RtGender: A Corpus for Studying Differential Responses to Gender

Rob Voigt[a], David Jurgens[b], Vinodkumar Prabhakaran[a], Dan Jurafsky[a], Yulia Tsvetkov[c]
[a]Stanford University, [b]University of Michigan, [c]Carnegie Mellon University
robvoigt@stanford.edu, jurgens@umich.edu, vinodkpg@stanford.edu, jurafsky@stanford.edu, ytsvetko@cs.cmu.edu

Abstract

Like many social variables, gender pervasively influences how people communicate with one another. However, prior computational work has largely focused on linguistic gender difference and communications *about* gender, rather than communications *directed to* people of that gender, in part due to lack of data. Here, we fill a critical need by introducing a multi-genre corpus of more than 25M comments from five socially and topically diverse sources tagged for the gender of the addressee. Using these data, we describe pilot studies on how differential responses to gender can be measured and analyzed and present 30k annotations for the sentiment and relevance of these responses, showing that across our datasets responses to women are more likely to be emotive and about the speaker as an individual (rather than about the content being responded to). Our dataset enables studying socially important questions like gender bias, and has potential uses for downstream applications such as dialogue systems, gender detection or obfuscation, and debiasing language generation.

Keywords: gender-annotated corpora, gender difference, gender bias, discourse, computational social science

1. Introduction

Language is a means for the construction of identity and social categories like gender; social issues such as gender bias, in turn, often take form in language. Linguistic datasets have been used both to debunk gender-biased myths — for example, contrary to stereotype women are not actually more talkative than men (Mehl et al., 2007) — and to identify social issues. For instance, women[1] journalists reach a smaller audience in terms of social media impressions (Matias and Wallach, 2012), and traditional gender stereotypes and unbalanced gender representation occur even in contemporary stories and movies (Fast et al., 2016; Sap et al., 2017).

Large datasets are particularly of use in this context due to the complex nature of differential responses to gender. However, previous computational work on language and gender has focused mainly on language *about* or *portraying* persons of a particular gender (Wagner et al., 2015; Flekova et al., 2016; Agarwal et al., 2015).

We thus present a large multi-genre dataset of online communication to enable research in a category of gender difference understudied in computational work: **responses to gender** in language. These include posts and talks labeled for the gender of the source,[2] along with comments given in response to the source texts. We collect such data from a variety of contexts, including:

- **Facebook (Politicians)**: Responses to Facebook posts from members of the U.S. House and Senate
- **Facebook (Public Figures)**: Responses to Facebook posts from other public figures, e.g., television hosts, journalists, and athletes
- **TED**: Responses to presentations from TED speakers
- **Fitocracy**: Responses to posts about fitness progress
- **Reddit**: Responses to Reddit comments across a variety of subreddits

These diverse datasets offer multiple perspectives on responses to gender. The first two sources (from Facebook and TED) represent the "broadcast" case, in which source texts (online posts and speech) from a small number of individuals (experts, authorities, and other public figures) receive a large number of responses which the source is unlikely to read and a discussion between the source and the responder is unlikely to continue. The second two (Fitocracy and Reddit) represent the "personal" case in which the responses are individualized, the source and responder may know one another and have an ongoing interaction afterwards.

2. Responses to Gender

Here we aim to encourage research on *responses to gender*. Contrasting with language *about* or *portraying* a given gender which address abstract representations of social categories, responses to gender are directed towards an individual person. We know that social characteristics of the addressee influence linguistic behavior; existing computational work has shown, for instance, that the gender of the interlocutor influences lexical choices of a speaker in spoken and written interactions (Boulis and Ostendorf, 2005; Jurgens et al., 2017; Prabhakaran and Rambow, 2017).

Looking at responses to gender also allows us to consider the important social issue of gender bias. Since important forms of bias (e.g., dehumanization or treating a person as their social category) often happen at the level of individual responses, responses to gender are an understudied but critical lens for studying gender bias.

The issue is related to that of abusive language (Xu et al., 2012; Clarke and Grieve, 2017), though often gender bias takes a less overt form than straightforward abuse. Social issues like gender bias are often not just about hostility but also behaviors such as stereotype-reinforcing benevolence (Eagly and Mladinic, 1989; Glick and Fiske, 1996;

[1]Throughout this paper we use the terms "woman" and "man" as labels for gender in preference to "female" and "male" since the latter terms are more commonly used as markers of sex.

[2]We use "source" to refer to the producer of the text being responded to (online posts and talk videos), and "responder" for the producer of the comment or response, regardless of its format.

	Dataset	Source Individuals		Source Text Count	Response Count	Response Word Count
BROADCAST	Facebook (Politicians)	M: 306	W: 96	399,037	13,866,507	376,114,950
	Facebook (Public Figures)	M: 41	W: 64	117,811	10,667,500	123,753,913
	TED Talks	M: 1,071	W: 349	1,671	190,425	15,549,984
PERSONAL	Fitocracy	M: 52,432	W: 47,498	318,535	318,535	6,606,087
	Reddit	M: 19,010	W: 11,116	1,453,512	1,453,512	44,537,612

Table 1: Basic statistics about the subcorpora within *RtGender*.

Jha and Mamidi, 2017). Nevertheless, biased responses to social categories like gender can lead to marginalization (Sue, 2010) and negatively impact a person's self-esteem and ability through mechanisms such as stereotype threat (Spencer et al., 1999). Perhaps most related to our work, Fu et al. (2016) analyze questions directed at men and women tennis players, finding that questions directed at men tend to be more about the game while questions directed at women are more likely to stray to topics about their appearance and off-court relationships. Tsou et al. (2014) similarly find comments on TED talks are more likely to be about the presenter than the content if the presenter is a woman.

In looking at responses to gender in our datasets (§3.), we note that some instances of gender bias may be overt, such as direct references to stereotypes ("Cool story babe, now make me a sandwich" - Facebook comment to television anchor Megyn Kelly) or inappropriate comments about physical appearances ("wow she is very sexy..."- TED comment to researcher Rachel Botsman); however, the larger problem is a subtle one in part because much of social bias is also not overt but rather implicit (Greenwald et al., 1985; Nosek et al., 2011). More commonly, many biases are exhibited through small but systematic differences in language which normally go unnoticed but, when viewed in aggregate, reveal large scale patterns in behavior towards a particular gender.

In the next section, we present **RtGender** – a corpus of responses to gender, compiled according to the following desired characteristics. First, it would be sufficiently large to allow for uncovering the subtle type of differentiation and bias mentioned above. Second, it would cover multiple genres and linguistic contexts to facilitate generalizable results. Third, it would allow for content and topic in the source texts to be controlled as much as possible so that researchers could know people are responding to the same types of sources, especially given existing research demonstrating gender-correlated clustering behavior by topic (Argamon et al., 2003; Bamman et al., 2014). Fourth, it would contain source texts from both authority figures and everyday persons, to facilitate the analysis of such subtle phenomena as implicit bias towards women authority figures (Rudman and Kilianski, 2000), while allowing for comparison to non-authority figures. Finally, it would ideally have gender labels for both the sources and the responders, to allow for gender-interaction analysis of interesting psychological phenomena like the propagation of self-favorable gender stereotypes (Rudman et al., 2001).

3. RtGender Datasets

We present five distinct datasets regarding *responses to gender* which fulfill many of the aforementioned desiderata. These data represent a variety of interactional contexts and relationships between the source and the responders.

The Facebook and TED "broadcast" datasets presented here contain many instances of responses to people in positions of authority or renown (politicians, topic experts, television personalities), and so can be analyzed with prior knowledge about the power differential between the source and the responders. The Fitocracy and Reddit "personal" datasets will allow research to contrast responses to gender in the public domain with more one-on-one interactions. In these datasets having interactional dyads of post-response also opens possibilities for studying normativity, for instance by asking whether comments on non-normative posts are more likely to exhibit elements of bias.

3.1. Facebook

Our largest dataset is comprised of top-level comments on Facebook posts from public figures, scraped from their public pages. We only include top-level comments (that is, comments directly responding to the post) to reduce the influence of comment-internal discussion so each comment is a response directly to the original poster. Each post is associated with the page of its relevant public figure, and includes metadata such as whether the post was text-only or included an image, video, or link.

The posts and responses in question are all public; however, to protect the anonymity of Facebook users in our dataset we remove all identifying user information as well as Facebook-internal information such as User IDs and Post IDs, replacing these with randomized ID numbers. Therefore users whose comments appear multiple times in our dataset may be compared, but without revealing their identity. We also only report commenter first names, since this is less identifying but still allows for running gender-identification algorithms. As a baseline for convenience we provide masculine/feminine ratios for these first names from Bergsma and Lin (2006).

We collect posts and their associated top-level comments for the categories of speakers described below. In each case we find the page for the speaker with a novel method for finding gender-labeled speakers from Wikipedia. Specifically, our method takes as input a Wikipedia category page such as `https://en.wikipedia.org/wiki/Category:American_female_tennis_players`, and for each name listed runs a search for public pages using Facebook's Graph API. If an exact match for the name appears in the top three results, and the category of the page matches a relevant category (for instance, "Public Figure" or "Athlete" in the case of female tennis players), and their gender is listed, and the page is "verified" with Facebook, we accept it as a member of that category and scrape the

Category	Example
Remarks on Appearance	Hot presenter.
Patronizing Tone	Stick to actually talking about the tenets of the topic and defer the blah blah blah to the politicians alone..
Doubting Expert Knowledge	I always thought the first approach to scientific study was to examine all evidence that dissproves an hypothesis.
Self-promotion is Perceived Negatively	After watching this, I know much more about Rachel Pike and what she does than the actual subject matter.

Table 2: Examples of categories of comments displaying potential forms of gender bias from the TED dataset. These categories were primarily observed in comments to women presenters.

relevant posts and comments.

Politicians. This subset contains all posts and associated top-level comments for all 412 current members of the United States Senate and House who have public Facebook pages meeting the requirements outlined above. This dataset inherently includes a strong control for content, since members of Congress tend to be talking about the same sorts of topics; each Congressperson is also labeled with their party affiliation to further facilitate controlling for cross-party stylistic and topical differences.

Public Figures. Beyond Congress, we consider other US public figures from the realms of journalism, fiction writing, television, film, and athletics. This subset contains posts and associated top-level comments for 105 such public figures, currently drawn from the following sets of Wikipedia categories:

- American television news anchors, American television journalists, American television talk show hosts, Political analysts
- American film actresses, American male film actors, American television actresses, American male television actors
- American male tennis players, American female tennis players, Olympic track and field athletes of the United States
- 21st-century American novelists

3.2. TED

TED Talks are influential videos from experts on a variety of topics ranging from education, business, science, tech and creativity.[3] The TED website also allows viewers to post comments in response to each video, which provides us an opportunity to study how these responses are impacted by the gender of the expert presenters. We include a dataset scraped from the TED website of 190,425 labeled for presenter gender. Gender labels were initially drawn from Mirkin et al. (2015) and the remaining labels were done manually. Talks that consisted solely of a dance or music performance, or talks presented by more than one speaker were excluded.

This domain has been previously explored in NLP: we know that TED presenters are more commonly male, and videos of talks by male speakers are more viewed and liked on YouTube (Sugimoto et al., 2013). Furthermore, considering responses to gender, Tsou et al. (2014) note that commenters are more emotional when the presenter is a woman. However, existing sources of this data such as `https://www.idiap.ch/dataset/ted` are not labeled for presenter gender.

3.3. Fitocracy

Fitocracy[4] is a social media fitness website in which users log and discuss their fitness-related activities. We include a dataset of 318,536 "status updates" and their corresponding top-level comments from Fitocracy which users have posted about their progress. We include only the first comment after a post because comments are not nested, so discussions can diverge as following comments may quickly become responses to previous comments; however, the first post is necessarily in direct response to the original post. Building upon the observations of Fu et al. (2016) on gender differences in questions posed to tennis players, we view Fitocracy as an ideal dataset for examining how gender stereotypes around fitness and sports play out in everyday interactions. In this dataset we have confident self-reported labels for the gender of most posters and commenters; over 91% of users of Fitocracy self-report both gender and age on their profile pages, and we include this information in the dataset.

3.4. Reddit

Posts on Reddit are a common source of data for computational linguistic analysis; in this corpus, we include a dataset of 1,453,512 Reddit post-response pairs for which we know the gender of the source poster. The data was gathered by finding gender-indicating flairs used on different subreddits (e.g,. "male" on /r/AskMen). For each of these users, we then find all of their posts in other subreddits and collect the first response to each post - as in other contexts, we take only the first response to guarantee it is directed towards the source poster. We also tag the responder for gender when we have that data available, which occurs for about 9.2% of our examples.

This dataset covers a wide variety of subreddits, so while the sources of our gender tags are from a relatively limited domain, ultimately researchers can control for content substantially by sampling the dataset in particular subreddits of interest.

4. Analysis and Challenges

In this section we discuss a preliminary qualitative and quantitative analysis regarding differential responses to gender in our new datasets, designed to illustrate the kind of studies they enable.

[3] `https://www.ted.com/talks`

[4] `https://www.fitocracy.com/`

Responder Gender		*Source Gender* WOMAN	MAN
	WOMAN	girl, gorgeous, can, if, yay, exercise, it, find, girlie, to, love, we, feel, ", do, each, mama, site, or, yoga, walk, help, started, go, healthy	HAPPY_EMOJI, thank, thanks, ..., haha, ?, no, well, problem, :p, mister, !, pleasure, follow, props, prop, lol, welcome, :d, bomb, handsome, very, for, my, course
	MAN	HAPPY_EMOJI, !, you, welcome, thank, your, follow, great, pp, pleasure, hope, love, back, following, very, luck, you're, girl, well, :d, are, awesome, thanks, for, young, beautiful, smile, hi, fun	man, bro, mate, dude, ., brother, buddy, [NUMBER], brah, bench, ,, bud, shit, 0x0, yeah, i, squat, press, lifts, sets, fuck, gains, 0kg, chest, last, strength, week, guy, this, ohp

Table 3: Top 30 words in comments in the Fitocracy dataset by log-odds based on the gender of the commenter and original poster.

Dataset	Source Gender Prediction Accuracy
Facebook (Political)	63.9%
Facebook (Public Figures)	80.3%
TED Talks	80.5%
Fitocracy	57.7%
Reddit	53.5%

Table 4: Cross-validation accuracy across contexts at predicting the gender of the source from the text of their post/talk.

Table 2 presents some qualitative examples of TED comments directed towards women presenters that exhibit possible gender bias. Some are overt, such as remarks on appearance, but others are more subtle. For authors of each gender, Table 3 gives the top 30 words in Fitocracy responses most associated with the responder's gender, computed using the weighted log-odds method of Monroe et al. (2008). The word preferences of responders show a substantial gender-correlated signal in this data. Men commenting on posts by men use many close terms of informal address ("bro," "dude," "brother," "buddy") and likewise for women commenting on posts by women ("girl," "girlie," "mama"). Cross-gender comments, however, are more emotive, with prominent use of emoticons, emoji, and exclamation points, as well as more playful and interactive language (talk of "following" each other and use of second person pronouns) and discussion of the addressee's appearance, e.g., "beautiful" (M→W) and "handsome" (W→M).

Each dataset in turn presents a unique challenge for researchers. The Facebook data is large and noisy: the comments are relatively unmoderated and may also be responding to photos and URLs in the source posts, rather than just the textual content of the post itself. The TED talks exemplify the challenge of separating gender difference from topical choice, since selection bias on the part of the TED organizers means there are more talks from men and talks from women are more likely to be about gendered topics. For Fitocracy, as Table 3 shows, the language used is often very positive overall, so a computational definition of bias must be able to also capture benevolent differential treatment. The Reddit data covers a very broad spectrum of topical content, and in the majority of subreddits gendered flair is not visible so the signal for differential responses to gender is much more subtle.

One important axis of variation across the linguistic environments of these contexts is to consider how differently men and women tend to speak in that context; a simple way to quantify this is to ask how well a predictive model can distinguish source posts written by men versus women. Table 3.3. shows ten-fold cross-validation accuracies of a simple unigram logistic regression model at predicting the gender of the source from the text in the source post. In the case of TED, accuracy is given at predicting gender from a sample of lines in the source transcript. Notice the wide diversity across contexts. While gender differences in the "personal" Reddit and Fitocracy contexts are relatively minimal, gender difference of the source is amplified in the "broadcast" contexts where posts by men and women are highly separable even with a simple unigram model.

5. Relevance and Sentiment Annotations

Our pilot analyses revealed that the RtGender datasets have the potential to offer interesting insights on differential responses to gender across diverse domains. To expand the possible range of questions that may be asked of this data, we conducted a crowdsourced annotation effort on a sample of the responses across our datasets.

Inspired by the annotation task for TED talk comments proposed by Tsou et al. (2014), we labeled over 15,000 post-response pairs with annotations for the relevance of the response to the source and the sentiment of the response. For this task we asked crowd workers on Amazon Mechanical Turk to read a post-response pair and mark whether it was relevant to the CONTENT ONLY, POSTER ONLY, CONTENT AND POSTER, or if it was IRRELEVANT. In the case of comments on TED talks since there was no "original post" we provided a list of the talk's keywords to help participants determine its relevance. We then asked about the sentiment of response (POSITIVE, NEGATIVE, MIXED, or NEUTRAL), regardless of its relevance. Our annotation interface is shown in Figure 1.

Crowd workers were paid $0.20 for completing one run of 5 post-response pairs. To control for annotation quality, on each run for a random one of the pairs we replaced the response with a snippet of text with a known expected response. These were drawn from the following sources:

- Random sentences from articles in the New York Times in 2007 (expected response: IRRELEVANT)

Figure 1: Screenshot of our relevance and sentiment annotation interface.

Dataset	Annotated Examples
Facebook (Politicians)	3,872
Facebook (Public Figures)	2,884
TED Talks	2,648
Fitocracy	2,900
Reddit	2,728

Table 5: Quantity of available post-response pairs labeled with relevance and sentiment annotations for each dataset.

- Random phrases with known polarity from the Stanford Sentiment Treebank (Socher et al., 2013) (expected response: IRRELEVANT or POSITIVE/NEGATIVE, respectively)
- Poster/speaker-directed utterances automatically generated with a heuristic method to have known polarity (expected response: relevant to POSTER ONLY or CONTENT+POSTER; known POSITIVE/NEGATIVE polarity as appropriate). Examples:
  - you are just fantastic, believe in yourself! (POSITIVE)
  - Stop trying. You are so garbage! (NEGATIVE)

Any runs for which these control questions were answered incorrectly were discarded, constituting 11.9% of total runs.

We performed basic analyses on these annotations to better understand the types of future research they might enable. Firstly, we ran a set of mixed-effects models predicting aspects of response relevance and sentiment, with gender as a fixed effect and dataset context as a random effect. Our overall results replicate Tsou et al. (2014), finding that in general responses to women were more likely to be about the source poster or speaker as an individual ($b=0.20$, $p<0.01$) and were more emotive (having non-neutral sentiment) ($b=0.17$, $p<0.01$) than responses to men. Interestingly, sentiment in responses to women was higher across the board; whether this represents "benevolent sexism" or genuine positive sentiment is an interesting and complex topic for future research.

However, the contexts represented by each dataset acted very differently. When we restrict the above analyses to only the "personal" Fitocracy and Reddit contexts we find no gender-based differences in response relevance ($b=0.01$, $p=0.87$), and the magnitude of the emotiveness difference is greatly reduced ($b=0.11$, $p=0.046$). This finding suggests a potential powerful effect of social distance in exacerbating gender bias, in line with classic social psychological findings on how group diffusion of responsibility can lead to increased dehumanization (Bandura et al., 1975).

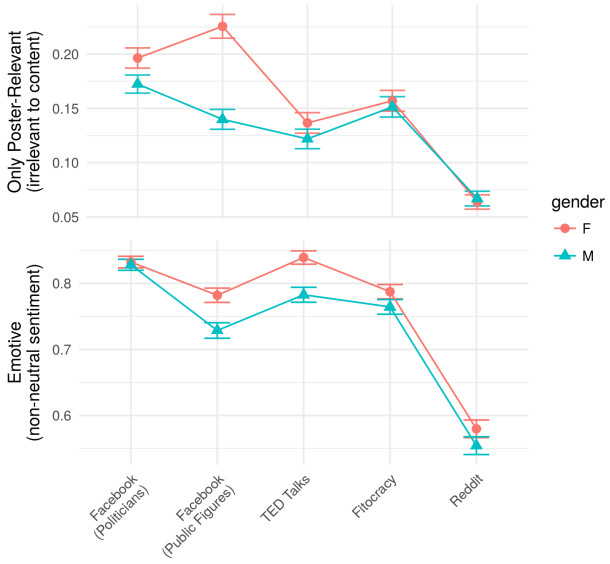

Figure 2: Cross-context characteristics of the responses per the relevance and sentiment annotations in RtGender.

6. Conclusion

Gender is a performative social phenomenon in which individual behavior is often shaped – subtly, over a lifetime – by the responses to that behavior (Lakoff, 1973; Butler, 1990). To encourage computational study in this area, in this paper we presented five large datasets in a corpus called RtGender that capture differential responses to gender online in a variety of genres, contexts, and social roles of the interacting participants, and publicly available for research purposes.[5] We labeled a sample of the responses in the corpus with annotations for relevance and sentiment, and gave some initial analyses of the data and resulting annotations. We found qualitative and quantitative evidence for gender bias in the responses, suggesting a need for future work in this area that we hope this corpus will facilitate. By studying responses to gender we can learn a great deal about the social construction of gender and other social categories in general.

Acknowledgments

This work was supported by the Stanford Data Science Initiative and the National Science Foundation through award IIS-1526745. The first author gratefully acknowledges the support of the Stanford Interdisciplinary Graduate Fellowship.

7. Bibliographical References

Agarwal, A., Zheng, J., Kamath, S. V., Balasubramanian, S., and Dey, S. A. (2015). Key Female Characters in Film Have More to Talk About Besides Men : Automating the Bechdel Test. *uman Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 830–840.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal writ-

[5] https://cs.cmu.edu/~ytsvetko/rtgender/

ten texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23(3):321–346.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender and variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, apr.

Bandura, A., Underwood, B., and Fromson, M. E. (1975). Disinhibition of aggression through diffusion of responsibility and dehumanization of victims. *Journal of Research in Personality*, 9(4):253–269, dec.

Bergsma, S. and Lin, D. (2006). Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pages 33–40, Morristown, NJ, USA. Association for Computational Linguistics.

Boulis, C. and Ostendorf, M. (2005). A quantitative analysis of lexical differences between genders in telephone conversations. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, (June):435–442.

Butler, J. (1990). *Gender Trouble - Feminism and the Subversion of Identity*, volume 53. Routledge.

Clarke, I. and Grieve, J. (2017). Dimensions of Abusive Language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10.

Eagly, A. H. and Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men, dec.

Fast, E., Vachovsky, T., and Bernstein, M. (2016). Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM '16)*, number Icwsm, pages 112–120.

Flekova, L., Carpenter, J., Giorgi, S., Ungar, L., and Preo\ctiuc-Pietro, D. (2016). Analyzing Biases in Human Perception of User Age and Gender from Text. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854.

Fu, L., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Tie-breaker: Using language models to quantify gender bias in sports journalism. *Proceedings of the IJCAI Workshop on NLP Meets Journalism*.

Glick, P. and Fiske, S. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*.

Greenwald, A. G., Krieger, L. H., Greenwaldt, A. G., Kriegertt, L. H., Eberhardt, J., Kang, J., Newkirk, T., and Rachlinski, J. (1985). Implicit Bias: Scientific Foundations. *California Law Review*, 47(445):35–421.

Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. (2017). Writer profiling without the writer's text. In *Proceedings of the 9th International Conference on Social Informatics (SocInfo)*.

Lakoff, R. T. (1973). *Language and Woman's Place*.

Matias, J. N. and Wallach, H. (2012). Working Paper : Modeling Gender Discrimination by Audiences of Online News.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., and Pennebaker, J. W. (2007). Are women really more talkative than men? *Science (New York, N.Y.)*, 317(5834):82, jul.

Mirkin, S., Nowson, S., Brun, C., and Perez, J. (2015). Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal, September. Association for Computational Linguistics.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Nosek, B. A., Hawkins, C. B., and Frazier, R. S. (2011). Implicit social cognition: from measures to mechanisms. *Trends in cognitive sciences*, 15(4):152–9, apr.

Prabhakaran, V. and Rambow, O. (2017). Dialog structure through the lens of gender, gender environment, and power. *D&D*, 8(2):21–55.

Rudman, L. and Kilianski, S. (2000). Implicit and explicit attitudes toward female authority. *Personality and social psychology*.

Rudman, L. A., Greenwald, A. G., and McGhee, D. E. (2001). Implicit Self-Concept and Evaluative Implicit Gender Stereotypes: Self and Ingroup Share Desirable Traits. *Personality and Social Psychology Bulletin*, 27(9):1164–1178.

Sap, M., Prasetio, M. C., Holtzman, A., Rashkin, H., and Choi, Y. (2017). Connotation Frames of Power and Agency in Modern Films. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2319–2324.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, 35(1):4–28, jan.

Sue, D. W. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons.

Sugimoto, C. R., Thelwall, M., Larivière, V., Tsou, A., Mongeon, P., and Macaluso, B. (2013). Scientists Popularizing Science: Characteristics and Impact of TED Talk Presenters. *PLoS ONE*, 8(4):e62403, apr.

Tsou, A., Thelwall, M., Mongeon, P., and Sugimoto, C. R. (2014). A community of curious souls: An analysis of commenting behavior on TED Talks videos. *PLoS ONE*, 9(4):e93609, apr.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier,

M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *Arxiv*, 1501.06307:1–10.

Xu, J., Jun, K., Zhu, X., and Bellmore, A. (2012). Learning from Bullying Traces in Social Media. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666.