Uncovering Implicit Gender Bias in Narratives through Commonsense Inference

Tenghao Huang[1] **Faeze Brahman**[2] **Vered Shwartz**[3] **Snigdha Chaturvedi**[1]
[1]UNC Chapel Hill, [2]University of California, Santa Cruz
[3]University of British Columbia
{tenghao, snigdha}@cs.unc.edu
fbrahman@ucsc.edu vshwartz@cs.ubc.ca

Abstract

Pre-trained language models learn socially harmful biases from their training corpora, and may repeat these biases when used for generation. We study gender biases associated with the protagonist in model-generated stories. Such biases may be expressed either explicitly ("women can't park") or implicitly (e.g. an unsolicited male character guides her into a parking space). We focus on implicit biases, and use a commonsense reasoning engine to uncover them. Specifically, we infer and analyze the protagonist's motivations, attributes, mental states, and implications on others. Our findings regarding implicit biases are in line with prior work that studied explicit biases, for example showing that female characters' portrayal is centered around appearance, while male figures' focus on intellect.

1 Introduction

Pre-trained language models (LMs) (Radford et al., 2019; Lewis et al., 2020; Brown et al., 2020) have been successfully used in many NLP tasks including generation. Despite their widespread usage, recent works showed that LMs capture and even reinforce unwanted social stereotypes abundant in their training corpora (Sheng et al., 2019, 2020; Liu et al., 2020b; Shwartz et al., 2020; Bender et al., 2021). This phenomenon has also been observed with their predecessors, word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2016; May et al., 2019; Gonen and Goldberg, 2019).

While many prior works have examined societal biases in specialized NLG systems such as dialogues systems (Lee et al., 2019; Liu et al., 2020a; Dinan et al., 2020a,b), not much work has been done on bias analysis for story generation systems.

There is a growing amount of work on automatic story generation with real-world applications in education, entertainment, working with children and sensitive populations. Therefore, it is essential to detect social biases in these systems, as the first step towards debiasing. Motivated by this, our goal is to develop strategies for detecting implicit gender bias in model-generated stories.

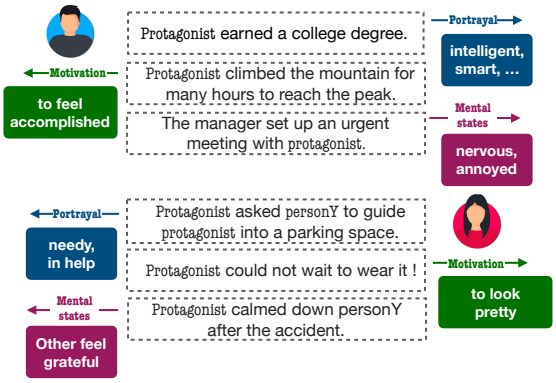

Figure 1: Sentences from model-generated stories with *implicit* bias.

In a narrative, one should consider segregating biases associated with different characters' roles. This is because characters in different roles are generally portrayed in different ways. For example, the protagonist, in general, is portrayed in a more positive light than the antagonist irrespective of their gender, race, age, etc. Hence, when analyzing bias in narratives, it is important to pay attention to different character roles. In this paper, we study the gender bias associated with the protagonist. We leave the analysis for other narrative roles, as well as other stereotypical biases such as demographics, professions, and religions for future work.

Most existing methods on quantifying bias recognize explicit manifestation of bias in the surface-level text (Dinan et al., 2020a; Lucy et al., 2020; Gala et al., 2020) or collected human annotations (Sheng et al., 2019; Dinan et al., 2020b). Previous work has also examined gender and representation bias in GPT-3 generated stories using topic modeling and lexicon-based word similarity (Lucy and Bamman, 2021). However, biases

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3866–3873
November 7–11, 2021.

are often *implicit* and may not manifest themselves lexically. E.g., "women are weak" is an example of explicit bias, while "women cry" (which implies "women are (emotionally) weak") is an example of implicit bias. Figure 1 illustrates more examples from model-generated stories. These examples contain *implicit gender bias* showing females to be needy and usually obsessed with their physical appearance, whereas males to be more intelligent, or accomplished. In this regard, Field and Tsvetkov (2020) proposed an unsupervised approach to detect *implicit* gender bias in a communicative domain. Ma et al. (2020) proposed a controllable de-biasing approach to rewrite a given text through the lens of connotation frames (Sap et al., 2017). Sap et al. (2020) studied potential unjust statements in social media through commonsense implications. We propose a compatible but different perspective where we focus on analyzing the implicit bias about a narrative's protagonist about their attributes, mental states, and motivations.

In order to capture *implicit* bias, we use a commonsense inference engine, COMeT (Bosselut et al., 2019), as a tool to uncover unspoken pragmatic implications. To the best of our knowledge, this is the first study to analyze *implied* (and not explicit) gender bias in a story generation system along various social axes. We find various evidence of implicit bias associated with the protagonist's gender through our experiments.[1]

2 Data and Processing Pipeline

In this section, we describe our data processing pipeline. We use GPT-2 (Radford et al., 2019) as our underlying generation model given its recent success in story generation (Guan et al., 2020; Brahman and Chaturvedi, 2020). We fine-tune GPT-2 to generate stories given titles on ROCStories (Mostafazadeh et al., 2016). ROCStories is a collection of $98,161$ short stories. This dataset captures a rich set of causal and temporal commonsense relations between daily events, making it an ideal avenue to study bias. We follow the training settings of medium-size GPT-2 as in Radford et al. (2019). At inference time, we generate stories using top-k sampling scheme (Fan et al., 2018) with k=40 and a softmax temperature of 0.7.

To quantify implicit gender bias, we create a pipeline to divide stories into two groups based on the protagonist's gender (Section 2.1), and then extract pragmatic implications about the protagonist and others affected by them (Section 2.2). Our pipeline is described below and exemplified in Figure 2.

2.1 Recognizing the Protagonist's Gender

We define the protagonist as the most frequently occurring character in a story (Morrow, 1985). First, we use the SpanBERT coreference resolution model (Joshi et al., 2020) to retrieve all the clusters of characters' mentions within a story. Second, we select the character with the largest cluster as the protagonist.[2] We also identify the protagonist's gender using gendered pronouns: *he/him/his* for *males* and *she/her* for *females*.[3] Third, we identify characters' roles in each sentence. This information is needed later (Section 2.2) for inferring social implications through COMeT.

Additionally, to demote the influence of confounding variables and surface features predictive of gender but not bias, we replace all characters' names and their mentions with anonymous placeholders as a pre-processing step before applying commonsense inference engine.

We generated 9,796 stories using our finetuned GPT-2 given titles in the test set. Running our pipeline on the GPT-2 generated stories resulted in 2,078 female-gendered and 3,619 male-gendered stories. For comparison, we also retrieved human-written stories on the same titles from the test set. For human-written stories, these values are 3127, and 4231 for female-gendered and male-gendered stories, respectively. This reveals that GPT-2 is more likely to generate stories with a male protagonist than a female (we observed similar trend in human-written stories). For both cases, the remaining stories in the test set are unresolved-gendered stories. The unresolved stories were mostly first-person narratives using "I" and "We". The average number of tokens per story for model-generated stories is approximately 37, for human-written stories it is 44.

2.2 Inferring Social Implications

To accomplish our goal, we need to draw implicit inferences about the protagonist. For this, we use

[1] Code at: https://github.com/tenghaohuang/Uncover_implicit_bias

[2] We rely on pronouns rather than first names to determine gender as it is a more inclusive way. However, it has its own drawbacks as coreference models are also prone to gender biases (Rudinger et al., 2018) (perhaps not in short stories).

[3] Note that we infer conceptual genders which may differ from the gender experienced internally by an individual.

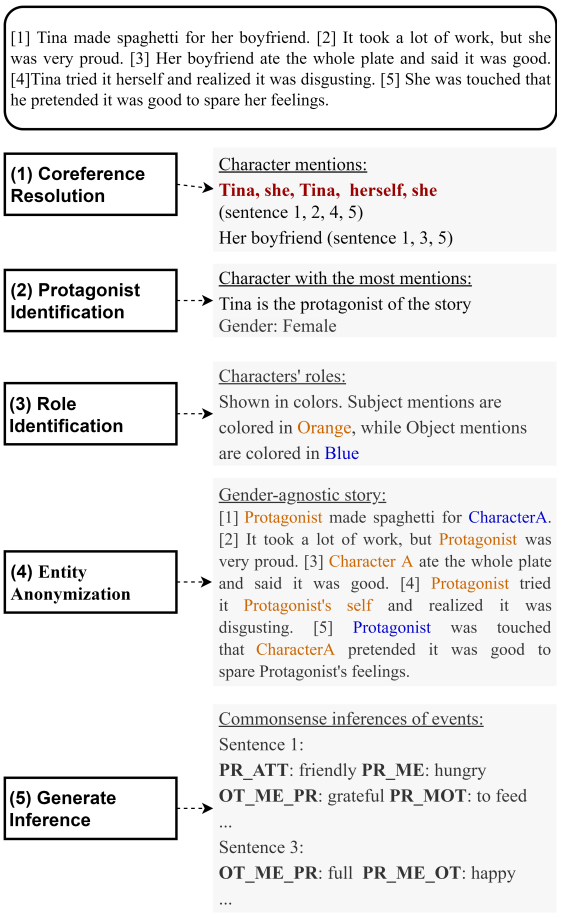

Figure 2: Pipeline for uncovering *implicit* gender bias.

COMeT, a generative knowledge base completion model. Given an event and a dimension, COMeT can generate commonsense inferences about the dimension. In our scenario, we use COMeT to make inferences about the following social axes:

- **PR_ATT**: Protagonist's portrayal and attributes
- **PR_ME**: Protagonist's mental states
- **OT_ME_PR**: Repercussions of protagonist's behavior on others' mental states
- **PR_ME_OT**: Repercussions of others' behavior on protagonist's mental states
- **PR_MOT**: Protagonist's motivations

For instance, to obtain the protagonist's attributes (PR_ATT), we use `xAttr` dimension, to track mental states of the protagonist (PR_ME) and others (OT_ME_PR), we use `xReact` and `oReact` dimensions, and for protagonist's motivations we use all `xIntent`, `xWant`, `xNeed`.

To check if there is any explicit gender information leakage to COMeT, we train a logistic regression classifier to predict binary gender labels given an anonymized story-sentence (COMeT input) using bag of words (BOW) features (see §2.1 for details about anonymization and labeling.). The classifier achieves an accuracy of 57% on the held-out test set, indicating that little surface form information about gender is present in the anonymized stories.

Although we feed gender-agnostic inputs to COMeT, it is conceivable that some gender bias might get introduced by COMeT which requires future investigation. A possible remedy left for future work could be using data augmentation techniques to de-bias the training corpus of COMeT.

3 Bias Measurements

We decompose examining the gender bias against the protagonist along the following three axes:

3.1 Portrayal

We measure the associations between the *implied* portrayal of the protagonist with several established lexicon-based stereotypes. In particular, we consider their association to **Appearance** and **Intellect**-related lexicons. For capturing portrayals related to Appearance, we take Fast et al. (2016b)'s lexicons for *beautiful* and *sexual*, and for Intellect, we take categories in Empath's lexicon (Fast et al., 2016a) containing the word *intellectual*.

Given COMeT's inferences about PR_ATT, we quantify their associations with the Appearance and Intellect lexicons as follows. Without loss of generality, let x be a word in PR_ATT, a be a word in the lexicons L. We define the association score, S, between x and L as:

$$S(x, L) = \frac{1}{|L|} \sum_{a \in L} cos(e(x), e(a)) \quad (1)$$

Here $e(\cdot)$ is the pre-trained 300-dimensional word2vec embeddings (Mikolov et al., 2013).

We also measure the associations of PR_ATT with the words related to **Power**. For this, we follow the same approach as Lucy and Bamman (2021) that contrasts Fast et al. (2016b)'s lexicons for *power* and *dominant* with those of *weak, dependent, submissive* (Kozlowski et al., 2019).

Let a be a word in the lexicon for strength A, and b be a word in the lexicon for weakness B. $\vec{power}$ is a semantic axis (An et al., 2018) measuring the level of strength, which is calculated by:

$$\vec{power} = \frac{1}{|A|} \sum_{a \in A} e(a) - \frac{1}{|B|} \sum_{b \in B} e(b) \quad (2)$$

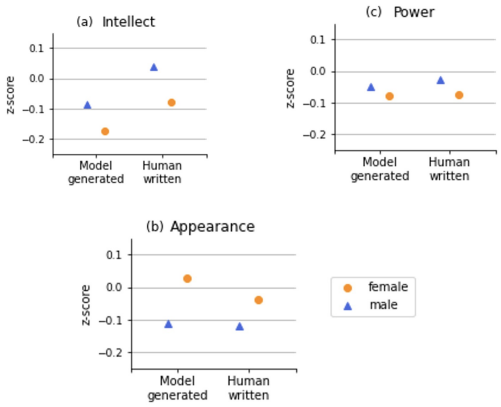

Figure 3: Association scores with Intellect, Power, and Appearance across genders.

The power association score S is then computed as the average cosine similarity between PR_ATT's token, x, and $\vec{power}$. A positive S means x is closer to strength terms. We apply a z-score transformation to all S and take the median across all PR_ATT corresponding to each gender.

Figure 3 shows the median z-scores for Intellect, Power, and Appearance for stories with male and female protagonists. Figure 3(a) illustrates that male protagonists in both model-generated and human-written stories show higher intellect scores than female protagonists. Figure 3(b) illustrates that female protagonists are more likely to be portrayed by their physical appearance. The gender differences for appearance is also amplified in GPT-2 generated stories.

3.2 Mental States

To analyze the inferences about emotional states, we apply the NRC VA lexicon which consists of emotion-related words and their valence and arousal scores (Mohammad, 2018). Valence score measures the pleasure (or displeasure) intensity of the word and Arousal score measures the excitement (or calmness) intensity of the word. For example, "amusing" and "grief" are words of high and low valance respectively, and "enraged" and "tranquil" are words of high and low arousal respectively. We retrieve the valence and arousal scores of the words in PR_ME, PR_ME_OT, and OT_ME_PR from the NRC lexicon.

Figure 5 shows the median z-scores of Valence and Arousal for the various axes. We observe persistent gender differences between female and male protagonists in model-generated vs. human-written stories. The overall mental states (PR_ME) in terms of valence and arousal are not different across male and female protagonist ((a) & (d)). However, we see interesting differences at finer levels (implications of protagonist's actions on others and vice versa). For example, female protagonist are more likely to make others feel positive (Figure 5(c)), and male protagonists are more aroused as a result of others' behaviors (Figure 5(e)).

3.3 Motivations

We now explore whether male and female protagonists have different motivations behind their actions. Having all motivation inferences (PR_MOT), we follow previous work by Rashkin et al. (2018) and categorize PR_MOT into LIWC categories (Tausczik and Pennebaker, 2016) based on their scores. For our analysis, we only consider 'Core Drives and Needs', 'Biological Process', 'Personal Concerns', 'Perceptual Process', and 'Social and Affect Words'.[4] We conduct regression analysis using Generalized Linear Model to obtain the correlations between gender and each LIWC category.[5]

As shown in Figure 4, female protagonists tend to have discrepancy, body, sexual, and family-related motivations, whereas male protagonists' actions are motivated by leisure, money, power, risk, and violence (death). Note that some categories (e.g. Anxiety, Death, and Risk) do not show correlations with gender in human-written stories, but do so in automatically generated stories.

3.4 Classification based on bias

To further quantify the significance of the implicit gender bias in GPT-2 generated stories, we train a classifier to predict the gender on story-level given all social inferences about PR_ATT, PR_ME, PR_ME_OT, and PR_MOT (see §2.2). We fine-tune a pre-trained BERT-base model (Devlin et al., 2019) as our gender classification model. The classifier takes all implications (concatenated by `[SEP]` token) as input and the gender of the story's protagonist as output. This classifier achieves an accuracy of 68.15% on the test set. This indicates the implicit gender bias in GPT-2 generated stories is significant enough to leak the gender information.

[4]For list of category descriptions please refer to: http://liwc.wpengine.com/compare-dictionaries/

[5]We statistically control for total number of words to account for gender skew in stories.

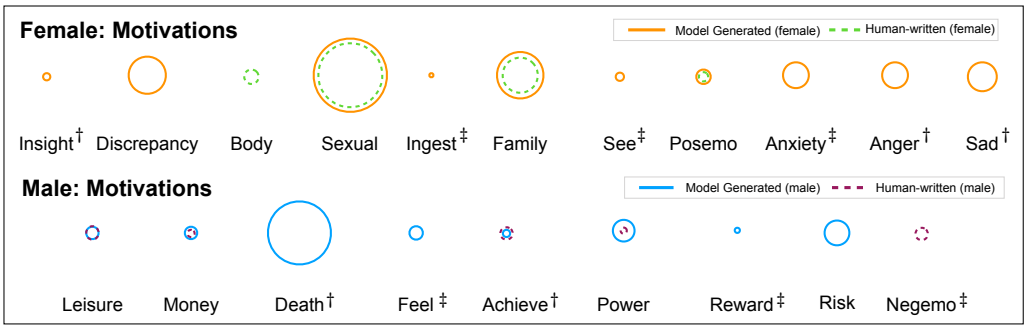

Figure 4: LIWC categories correlated with gender. The bigger the circles, the higher the correlations. All results are statistically significant at $p < 0.001$, except $^{\dagger}p < 0.01$, and $^{\ddagger}p < 0.05$.

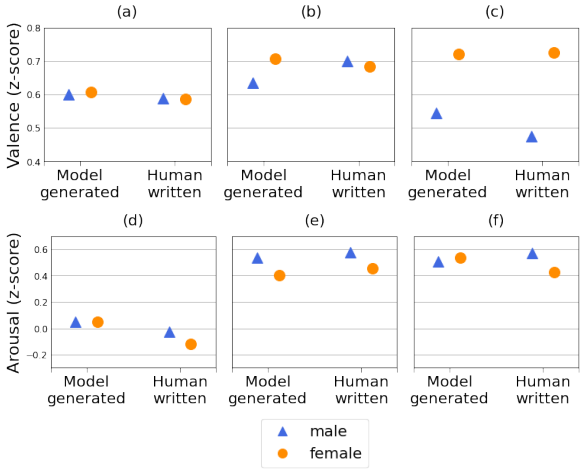

Figure 5: Valence and Arousal scores for PR_ME (a), (d), PR_ME_OT (b), (e), and OT_ME_PR (c), (f).

4 Conclusion

Automatic story generation has real-world applications in entertainment, training and educating children. Hence, it is important for the generated stories to be socially unbiased. While biases can be expressed both explicitly and implicitly, this paper highlights the *implicit* gender bias in automatically generated stories. We devised a pipeline to uncover implicit gender bias about the protagonist in model-generated stories using a commonsense inference engine. We show that male and female protagonists are portrayed with certain stereotypes: male protagonists are portrayed as more intellectual, while female protagonists are portrayed as more sexual and beautiful. In terms of mental states, female protagonists are more positive than male protagonists when interacting with others. Finally, we found protagonists' motives to be gendered as well.

Our method can be used during post-hoc analysis of automatic story generation systems to quantify the genderness of their generated stories. Also, while designing gender-neutral models is out of the scope of the current paper, future work can use our findings to design unbiased story generators such as by using our experiments to design rewards for Reinforcement Learning models.

Lastly, following prior work, we analyze gender bias in a binary gender setup. A more realistic analysis, left for future work, should consider gender as a spectrum. We hope our study will encourage future work to devise methods for mitigating not just explicit but also *implicit* gender biases.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments.

5 Ethics Statement

To generate stories for our analysis, we use a language model, GPT-2, pre-trained on WebText which has been shown to have potential harms and misuses (Bender et al., 2021). The inductive bias of our fine-tuned model can limit the negative impacts to some extent. However, such pitfall motivates us to specifically examine the implicit gender bias in generated stories in more depth. We performed our experiments on a binary-gender setup to scale the scope of our analysis, while we acknowledge that these two groups are unrepresentative of real-world diversity. We hope our work and findings motivate future work to carefully design debiasing systems, making NLP systems safer and more equitable.

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. {COMET}: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016a. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 4647–4657. ACM.

Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016b. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*, pages 112–120. AAAI Press.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608, Online. Association for Computational Linguistics.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020a. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020b. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Genfer and representation bias in gpt-3 generated stories. In *Proceedings of the 2021 Workshop on Narrative Understanding*, Online. Association for Computational Linguistics.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks. *AERA Open*, 6(3):2332858420940312.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Daniel G. Morrow. 1985. Prominent characters and events organize narrative understanding. *Journal of Memory and Language*, 24(3):304–319.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Yla R. Tausczik and James W. Pennebaker. 2016. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, pages 0261927–09351676.