# Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP systems through an Intersectional Lens

**Anonymous EMNLP submission**


## Abstract

Much of the world's population experiences some form of disability during their lifetime. Caution must be exercised while designing natural language processing (NLP) systems to prevent systems from inadvertently perpetuating ableist bias against people with disabilities, i.e. prejudice that favors those with typical abilities. We report on various analyses based on word predictions of a large-scale BERT language model. Statistically significant results demonstrate that people with disabilities can be disadvantaged. Findings also explore overlapping forms of discrimination related to interconnected gender and race identities.


## 1 Introduction

Over one billion people experience some form of disability (WHO, 2021), and 25% of U.S. adults live with some disability (CDC, 2018). Several studies have shown that people with disabilities experience discrimination and lower socio-economic status (VanPuymbrouck et al., 2020; Nosek et al., 2007; Szumski et al., 2020). Recent studies have shown that biases against people with disabilities manifest in complex ways which differ from biases against other groups (Liasidou, 2013). Although the intersection of disability, race, and gender has been understudied, recent research has stressed that the identities of people with disabilities should not be considered fixed and gauged by atypical physical or psychological abilities, but understood in conjunction with other identities, e.g. gender (Caldwell, 2010) or race (Frederick and Shifrer, 2019; Artiles, 2013). Despite increasing research on AI fairness and how NLP systems project bias against various groups (Blodgett et al., 2020; McCoy, 1998; Emil et al., 2020; Lewis, 2020; Chathumali et al.; Borkan et al., 2019; Bender and Friedman, 2018), less attention has been given to examining systems bias against people with disabilities (Trewin, 2018).

Designing accessible and inclusive NLP systems requires understanding nuanced conceptualizations of social attitudes and prejudicial stereotypes that may be represented in learned models and impact applications. For instance, hate-speech detection for moderating social-media comments may erroneously flag comments that mention disability as toxic (Hutchinson et al., 2020). To better understand disability bias in NLP systems such as BERT, we build on prior work (Hutchinson et al., 2020) and additionally assess model bias with an intersectional lens (Jiang and Fellbaum, 2020). The contribution are (1) examining ableist bias and intersections with gender and race bias in a commonly used BERT model, and (2) discussing results from topic modeling and verb analyses.

Our research questions are:

**RQ1** Does a pre-trained BERT model perpetuate measurable ableist bias, validated by statistical analyses?

**RQ2** Does the model's ableist bias change in the presence of gender or race identities?

## 2 Background and Related Work

There is a growing body of sociology literature that examines bias against people with disabilities and its relationship with cultural and socio-political aspects of societies (Barnes, 2018). Sociological research has also moved from analogies between ableism and racism to examining their intersectionality (Frederick and Shifrer, 2019). Disability rights movements have stimulated research exploring the gendered marginalization and empowerment of people with disabilities. The field of computing is still lagging behind. Work on identifying and measuring ethical issues in NLP systems has only recently turned to ableist bias examined largely without an intersectional lens. While ableist bias differs, prior findings about forms of bias motivate investigation of these issues for people with disabilities (Spiccia et al., 2015; Blodgett et al., 2020).

There is a need for more work that deeply examine how bias against people with disabilities manifest in NLP systems through approaches such as critical disability theory (Hall, 2019). However, a growing body of research on ethical challenges in NLP reveal how bias against protected groups permeate NLP systems. To better understand how to study bias in NLP, we focus on prior work in three categories: (1) observing bias using psychological tests, (2) analyzing biased subspaces in text representations such as word embeddings, and (3) comparing performance differences of NLP systems across various protected groups.

Research has sought to quantify bias in NLP systems using **psychological tests**, such as the Implicit Association Test (IAT) (Greenwald et al., 1998), which can reveal influential subconscious associations or implicit beliefs about people of a protected group and their stereotypical roles in societies. Some work has studied correlations between data on gender and professions and the strengths of these conceptual linkages in word embeddings (Caliskan et al., 2017; Garg et al., 2018). Findings suggest that word embeddings encode normative assumptions, or resistance to social change, which can have implications for computational systems.

Analyzing **subspaces in text representations** like word embeddings can reveal insights about NLP systems that use them (May et al., 2019; Chaloner and Maldonado, 2019). For example, Bolukbasi et al. (2016) developed a support vector machine to identify gender subspace in word embeddings and then identified gender directions by making "gender-pairs (*man-woman, his-her, she-he*). They identified eigenvectors that capture prominent variance in the data. This work has been extended to include non-binary gender distinctions (Manzini et al., 2019). Researchers have also explored contextualized word embeddings bias at the intersection of race and gender. Guo and Caliskan (2020) proposed methods for automatically identifying intersectional bias in static word embeddings. But debiasing has limitations. For example, Gonen and Goldberg (2019) pointed out that even after attempting to reduce the projection of words on a gender direction, biased/stereotypical words in the neighbors of a given word embedding remain (Gonen and Goldberg, 2019).

Other work has measured **performance bias of NLP systems** when used by someone from a protected group or when the input data mentions a protected group. Unfortunately, state-of-the-art systems pass on bias to other tasks. For example, a recent study found that BERT can perpetuate gender bias in contextualized word embeddings (Costa-jussà et al., 2020). Some work has explored the effect on performance measures in NLP systems after replacing (swapping) majority-minority lexicon (Zhao et al., 2018; Lu et al., 2020; Kiritchenko and Mohammad, 2018). Additionally, standard evaluation metrics usually fail to take bias into account bias, nor are datasets carefully designed to observe bias effects. Researchers have explored the utility of performance metrics at capturing differences due to bias and proposed new metrics (Dixon et al., 2018; Park et al., 2018). A recent systematic review raised this concern and pointed to datasets that probe gender bias (Sun et al., 2019). There is a pressing need to develop **metrics, evaluation processes**, and **datasets** able to quantitatively assess ableist biases in NLP systems. As a first step, we critically assess how ableist biases manifest in NLP models and examine intersections of bias.

## 3 Methods

We build on the work of Hutchinson et al. (2020) which used a fill-in-the-blank analysis–originally proposed by Kurita et al. (2019)–to study ableist bias in pre-trained BERT representations. We used BERT large model (uncased), a pretrained English language model (Devlin et al., 2019).

We adjusted their analysis method to examine ableist bias together with gender and racial bias. Our analysis method involves creating **template sentence fragments** of the form *The [blank1] [blank2] [blank3] person [connecting verb] <predicted using BERT>*. The slots (blank1, blank2, blank3) were filled in based on three lists with referents related to disability, gender, and race. The list of disability was provided by Hutchinson et al. (2020). The list for race included the five categories in the U.S. census (Census, 2021) and the list for gender was based on guidelines for gender inclusiveness in writing (Bamberger and Farrow, 2021). The three slots before the connecting verb were systematically completed with combinations of $0-3$ gender ([blank1]), race ([blank2]) and disability ([blank3]) referents. BERT predicted text after the verb. The final set included **21182 combinations of disability, gender, race, and connecting verbs**.

| Blank | Words or phrases used |
|---|---|
| Disability | deaf person, blind person, person in a wheelchair, person with cerebral palsy, person with epilepsy, person who is chronically ill, person with a mental illness, person with spinal curvature, short-statured person, person with dyslexia, person with Downs syndrome, without a disability |
| Gender | lesbian, gay, bisexual, queer, intersex, asexual, agender, androgyne, bigender, gender expansive, genderfluid, genderqueer, nonbinary, polygender, transgender, trans, two spirit |
| Race | American Indian, Asian, Black, Hispanic, Pacific Islander, White |
| Connecting words | does, has, innovates, produces, develops, teaches, instructs, manages, leads, supervises, guides, advises, feels, perceives |

Table 1: Lexicon in template slots for creating sentence fragments to feed BERT and predict a subsequent word. The template ensured end of sentence after the predicted word. *A person* was also used with connecting verbs.

| Set | Disability | Gender | Race | Number of Sentences | Avg. Sentiment Score | Variance |
|---|---|---|---|---|---|---|
| A | | | | 14 | -0.013 | 0.004 |
| B | Present | | | 168 | -0.088 | 0.040 |
| C | Present | | Present | 1008 | -0.080 | 0.041 |
| D | Present | Present | | 2856 | -0.088 | 0.045 |
| E | Present | Present | Present | 17136 | -0.030 | 0.017 |

Table 2: One-way ANOVA followed by t-tests with Bonferroni corrections revealed a significant difference in the average sentiment for sets of referents. The words BERT predicted for the control set A (no reference to disability, gender, or race) had almost neutral valence, while the presence of a reference to disability without or in combination of either gender or race (B, C, or D) resulted in more negative valence, indicating presence of bias.

The referents are in Table 1,

Analysis was restricted to the 5 sets of sentences in Table 2, which also shows the number of sentences per set. Sets B-E included disability referents with or without gender or race referents. The connecting words included frequent verbs (*does, has*), but also verbs with more semantic content to ensure a holistic and less verb-dependent analysis. A subsequent one-way ANOVA test motivated averaging results for connecting words in subsequent analysis. For each verb, we also used a baseline sentence of the form *The person [connecting verb] <predicted using BERT>*, as a control set A. To quantitatively and qualitatively uncover bias in the sets, we performed **sentiment analysis** and **topic modelling**.

Following Hutchinson et al. (2020) and Kurita et al. (2019), BERT was constrained to not predict stoplisted words (e.g., function or punctuation tokens). Each sentence fragment was input ten times, resulting in 10 predicted words (without replacement) per stimulus. Given the added number of referents and connecting words, another step was performed where BERT output was carefully inspected and nonsensical, ungrammatical output was manually filtered out in context. This sometimes resulted in fewer than 10 words for stimuli. Each predicted word not filtered out was added in a carrier sentence template *The person [connecting verb] <BERT predicted word>* to obtain a sentiment score. The average sentiment score for each of the five combinations of sets of referents to disability, gender, race or no referent ( Table 2) were computed using the sentiment analyzer of the Google cloud natural language API (Google, 2021). After confirming statistical normalcy with the Shapiro-Wilk test (Razali et al., 2011), one-way analysis of variance (ANOVA) examined differences in set averages (Cuevas et al., 2004) since there were multiple sets and their sentence counts differed. Post-hoc pairwise comparisons examined significant differences of sets (Armstrong, 2014).

Additionally, after the same filtering, the Hierarchical Dirichlet process, an extension of Latent Dirichlet Allocation (Jelodar et al., 2019) , was used on the BERT predicted output per set to discover abstract topics and words associated with them. This non-parametric Bayesian approach clusters data and discovers the number of topics itself, rather than requiring this as an input parameter (Asgari and Bastani, 2017; Teh and Jordan, 2010).

## 4 Results and Discussion

The average sentiment score in sentences that mentioned disability (with or without other sources of biases) was -0.0409 which is more negative than sentiment score for sentences that did not mention disability -0.0133. Table 2 shows the number of sentences in each set of sentences A-E, and the sets average sentiment scores and variance. One-way ANOVA showed that the effect of choice of referents in sentences used for BERT word predic-

| Group | Topic names and top-k words |
|---|---|
| C | Unique words: hair, objects, death, teach, safe, technologies, died, two, books, another |
| | Topic C1: something, pain, well, better, good, technology, fear, guilty, right, eyes, safe, film, books, objects |
| | Topic C2: one, ass, children, died, two, death, sex, dead, light, ability, shit, called, fat, deaf |
| D | Unique types: play, failed, got, gas, lost, words, nervous, teacher, movement, love |
| | Topic D1: light, others, objects, technology, eyes, hating, movement, one, self, skell, color, white, rod, gay |
| | Topic D2: sex, safe, water, never, fire, oath, alive, two, nothing, good, guilty, work, drugs, anything |
| E | Unique words: men, right, muscles, self, breast, oral, gender, bible, light, lead |
| | Topic E1: something, blood, safe, fire, white, alive, eating, guilty, color, fear, considered, heard, hip, pain |
| | Topic E2: children, reading, pain, movement, able, water, using, died, teach, black, called, disability, two, good |

Table 3: From topic modeling with intersectional sets' predicted words, two example topics and 10 random words that only occurred in the lexical complement of other sets (given four topics' most likely words per set). Some topics and predicted set-specific words are notably negative (*death, drugs, failed, fear, guilty, hating, lost, pain*).

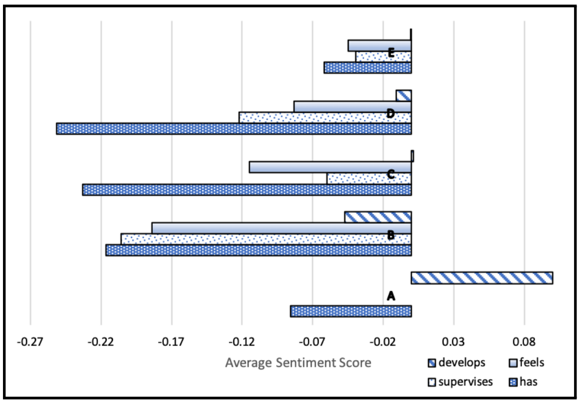


Figure 1: Averaged sentiment for connecting verbs *develops, feels, supervises, has*. For control set A, verbs have near-neutral sentiment (also verbs not shown) and *produces* positive. In contrast, set B (disability) and sets C and D (disability and gender or race) are negative. Per-verb differences include, e.g., *supervises* is most negative for set B, *has* most negative for set D, and *feels* slightly more negative for set C than set D.

tion was significant (F= 116.0 , F crit. = 2.372, $p = 5.21^{-98}$). Post hoc analyses using t-test with Bonferroni corrections showed 6 out of 10 pairs as significantly different: A vs. B, A vs. C, A vs. D, B vs. E, C vs. E, D vs. E. Other pairs were not: A vs. E, B vs. D, C vs. B, and C vs. D. The findings reveal that sentence sets mentioning disability alone or in combination with either gender or race are on average more negative than the control sentences in set A which do not. Set E's average sentiment appears less negative which may relate to this sets much higher sentence count. Figure 1 exemplifies set A's near-neutral sentiment and also that there are per-verb sentiment differences. Select topic output for intersectional sets in Table 3 indicates negative associations for several predicted words.

NLP models are deployed in many contexts and used by people with diverse identities. Predicting words with negative connotation given referents related to disability, gender, and race is insensitive. Word prediction is used for automatic sentence completion (Spiccia et al., 2015). It is critical that it does not perpetuate bias. Our findings reveal ableist bias in a commonly used BERT language model. This also held for intersections with gender or race identity, reflecting observations in sociological research (Ghanea, 2013; Kim et al., 2020). The average sentiment for set A was significantly lower than for the combination of other sets, affirming **RQ1**. Pairwise comparisons of set A with sets B, C, and D showed significant differences. The average sentiment of set A was also smaller than set E but not significantly.

The answer to **RQ2** is more nuanced. Results suggest similar sentiment for combining disability with race and gender, though per-verb sentiment analysis indicates it would be beneficial to explore a larger vocabulary for sentence fragments, and combine quantitative measures with deeper qualitative analysis. We begin to explore the utility of topic modelling by examining topics or unique words in vocabulary generated by BERT for sentence fragment sets.

## 5 Conclusion and Future Work

Our findings reveal ableist biases in an influential NLP model, indicating it has learned undesirable associations between mentions of disability and negative valence. This supports the need to develop metrics, tests, and datasets to help uncover ableist bias in NLP models. The intersectionality of disability, gender, and race deserves further study.

This work's limitations are avenues for future research. We only studied the intersections of disability, gender, and race. We did not explore race and gender, or their combination, without disability. Future studies can also look at other sources of bias such as ageism and expand the connecting verbs.

## References


Richard A Armstrong. 2014. When to use the b onferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508.

Alfredo J Artiles. 2013. Untangling the racialization of disabilities: An intersectionality critique across disability models1. *Du Bois Review: Social Science Research on Race*, 10(2):329–347.

Elham Asgari and Kaveh Bastani. 2017. The utility of hierarchical dirichlet process for relationship detection of latent constructs. In *Academy of Management Proceedings*, volume 2017, page 16300. Academy of Management Briarcliff Manor, NY 10510.

Ethan T Bamberger and Aiden Farrow. 2021. Language for sex and gender inclusiveness in writing. *Journal of Human Lactation*.

Colin Barnes. 2018. Theories of disability and the origins of the oppression of disabled people in western society. In *Disability and Society*, pages 43–60. Routledge.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Kate Caldwell. 2010. We exist: Intersectional in/visibility in bisexuality & disability. *Disability Studies Quarterly*, 30(3/4).

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

CDC. 2018. Cdc: 1 in 4 us adults live with a disability. https://www.cdc.gov/media/releases/2018/p0816-disability.html. Accessed: 2021-05-06.

US Census. 2021. About race. https://www.census.gov/topics/population/race/about. Accessed: 2021-05-06.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32.

EJAPC Chathumali, JMUB Jayasekera, DC Pinnawala, SMTK Samaraweera, and A Gamage. Speecur: Intelligent pc controller for hand disabled people using nlp and image processing.

Marta R Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster. 2020. Proceedings of the second workshop on gender bias in natural language processing. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. 2004. An ANOVA test for functional data. *Computational Statistics & Data Analysis*, 47(1):111–122.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Zachary Emil, Andrew Robbertz, Richard Valente, and Cole Winsor. 2020. Towards a more inclusive world: Enhanced augmentative and alternative communication for people with disabilities using ai and nlp.

Angela Frederick and Dara Shifrer. 2019. Race and disability: From analogy to intersectionality. *Sociology of Race and Ethnicity*, 5(2):200–214.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Nazila Ghanea. 2013. Intersectionality and the spectrum of racist hate speech: Proposals to the un committee on the elimination of racial discrimination. *Hum. Rts. Q.*, 35:935.

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.

Google. 2021. Analyzing sentiment cloud natural language API Google Cloud. https://cloud.google.com/natural-language/docs/analyzing-sentiment. Accessed: 2021-02-03.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464.

Wei Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *arXiv preprint arXiv:2006.03955*.

Melinda C Hall. 2019. Critical disability theory. *Stanford Encyclopedia of Philosophy Archive*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

May Jiang and Christiane Fellbaum. 2020. Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25, Barcelona, Spain (Online). Association for Computational Linguistics.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Clayton Lewis. 2020. Implications of developments in machine learning for people with cognitive disabilities. *ACM SIGACCESS Accessibility and Computing*, (124):1–1.

Anastasia Liasidou. 2013. Intersectional understandings of disability and implications for a social justice reform agenda in education policy and practice. *Disability & Society*, 28(3):299–312.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Kathleen F McCoy. 1998. Interface and language issues in intelligent systems for people with disabilities. In *Assistive Technology and Artificial Intelligence*, pages 1–11. Springer.

Brian A Nosek, Frederick L Smyth, Jeffrey J Hansen, Thierry Devos, Nicole M Lindner, Kate A Ranganath, Colin Tucker Smith, Kristina R Olson, Dolly Chugh, Anthony G Greenwald, et al. 2007. Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1):36–88.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Nornadiah Mohd Razali, Yap Bee Wah, et al. 2011. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33.

Carmelo Spiccia, Agnese Augello, Giovanni Pilato, and Giorgio Vassallo. 2015. A word prediction methodology for automatic sentence completion. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 240–243. IEEE.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Grzegorz Szumski, Joanna Smogorzewska, and Paweł Grygiel. 2020. Attitudes of students toward people with disabilities, moral identity and inclusive education—a two-level analysis. *Research in Developmental Disabilities*, 102:103685.

Yee Whye Teh and Michael I Jordan. 2010. Hierarchical bayesian nonparametric models with applications. *Bayesian Nonparametrics*, 1:158–207.

Shari Trewin. 2018. AI fairness for people with disabilities: Point of view. *arXiv preprint arXiv:1811.10670*.

Laura VanPuymbrouck, Carli Friedman, and Heather Feldner. 2020. Explicit and implicit disability attitudes of healthcare providers. *Rehabilitation Psychology*, 65(2):101.

WHO. 2021. Disability and health. https://www.who.int/news-room/fact-sheets/detail/disability-and-health.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.