Examining Stereotypes in News Articles

Damin Zhang and Julia Taylor Rayz
Department of Computer and Information Technology
Purdue University, USA
zhan4060, jtaylor1@purdue.edu

Abstract
Gender biases or stereotypes have been studied in short text and manually labeled corpora, but little work has been done in real-world unlabeled text corpora like news articles. This work investigated news articles from mainstream U.S. media outlets ranging from 2013 to early 2020. We used structural topic modeling to estimate gender prevalent topics, compared the results with topic modeling embedding, and incorporated qualitative and quantitative analyses to understand the appearance of gender stereotypes in news articles from each gender group. The structural topic modeling results showed that gender prevalent topics align with stereotypical representations of either gender group and media outlets with imbalanced gender distribution are more influential on stereotypical representations. The topic modeling embedding results support prior results and provide additional information supporting the conclusion.

Introduction
As machine learning applications are ubiquitous in humans’ daily lives, the fairness issues within these models attract many social concerns. While online news articles have become a major source of people obtaining information, the implicit human biases or stereotypes within the natural language could affect the reader’s perception of the target. In the domain of fairness issues in artificial intelligence, there is a growing amount of work focused on the detection of human biases or stereotypes with annotated datasets (Dixon et al. 2018; Zhao et al. 2018; Nadeem, Bethke, and Reddy 2020). Many recent works explored the human biases or stereotypes with manually selected corpora but had different definitions of biases or stereotypes depending on the tasks. Moreover, few connect well with related fields like Social Sciences (Blodgett et al. 2020). We follow the proposed social cognitive theory (Bandura, Ross, and Ross 1963) and choose news articles to reflect the real-world setting. According to Bandura, Ross, and Ross (1963), media can be used to discover the human behavioral patterns, which can be treated as a medium of generating and reproducing biases or stereotypes (Foucault 1990). Such biased or stereotypical representations would spread and affect the reader’s understanding of either gender group through natural language (del Teso-Craviotto 2006).

Hypothesizing that the news articles may contain gender biases or stereotypes, we explored if the author’s gender contributes to the stereotypical representation of either gender group. Our work seeks to answer are the gendered topics associated with the stereotypical representations. Instead of identifying the existence of an expected bias or stereotype, we use Structural Topic Modeling (Roberts, Stewart, and Tingley 2019) and author’s gender as a covariate, combined with quantitative and qualitative analyses to understand the stereotype appearance by exploring the association between the author’s gender and the implicit stereotypes within news articles. We also use topic modeling embedding technique on the same corpora for comparisons. Following Butler (2011), people’s understanding of gender categories is constructed through repeated behaviors over time rather than simply the static gender characteristics of women and men, and therefore words associated with women and men contributed to reproducing the ideas of femininity and masculinity even though they do not necessarily reflect real-world settings. Our results show certain topics are more prevalent in either gender group and correspond to those that are considered to be stereotypical representations of females and males (Ellemers 2018). We do not include non-binary gender in this research as we did not find suitable word representations for this group. This work takes a small step forward analyzing biases or stereotypes in a context-dependent setting with topic modeling and extends further by using embeddings on unlabeled textual data.

Related Work
The two general directions of analyzing biases or stereotypes in models are either focusing on the model architecture or the data quality.

Measuring and mitigating biases or stereotypes in embedding space have been tried. Word Embedding Association Test (WEAT) was proposed to measure fairness at the word level (Caliskan, Bryson, and Narayanan 2017). Sentence Encoder Association Test (SEAT) proposed by May et al. (2019) measures fairness at the sentence level. Additional frameworks have been proposed as the extensions of WEAT, for example, contextual word-level embedding association test (Tan and Celis 2019) and Word Embedding...
Fairness Evaluation (WFE) (Badilla, Bravo-Marquez, and Pérez 2020). Bias detection and mitigation have been tried in embedding spaces (Bolukbasi et al. 2016; Zhao et al. 2018), but later works indicate not all types of biases are removed from the embeddings (Gonen and Goldberg 2019; Bordia and Bowman 2019; Nissim, van Noord, and van der Goot 2020).

Some works explored the feasibility of topic modeling technique in analyzing biases or stereotypes. Kozlowski et al. (2020) use topic modeling to investigate gender bias in English corpora. Different types of topic modeling frameworks have been tried for further analysis, for example, semi-supervised topic modeling (Andrzejewski and Zhu 2009) has been tried on English and Swedish corpora (Devlinney, Björklund, and Björklund 2020) and Davidson and Bhattacharya (2020) examine racial bias and hate speech with structural topic modeling (STM) (Roberts, Stewart, and Tingley 2019).

Methodology

We used structural topic modeling on unlabeled corpora with author gender as the covariate to examine the gender-prevalent topics in order to understand if gender contributes to the stereotypical representations. For each media outlet, we estimated an STM to identify the latent topics. We also split each media outlet corpus into two gender subsets and trained two topic modeling embedding using Top2Vec (Angelov 2020). Due to usually accepted stereotypes in the news articles, we hypothesize that certain topics will be gender prevalent.

Data

We used an online dataset containing 2.7 million English news articles from 26 American publications in various domains, including politics, business, and entertainment, ranging from 2013 to early 2020. The dataset contains 221,728 unique author names.

We used the following rules to achieve high confidence in the author’s gender identification:

- Exclude names that do not list a person’s name, for example, “Staff”.
- Exclude gender-neutral names, for example, “Alex” or “Kyle”.

These rules exclude both affiliation indicators and gender-neutral names, resulting in 140,155 unique names.

In order to identify the author’s gender, we used the gender-guesser package in Python which takes first names and predicts possible gender. The package has six output values based on its prediction confidence: unknown, androgynous, male, female, mostly_male, and mostly_female. For this work, we only accepted author names with “male” or “female” predictions and discarded author names with other output values, such as unknown, androgynous, mostly_male, or mostly_female. After the gender prediction, we retained 47,713 unique author names with high confidence in the prediction.

Our goal is to discover and understand how the author’s gender affects biased or stereotypical content. While preprocessing the documents, we removed unnecessary information Date, Year, Month, Day, Title, and URL, and kept Article content, Author names, and Author gender. All documents were tokenized and each token was converted into lowercase, punctuations and numbers were removed, and the words were stemmed. For the purpose of training STM, we also removed stopwords to improve the performance of the model. We did not remove stopwords while training topic modeling embedding with Top2Vec since they are equidistant from all topics in almost all documents.

Structural Topic Model

While estimating STM, we set the model lower threshold to 5 to filter out words that occur in fewer than 5 documents, as infrequent terms provide less useful information (Davidson and Bhattacharya 2020). In order to mitigate the influence that media outlets with more documents have on the ones with fewer documents, we tested different values from 10 to 30 to find the appropriate number of topics $K$ with STM. We utilized the $stm$ R package to evaluate the results and found that the model converges well with high semantic coherence, low residuals overdispersion, and better prediction performance when $K = 20$. Therefore, we used the top 20 topics from each model.

We used estimateEffect from the $stm$ package to estimate the gender prevalence of each generated topic within each corpus. We denoted $K$ as the number of topics and $M$ as the number of media outlets. Additionally, we defined $K \times M$ matrix $A$ to represent the topic association level in each model by dividing the weights from topic prevalence result into three categories:

$$A_{k,m} = \begin{cases} weak & \text{if } 0 \leq \frac{\mu_{k,m}}{\mu_{m}} < 1 \\ moderate & \text{if } 1 \leq \frac{\mu_{k,m}}{\mu_{m}} < 2 \\ strong & \text{if } 2 \leq \frac{\mu_{k,m}}{\mu_{m}} \end{cases}$$

$$\mu_{k,m} = X_m \gamma_k$$

where $\mu_{k,m}$ is the topic prevalence value of topic $k$ within median outlet $m$. $X_m$ denotes the prevalence metadata and $\gamma_k$ denotes the metadata weights. Weak association indicates the number of articles about a certain topic from the group is less or equal to the other group. Moderate association indicates the number is slightly more than the other group. Strong association indicates the number is much more than the other group.

We also defined a $K \times M$ matrix $E$ to represent the topic-gender prevalence in each model:

$$E_{k,m} = \begin{cases} male & \text{if } \mu_{k,m} < 0 \\ female & \text{if } \mu_{k,m} > 0 \end{cases}$$

where negative $\mu_{k,m}$ value means the topic was male author prevalent and positive $\mu_{k,m}$ value means the topic was female author prevalent.
**Topic Modeling Embedding**

We used Top2Vec (Angelov 2020) to train topic modeling embedding for each media outlet gender group. Top2Vec uses Distributed Bag of Words (DBOW) Doc2Vec (Le and Mikolov 2014) to generate semantic space where the distance between vectors indicates the semantic similarity. The generated semantic space consists of joint word and document vectors to form a continuous representation of topics where the dense areas indicate the topics.

We used the same threshold as STM to filter out infrequent terms and the same K value to estimate topics. Unlike estimating with the whole media outlet corpus in STM, we divided each corpus into two subsets based on the author’s gender and trained two topic modeling embeddings respectively.

**Results and Discussion**

We explored the generated topics from STMs and topic modeling embeddings and found that the male author prevalent topics are different from the female author prevalent topics. Additionally, we examined if imbalanced gender distribution affects the stereotypical representations of each media outlet.

**Structural Topic Model**

Since the corpus contains a large number of news articles from various fields, we expected to see a wide range of topics. Of the 36 generated topics from all media outlets, We were able to identify 20 topics by checking if they align with the news articles: International Politics; U.S. Politics; Law and Crime; Immigration; Market and Finance; Economy; Education; Technology; Health and Medication; Science; Social Media; Sports; Art; Movie; Music; Literature; Environment; Family and Relationships; Entertainment; and Beauty.

Among all media outlets, we found that topics strongly associated with the male author were grouped in the stereotypical masculine sphere, including International Politics (state, russian, taliban, iran, nuclear), U.S. Politics (presid, republican, democrat, senat, elect), Law and Crime (police, kill, shoot, death, court), and Sports (team, game, player, season, playoff). We also found that such associations were not exclusive as the female authors also contributed a lot of views in International Politics, U.S. Politics, and Law and Crime. On the other hand, female authors had fewer outputs compared to male authors in Sports, and this showed a stereotypical representation of Sports being masculine (as evidenced by previously reported unequal treatment of male and female athletes).

The female author prevalent topics were mostly grouped within the stereotypical feminine sphere: Music (song, album, record, band, lyric), Literature (literati, poem, book, harri, potter), Family and Relationships (mom, dad, home, kid, home), Entertainment (show, episod, podcast, disney, comedil), and Social Media (youtub, snapchat, instagram, stream, post). Similarly, we found that Music and Family and Relationships were not exclusive to the female group as some male authors also had outputs about this topic, but such cases were too rare to constitute quantifiable impact.

In order to gain the knowledge of the overall topic prevalence of all media outlets, we combined the topic association level matrix $A$ and the topic-gender prevalence matrix $E$. The co-occurrence of $A_{k,m}$ and $E_{k,m}$ was shown in Figure 1 with each cell indicating the frequency of certain topic-gender prevalence occurs at certain topic association levels among all media outlets. We found that female authors were less influential on political topics. In International Politics, the female group was shown to have more weak associations and less strong and moderate associations, meaning that although many news articles published by female authors mentioned International Politics, the same information was also covered by male authors. As there were more strong and moderate associations between male authors and International Politics, the male group was shown to be more sensitive to international affairs and was able to have more distinct points of view than the female group. Such difference was mitigated in U.S. Politics. Male authors were shown to have less strong associations than female authors, but they had more moderate and weak associations than the female group. This indicates that male authors led the U.S. Politics discussion, but there was a growing number of female authors to publish distinct views in this field that contributed to mitigating the stereotypical representation. Similarly, the difference was also minimized in Law and Crime. Overall, we found that International Politics, U.S. Politics, and Law and Crime aligned with stereotypical masculine representations, and such representations were less significant in U.S. Politics and Law and Crime.

We also identified certain topics that did not align with the expected stereotypical representations. As being closely mentioned with political topics, Immigration was expected to show a similar stereotypical structure. However, the results showed female authors led the discussion as they had more moderate and weak associations than male authors. The topic words of Immigration indicated the topic correlations mitigate the stereotypical representation. Among the news articles that mentioned male author prevalent Immigration topic, the contents were about U.S.-Mexico border control and conflicts, whereas female author prevalent Immigration news articles mostly discussed not only the border control, but also VISA, refugees, immigration regulations, and social impact. The top words extracted from female author prevalent Immigration topics contained terms like family, child, home, provide a potential explanation that Immigration correlated with Family and Relationships, a female author prevalent topic, so that explain female authors had more outputs than male authors on Immigration. In the domain of business, we found that both gender groups had relatively similar views of Market and Finance, but female authors had more outputs about the related topic Economy. News articles content about Market and Finance mentioned more about companies and regulations, whereas the content about Economy concerned more with social events, for example, the employment rate. This showed that Market and Finance correlated with law topics, where the latter was male author prevalent. Female authors, on the other hand, were
Figure 1: Topic-gender prevalence at each association level in structural topic models and topic modeling embeddings. Colors are used to differentiate association levels. Darker colors indicate more occurrences and lighter colors mean less occurrences.

Concerned more about social events, thus had more outputs on Economy than on Market and Finance. Technology was a male author prevalent topic, but the contributions from female authors significantly change the stereotypical representation by providing more distinct views than male authors. Additionally, unlike most male author prevalent technology topics concerned with technical aspects, for example, autonomous vehicles and robots, female authors also discussed the applications of technology in psychology and education. Figure 2 shows a topic prevalence sample from TechCrunch, a media outlet that mostly reports technology-related news. Topics 4, 9, 10, 12, and 16-20 were all about Technology, and female author prevalent topics 16-20 were distinct from male author prevalent topics 4, 9, 10, and 12 based on the context. Such a pattern could also be verified by the byproduct of Technology, Social Media, where female authors produced the majority of outputs in this field.

Further analysis examined if imbalanced gender distribution contributes to the stereotypical representations within news articles. We used the topic-gender prevalence matrix $E$ and selected the 5 most stereotypical topics from the prior analysis: International Politics; U.S. Politics; Law and Crime; Entertainment; and Social Media. Results were shown in Table 1. Previous analysis showed that International Politics, U.S. Politics, and Law and Crime were male author prevalent topics, but we found that there were more female authors who talked about International Politics than male authors in 4 out of 11 media outlets, more female authors writing about U.S. Politics than male authors in 7 out of 17 media outlets and more female authors published articles about Law and Crime than male authors in 5 out of 10 media outlets. Of Social Media, there were more articles from male authors than from female authors in 3 out of 7 media. We hypothesized the imbalanced gender distribution of each media could affect the topic-gender prevalence as imbalanced gender distributions would favor one gender group over the other gender group on topic prevalence.

Additionally, we looked at the gender prevalence of each media to assist further interpretation on topic-gender distribution. Results was shown in Table 2. 85.7% of the media with more female authors discussed U.S. Politics were female author prevalent; 50% of the media outlets with more female authors talked about International Politics were female author prevalent; 60% of the media with more female authors covered Law and Crime were female author prevalent. Similarly, we found all the media outlets having more male authors who wrote about Social Media were male author prevalent. Additionally, 70% of the media outlets that were male author prevalent in U.S. Politics were male author prevalent. In the media that were female author prevalent in U.S. Politics, 66.7% of them were female author prevalent. In the media that were female author prevalent in Social Media, 75% of them were female author prevalent. This suggested a correlation between topic-gender prevalence and gender distribution in each media outlet: there were 15 out of 20 media outlets showed that male author prevalent distribution would have male prevalence on the selected topics whereas female author prevalent distribution would have female prevalence or equal prevalence on the selected topics. Specifically, in male author prevalent media outlets, most articles about U.S. Politics, International Politics, and Law and Crime were published by male authors. Similarly in female author prevalent media outlets, most articles about Entertainment and Social Media were published by female authors.
Table 1: Topic-gender prevalence on selected topics in each media outlet in structural topic models. “B” stands for Both gender groups, “−” means the topic did not appear. “M” and “F” are male group and female group respectively. 1 represents U.S. Politics. 2 is International Politics. 3 shows Law and Crime. 4 corresponds Entertainment. 5 stands Social Media.

| Media            | 1 | 2 | 3 | 4 | 5 |
|------------------|---|---|---|---|---|
| The Hill         | M | M | - | M | - |
| Reuters          | M | B | F | - | - |
| N.Y.T.           | F | F | F | - | - |
| People           | M | - | M | - | F |
| CNN              | M | M | M | F | - |
| Vice             | M | - | - | - | M |
| Mashable         | F | - | M | M | - |
| Refinery 29      | F | - | - | B | - |
| TechCrunch       | - | - | M | - | M |
| The Verge        | F | - | - | B | F |
| Vox              | F | B | - | - | - |
| Axios            | F | M | F | - | - |
| Buzzfeed News    | B | M | F | - | F |
| Fox News         | M | B | M | M | F |
| New Republic     | M | F | - | - | - |
| Business Insider | M | F | F | F | - |
| New Yorker       | F | F | - | F | - |
| Wired            | M | - | - | - | M |

Table 2: Gender prevalence by media outlet

| Media            | Male | Female |
|------------------|------|--------|
| The Hill         | 10   | 9      |
| Reuters          | 11   | 8      |
| N.Y.T.           | 7    | 8      |
| People           | 10   | 8      |
| CNN              | 8    | 10     |
| Vice             | 7    | 5      |
| Mashable         | 8    | 9      |
| Refinery 29      | 10   | 5      |
| TechCrunch       | 11   | 8      |
| The Verge        | 5    | 12     |
| Vox              | 9    | 9      |
| Axios            | 7    | 11     |
| Buzzfeed News    | 7    | 9      |
| Fox News         | 9    | 10     |
| New Republic     | 8    | 7      |
| Business Insider | 9    | 8      |
| New Yorker       | 6    | 9      |
| Wired            | 11   | 8      |

Topic Modeling Embedding

In addition to the 20 topics found in structural topic model results, we were able to agree upon another 2 topics in the topic modeling embedding results: Fashion and Society. Figure 1 contains the results from topic modeling embedding.

We found the topics mostly mentioned by male authors were International Politics (syria, missile, brexit, icbm, korean), Law and Crime (gunshot, harassment, suspect, lawsuit, court), Market and Finance (ipo, capital, firms, sponsorship, nasdap), Sports (championship, nfl, team, score, coach), and U.S. Politics (republican, nominee, hillary, bernie, trump). These findings aligned with previous STM results, indicating a stereotypical masculine sphere consisting of International Politics, Law and Crime, Market and Finance, Sports, Technology, and U.S. Politics.

The topics mostly discussed by female authors also aligned with previous findings: Economy (inflation, unemployment, wages, tax, debt), Entertainment (comedy, star, broadway, drama, celebrities), Family and Relationships (dad, son, friends, mom, puppies), Immigration (shelters, border, patrol, customs, refugee), Literature (poem, dickinson, novelist, protagonist, narrator), Music (album, song, lyrics, symphony, chords), Social Media (snapshot, retweet, instagram, post, messenger). These topics were mostly linked with the stereotypical feminine sphere. Additionally, topic modeling embedding also identified another two latent topics that were strongly associated with female authors: Fashion (prada, balenciaga, designer, celine, collaborations) and Society (equality, gender, supremacist, racism, minorities).

We also observed the same correlation between topic-gender prevalence and gender distribution in the topic modeling embedding results. Among the male author prevalent media outlets, 70% were male author prevalent on U.S. Politics and 57% were male author prevalent on International Politics. Similarly, 87.5% of the female author prevalent media outlets were also female author prevalent on Entertainment and 71.4% of the female author prevalent media outlets were female author prevalent on Social Media.

Compared with STM, topic modeling embedding is able to discover more latent topics, for example, Fashion and Society were not identified in STM results. Moreover, it provides more hidden information for analysis by visualizing the semantic embedding. A sample semantic embedding of Business Insider is shown in Figure 3 where a clear boundary divides the female author prevalent topics from male author prevalent topics.

Conclusion

We explored how to use structural topic modeling and topic modeling embedding to understand stereotypes in large unlabeled corpus. With the author’s gender as the covariate, we were able to examine how certain topics are disproportionately associated with each gender group. Additionally, we used topic modeling embedding to compare the results. We found that both female and male author prevalent topics aligned with the stereotypical representation of either gender group. We further found that gender imbalance powers the stereotypical representations. The results from topic modeling embedding validated our analysis. Overall, this work demonstrates that unsupervised learning methods can be useful to uncover and assist to understand the implicit stereotypes in unlabeled text corpus.

Limitations

First, we used gender-guesser package to predict the author’s gender, kept gender-specific names, and removed gender-neutral names based on the results. Although we only ac-
cepted the most confident predictions, it is possible that the algorithm could misclassify. Second, it is likely that removing gender-neutral names removes valuable information from the analysis. Finally, we have only examined the gender stereotypes and it is possible that other types of bias exist in the dataset. We hope future works can address these limitations.

References

Andrzejewski, D., and Zhu, X. 2009. Latent dirichlet allocation with topic-in-set knowledge. In Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, 43–48.

Angelov, D. 2020. Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Badilla, P.; Bravo-Marquez, F.; and Pérez, J. 2020. Wefe: The word embeddings fairness evaluation framework. In IJCAI, 430–436.

Bandura, A.; Ross, D.; and Ross, S. A. 1963. Imitation of film-mediated aggressive models. The Journal of Abnormal and Social Psychology 66(1):3.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (technology) is power: A critical survey of” bias” in nlp. arXiv preprint arXiv:2005.14050.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. arXiv preprint arXiv:1607.06520.

Bordia, S., and Bowman, S. R. 2019. Identifying and reducing gender bias in word-level language models. arXiv preprint arXiv:1904.03035.

Butler, J. 2011. Gender trouble: Feminism and the subversion of identity. routledge.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186.

Davidson, T., and Bhattacharya, D. 2020. Examining racial bias in an online abuse corpus with structural topic modeling. arXiv preprint arXiv:2005.13041.

del Teso-Craviotto, M. 2006. Words that matter: Lexical choice and gender ideologies in women’s magazines. Journal of pragmatics 38(11):2003–2021.

Devinney, H.; Björklund, J.; and Björklund, H. 2020. Semi-supervised topic modeling for gender bias discovery in english and swedish. In GeBNLP2020, COLING’2020–The 28th International Conference on Computational Linguistics, December 8-13, 2020, Online, 79–92. Association for Computational Linguistics.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 67–73.

Ellemers, N. 2018. Gender stereotypes. Annual review of psychology 69:275–298.

Foucault, M. 1990. The history of sexuality: An introduction, volume i. Trans. Robert Hurley. New York: Vintage 95.

Gonen, H., and Goldberg, Y. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862.

Kozlowski, D.; Lozano, G.; Felcher, C. M.; Gonzalez, F.; and Altszyler, E. 2020. Gender bias in magazines oriented to men and women: a computational approach. arXiv preprint arXiv:2011.12096.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In International conference on machine learning, 1188–1196. PMLR.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456.

Nissim, M.; van Noord, R.; and van der Goot, R. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. Computational Linguistics 46(2):487–497.

Roberts, M. E.; Stewart, B. M.; and Tingley, D. 2019. Stm: An r package for structural topic models. Journal of Statistical Software 91(1):1–40.

Tan, Y. C., and Celis, L. E. 2019. Assessing social and intersectional biases in contextualized word representations. arXiv preprint arXiv:1911.01485.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:1804.06876.