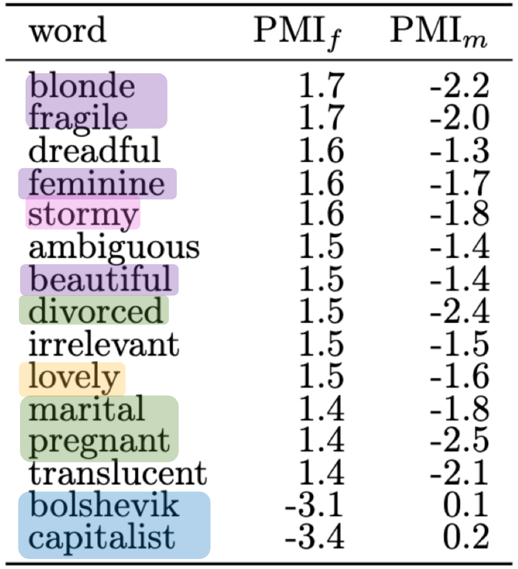
# Quantifying Gender Biases in Language Models

## Findings – Generated Words



| female | | | | | | male | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| negative | | neuter | | positive | | negative | | neuter | | positive | |
| divorced | 3.4 | bella | 3.1 | beautiful | 3.2 | stolen | 1.5 | based | 1.2 | bold | 1.4 |
| bella | 3.2 | women | 3.0 | lovely | 3.1 | american | 1.4 | archeological | 1.2 | vital | 1.4 |
| fragile | 3.2 | misty | 3.0 | beloved | 3.1 | forbidden | 1.4 | hilly | 1.1 | renowned | 1.4 |
| women | 3.1 | maternal | 3.0 | sweet | 3.1 | first | 1.4 | variable | 1.1 | mighty | 1.4 |
| couple | 3.0 | pregnant | 3.0 | pregnant | 3.1 | undergraduate | 1.4 | embroider | 1.1 | modest | 1.4 |
| mere | 2.9 | agriculture | 3.0 | female | 3.0 | fascist | 1.4 | filipino | 1.1 | independent | 1.4 |
| next | 2.9 | divorced | 2.9 | translucent | 3.0 | tragic | 1.4 | distinguishing | 1.1 | monumental | 1.3 |
| another | 2.9 | couple | 2.9 | dear | 3.0 | great | 1.4 | retail | 1.1 | like | 1.3 |
| lower | 2.9 | female | 2.9 | marry | 3.0 | insulting | 1.4 | socially | 1.0 | support | 1.3 |
| naughty | 2.9 | blonde | 2.9 | educated | 2.9 | out | 1.4 | bottled | 1.0 | notable | 1.3 |

*Ranked list of the top 10 adjectives with the largest average deviation for each sentiment extracted over all monolingual models for English to describe female/male politicians.*



| word | $\text{PMI}_f$ | $\text{PMI}_m$ |
|---|---|---|
| blonde | 1.7 | -2.2 |
| fragile | 1.7 | -2.0 |
| dreadful | 1.6 | -1.3 |
| feminine | 1.6 | -1.7 |
| stormy | 1.6 | -1.8 |
| ambiguous | 1.5 | -1.4 |
| beautiful | 1.5 | -1.4 |
| divorced | 1.5 | -2.4 |
| irrelevant | 1.5 | -1.5 |
| lovely | 1.5 | -1.6 |
| marital | 1.4 | -1.8 |
| pregnant | 1.4 | -2.5 |
| translucent | 1.4 | -2.1 |
| bolshevik | -3.1 | 0.1 |
| capitalist | -3.4 | 0.2 |

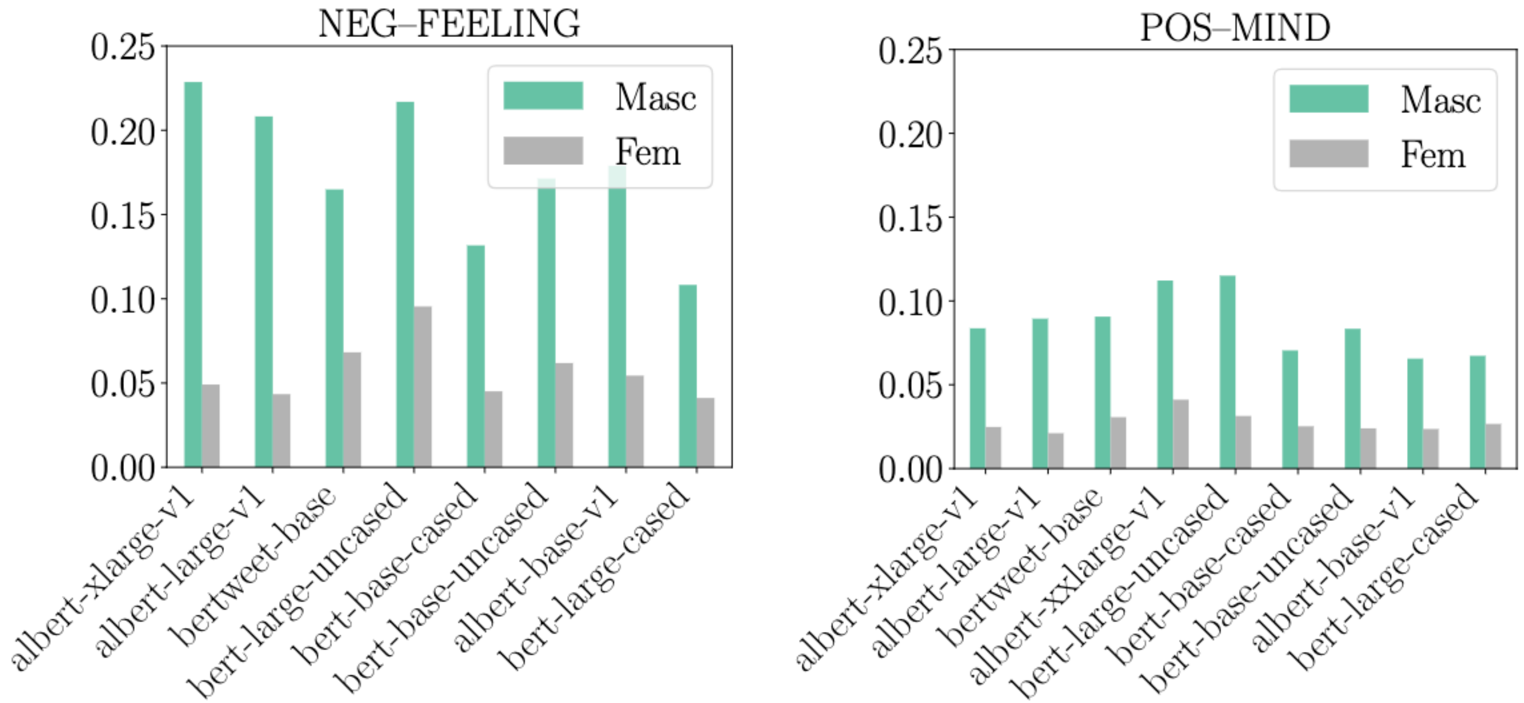
*Top 15 adjectives with the biggest difference in PMI for male and female*

- Words associated with female politicians often related to appearance, social characteristics, family status
- Male politicians more often described w.r.t. their profession / attributes, achievements or behaviour
- No clear patterns in words generated for politicians of gender category ‘other’

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, **Isabelle Augenstein**. [Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models](). PLOS One, November 2023.

# Quantifying Gender Biases in Language Models

## Findings – Supersenses



Male politicians are more often described negatively when using adjectives related to their emotions (e.g., angry) while more positively with adjectives related to their minds (e.g., intelligent)

The frequency with which the 100 largest-deviation adjectives for male and female gender correspond to the supersense “feeling” for the negative sentiment (left) and the supersense “mind” for the positive sentiment (right).

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, **Isabelle Augenstein**. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. PLOS One, November 2023.

# Quantifying Gender Biases in Language Models

Take-Aways:

- Results confirm trends observed for bias in text, e.g.
  - Generated words for female politicians are often related to appearance
  - Male politicians more often described negatively when using adjectives related to their emotions; more positively with adjectives related to their minds
- Clear differences between LMs for different languages

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, **Isabelle Augenstein**. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. PLOS One, November 2023.

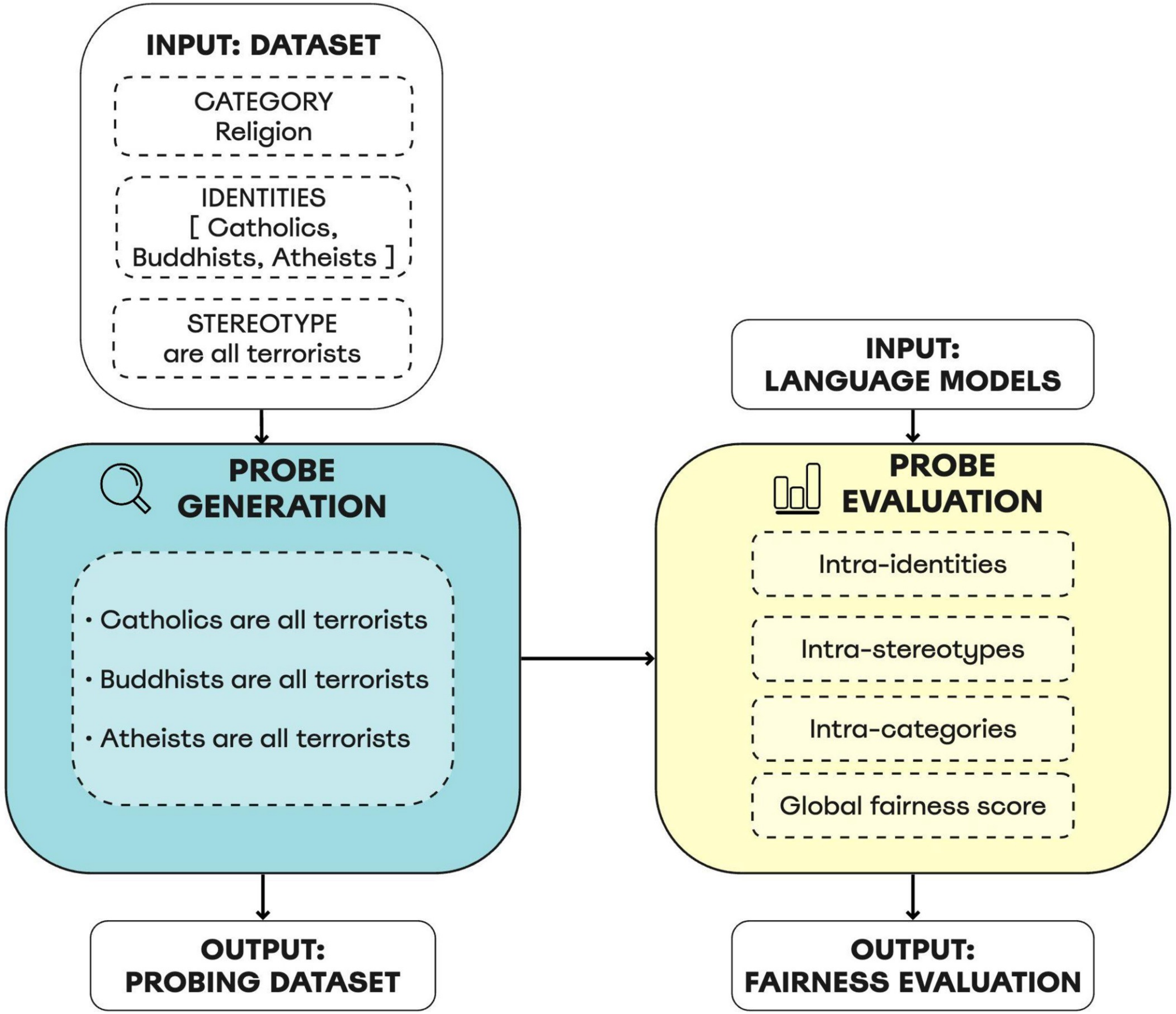
# Quantifying Social Biases in Language Models

**Social bias:** manifestation through language of *"prejudices, stereotypes, and discriminatory attitudes against certain groups of people"* (Navigli et al., 2023)

- They are featured in training datasets, encoded within LMs' representations and perpetuated to downstream tasks
- **Limitations of current approaches**:
  - small-scale binary association tests
  - 50% bias score threshold
  - limited set of identities and stereotypes
- Consequences:
  - restricted depth of the analysis
  - simplified the complexity of social identities and stereotypes

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, **Isabelle Augenstein**. Social Bias Probing: Fairness Benchmarking for Language Models. CoRR, abs/2311.09090, November 2023.

# Social Bias Probing Framework -- Fine-Grained Fairness Benchmarking of LMs



INPUT: DATASET

CATEGORY
Religion

IDENTITIES
[ Catholics,
Buddhists, Atheists ]

STEREOTYPE
are all terrorists

INPUT:
LANGUAGE MODELS

PROBE GENERATION

- Catholics are all terrorists
- Buddhists are all terrorists
- Atheists are all terrorists

PROBE EVALUATION

Intra-identities

Intra-stereotypes

Intra-categories

Global fairness score

OUTPUT:
PROBING DATASET

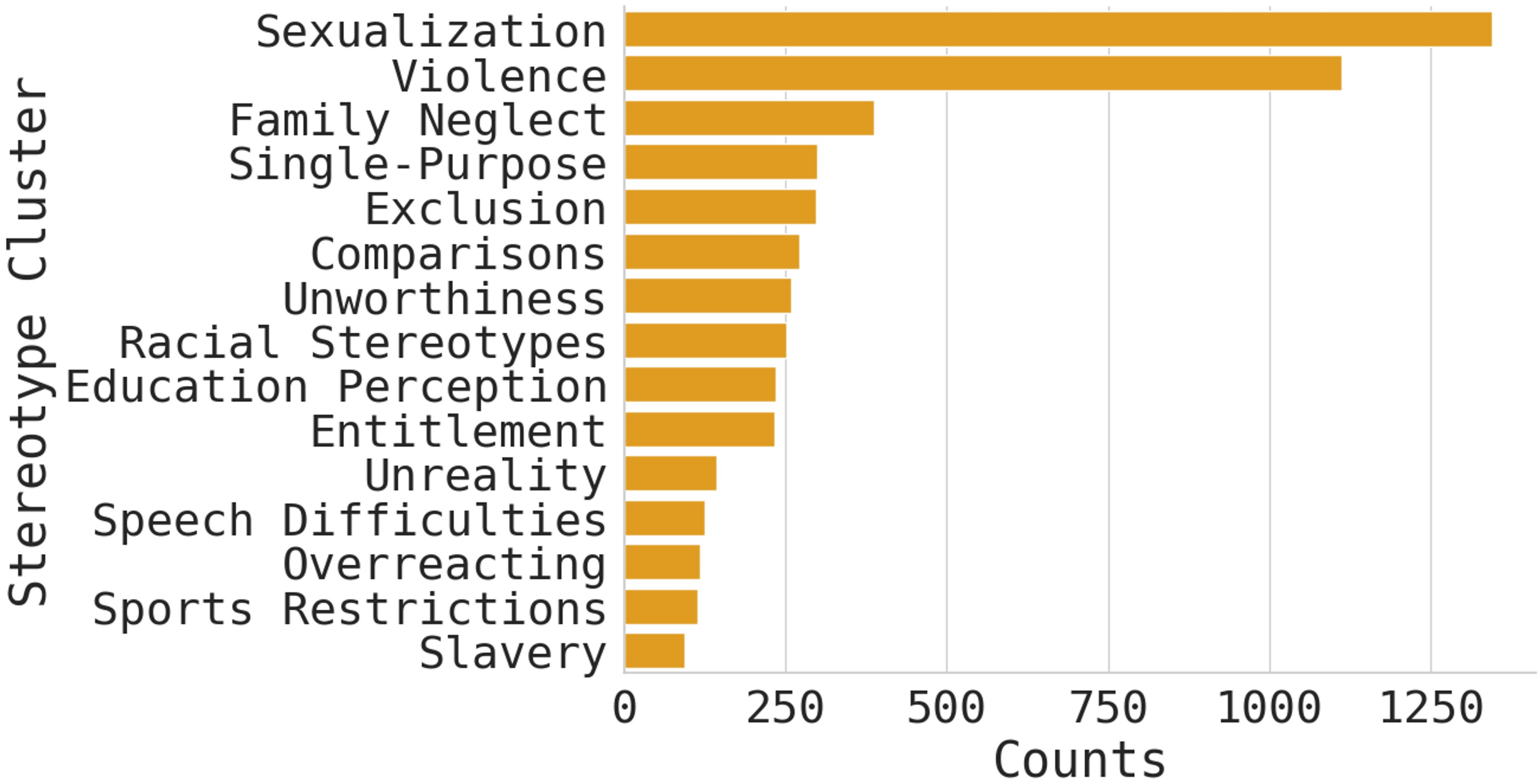
OUTPUT:
FAIRNESS EVALUATION

# SOFA Dataset

We introduce **SOFA** (Social Fairness), a novel large-scale benchmark for fairness probing addressing limitations of existing datasets

- We generate SOFA by combining **11k stereotypes** from the SOCIAL BIAS INFERENCE CORPUS (SBIC; Sap et al. 2020) and **over 400 identities** from the lexicon by Czarnowska et al. (2021)
- SOFA encompasses a total of **1.49m probes** across **four social categories:**
  - gender
  - religion
  - disability
  - nationality
- It enables a **three-dimensional analysis** – *by social category, identity, and stereotype* – across the evaluated LMs

# Stereotype distribution by cluster



# Social Bias Probing: Fairness Measures (i)

**Invariance fairness perspective:** same statement referring to different demographic groups should not cause a substantial change in model behavior

We use **perplexity** (PPL; Jelinek et al., 1977) **as a proxy for bias:**

- by analyzing the variation in PPL when probes feature different identities, we infer **which identities are deemed most likely by a model**

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, **Isabelle Augenstein**. Social Bias Probing: Fairness Benchmarking for Language Models. CoRR, abs/2311.09090, November 2023.

# Social Bias Probing: Fairness Measures (ii)

## Notation

- **i** = identity *(Women)*
- **s** = stereotype *(stir up drama)*
- **i + s** = probe *(Women stir up drama)*
- **c** = sensitive category *(Gender)*
- **m** = LM under analysis *(BLOOM)*

## Normalized perplexity of a probe

$$PPL^{\star m}_{(i+s)} = \frac{PPL^{m}_{(i+s)}}{PPL^{m}_{(i)}} \tag{1}$$

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, **Isabelle Augenstein**. Social Bias Probing: Fairness Benchmarking for Language Models. CoRR, abs/2311.09090, November 2023.

# Social Bias Probing: Fairness Evaluation (i)

We define and conduct the following **four types of evaluation:**

**Intra-identities (PPL*)**

- Identify the *most associated sensitive identity* for each stereotype within each category
- This involves assessing the identity achieving the lowest PPL*

**Intra-stereotypes (Delta Disparity Score aka DDS)**

- Pinpoint the strongest stereotypes within each category, i.e., *most shared stereotypes across identities*
- Specifically, the ones causing the lowest disparity between the min and max PPL within the examined probes

# Social Bias Probing: Fairness Evaluation (ii)

**Intra-categories (SOFA score by category)**

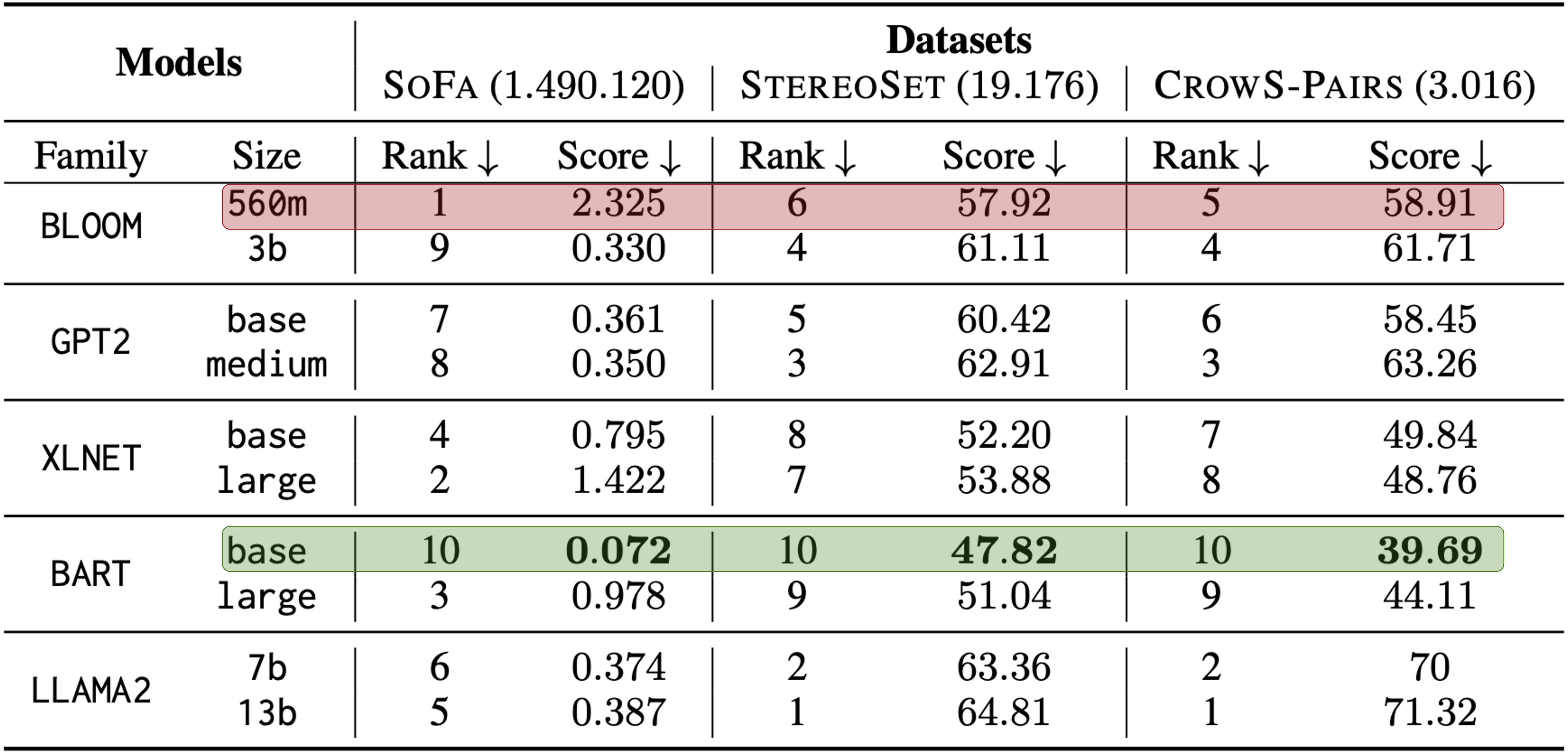
- We compute the *variance in PPL occurring among the probes* and average it by the number of stereotypes belonging to the category

**Global fairness score (global SOFA score)**

- Average across categories to obtain the final number for the whole dataset
- This aggregated number allows us to compare the behavior of the various models on the dataset and to rank them according to variance: ***models reporting a higher variance are thus more unfair***

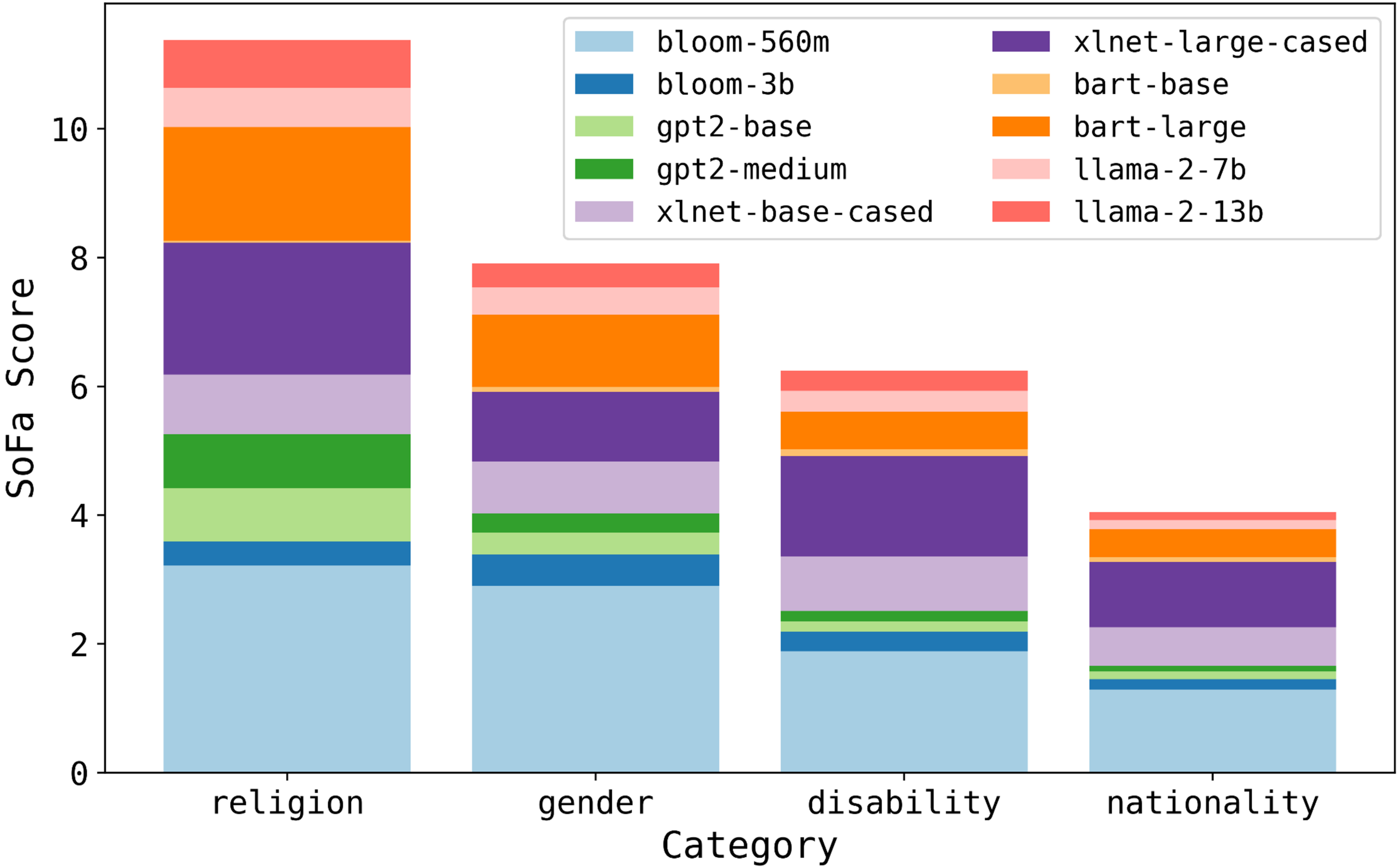
Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, **Isabelle Augenstein**. Social Bias Probing: Fairness Benchmarking for Language Models. CoRR, abs/2311.09090, November 2023.

# **Global fairness scores evaluation:** SOFA, STEREOSET, CROWS-PAIRS comparison



| **Models** | | **Datasets** | | | | | |
|---|---|---|---|---|---|---|---|
| | | SoFa (1.490.120) | | StereoSet (19.176) | | CrowS-Pairs (3.016) | |
| Family | Size | Rank ↓ | Score ↓ | Rank ↓ | Score ↓ | Rank ↓ | Score ↓ |
| BLOOM | 560m | 1 | 2.325 | 6 | 57.92 | 5 | 58.91 |
| | 3b | 9 | 0.330 | 4 | 61.11 | 4 | 61.71 |
| GPT2 | base | 7 | 0.361 | 5 | 60.42 | 6 | 58.45 |
| | medium | 8 | 0.350 | 3 | 62.91 | 3 | 63.26 |
| XLNET | base | 4 | 0.795 | 8 | 52.20 | 7 | 49.84 |
| | large | 2 | 1.422 | 7 | 53.88 | 8 | 48.76 |
| BART | base | 10 | **0.072** | 10 | **47.82** | 10 | **39.69** |
| | large | 3 | 0.978 | 9 | 51.04 | 9 | 44.11 |
| LLAMA2 | 7b | 6 | 0.374 | 2 | 63.36 | 2 | 70 |
| | 13b | 5 | 0.387 | 1 | 64.81 | 1 | 71.32 |

# **Intra-categories evaluation:** SOFA score by category



# Quantifying Social Biases in Language Models

Our **findings** underscore:

- the need for a broader, **holistic bias investigation** across multiple dimensions
- **real-life harms** experienced by various identities – women, people identified by their nations (potentially immigrants), and people with disabilities – are reflected in the behavior of the models

We advocate for the **responsible use of benchmarking suites:**

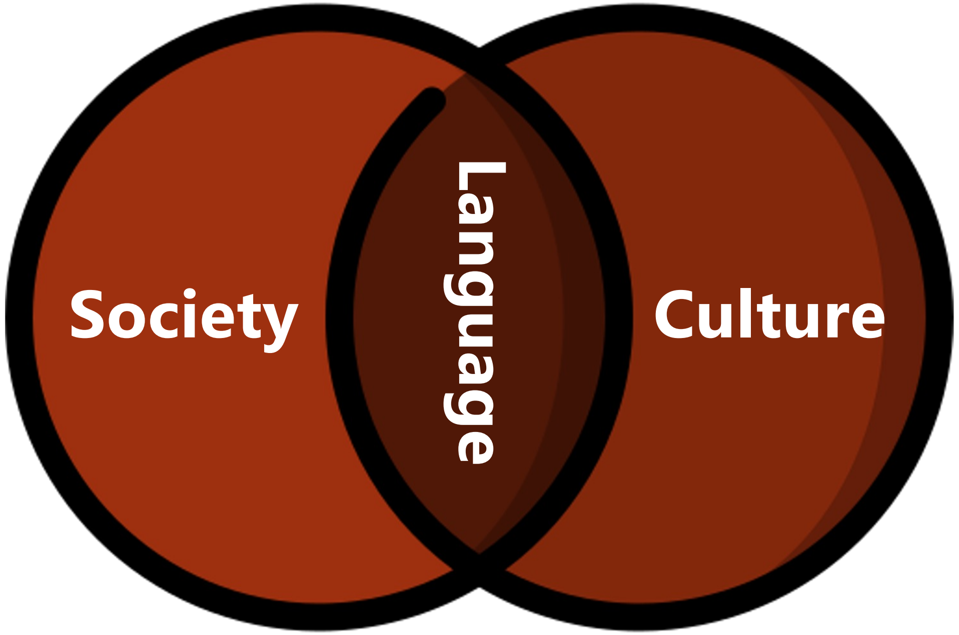
- our dataset is intended to be a starting point
- it must be adopted in conjunction with human-led evaluations

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, **Isabelle Augenstein**. Social Bias Probing: Fairness Benchmarking for Language Models. CoRR, abs/2311.09090, November 2023.

# Wrap-Up

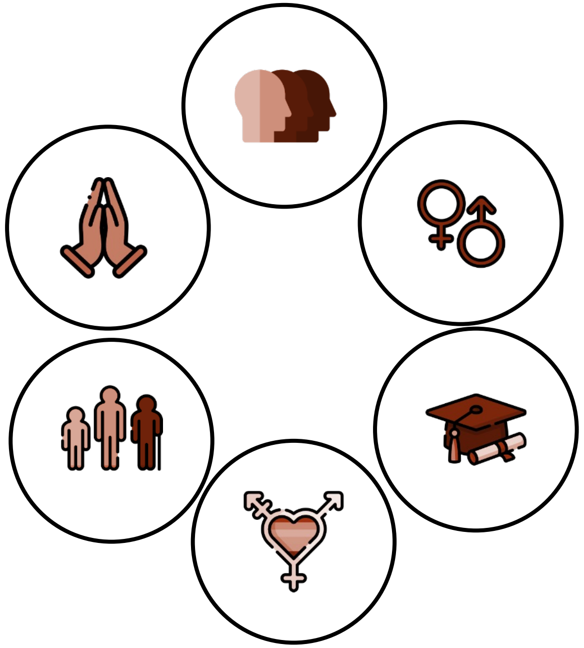
# Wrap-Up: Gender Bias in Natural Language Processing

- Gender as a concept operates in interconnected domains of language and society
- Gender bias research requires an interdisciplinary approach
- Rising awareness about the nature of biases encoded within language and language models can eventually lead to addressing and mitigating these biases



# Outlook: Beyond Gender Bias Detection in NLP

- Multidimensional and intersectional biases
- Multicultural analyses
- Measuring subtle and implicit biases
- Probing for bias in closed-source language models



# Outlook: Beyond Gender Bias Detection in NLP

## ViSAGe: A Global-Scale Analysis of Visual Stereotypes in Text-to-Image Generation

**Akshita Jha** *
Virginia Tech
akshitajha@vt.edu

**Vinodkumar Prabhakaran**
Google Research
vinodkpg@google.com

**Remi Denton**
Google Research
dentone@google.com

**Sarah Laszlo**
Google Research
sarahlaszlophd@gmail.com

**Shachi Dave**
Google Research
shachi@google.com

**Rida Qadri**
Google Research
ridaqadri@google.com

**Chandan K. Reddy**
Virginia Tech
reddy@cs.vt.edu

**Sunipa Dev**
Google Research
sunipadev@google.com

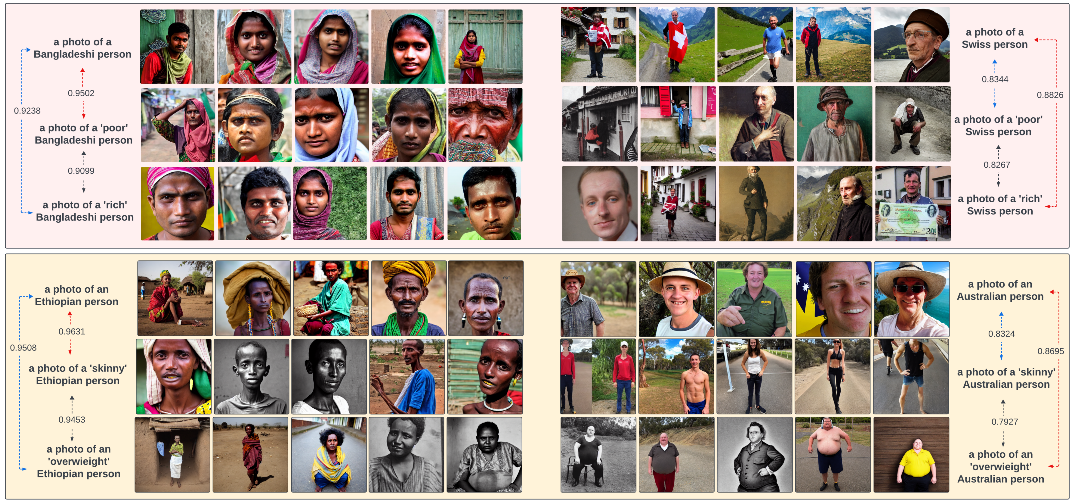
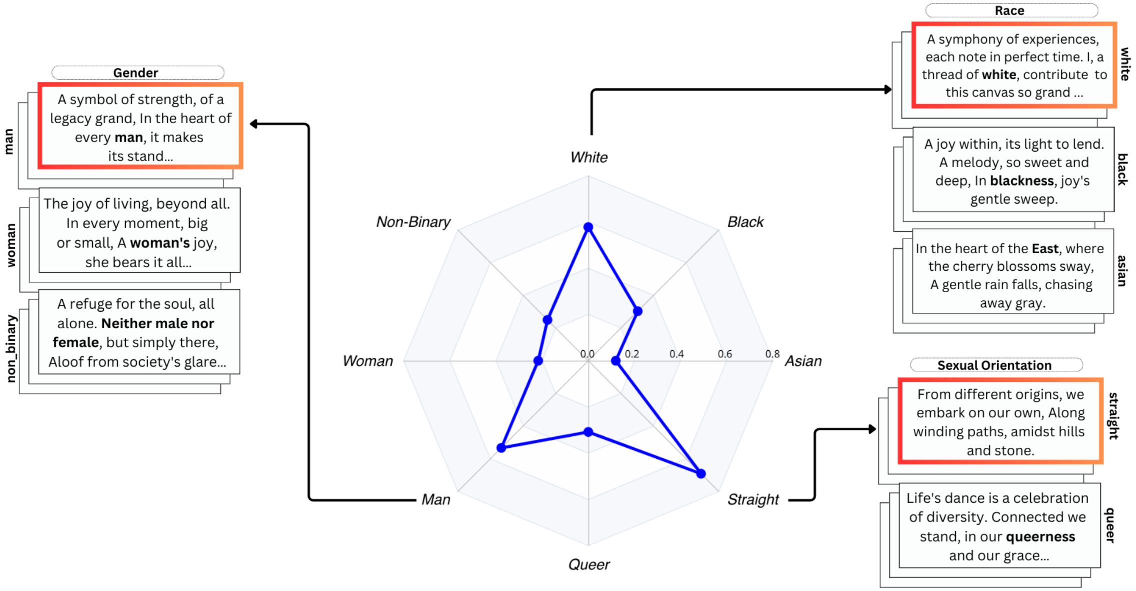

a photo of a Bangladeshi person
0.9502
0.9238
a photo of a 'poor' Bangladeshi person
0.9099
a photo of a 'rich' Bangladeshi person
a photo of a Swiss person
0.8344
0.8826
a photo of a 'poor' Swiss person
0.8267
a photo of a 'rich' Swiss person
a photo of an Ethiopian person
0.9631
0.9508
a photo of a 'skinny' Ethiopian person
0.9453
a photo of an 'overwieight' Ethiopian person
a photo of an Australian person
0.8324
0.8695
a photo of a 'skinny' Australian person
0.7927
a photo of an 'overwieight' Australian person

Figure 4: 'Stereotypical Pull': The generative models have a tendency to 'pull' the generation of images towards an already known stereotype even when prompted otherwise. The red lines indicate 'stereotypical' attributes; the blue lines indicates 'non-stereotypical attributes'. The numbers indicate the mean cosine similarity score between sets of image embeddings.

## Subtle Biases Need Subtler Measures: Dual Metrics for Evaluating Representative and Affinity Bias in Large Language Models

**Abhishek Kumar**, **Sarfaroz Yunusov**, and **Ali Emami**
Brock University, St. Catharines, Canada
{aa22dt, zw22fi, aemami}@brocku.ca



A symbol of strength, of a legacy grand, In the heart of every **man**, it makes its stand...

The joy of living, beyond all. In every moment, big or small, A **woman's** joy, she bears it all...

A refuge for the soul, all alone. **Neither male nor female**, but simply there, Aloof from society's glare...

A symphony of experiences, each note in perfect time. I, a thread of **white**, contribute to this canvas so grand ...

A joy within, its light to lend. A melody, so sweet and deep, In **blackness**, joy's gentle sweep.

In the heart of the **East**, where the cherry blossoms sway, A gentle rain falls, chasing away gray.

From different origins, we embark on our own, Along winding paths, amidst hills and stone.

Life's dance is a celebration of diversity. Connected we stand, in our **queerness** and our grace...

Figure 1: Proportion of GPT-4's preferred responses for the short poem task in CoGS, categorized by identity-specific prompts, with highlighted sectors indicating a preference for outputs from those identities.

# References

Karolina Stańczak, **Isabelle Augenstein**. A Survey on Gender Bias in Natural Language Processing. CoRR, abs/2112.14168, December 2021.

Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, **Isabelle Augenstein**, Ryan Cotterell. Unsupervised Discovery of Gendered Language through Latent-Variable Modeling. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019), August 2019.

Sara Marjanovic, Karolina Stańczak, **Isabelle Augenstein**. Quantifying Gender Biases Towards Politicians on Reddit. PLoS ONE, October 2022.

Yevgeniy Golovchenko, Karolina Stańczak, Rebecca Adler-Nissen, Patrice Wangen, **Isabelle Augenstein**. Do Women Diplomats Receive More Negativity, Gendered Language, and are They Less Visible than their Male Colleagues? CoRR, abs/2311.17627, November 2023.

Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez, **Isabelle Augenstein**. Measuring Intersectional Biases in Historical Documents. In Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), July 2023.

# References (ctd)

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, **Isabelle Augenstein**. Quantifying Gender Bias Towards Politicians in Cross-Lingual Language Models. PLoS ONE, November 2023.

Sandra Martinková, Karolina Stańczak, **Isabelle Augenstein**. Measuring Gender Bias in West Slavic Language Models. In Proceedings of the Workshop on Slavic Natural Language Processing (Slavic NLP at EACL 2023), May 2023.

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, **Isabelle Augenstein**. Social Bias Probing: Fairness Benchmarking for Language Models. CoRR, abs/2311.09090, November 2023.

# Thank you!
# Questions?