Applying the Stereotype Content Model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies

Brienna Herold and **James Waller** and **Raja S. Kushalnagar**
Gallaudet University,
800 Florida Ave NE, Washington, DC 20002 USA
brienna.herold@gmail.com, james.waller@gallaudet.edu, raja.kushalnagar@gallaudet.edu

Abstract

Stereotypes are a positive or negative, generalized, and often widely shared belief about the attributes of certain groups of people, such as people with sensory disabilities. If stereotypes manifest in assistive technologies used by deaf or blind people, they can harm the user in a number of ways—especially considering the vulnerable nature of the target population. AI models underlying assistive technologies have been shown to contain biased stereotypes, including racial, gender, and disability biases. We build on this work to present a psychology-based stereotype assessment of the representation of disability, deafness, and blindness in BERT using the Stereotype Content Model. We show that BERT contains disability bias, and that this bias differs along established stereotype dimensions.

1 Introduction

Pre-trained natural language processing (NLP) models are becoming more commonly deployed in pipelines for consumer tools, including those that fall under the umbrella of assistive technologies. Models such as BERT are used in tools that utilize automatic text simplification (ATS) for reading assistance (Lauscher et al., 2020), where complex words get replaced with simpler alternatives. BERT is also used in natural language understanding tools such as automatic speech recognition (Chuang et al., 2020).

In addition to a continuing increase in the use cases and complexity of AI-based assistive technologies, there is also growing interest in using them. Alonzo et al. (2020) found that the deaf community expressed strong interest in ATS-based reading assistance tools. To achieve fair and inclusive experiences for deaf and blind people, it is important to understand how they may be represented by the models underlying the assistive technologies that are designed for them (Kafle et al., 2019).

If an AI-based consumer tool perpetuates existing biases and stereotypes in society, it can inadvertently cause and reinforce structural stigma, or "societal level conditions, cultural norms, and institutional policies that constrain the opportunities, resources, and well-being of the stigmatized" (Hatzenbuehler, 2016). The bias against deafness—or audism—is prevalent in both mainstream society (Humphries, 1977) and in the deaf community (Gertz, 2003). Audism has been linked to discrimination in multiple real-world scenarios, including the job application process (Task Force Members and Contributors, 2012). In Szymanski (2010), 100% of highly qualified psychology internship applications that mentioned deafness were rejected, whereas 100% of those that didn't mention deafness were invited for an interview.

Causing or reinforcing structural stigma can lead to allocational and representational harms (Blodgett et al., 2020). Allocational harms arise if assistive technologies distribute resources or opportunities unfairly to disabled people. With representational harms, if assistive technologies represent these people unfairly, disabled people may experience alienation, decreased quality of service, stereotypes, denigration and stigmatization, erasure, and/or decreased public participation.

Despite recent ballooning of research in NLP fairness (Sheng et al., 2020; Blodgett, 2021), there has been little investigation into how AI models represent disabled people, who comprise at least 12.5% of the global population (WHO, 2021). There has been even less of a focus on how people with sensory disabilities are represented in NLP models. Hutchinson et al. (2020) provided preliminary evidence that disability-mentioning text may be accidentally flagged as toxic. Hassan et al. (2021) detected signs of disability bias in BERT using sentiment analysis, and they investigated how this bias might shift when applying an intersectional lens to the analysis.

Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022), pages 58 - 65
May 27, 2022 ©2022 Association for Computational Linguistics

To further investigate sensory disability bias in NLP models, we build upon prior work in association bias in BERT. Our contributions include adapting Kurita et al. (2019)'s sentence templates to examine associations between disability qualifiers and stereotype traits, drawing from the Stereotype Content Model (SCM), an established approach in social psychology to defining stereotyped bias (Fiske et al., 2002).

Specifically, we answer these research questions:

- **RQ1**. In BERT, is there evidence of bias in how the model perceives disability, compared to ability?
- **RQ2**. Do BERT's representations of ability and disability differ across various stereotype dimensions?

2 Related Work

We review previous work in examining stereotypes in NLP models, and then we briefly describe the SCM and its relevance to measuring bias.

2.1 Stereotypes in NLP models

Bolukbasi et al. (2016) first observed that gender stereotypes are present in static word embeddings (e.g. word2vec and GloVe) using subspace analysis. Caliskan et al. (2017) found that word embeddings capture a spectrum of implicit biases, using lexicons developed for the Implicit Association Test, or the IAT (Greenwald et al., 1998), and calculated associations within static word embeddings. Kurita et al. (2019) extended this approach to work with contextualized embedding models such as BERT.

However, using word lists pulled from the IAT is limiting when it comes to assessing disability bias, since the relevant tests incorporate images instead of words. For this reason, there has been more work in downstream tasks such as sentiment analysis and topic modelling (Hutchinson et al., 2020; Hassan et al., 2021), and less in direct association analysis.

2.2 Stereotype Content Model (SCM)

Stereotypes have been studied in social psychology for decades (Asch, 1946; Greenwald et al., 1998; Fiske et al., 2007). To concisely summarize the current knowledge about stereotypes, Fiske et al. (2002) proposed the SCM, which postulates that stereotypes can be aligned along two dimensions: competence and warmth. When we meet someone new, our first psychological response is to subconsciously evaluate whether they are a friend or a foe. This is a judgement along the warmth dimension. Immediately after we make this evaluation, we go on to evaluate how well they may be able to act in accordance to our perception of their warmth. Abele et al. (2016); Nicolas et al. (2021) suggested that these dimensions can be further split into two subdimensions. Warmth is comprised of Morality and Sociability, and competence is comprised of Agency and Ability.

Researchers working under the SCM framework also propose a causal link between stereotypes and structural stigma (Fiske et al., 2007). People perceived as warm and competent evoke feelings of pride and admiration, whereas people perceived as cold and incompetent evoke feelings of disgust and contempt. Ambivalent perceptions involving warmth and incompetence typically elicit pity and sympathy. Coldness and competence evokes envy and jealousy. These biases, whether explicit or implicit, can lead to harms if they are perpetuated in AI-based assistive technologies.

To the best of our knowledge, Fraser et al. (2021) is the only work to date that has applied the SCM to analyze stereotypes in text. The SCM has not yet been used to investigate stereotypes in NLP models.

3 Methods

Following Kurita et al. (2019) and Bartl et al. (2020), we measured association bias in BERT using a fill-in-the-blank task, and synthetic, semantically bleached sentence templates. Our goal was to directly examine representations in the model, without potential interference from unexpected context or downstream input, which may occur when using natural sentence templates or with tasks such as sentiment analysis and topic modelling.

3.1 Data

Table 1 displays the targets, stereotype attribute dimensions, and sentence templates used in our study. For the targets, we used three abled/disabled antonym pairs to represent the concepts of ability and disability for general ability, deafness, and blindness. We recognize that some words such as "hearing" may not be commonly used in mainstream society, and in turn may not appear often as a person-describing qualifier in the Wikipedia and Books Corpus, which BERT was pre-trained on.

Targets	
disabled	abled
deaf	hearing
blind	sighted

Stereotype Dimension	Subdimension	Attributes
Warmth	Sociable	155
	Unsociable	156
	Moral	159
	Immoral	334
Competence	Able	153
	Unable	127
	Independent	156
	Dependent	109

	Templates
1	A `[TARGET]` person is `[ATTRIBUTE]`.
2	`[TARGET]` people are `[ATTRIBUTE]`.
3	A person who is `[TARGET]` is `[ATTRIBUTE]`.
4	People who are `[TARGET]` are `[ATTRIBUTE]`.

Table 1: Targets, stereotype attribute dimensions, and semantically bleached templates. The syntactic structure of templates 1 and 2 is typical of identity-first language, whereas templates 3 and 4 use person-first language.

However this word represents how the members of the deaf community describe those who hear. It is important to explore how a model may represent a word that has different usage in certain communities, if the model is used in end-applications by those communities.

Taking inspiration from Fraser et al. (2021), we constructed the stereotype subdimensions using the extended lexicon created by Nicolas et al. (2021), with the four subdimensions of Morality, Sociability, Agency, and Ability. In this lexicon, words are annotated with either +1 or -1 to indicate a positive or negative association with the given subdimension. We removed words that were not labelled with either valence value. We represent each valence pole of these subdimensions as their own subdimension, e.g. words with a negative association to Morality represent the Immoral subdimension. We expect these 8 subdimensions to provide a more granular understanding of stereotyped representations in BERT.

We used four semantically bleached sentence templates, which are shown in Table 1. We adapted them from Kurita et al. (2019) and Hutchinson et al. (2020). The first two templates use identity-first language, in which `[TARGET]` precedes "person." Despite removing context, the syntactic structure of the sentence itself is known to carry cultural connotations (Beukeboom and Burgers, 2019; Shakespeare, 2016). Members of the deaf community often prefer to use identity-first language, whereas the person-first language is usually found in a medical lens. To get a general picture of associations, we also include two templates that use person-first language, in which `[TARGET]` follows "person."

We removed words that would not fit the grammar of our selected templates. We kept adjectives, as identified by WordNet part-of-speech labelling. This leaves 1,256 unique words in this lexicon. Most belong to one subdimension, while 87 words belong to two subdimensions (e.g. "negligent" belongs to both the Immoral and Unable subdimensions), and 3 words belong to three subdimensions (e.g., "ingenuous" belongs to the Sociable, Immoral, and Unable subdimensions).

To further reduce possible causes of variation, we also removed all multi-word attributes. Although we are able to mask a couple of words in a sentence when feeding it to BERT, as done in Bartl et al. (2020), it is not possible to predict the probability of a multi-word phrase, only a single subtoken. Most of our targets are whole tokens, except for "abled," which is a multi-token word: "able" + "ed". We multiplied the probabilities for the subtokens that make up this word, since it is implicit that these subtokens are associated.

The final dataset consisted of 30,144 combinations of targets, attributes, and templates.

3.2 Measuring Bias in BERT

We used the `PyTorch` implementation of the `transformers` library from HuggingFace, a widely used hub for the distribution of pre-trained Transformer models (Wolf et al., 2020). We downloaded `bert-base-uncased`, the most popular version of BERT according to download count, along with a language modeling head on top and its tokenizer.

Below we outline our methodology to measure bias in BERT, which we adapted from Kurita et al. (2019).

1. Prepare semantically bleached template sentences. For example,

   A [TARGET] person is [ATTRIBUTE].

2. For each combination of target, attribute, and template,

   (a) Fill in the template.
   "A deaf person is eligible."

   (b) Mask the target.
   "A `[MASK]` person is eligible."

   (c) Compute the target's probability, given the context provided by the attribute.

   $p_x = P(\texttt{[MASK]}=\text{"deaf"} \mid sentence)$

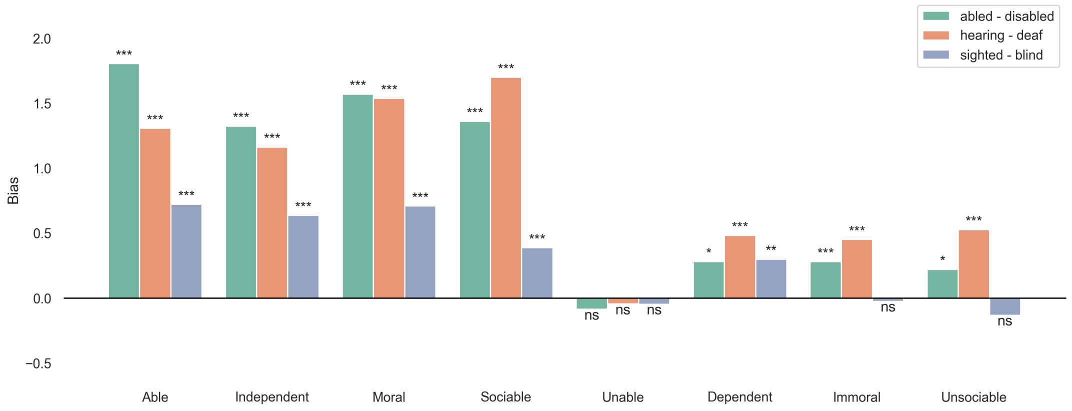

Figure 1: Bias scores for pairs of targets, when the target is predicted in the presence of the attribute. Each bias score is annotated with statistical significance where $n.s.$ means the bias is not significant at $p > 0.05$, $*$ is $p \leq 0.05$, $**$ means $p \leq 0.01$, and $***$ is highly significant at $p \leq 0.001$. The further the score gets from zero, the more unequal the representations of ability and disability. Scores above zero indicate that BERT more closely associates the abled target with the corresponding stereotype subdimension, whereas scores below zero indicate a bias where the model prefers the disabled target more, given the stereotype context. These results show evidence of significant, nuanced bias in how BERT represents disability, compared to ability.

(d) Mask both the target and attribute.
"A [MASK] person is [MASK]."

(e) Compute the target's prior probability, given no context.

$$p_{prior} = P(\texttt{[MASK]}=\text{"deaf"} \mid masked\_sentence)$$

(f) Compute the association (a) between the target (x) and attribute (m).

$$a_{x,m} = \log\left(\frac{P_x}{P_{prior}}\right)$$

(g) Compute the mean association score (A) between the target (x) and the attribute subdimension (M).

$$A_{x,M} = mean_{m \in M}\, a_{x,m}$$

(h) Compute the bias score for the attribute subdimension (M) as the difference between the mean association scores for two targets.

$$bias_M = A_{y,M} - A_{x,M}$$

If the association is negative, this means that the target's probability is lower than its prior probability. In other words, the attribute's context *decreased* the probability that BERT predicts the target. Likewise, if the association is positive, the context *increased* the target's probability of being predicted.

In all bias calculations, the minuend is the abled target's association score, and the subtrahend is the disabled target's association score. Thus, if the bias is positive, the association between the abled target and the attribute subdimension is stronger. If the bias is negative, the disabled target is more strongly associated to the attribute subdimension. If the bias is zero, there is no difference in the probability of predicting either target, given the context.

We measured statistical significance via a paired-attribute permutation test over $A_{y,M}$ and $A_{x,M}$.

We also performed the inverse analysis, where we explored the representation of stereotype content given the presence of ability or disability. To carry out this analysis, we essentially treated attributes as targets, meaning that we masked the attribute and computed its probability, given the context provided by the target. Aside from this swap, the overall methodology remains the same.

4 Results and Discussion

Figure 1 displays the bias score between each pair of targets (abled/disabled antonyms, e.g. "hearing" and "deaf") for each stereotype subdimension in the SCM. Here we can see certain patterns in how disability is represented in BERT, compared to ability.

The first takeaway from this figure is that there is a bias, or a difference, in the representations, confirming **RQ1**. The bias is significant at varying levels across all subdimensions except the Unable subdimension. Correlation in language usage may have contributed to the lack of bias in the Unable subdimension. Mentions of disability are often accompanied by words referring to ability, and often in a negative, medical context where disability is

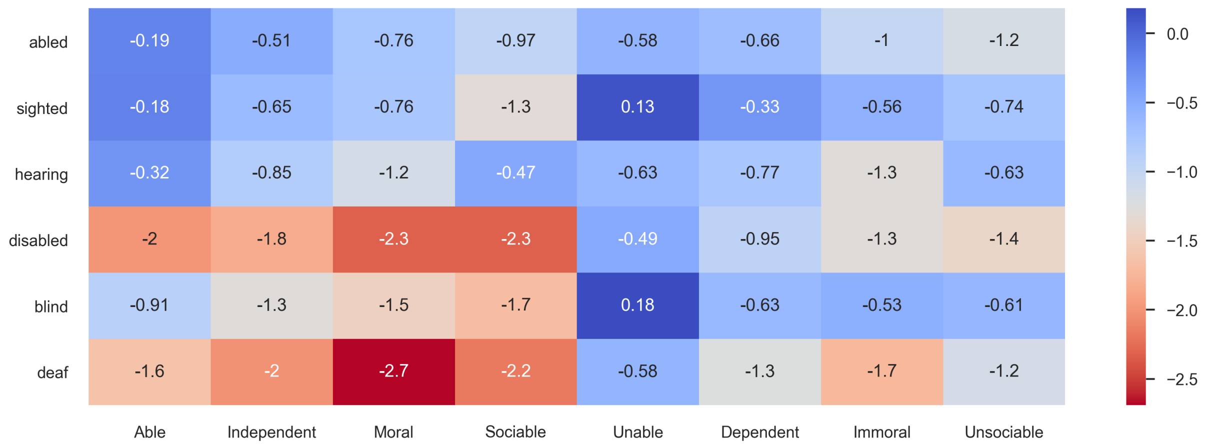

Figure 2: Mean association scores for each combination of target and stereotype subdimension. The further the score is from zero, the stronger the association is in BERT. If the score is above zero, this means that BERT positively associates the target with the stereotype subdimension. Conversely, if the score is below zero, BERT negatively associates the target with the stereotype subdimension. These results reveal patterns in how BERT's representations of ability and disability align to known stereotype subdimensions.

framed as a problem on the body, rather than on society (Shakespeare, 2016).

The second takeaway is that BERT is generally more likely to associate the abled target to all stereotype subdimensions, except the Unable subdimension for all three pairs of targets, and the Immoral and Unsociable subdimensions for blindness. This partiality toward ability may been caused by higher frequencies of abled targets in the training data (Schick and Schütze, 2020). People with disabilities are an underrepresented population and are thus mentioned less in mainstream text; there is an ongoing project to improve one of the training datasets to create more text related to disability (Wikipedia contributors, 2022). It is also less common to use an abled target to describe a person without a disability (Beukeboom and Burgers, 2019), and this in addition to these words' increased frequency may have led BERT to "understand" them better but in different contexts.

The third takeaway is that the bias is stronger if the sentence includes a positive warmth (Moral, Sociable) or competence (Able, Independent) context, presenting a high-level insight into **RQ2**. Given a positive stereotype context, BERT is more likely to predict the abled target than the disabled target in the fill-in-the-blank task. In other words, BERT is less likely to associate disability to warmth and competence. This bias is significant for ability, deafness, and blindness at $p \leq 0.001$.

On the other hand (or the other side of the figure), the bias between abled/disabled antonym target pairs is weaker if the sentence includes a negative warmth (Immoral, Unsociable) or competence (Dependent) context. This smaller difference in representation is still significant for deafness at $p \leq 0.001$, significant for general ability at varying levels, and significant for blindness with only the Dependent subdimension at $p \leq 0.01$.

To investigate **RQ2** in more depth, we show in Figure 2 the mean association scores for each combination of target (an abled or disabled antonym) and stereotype subdimension. This figure reveals more nuanced patterns in BERT's representation of disability and how this representation aligns to stereotype subdimensions from the SCM.

One pattern that stands out is that almost all of the mean association scores are negative, regardless of target or subdimension. A negative association score indicates that BERT is less likely to predict the target given the stereotype content *and* the syntactic structure of the sentence template. These negative association scores provide further support for BERT having limited knowledge about abled targets' range of usage, and/or the under-representation of disabled targets in the model.

Figure 2 also sheds additional light on the weaker bias shown in Figure 1 for negative subdimensions. Although BERT may have an overall preference for abled targets, the disabled targets' associations to these negative subdimensions are strong enough to appear nearly on par with the abled targets' associations to the same subdimen-

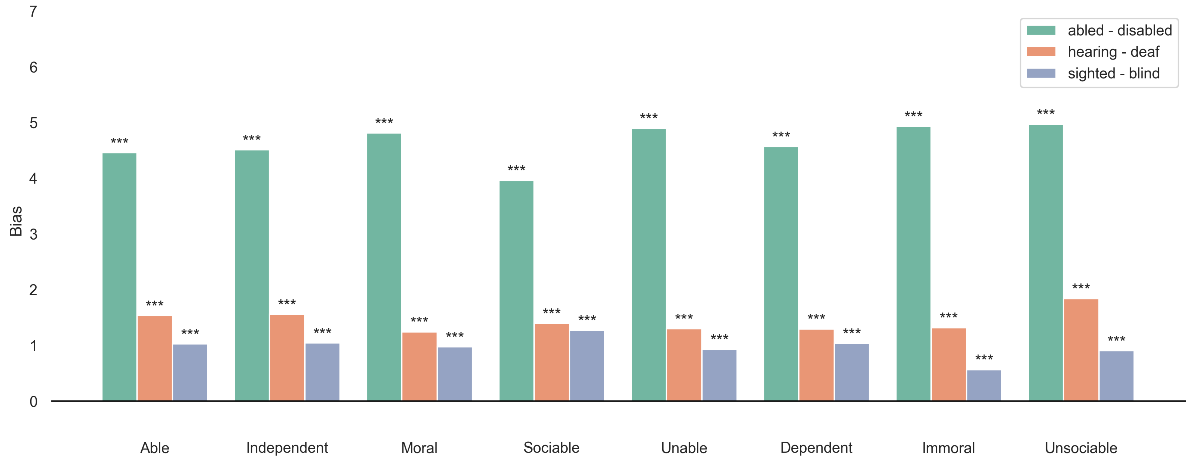

Figure 3: Bias scores for pairs of targets, when the attribute is predicted in the presence of the target. For interpretation details, please refer to Figure 1. These results show evidence that BERT is less likely to predict any attribute given an accompanying disability context. BERT contains significantly stronger associations between all stereotype attribute subdimensions and the abled target.

sions.

A third takeaway from Figure 2 is that disabled targets are less associated with Able, Independent, Moral, and Sociable contexts, compared to all other associations. This is especially pronounced with "disabled" and "deaf".

In Figure 4, the bias scores from the inverse analysis present evidence that predicting different attributes given the same target do not lead to different biases. Different stereotype subdimensions are not any closely combined with different targets, when the target context is already present in the sentence. However, BERT shows a general preference for predicting any attribute in the presence of abled targets, since the bias scores are all significantly positive, especially for ability.

We want to note that, despite semantic bleaching, syntactic differences in the sentence templates affected the strength of the association scores, but not the patterns. When using identity-first templates to predict a target given stereotype content, BERT more strongly associated "abled" and "hearing" to all subdimensions, whereas "sighted", "disabled", "blind", and "deaf" had stronger associations to all attribute subdimensions using person-first templates.

This is interesting, because identity-first and person-first language are known to carry cultural connotations. Furthermore, some common identity-first disability qualifiers, such as "disabled" and "deaf", and "blind" are used in contexts outside of social identity categories, e.g. as metaphors: "deaf as a post," "deaf and blind to [insert situation]". This may have impacted how they were understood by the model, and subsequently how they are predicted in identity-first or person-first language contexts.

5 Conclusions and Future Work

Regardless of how biases manifest, the first step toward ensuring harmless use of AI-based assistive technologies is to understand how target users are represented in the underlying models. By applying the Stereotype Content Model to evaluate representational differences, we present evidence of disability association bias in a popular pre-trained NLP model that is used in state-of-the-art AI-based assistive technologies such as text simplification and speech recognition.

We also present a breakdown of this bias along stereotype dimensions, which uncovers nuanced patterns in undesirable associations between disability and stereotypes, the most notable being that disabled people are significantly less likely to be associated to warmth and competence. Our results emphasize the need to work toward more fair and inclusive assistive technologies, especially since disabled people are the target population for these tools.

There are a number of limitations with our study. First, we explored these associations through a broad lens, looking at only ability versus disability. It is important to recognize that disability is not a siloed, unitary concept (Peña et al., 2016). Future work should investigate the associations through an intersectional lens (Crenshaw, 1989), to better understand how disability bias is affected by the interconnected nature of social categorizations.

A second limitation of our study is our usage of sentence templates. Despite attempts to semantically strip a sentence to provide a neutral context, BERT still draws on the syntactic structure of the sentence itself to help make its predictions (Devlin et al., 2019). We took this into consideration by varying the structure. However, we observed that association strengths appear to be influenced to a degree by syntactic differences. Future work can investigate stabilizing the bias evaluation metrics by including more templates and a wider range of sentence structure, or randomly sampling a natural sentence dataset. It would also be interesting to further differentiate between identity-first and person-first language, as well as to explore question-answering templates.

Third, we examined a limited number of targets and only in one model, BERT. Future work can extend our approach to evaluate additional disabled targets in additional models, such as GPT-2 (Radford et al., 2018) and GPT-3 (Radford et al., 2019), to get a fuller picture of disability representation in a wider range of popular pre-trained NLP models underlying AI-based assistive technologies.

Future work can also draw on debiasing approaches to mitigate bias in these models. We want to note that it is important in this work to also take into consideration the specific model deployment context, because enforcing fairness in an inappropriate context can result in the unintended erasure of a marginalized population (Blodgett, 2021). We provided an array of possible causes of the stereotype patterns that we observed, and these can be avenues for exploring debiasing solutions.

References

Andrea E. Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the Fundamental Content Dimensions: Agency with Competence and Assertiveness—Communion with Warmth and Morality. *Frontiers in Psychology*, 7.

Oliver Alonzo, Lisa Elliot, Becca Dingman, and Matt Huenerfauth. 2020. Reading Experiences and Interest in Reading-Assistance Tools Among Deaf and Hard-of-Hearing Computing Professionals. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–13, Virtual Event Greece. ACM.

S. E. Asch. 1946. Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3):258–290.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. *arXiv:2010.14534 [cs]*.

Camiel J. Beukeboom and Christian Burgers. 2019. How Stereotypes Are Shared Through Language: A Review and Introduction of the Social Categories and Stereotypes Communication (SCSC) Framework. *Review of Communication Research*, 7(1):1–37.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]*.

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:1607.06520 [cs, stat]*.

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

Yung-Sung Chuang, Chi-Liang Liu, Hung-yi Lee, and Lin-shan Lee. 2020. SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering. In *Interspeech 2020*, pages 4168–4172. ISCA.

Kimberlé Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. page 31.

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.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902.

Susan T. Fiske, Amy J.C. Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and Countering Stereotypes: A Computational Approach to the Stereotype Content Model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Eugenie Nicole Gertz. 2003. *Dysconscious Audism and Critical Deaf Studies: Deaf Crit's Analysis of Unconscious Internalization of Hegemony within the Deaf Community*. Ph.D. thesis.

Anthony G Greenwald, Debbie E McGhee, and Jordan L K Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. page 17.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the Interdependent Systems of Discrimination: Ableist Bias in NLP Systems through an Intersectional Lens. *arXiv:2110.00521 [cs]*.

Mark L Hatzenbuehler. 2016. Structural stigma: Research evidence and implications for psychological science. *American Psychologist*, 71(8):742.

Tom Humphries. 1977. *Communicating across cultures (deaf-hearing) and language learning*. Union Institute and University.

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.

Sushant Kafle, Abraham Glasser, Sedeeq Al-khazraji, Larwan Berke, Matthew Seita, and Matt Huenerfauth. 2019. Artificial intelligence fairness in the context of accessibility research on intelligent systems for people who are deaf or hard of hearing. *SIG ACCESS*, (125).

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.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gandalf Nicolas, Xuechunzi Bai, and Susan T. Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Edlyn Vallejo Peña, Lissa D. Stapleton, and Lenore Malone Schaffer. 2016. Critical Perspectives on Disability Identity. *New Directions for Student Services*, 2016(154):85–96.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. page 12.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

Timo Schick and Hinrich Schütze. 2020. Rare Words: A Major Problem for Contextualized Embeddings and How to Fix it by Attentive Mimicking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8766–8774.

Tom Shakespeare. 2016. The social model of disability. In Lennard J Davis, editor, *The Disability Studies Reader*, fifth edition, chapter 13, pages 190–199. Routledge.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*.

Christen Szymanski. 2010. An open letter to training directors regarding accommodations for deaf interns. *AAPIC-E Newsletter*, 3(2):16–17.

Task Force Members and Contributors. 2012. Final report of the task force on health care careers for the deaf and hard-of-hearing community.

WHO. 2021. Disability and health. [Online; accessed 03-March-2022].

Wikipedia contributors. 2022. Wikipedia:wikiproject disability. [Online; accessed 03-March-2022].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*.