THE UNIVERSITY OF CHICAGO

MEASURING PERCEPTIONS AND MITIGATING BIAS IN TEXT AND VOICE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
PHD

DEPARTMENT OF COMPUTER SCIENCE

BY
JENNA CRYAN

CHICAGO, ILLINOIS
JUNE 2022

TABLE OF CONTENTS

ABSTRACT

Our ability as humans to effectively communicate depends heavily on the language we use and the way we speak to one another. The values of our society are both reflected in and reinforced by our use of language. Detecting how this language could convey bias needs to remain effective as these values evolve over time. Specifically, gendered language in text often affirms gender stereotypes and often perpetuate bias and discrimination. As we as readers absorb written content, gendered language used settings such as biographies, recommendation letters, and job advertisements can negatively impact the subjects and audiences alike. Gender stereotypes have been studied extensively, however, the current methods used today still rely on word banks from nearly 50 years ago. Since then, societal views have continued to evolve and it's important to be able to reflect these changes. Additionally, significant advances have been made in developing new methods for analyzing how words are used in text. To address this, the methodology of this work proposes updating existing gender lexicons to reflect modern language use, and applying machine learning to detect gendered language more efficiently. In addition to written text, efficient and unbiased communication depends upon not only the content, but the manner in which it is presented. The tone of voice of a speaker can heavily influence how they are perceived. Particularly, changes in emotion tone of voice can reduce these inherent biases in the listener, thereby providing more effective communication. This work explores ways to improve methods for measuring perceptions of gendered language in text and emotion tone in voice, and ways to mitigate resulting biases.

CHAPTER 1

INTRODUCTION

Through written text and spoken language, biases influence our perceptions of each other and society. These biases often hamper our ability for effective communication, and may alter our behaviors as a result. However, measuring bias remains an ongoing task that requires iterative improvements as societal perceptions and measurement methods change and improve overtime. This work attempts to update the data and methods used to measure and mitigate biases through leveraging large corpora of existing text and audio data, crowd-sourcing perspectives of real people, and using recently developed machine learning models to analyze the data.

For instance, studies show that certain job postings contain potentially gender biased language, through language that frames a position as intended for a more masculine or feminine person, thereby potentially deterring many people from even applying to such jobs. Existing tools attempting to detect and correct gender biased language rely on a very simplistic "bag-of-words" lexicon approach. Further, the word inventories used are based on psychology studies from nearly 50 years ago, and may no longer be relevant today. Recent advances in natural language processing and machine learning methods provide us with updated tools that can be used to more efficiently detect gender bias in language. These tools, though, require an up-to-date corpus to learn how gender bias presents itself in the language of today. This work proposes improving methods of detecting gender biased language by crowdsourcing a large corpus of labeled text to train an end-to-end classification model using deep learning. Through crowdsourcing the data and labels, it's possible to generate a modern corpus of gendered language that reflects the views of real people and based on existing written text.

First, I discuss an overview of existing lexicons of gendered words, how they have been used in the past, and their shortcomings of representing modern perspectives of gender. I will show how these lexicons can be updated to reflect current language use with a large corpora of data from Wikipedia. After identifying the adjectives and verbs most commonly used to describe males and females, crowdsourcing is used to label the new lexicon, by having users indicate if they perceive the words as consistent or contradictory to an association with males or females. The updated gender word bank is then used to train traditional machine learning models, which can in turn be used to label entire articles of text with a "gender score". Then, additional crowdsourcing to gather articles of labeled data, which is used to train a deep learning model that can label unseen articles in an end-to-end approach. By leveraging the labeled articles, the BERT language representation tool is extended to analyze whether a body of text contains masculine or feminine language. Both approaches are evaluated through a user study to determine which produces labels that more closely coincide with the perspectives of real people. While the updated lexicon achieves higher accuracy than using the traditional lexicons, the end-to-end approach proves to be significantly more efficient and accurate than any of the lexicon methods at detecting gender stereotypes in language. By reexamining traditional methods for identifying gendered language, the new end-to-end approach can more accurately identify gender stereotypes in text, and inform authors and researchers when such instances occur.

Lastly, I will discuss how biases affect our ability to communicate through speech. The way speech is perceived has been studied for decades in psychology and socio-linguistics, especially in reference to intonation, accent, and perceived emotion. By manipulating emotional tone of voice, we can increase likeability and reduce potential effects of inherent bias, thereby improving the experience for both the listener and speaker. This work seeks to explore how emotional tone is perceived for different people, how it can be manipulated through manual software and machine learning methods, and if changing tone of voice can change the way the speaker is perceived by others.

CHAPTER 2

MEASURING AND ALTERING GENDER BIAS IN TEXT

2.1 Lexicon Approach

Existing Lexicons. Stereotypes can be captured by gender word inventories – precompiled lists of items describing social traits and behaviors that differentiate males and females [4, 33]. Gender word inventories are historically extracted from self-reported characteristics through questionnaires given to college students to measure their self-concept and valuation of feminine and masculine characteristics. The Personal Attributes Questionnaire (PAQ [35]) and Bem Sex Role Inventory (BSRI [4]) are two of the most representative questionnaires in early studies. The items extracted for the BSRI and PAQ typically associate females with more communal attributes (i.e., gentle, warm) and men with more agentic attributes (i.e., aggressive, competitive), which are highly consistent with traditional perceptions regarding gender stereotypes. Other studies generalized these words into expressive and instrumental traits [34]. Tying these together, aggregated lists of masculine and feminine characteristics have been compiled from previous studies, particularly through gendered wording in job advertisements [19].

However, perceptions captured by BSRI are less endorsed by women in recent years [15, 40]. These works reviewed a large collection of studies that apply BSRI, and tracked how user responses change over a long period of time. Women's femininity scores have decreased significantly over the years, indicating that societal gender norms may require an update of masculine and feminine stereotyped characteristics. Given previous results showing that existing gender word inventories may not properly reflect these concepts in the modern world, it's necessary to develop a lexicon that captures people's perceptions of gender stereotypes in contemporary society.

Generating Updated Lexicon. The lexicon-based approach first analyzes how people associate particular words with common gender stereotypes, and then aggregates these scores to derive a gender score for the entire article. Currently, no databases exist that reflect *modern* perspectives on gender stereotypes in language. Thus, it's necessary to build novel datasets using an existing (unlabeled) data repository and crowdsourcing (to label data from the repository, and provide additional data from online sources). The goal is to create datasets that represent current language use, with minimal bias, and can easily scale up when additional resources become available. This goal is approached in three steps. First, a large corpora (i.e., Wikipedia DataDump) of existing text samples is leveraged to reflect typical use of language. Second, human crowdsourcing is used to label the data. Finally, these datasets can be iteratively updated and expanded through the methods described, thereby providing practical scalability.

Like previous studies [4, 35], our survey asks the participants to rate the extent to which they associate each word with a typical man or woman. Specifically, the participants are shown a list of words, and asked to evaluate the statement "I feel that _______ is commonly associated with the characterization of a typical man in US society" or "of a typical woman in US society." The evaluation uses a 7-point Likert Scale, from "strongly disagree (1)" to "strongly agree (7)." The ground truth score of a word is measured by the difference between the ratings associating the word with men and the ratings associating the word with women. Except a few words related to appearance (e.g., hairy, beautiful), the highly stereotypical words are consistent with recent work demonstrating that stereotypically men are perceived as strong, active and violent, and women are perceived as weak, emotional and kind [17]. Using the ground truth dataset, we first examine whether existing (unsupervised) language models (i.e., word embedding) can be used to automatically label gender lexicon without human input. After testing 3 commonly used word embeddings: word2vec [27], GloVe [30], and FastText [7], the results show that automated lexicon labeling via word

Word Embedding Method	Adjectives	Verbs
Odds ratio	0.09	0.29
Distance + word2vec	0.44	0.37
Distance + GloVe	0.47	0.41
Distance + FastText	0.47	0.41
Gender direction	0.40	0.33
Gender dimension	0.20	0.08

Table 2.1: Pearson Correlation of gender scores between predictions from word embedding methods and ground truth.

Supervised Learning Method	Adjectives	Verbs
LR + word2vec	0.63	0.57
SVM + word2vec	0.62	0.57
LR + GloVe	0.53	0.52
SVM + GloVe	0.58	0.55
LR + FastText	0.52	0.45
SVM + FastText	0.58	0.53

Table 2.2: Pearson Correlation of scores calculated by supervised learning methods and ground truth.

embedding produces gender scores with mediocre results. Since the labeled training dataset only contains around 2000 words, we cannot use deep neural network models that require large training datasets. Instead we use two classical machine learning models: Support Vector Machine (SVM) and Linear Regression (LR). Our models use word embeddings of each word as features, and the pre-trained word2vec, GloVe and FastText as model inputs. Results in Table 2.2 are higher than those produced by word embedding (Table 2.1).

Finally, the gender lexicon is used to detect gender stereotypical language in articles. Similar to previous work [39], a gender score is assigned to an article based on word usage, first by extracting all verbs and adjectives in the article, then adding the scores of these words together to get an overall gender score. Then, this approach is evaluated against the performance of an end-to-end deep learning model.

2.2 End-to-End Approach

Different from the lexicon approach, the end-to-end approach operates directly on paragraphs without breaking them down to individual words. Here we take a supervised learning ap-

Domain	Number	Domain	Number	Domain	Number
wikipedia.org	385	npr.org	58	huffpost.com	46
nytimes.com	178	forbes.com	57	washingtonpost.com	45
theguardian.com	78	dailymail.co.uk	54	biography.com	39
cnn.com	78	foxnews.com	48	cnbc.com	37
people.com	63	time.com	47	vogue.com	37

Table 2.3: Top domains and number of articles from each domain.

	Consistent	Contradict
Masculine	championship, ceo, gun, league, player, businessman, top, service, mountain, fight, basketball, win, drive	gay, makeup, gender, singer, fashion, comfortable, mom, youtube, cosmetic, dress, feel, wear, caregiver, beauty, sexuality
Feminine	cook, child, home, beautiful, beauty, care, clean, fighter, daughter, makeup, family, mother, dress, kid, mom	field, champion, history, sport, athlete, fight, martial, force, training, team, technology, institute, lesbian, rank, tech

Table 2.4: Top keywords that distinguish consistent and contradicting stereotypes.

proach: first gathers human perceptions of gender stereotypes at the paragraph level to build a moderately sized training dataset, then uses it to train a deep learning classification model based on the BERT language representation tool [14]. BERT is a unsupervised language representation tool that converts text articles into vectors. Since BERT does not target any particular language task, we can use it to examine and search for common patterns of language use that may be associated with gender stereotypes. Specifically, the training dataset is used to fine-tune BERT by adding one additional output layer to implement the above mentioned binary classification tasks. The resulting classification model can detect gender stereotypes on arbitrary paragraphs and articles.

The survey asked users to search the Internet, and copy & paste articles (or a subsection of paragraphs from an article) that describe a man (or woman) with a description consistent with (or contradictory to) common gender stereotypes. We received 4360 articles (4 per participant), primarily from biography pages (e.g., Wikipedia) or news sites (e.g., New York Times). Table 2.3 lists the most frequently used domains.

To understand the content of these articles, top keywords are extracted to distinguish articles for each category (i.e., consistent or contradictory). Then, Chi-square statistics are calculated for masculine stereotypes and feminine stereotypes separately, and list the top keywords in Table 2.4. The survey results show participants commonly choose sports and business related terms for men and domestic related terms for women as exemplifying gender stereotypes. Further, some similarities appear between men who contradict stereotypes and women who are consistent with stereotypes (and vice versa).

2.3 Evaluation

The classification model ran two tasks: determining whether the description of a man is consistent with masculine stereotypes, and whether the description of a woman is consistent with feminine stereotypes. Overall, the study shows that the end-to-end approach largely outperforms the lexicon approach, in terms of detection accuracy and robustness. To ensure that our evaluation (using the testing dataset) is sound, we performed another user study to understand whether the per-user contributed labels in the test dataset can accurately capture public perception of gender stereotypes. The new ratings based on at least 4 user responses are reasonably consistent with the original rating, indicating that our testing dataset offers a consistent, public view of gender stereotypes.

We see that the use of full set lexicon effectively improves the detection accuracy, but still cannot match that of the end-to-end approach. Although the two approaches are trained on different data, both datasets are curated from commonly used language in current bodies of text, then evaluated by multiple crowdworkers to generate ground truth labels.

To understand why the lexicon approach generates less satisfactory prediction results, we manually examine all the incorrect predictions the lexicon approach makes in the test set. The possible reasons behind the misclassifications along with examples are summarized in Table 7. The updated lexicon was also compared to those from previous works (PAQ, BSRI, Gaucher), and the overlap with previous lexicons is less than half. Many of the terms are not often found in current language, and the sparsity of their occurrence in our data makes any comparison of results marginally meaningful. Although some of the terms appear to exemplify strong gender connotations (e.g., “feminine,” “masculine”), such words do not

Reason	Lexicon Wrong	Also E-to-E Wrong	Example
Lexicon Coverage	8	0	The first woman I invited to co-author a **publication** was in 2015, four years after completing my **PhD**.
Phrase	10	0	... who **paints his fingernails**, **braids his hair** and **poses for gay magazines**...
Non-human	6	0	Katie Bouman has already worked on looking around corners by analyzing **tiny** shadows ...
Consistent and contradictory	27	4	Even as I regularly work out and **lift weights**, I am a rather **fragile** excuse for a woman, constantly getting sick...
Multiple people	10	3	**My wife** had more earning potential and so I volunteered to concentrate on family and home.
Subtle stereotype, insufficient information	50	123	*American actor Peter Dinklage is labeled as contradicting masculine stereotypes because he is a dwarf, which is not discussed.*
Data noise	30	18	*Random response or failure to meet task requirement.*

Table 2.5: Reasons for lexicon approach making wrong classification. The "Lexicon Wrong" column is the number of cases when the lexicon approach makes a wrong prediction, and the "and E-to-E Wrong" column is the number of cases the end-to-end approach is also wrong among these cases. Bold words are words that are closely related to the reasons provided by the survey participants. Italic words are not exact content from our data, but summarize participant explanations.

often appear in descriptive language and therefore are not necessary to include in the lexicon. Being data-driven, this work is able to evaluate most commonly used verbs and adjectives in current bodies of text.

Applying to Job Postings. Based on a previous study [39], we know that job postings can often contain gendered language. To convert our male and female stereotype detection models into a single gender bias indicator, we take the job posting text and use it as input in both the masculine stereotype classifier and feminine stereotype classifier. We take the probability output from both classifiers, and calculate the difference in the two probabilities.

	Textio	Unitive	BERT fine-tune
% of females	0.59	0.54	0.77
Attractiveness to female applicants	0.64	0.54	0.80

Table 2.6: Pearson correlation between user responses and gender bias scores.

We calculate the score for all the job advertisements, along with two state-of-the-art services cited by prior work: *Textio* and *Unitive* (later renamed Talent Sonar), both of which are specifically designed to detect gender bias in job posts using a lexicon-based approach [39]. Table 2.6 shows the correlation between the scores and user responses. This shows that although the models in our end-to-end approach are not specifically trained for job advertisements, they still outperform the best lexicon approaches designed for this task.

2.4 Discussion

This work seeks to reconcile the traditional lexicon-based approaches for detecting gender stereotypes in language, with modern natural language processing tools almost entirely based on end-to-end deep learning models. The high level question is: what approach should researchers and practitioners take moving forward, an updated version of lexicon-based models (which we developed in this work), or an end-to-end deep learning model built on existing language models (BERT) and further trained with paragraph-length text samples? Despite our best efforts to update and strengthen the lexicon-based models, end-to-end models based on BERT provide substantially stronger results, even when trained on our moderately-sized, crowdsourced dataset. In fact, when applied to the context of gender bias in job listings, the end-to-end model significantly outperforms models used by industry services.

CHAPTER 3

MEASURING AND ALTERING PERCEPTIONS OF SPEECH

The way humans express themselves through language includes not only the words spoken, but also *how* they say it. Content, emotion, and intention influence how speakers express themselves, but these are not always perceived as desired by the speaker. The way speech is perceived has been studied for decades in psychology and socio-linguistics, especially in reference to intonation, accent, and perceived emotion. Computational linguistics and natural language processing have made advancements to manual and machine learning methods to manipulate emotion expression in voice. By manipulating voice, we can reduce potential effects of inherent bias and improve the experience for both the listener and speaker.

3.1 Methodology

Existing tools like DAVID [31] and Praat [6] allow for manual and automatic manipulation of audio features (e.g., pitch, frequency) that can be used for fine-grain adjustments at the individual frame level, and even emotion transformations (e.g., happy, sad, afraid). Aucouturier et al. [1] found that DAVID produces subtle manipulations of emotional tone in one's voice, and the changes correlate with changes in the emotion of the listener. In recent years, computational linguistics has also used a variety of machine learning architectures to manipulate emotion in speech, such as GANs, variational autoencoders, and sequence-to-sequence text-to-speech models. These machine learning models capture linguistic frequency and prosody patterns in order to This work focuses on a recently developed emotion conversion model based on a VAW-GAN architecture that makes use of continuous wavelet transformation for prosody conversion [45]. This model allows for achieving high quality voice conversions with the possibility to convert unseen speakers.

3.2 Project Plans

This work seeks to answer the following questions:

- Can we systematically classify, manipulate human speech?
- Do listeners consistently perceive different emotions/ characteristics from speech?
- How do these manipulations change perceptions in the listener about the speaker?

To answer these questions, I will conduct user studies to compare user perceptions of raw speech, manipulated speech via DAVID, and speech converted with a VAW-GAN machine learning model. For each emotion (i.e., happy, sad, angry, surprise), the tools will be used to convert the emotion to a neutral tone. Several series of questions will determine if the tools effectively change the perceived emotion tone, which emotion tone(s) are associated with various characteristics, and which emotion tone(s) are preferred for various real-life scenarios. First, users will be asked to rate the quality of the speech, the perceived emotion tone, and extent of emotional tone. Then, users will be asked to compare pairs of speech samples (one with emotion tone and one neutral / converted-to-neutral tone) for several personal characteristics (i.e., trustworthy, warm, competent, anxious, polite, positive, negative). Finally, users will be asked to listen to the same pairs and indicate their preference for each voice in several scenarios (i.e., speaking with a telephone operator, being interviewed for a job, getting a surgery, interviewing a personal assistant, listening to a political debate). Collectively, these tasks will show the ability for these tools to alter perceptions of voices to increase effectiveness and preferability in real life scenarios.

CHAPTER 4

EXPECTED TIMELINE

Summer 2022

- Candidacy Exam
- Continue work on emotion voice perceptions project
- Submit emotion voice perceptions project to CHI (September deadline)

Fall 2022

- Begin work on computer-generated voice perceptions project

Winter 2023

- Write computer-generated voice perceptions paper for submission to CSCW
- Work on Dissertation paper

Spring 2023

- Complete Dissertation paper
- Thesis Defense

REFERENCES

[1] Jean-Julien Aucouturier, Petter Johansson, Lars Hall, Rodrigo Segnini, Lolita Mercadié, and Katsumi Watanabe. Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 113(4):948–953, 2016.

[2] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[3] Mahzarin R Banaji and Curtis D Hardin. Automatic stereotyping. *Psychological science*, 7(3):136–141, 1996.

[4] Sandra L Bem. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2):155–162, 1974.

[5] Jayadev Bhaskaran and Isha Bhallamudi. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proc. of ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[6] Paul Boersma and Vincent Van Heuven. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347, 2001.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proc. of NIPS*, pages 4349–4357, 2016.

[9] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(14):183–186, April 2017.

[10] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[11] Amy JC Cuddy, Susan T Fiske, and Peter Glick. When professionals become mothers, warmth doesn't cut the ice. *Journal of Social issues*, 60(4):701–718, 2004.

[12] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[13] M Lee Dean and Charlotte Chucky Tate. Extending the legacy of sandra bem: Psychological androgyny as a touchstone conceptual advance for the study of gender in psychological science. *Sex Roles*, 76(11-12):643–654, 2017.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[15] Kristin Donnelly and Jean M Twenge. Masculine and feminine traits on the bem sex-role inventory, 1993–2012: a cross-temporal meta-analysis. *Sex Roles*, pages 1–10, 2016.

[16] Alice H Eagly, Wendy Wood, and Amanda B Diekman. Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender*, 12:174, 2000.

[17] Ethan Fast, Tina Vachovsky, and Michael S Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proc. of ICWSM*, 2016.

[18] Joel Escude Font and Marta R. Costa-jussa. Equalizing gender biases in neural machine translation with word embeddings techniques. In *Proc. of ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[19] Danielle Gaucher, Justin Friesen, and Aaron C Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1):109, 2011.

[20] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proc. of ECCV*, pages 793–811, 2018.

[21] Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proc. of ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[22] Rie Johnson and Tong Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. In *International Conference on Machine Learning*, pages 526–534, 2016.

[23] Anne Maass and Luciano Arcuri. Language and stereotyping. *Stereotypes and stereotyping*, pages 193–226, 1996.

[24] Juan M Madera, Michelle R Hebl, and Randi C Martin. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591, 2009.

[25] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proc. of NAACL*, 2019.

[26] Michela Menegatti and Monica Rubini. Gender bias and sexism in language, 2017.

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[28] Saif Mohammad. 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, 2018.

[29] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proc. of EMNLP*, pages 2799–2804, 2018.

[30] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[31] Laura Rachman, Marco Liuni, Pablo Arias, Andreas Lind, Petter Johansson, Lars Hall, Daniel Richardson, Katsumi Watanabe, Stéphanie Dubal, and Jean-Julien Aucouturier. David: An open-source platform for real-time emotional speech transformation: With 25 applications in the behavioral sciences. *bioRxiv*, page 038133, 2016.

[32] Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. Linguistic analysis of differences in portrayal of movie characters. In *Proc. of ACL*, volume 1, pages 1669–1678, 2017.

[33] Paul Rosenkrantz, Susan Vogel, Helen Bee, Inge Broverman, and Donald M Broverman. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 32(3):287, 1968.

[34] Stephen A Schullo and Burton L Alperson. Interpersonal phenomenology as a function of sexual orientation, sex, sentiment, and trait categories in long-term dyadic relationships. *Journal of Personality and Social Psychology*, 47(5):983, 1984.

[35] Janet T. Spence, Robert L. Helmreich, and Joy Stapp. The personal attributes questionnaire: A measure of sex role stereotypes and masculinity-femininity. *JSAS Catalog of selected documents in psychology*, 4(43), 1974.

[36] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[37] Latany Sweeney. Discrimination in online ad delivery. In *arXiv:1301.6822*, 2013.

[38] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[39] Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J. Metzger, Haitao Zheng, and Ben Y. Zhao. Gender bias in the job market: A longitudinal analysis. In *Proc. of CSCW*, 2017.

[40] Jean M Twenge. Changes in masculine and feminine traits over time: A meta-analysis. *Sex roles*, 36(5-6):305–325, 1997.

[41] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proc. of ICWSM*, 2015.

[42] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, page 35, 1997.

[43] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proc. of NAACL*, volume 2, 2018.

[44] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proc. of EMNLP*, pages 4847–4853, 2018.

[45] Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li. Converting anyone's emotion: Towards speaker-independent emotional voice conversion. *arXiv preprint arXiv:2005.07025*, 2020.

REFERENCES

[1] Jean-Julien Aucouturier, Petter Johansson, Lars Hall, Rodrigo Segnini, Lolita Mercadié, and Katsumi Watanabe. Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 113(4):948–953, 2016.

[2] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[3] Mahzarin R Banaji and Curtis D Hardin. Automatic stereotyping. *Psychological science*, 7(3):136–141, 1996.

[4] Sandra L Bem. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2):155–162, 1974.

[5] Jayadev Bhaskaran and Isha Bhallamudi. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In *Proc. of ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[6] Paul Boersma and Vincent Van Heuven. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347, 2001.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proc. of NIPS*, pages 4349–4357, 2016.

[9] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(14):183–186, April 2017.

[10] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[11] Amy JC Cuddy, Susan T Fiske, and Peter Glick. When professionals become mothers, warmth doesn't cut the ice. *Journal of Social issues*, 60(4):701–718, 2004.

[12] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[13] M Lee Dean and Charlotte Chucky Tate. Extending the legacy of sandra bem: Psychological androgyny as a touchstone conceptual advance for the study of gender in psychological science. *Sex Roles*, 76(11-12):643–654, 2017.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[15] Kristin Donnelly and Jean M Twenge. Masculine and feminine traits on the bem sex-role inventory, 1993–2012: a cross-temporal meta-analysis. *Sex Roles*, pages 1–10, 2016.

[16] Alice H Eagly, Wendy Wood, and Amanda B Diekman. Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender*, 12:174, 2000.

[17] Ethan Fast, Tina Vachovsky, and Michael S Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proc. of ICWSM*, 2016.

[18] Joel Escude Font and Marta R. Costa-jussa. Equalizing gender biases in neural machine translation with word embeddings techniques. In *Proc. of ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[19] Danielle Gaucher, Justin Friesen, and Aaron C Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1):109, 2011.

[20] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proc. of ECCV*, pages 793–811, 2018.

[21] Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proc. of ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[22] Rie Johnson and Tong Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. In *International Conference on Machine Learning*, pages 526–534, 2016.

[23] Anne Maass and Luciano Arcuri. Language and stereotyping. *Stereotypes and stereotyping*, pages 193–226, 1996.

[24] Juan M Madera, Michelle R Hebl, and Randi C Martin. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591, 2009.

[25] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proc. of NAACL*, 2019.

[26] Michela Menegatti and Monica Rubini. Gender bias and sexism in language, 2017.

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[28] Saif Mohammad. 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, 2018.

[29] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proc. of EMNLP*, pages 2799–2804, 2018.

[30] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[31] Laura Rachman, Marco Liuni, Pablo Arias, Andreas Lind, Petter Johansson, Lars Hall, Daniel Richardson, Katsumi Watanabe, Stéphanie Dubal, and Jean-Julien Aucouturier. David: An open-source platform for real-time emotional speech transformation: With 25 applications in the behavioral sciences. *bioRxiv*, page 038133, 2016.

[32] Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. Linguistic analysis of differences in portrayal of movie characters. In *Proc. of ACL*, volume 1, pages 1669–1678, 2017.

[33] Paul Rosenkrantz, Susan Vogel, Helen Bee, Inge Broverman, and Donald M Broverman. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 32(3):287, 1968.

[34] Stephen A Schullo and Burton L Alperson. Interpersonal phenomenology as a function of sexual orientation, sex, sentiment, and trait categories in long-term dyadic relationships. *Journal of Personality and Social Psychology*, 47(5):983, 1984.

[35] Janet T. Spence, Robert L. Helmreich, and Joy Stapp. The personal attributes questionnaire: A measure of sex role stereotypes and masculinity-femininity. *JSAS Catalog of selected documents in psychology*, 4(43), 1974.

[36] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[37] Latany Sweeney. Discrimination in online ad delivery. In *arXiv:1301.6822*, 2013.

[38] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[39] Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J. Metzger, Haitao Zheng, and Ben Y. Zhao. Gender bias in the job market: A longitudinal analysis. In *Proc. of CSCW*, 2017.

[40] Jean M Twenge. Changes in masculine and feminine traits over time: A meta-analysis. *Sex roles*, 36(5-6):305–325, 1997.

[41] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proc. of ICWSM*, 2015.

[42] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, page 35, 1997.

[43] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proc. of NAACL*, volume 2, 2018.

[44] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proc. of EMNLP*, pages 4847–4853, 2018.

[45] Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li. Converting anyone's emotion: Towards speaker-independent emotional voice conversion. *arXiv preprint arXiv:2005.07025*, 2020.