Automatically Neutralizing Ableist Language in Text

Stanford CS224N Custom Project

Tiffany Liu
Symbolic Systems
Stanford University
ttliu00@stanford.edu

Tyler Shibata
Mathematics
Stanford University
yoyo2000@stanford.edu

Abstract

Ableism involves the systemic oppression or discrimination against people with disabilities. It is often reinforced through language that perpetuates harmful biases and stigmatizes those with disabilities. However, such language can often be difficult to detect due to its pervasiveness in mainstream media. To address this issue, we introduce the first parallel corpus of ableist language, as well as a model for natural language generation that automatically brings ableist text into a neutral point of view. Our corpus contains 1500 sentence pairs that originate from movie scripts, news articles, and speech transcripts. Our language generation model is a CONCURRENT system that utilizes a BERT encoder to identify and replace ableist words and phrases as part of the language generation process. In addition, we contribute a self-training pipeline that can generate more training data for the task of neutralizing ableism, as well as a novel evaluation method to more quantitatively assess a model's prowess at reducing bias. Human evaluation and our novel evaluation method suggest that these data and models are a first step towards the automatic identification and reduction of ableism in text.

1 Introduction

Ableism involves the systemic oppression or discrimination against people with disabilities, and is often reinforced through language [1]. Ableist language uses terms associated with disability to mock, insult, or degrade, perpetuating harmful biases and stigmatizing those with disabilities. For example, the sentence "He is confined to a wheelchair" is ableist, as it implicitly associates wheelchair usage with confinement; a non-biased sentence would use a verb like "uses" rather than "is confined" so as not to presuppose the ableist view that wheelchair usage is a form of being trapped. While previous research has investigated unintended biases in NLP systems against other historically marginalized groups [2] [3] [4], bias against disability has yet to be fully explored in NLP literature.

In this project, we aim to develop an NLP system that not only identifies but automatically reduces bias against disability in text. in particular, we hope to neutralize text by suggesting edits that would make it less ableist. We develop a corpus of ableist and neutralized text, adopt a pre-trained BERT model to neutralize ableist text, and devise a self-training data generation pipeline to improve our model's performance. We also propose a novel evaluation method, which more quantitatively measures a given model's proficiency in bias reduction beyond currently-standard approaches of human evaluation.

Using these approaches, we discover that our models are capable of generating neutralized text from ableist text, and that using training data generated by our self-training pipeline improves model performance. According to our bias-mitigation evaluation method, as well as various quantitative and qualitative evaluation metrics, our models appear to successfully reduce bias against disability in text.

2 Related Work

2.1 Capturing Bias in Disability Rhetoric

Language has long been used to marginalize and devalue those with disabilities [1]. However, current research on building equitable and inclusive NLP systems largely focuses on algorithmic bias in gender and race, while the intersection of NLP systems and disability has been largely untouched. The only quantitative research in this area explores undesirable biases in mentions of disability within two English-language models: toxicity prediction and sentiment analysis [5]. This research discovered that representations encoded in NLP models often inadvertently perpetuate undesirable social biases against those with disabilities, due to biased training data.

From a qualitative lens, previous works in disability studies have explored the use of language as a tool of perpetuating ableism, which manifests in metaphors, jokes, and euphemisms that institutionally devalue bodies and minds that are deemed deviant, abnormal, and defective.[6][1]. However, the task of identifying ableism can often be difficult, as ableist language is pervasive in mainstream media and no standard frameworks exist for identifying and classifying ableist *language*. In fact, one of the only standard measurements of ableism is the Symbolic Ableism Scale [7], which measures explicit disability attitudes rather than language. Other psychological frameworks of identifying ableism elucidate disability myths and stereotypes [8] [9], but do not account for linguistic ableism in particular.

In our project, we aim to combine previous research in disability studies, sociolinguistics, and NLP to develop a more concrete method of measuring and mitigating ableist bias in language.

2.2 Automatically Neutralizing Subjective Bias in Text

Pryzant, Martinez, and Daas [10] created the first *generative* model which neutralizes biased text, and added three useful tools and frameworks to the conversation: the Wiki Neutrality Corpus (WNC), which is a corpus of 180,000 sentence pairs of subjective and neutralized text from Wikipedia, and two generative models which were trained on the WNC and used to: (1) identify subjective bias in text and (2) propose edits to the text to neutralize it. Though there have been numerous explorations of bias identification and text generation separately, the *join embedding* architecture used to integrate both tasks is groundbreaking. This paper appears to be the first to be able to both identify bias in text and utilize the identification algorithm to directly fine-tune a generative algorithm. Additionally, the construction methodology of the Wiki Neutrality Corpus can be used as a framework for constructing other types of bias-related corpora.

Their work, however, is constrained exclusively to subjective bias, though their methodology lays ripe groundwork for exploring the process of mitigating other forms of bias. In our project, we extend the application of their model to address ableism, while developing a more robust training data generation pipeline using self-training. We also introduce a novel evaluation metric beyond human evaluation that more quantifiably captures a model's proficiency in reducing bias in text.

3 Approach

3.1 Model Architecture

Our models fine-tune a CONCURRENT system, as proposed by Pryzant, Martinez, and Daas [10], which leverages a BERT encoder and a token-weighted loss function to identify ableism and generate neutralized text. The CONCURRENT model is a an encoder-decoder neural network, in which the encoder is BERT while the decoder is an LSTM decoder that generates text one token at a time by repeatedly attending to the hidden states and producing probability distributions over the vocabulary. Once the detection and editing modules have been pre-trained, they are joined and fine-tuned together as an end-to-end system for sentence translation.

This is done with a novel *join embedding* mechanism that lets the detector control the editor. The *join embedding* is a vector $\mathbf{v} \in \mathcal{R}^h$ that is added to each encoder hidden state in the editing module. This operation is gated by the detector's output probabilities $\mathbf{p} = (p_1, \ldots, p_n)$. Note that the same $\mathbf{v}$ is applied across all timesteps.

$$\mathbf{h}'_i = \mathbf{h}_i + p_i \cdot \mathbf{v}$$

The decoder is then conditioned on the new hidden states $\mathbf{H}' = (\mathbf{h}'_1, \ldots, \mathbf{h}'_n)$ which have varying amounts of v in them. Intuitively, $\mathbf{v}$ is enriching the hidden states of words that the detector identified as subjective. This tells the decoder what language should be changed and what is safe to be be copied during the neutralization process.

3.2 Procedure

No parallel corpus of ableist language currently exists. Thus, we put forth the Ableism Neutrality Corpus (ANC): an original parallel corpus of 1,500 ableist and neutralized sentence pairs that can be used as training data for the task of automatically reducing ableism in text.

Upon pre-training the CONCURRENT model on the Wiki Neutrality Corpus (WNC), we fine-tune on 1,200 sentence pairs of our ANC to construct our first model checkpoint, **FineTune**. To account for training data scarcity, we then employ a self-training pipeline using FineTune to construct **SelfTrain**. To do so, we crawl an additional 4,500 ableist source sentences from the Internet and use FineTune to generate corresponding neutralized sentences. We prune the FineTune-generated sentences based on perplexity, which is defined as the exponential of the cross entropy of the generated texts and the target texts. Namely, for each model-generated sentence, we calculate the average perplexity score measured against a huggingface BERT. We then compile the training dataset for SelfTrain, which consists only of the FineTune-generated sentences with an average perplexity score less than 5.0.

As a baseline, we employ Pryzant, Martinez, and Daas's original debiaser model [10]. Because this model was trained to reduce overall subjective bias in text without a specific focus on ableism, it serves as a cogent baseline to assess our model's performance in neutralizing ableist text in particular.

We employed three original evaluation methods: a binary classifier to evaluate *identification* of ableism in text, human evaluation to assess the *quality* of neutralized text generation, and word embeddings visualized by a Word2Vec model for evaluating *proficiency* in reducing bias in text.

4 Experiments

4.1 Data

Drawing from disability studies literature that outlines mainstream disability myths [8], alternatives for ableist phrases [6], and standard language about disability [11], we propose a new framework to categorize types of ableist language: derogatory depictions of disability, equating disability to pathology, trivializing disability, euphemizing disability, using language that is non-inclusive of disability, using disability as metaphor, and using disability as idiom (see Table 1 for examples of each type of ableism).

Given the lack of corporal data surrounding ableism, we developed the Ableism Neutrality Corpus (ANC), the first parallel corpus of ableist text (Table 1). This dataset consists of 1,500 ableist and neutralized sentence pairs along with metadata. To construct our corpus, we referenced the Disability Language Style Guide from the National Center on Disability and Journalism, a language guide from People with Disability Australia, the Caltech Center for Inclusion and Diversity's Ableist Terms and Phrases, and disability rights activist Lydia Brown's Glossary of Ableist Terms and Phrases, in order to better delineate types of ableist language and discover alternative terms and phrases to use. We crawled 1,500 sentences that contain linguistic ableism from the Internet, such as headlines from news aggregator sources like Google News, lines from movies through the movie script database QuoDB, and sentences from speeches through Truman State University's speech transcript database. We extracted the sentences that contained ableist language and neutralized them by hand to form our corpus sentence pairs.

The full corpus does not include labels for all subcategories, but we hand-labeled a random sample of 300 examples to approximate the distribution of the 7 types of ableism. As shown by Table 2, all types of ableism are reflected in the data, with use of disability as metaphor and derogatory depictions of disability appearing most frequently.

Source	Target	Subcategory of ableism
She raises our three kids. Two of them are autistic.	She raises our three kids. Two of them are autistic.	non-ableist use
And I know about your experiments with the inmates of your **nut house**.	And I know about your experiments with the inmates of your **psychiatric hospital**.	derogatory depiction of disability
We can no longer **turn a blind eye to** the damage done to our seas.	We can no longer **feign ignorance about** the damage done to our seas.	using disability as idiom
Indeed, when communism constituted one of the two poles in the previous **bipolar** world order, terrorist acts were few and far between.	Indeed, when communism constituted one of the two poles in the previous **rapidly-changing** world order, terrorist acts were few and far between.	trivializing disability
Apparently, she's **confined to a wheelchair**.	Apparently, she **uses** a wheelchair.	equating disability to pathology
There may be a prophet hidden inside each of us, but we tend to be **deaf to** such warnings.	There may be a prophet hidden inside each of us, but we tend to **deliberately ignore** such warnings.	using disability as metaphor
Their son is a **special needs case**.	Their son is **disabled**.	euphemizing disability
But neither of the victims, he concedes, were the most honest and **upstanding** of people.	But neither of the victims, he concedes, were the most honest and **respectable** of people.	using non-disability-inclusive language

Table 1: Samples from the Ableism Neutrality Corpus (ANC). 300 sentence pairs are annotated with the type of ableism they portray.

Subcategory	**Percent**
derogatory depiction of disability	18.7
euphemizing disability	4.0
using disability as idiom	12.0
using disability as metaphor	25.0
equating disability to pathology	4.0
using disability non-inclusive language	8.3
trivializing disability	8.3
non-ableist language	19.7

Table 2: Proportions of ableism subcategories in the test set.

4.2 Evaluation Methods

Bag of Words: To evaluate performance on the identification step, we construct a Bag of Words vocabulary to perform a simple logistic regression. We mark each of our original 1,500 source sentences from the ANC as either ableist or non-ableist based on whether `source sequence != target sequence`. We then shuffle and split the 1,500 sentences in a 80-20 split for training and testing on a logistic regression model, and leverage the built-in scikit-learn logistic regression model to construct a simple binary classifier.

Human Evaluation: To establish evaluate performance on the generative step, we perform qualitative analysis of the model-generated sentences with respect to the hand-annotated target sentences in the test set.

Word Vectorization: To establish a more quantifiable evaluation of the overall reduction of ableist bias in our CONCURRENT model, we leverage the built-in gensim Word2Vec model, which we pre-trained with word vectors learned by a GloVe model on a Wikipedia corpus. We assess the word embedding spaces of five separate vocabularies – namely, the vocabularies of the source sequences, human-

annotated gold sequences, the baseline model-generated sequences, the FineTune model-generated sentences, and the SelfTrain model-generated sentences of the 300-sentence pair test set. Thus, we create five copies of this pre-trained model and fine-tune each model on their respective vocabularies. As such, each of the fine-tuned models has a unique vocabulary of word vectors based on the set it was fine-tuned on. For each fine-tuned model, we then project the word embedding spaces on a two-dimensional plane via PCA on the word vectors. In doing so, we visualize changes in word representations of ableist terms across model checkpoints. Namely, we observe how terms associated with disability become closer to or further from negative words in the embedding space. As a more quantifiable metric, we also observe the cosine similarities between the word vectors of ableist terms and their neutralized counterparts to assess if the CONCURRENT model indeed generated a dataset with reduced ableism.

4.3 Experimental Details

From the ANC, we sampled 12,000 sentence pairs (80%) as training data, setting aside 300 pairs for testing (20%). Following Pryzant, Martinez, and Daas's model-training procedure [10], **FineTune** was implemented with Pytorch [12] and optimized using Adam [13] [14] with a learning rate of 5e-5. We used a batch size of 16, with all vectors of length h = 512. We used gradient clipping with a maximum gradient norm of 3 and a dropout probability of 0.2 on the inputs of each LSTM cell [15]. We initialized the BERT component of the tagging module with bert-base-uncased parameters. All other parameters were uniformly initialized in the range [0.1, 0.1], and the model was trained with the maximum number of 20 epochs. Using Pryzant et al's pretrained debiaser model checkpoint, we fine-tuned FineTune on the 1,200 sentence pairs from the ANC. Training FineTune took approximately 2 hours.

To train **SelfTrain**, we used the same parameters as we did for training FineTune. Training SelfTrain took approximately 5 hours. Following the training of FineTune and SelfTrain, we used beam search and a beam width of 5 to perform inference on the 300-pair test set from the ANC.

5 Results and Analysis

5.1 Neutralized Text Generation Result Types

Through human evaluation of our model's generated sentences, we identified 7 types of results (as shown in Table 3), such that the best sentences were both neutralized and fluent if they were originally ableist, or unchanged if they were originally not ableist.

Source Sequence	Gold Sequence	SelfTrain-Generated Sequence	Type of Result
Unlike the stuffy, overpriced hotel restaurants of yesteryear, the new dining options can **stand up to** the best epicurean eateries.	Unlike the stuffy, overpriced hotel restaurants of yesteryear, the new dining options can **take on** the best epicurean eateries.	Unlike the stuffy, overpriced hotel restaurants of yesteryear, the new dining options can **take on** the best epicurean eater.	Correctly Identified & Reduced Ableism; Fluent
Now I get to fight a **retard** and a old man.	Now I get to fight a **fool** and a old man.	Now I get to fight a **re foolish** and a old man.	Correctly Identified & Reduced Ableism; Not Fluent
Some situations call for **blind** trust.	Some situations call for **unquestioning** trust.	Some situations call for **anonymous** trust.	Correctly Identified Ableism, but Wrong Type
she's been **afflicted with deafness** since she was very young.	she's been **deaf** since she was very young.	she's been **afflicted with deafness** since she was very young.	Did Not Correctly Identify Ableism
She's over 90 and has been partially **paralyzed** for over 10 years.	She's over 90 and has been partially **paralyzed** for over 10 years.	She's over 90 and has been partially **paralyzed** for over 10 years.	Correctly Identified Non-Ableist Text; Fluent
OCD is a long-term condition that can get in the way of important daily activities	**OCD** is a long-term condition that can get in the way of important daily activities	**OCD** is a long-term condition that can get in the way of important daily activities activities daily condition that can get in the way of'	Correctly Identified Non-Ableist Text; Not Fluent
Well am I gonna be **mad**?	Well am I gonna be **mad**?	Well am I gonna be **wild**?	Did Not Correctly Identify Non-Ableism

Table 3: Human evaluation of SelfTrain's results produces 7 types of performance results.

5.2 Quantifying Bias Reduction

Implementing our novel evaluation method to quantify our model's reduction of ableist bias, we sample words associated with disability and plot their associations to both ableist and non-ableist related words across the source, gold, baseline model-generated, FineTune-generated, and SelfTrain-generated vocabularies.

As shown in Figure 1, we plotted the word "wheelchair" alongside ableist connotations ("confined", "bound") as well as a non-ableist connotation ("user"). We observe that our model's vocabulary denotes the word "wheelchair" as further away in Euclidean distance from the words "confined" and "bound," and closer to the neutral word "user." In contrast, the source sequence vocabulary and baseline model-generated vocabulary associates "wheelchair" more closely with "confined" and "bound" rather than "user." This indicates that in our model, the word "wheelchair" is less associated with negative words and ableist representations. In the same vein, as shown in Figure 2, the word "nuts" is more associated with the word "peanuts" in our models, whereas "nuts" is closer to "wild" in the baseline-generated and source sequence vocabularies. This indicates that "nuts" has been neutralized in our models, such that it is no longer as associated with the ableist connotation of "nuts" meaning "wild" or "crazy."

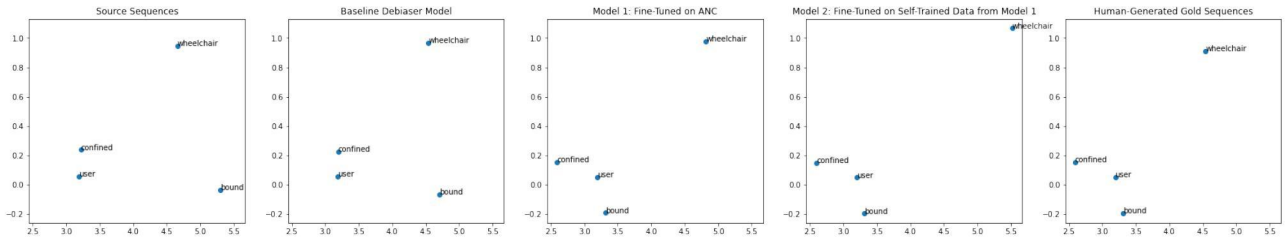

Figure 1: Visualization of "wheelchair" embedding w.r.t. related words under PCA

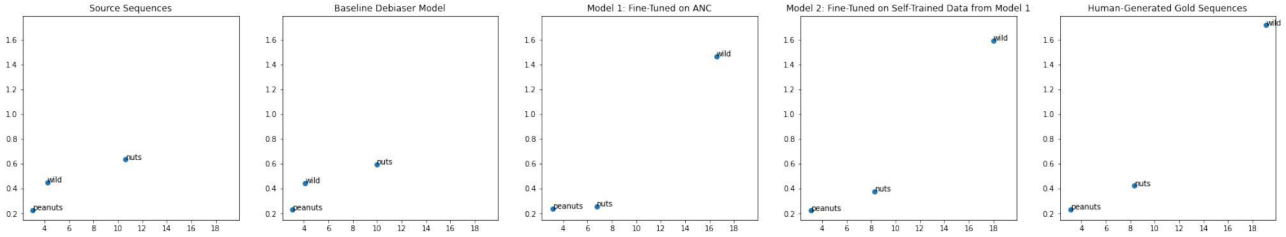

Figure 2: Visualization of "nuts" embedding w.r.t. related words under PCA

In addition, we selected various ableist terms and their neutral counterparts (i.e. "lame" <-> "uncool," "nutjob" <-> "wild card," "freaked out" <-> "spooked") and evaluated their pairwise cosine similarities across the five different vocabularies.

As shown in Table 4 for "lame" <-> "uncool," the corresponding cosine similarity between the word "lame" and "uncool" is higher for our models' vocabularies (0.26 for both SelfTrain and FineTune) when compared to that of the vocabularies of the source sequences (0.08) and baseline model-generated sentences (0.13). This aligns with our expectations for our model's behavior: as the model learns to generate the term "uncool" to replace "lame" in the ableist source sentences, the cosine similarity between the two words ought to increase. Our model appears to successfully learn to replace key ableist terms with neutralized replacement terms, albeit with the possibility that it may be over-fitting to do so.

Thus, according to our embeddings evaluation with cosine similarities and Euclidean distances, our model appears to reduce ableist associations within texts.

5.3 Performance in Identifying and Reducing Ableism

As presented in Table 5, both FineTune and SelfTrain perform significantly better in reducing ableism in text (68.2% and 72.7% accuracy, respectively) than the baseline debiaser model (28.1%

Vocabulary	Cosine Similarity
Source Sequences	0.08
Baseline Debiaser Model	0.13
FineTune: Fine-Tuned on ANC	0.26
SelfTrain: Fine-Tuned on Self-Trained Data from FineTune	0.26
Human-Generated Gold Sequences	0.32

Table 4: Cosine similarity between "lame" and "uncool" for Word2Vec Models fine-tuned on each vocabulary

accuracy). SelfTrain (trained first on the ANC and then on self-trained data) achieved a massive 215% improvement in ableism reduction from the baseline model and a 10.6% improvement from FineTune (trained just on the ANC). Of the three models, SelfTrain generated the most sentences that were both completely fluent and reduced ableism correctly. SelfTrain also had the lowest instances of not correctly identifying ableism in ableist text; this occurred in 12.3% of the test set, as compared with 14.7% of the test set and 56% of the test set for FineTune and the baseline model, respectively.

	Percentage of Test Set		
Result Type	**Baseline**	**FineTune**	**SelfTrain**
Correctly Identified & Reduced Ableism; Fluent	0.13	0.34	0.37
Correctly Identified & Reduced Ableism; Not Fluent	0.10	0.21	0.22
Correctly Identified Ableism, but Wrong Type	0.01	0.09	0.08
Didn't Ientify Ableism	0.56	0.15	0.12
Total Ableism Reduced	**0.28**	**0.68**	**0.73**

Table 5: Comparison of model performance on the ableist text in the test set.

Note that the binary classifier built on a simple Bag of Words vocabulary of the test set performed with 83% accuracy in *identifying* ableist text. This is most likely due to the fact that the task of identifying ableism in a sentence is contingent upon a handful of ableist words that a binary classifier can simply flag and classify; for example, any sentence that contains words like "freak" or "nutjob" can immediately be identified as ableist. Though our model performs with slightly lower accuracy, our model's additional functionality of incorporating the ableism identification algorithm into a generative one makes this a reasonable trade-off.

As shown in Table 5, FineTune and SelfTrain share similar performance quality in correctly identifying ableism but mistaking the specific subcategory of ableism, as well as incorrectly identifying ableism in non-ableist text. However, the baseline model performs the best in leaving non-ableist text unchanged (93.2% accuracy compared to 81.4% and 79.7% for FineTune and SelfTrain, respectively). We hypothesize that this is because the baseline model is not trained on a corpus that is focused on ableism and thus, it may default to leaving sentences unchanged. SelfTrain performs the worst in correctly identifying non-ableism, possibly because it is fine-tuned on self-trained data. This may make the model prone to amplifying its existing mistakes and wrongly identifying ableism where there is none.

	Percentage of Test Set		
Result Type	**Baseline**	**FineTune**	**SelfTrain**
Correctly Identified Non-Ableist Text; Fluent	0.18	0.13	0.11
Correctly Identified Non-Ableist Text; Not Fluent	0.01	0.03	0.04
Did Not Correctly Identify Non-Ableism	0.02	0.06	0.05
Total Non-Ableism Identified Correctly	**0.93**	**0.81**	**0.80**

Table 6: Comparison of model performance on the non-ableist text in the test set.

Though both FineTune and SelfTrain perform much better at reducing ableism than the baseline model in human evaluation, the baseline model achieves a highest BLEU Score (see Table 7). We

hypothesize that this is possibly because the baseline model was trained on 180,000 sentence pairs, whereas our models were fine-tuned on only 1,200 sentence pairs. Thus, the baseline model has potentially learned more about fluent sentence construction than our models, although many studies have noted the weak association between BLEU Score and human evaluation scores [16].

	Baseline	**FineTune**	**SelfTrain**
BLEU Score	81.41	80.16	71.33

Table 7: BLEU Scores for the three models.

5.4 Model Performance on Subcategories of Ableism

Honing in on the SelfTrain checkpoint, we observe that the model appears to perform best on text that uses disability as metaphor (the model reduces ableism on 85.4% of this ableism type), as well as text with derogatory depictions of disability (75.0% accuracy in correctly reducing ableism) (see Table 8). It performs most poorly on identifying language that is non-inclusive toward those with disabilities, as well as text that trivializes disability. We hypothesize that this is because using disability as metaphor and derogatory depictions of disability tend to have relatively straightforward fixes: replacing the problematic word or phrase (maniac, nutjob, etc.) with non-ableist substitutes (wild card, fiend, etc.). In contrast, language that is non-inclusive of disability and text that trivializes disability are often more context-dependent and thus more complex to handle.

Ableism Subcategory	**Model Performance (%)**			
	Reduced Ableism Correctly		**Did Not Reduce Ableism Correctly**	
	Fluent	**Not Fluent**	**Identified Incorrect Type of Ableism**	**Did Not Identify Ableism at All**
derogatory depiction of disability	0.43	0.32	0.09	0.13
euphemizing disability	0.25	0.25	0.33	0.17
using disability as idiom	0.36	0.36	0.14	0.14
using disability as metaphor	0.63	0.23	0.08	0.04
equating disability to pathology	0.25	0.42	0	0.33
using disability non-inclusive language	0.4	0.16	0.04	0.36
trivializing disability	0.36	0.2	0.12	0.28

Table 8: SelfTrain's performance on various subcategories of ableism represented in the test set.

Overall, we observe that the results obtained from our quantitative metrics and human evaluation largely align in corroborating our model's proficiency at automatically neutralizing ableism in text.

6 Conclusion

Developing algorithms to identify and reduce ableist language in texts like articles, books, and speeches can help reduce unconscious biases against people with disabilities, as well as mitigate the negative connotations associated with disability. Identifying ableist language can be difficult for humans because ableist bias is often subtle and implicit, and assessing language that is biased against those with disabilities is still a relatively new exploration. Thus, this project represents a first step towards automatically detecting and managing ableism in the real world. We contribute the first annotated corpus of ableist text, and our results indicate that our models are proficient in suggesting edits to reduce ableism in real-world text from movies, speeches, news articles, and more. Finally, we also contribute a novel evaluation method that provides a more quantitative benchmark for measuring a model's skill in reducing bias. Nonetheless, we still encounter a data scarcity problem, as our hand-labeled corpus only consists of 1500 sentence pairs, and training data generated via self-training can amplify the model's errors. In addition, our hand-labeled corpus mainly accounts for single-word, straightforward edits. Further research can tackle more complex and nuanced instances of ableism across multiple sentences and even across multiple languages. Finally, future work ought to involve disability communities, disability studies and sociolinguistics researchers, and other affected stakeholders as we strive toward addressing ableism in NLP systems.

References

[1] People with Disability Australia. What is ableist language and what's the impact of using it? In *pwd.org.au/resources/disability-info/language-guide/ableist-language/*.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

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

[4] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[5] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities. In *ACL 2020*.

[6] Lydia Brown. Ableism/language. In *Autistic Hoya*, 2021.

[7] Carli Friedman and Aleksa L. Owen. Defining disability: Understandings of and attitudes towards ableism and disability. In *Disability Studies Quarterly*, 2017.

[8] Jay Timothy Dolmage. Disability rhetoric. In *Syracure University Press*, 2014.

[9] Michelle R. Nario-Redmond, Alexia A. Kemerling, and Arielle Silverman. Hostile, benevolent, and ambivalent ableism: Contemporary manifestations. In *Journal of Social Issues*, 2019.

[10] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Dan Kurohashi, Sadao Jurafsky, and Diyi Yang. Automatically neutralizing subjective bias in text. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 480-489*, 2020.

[11] National Center on Disability and Journalism. Disability language style guide. In *Arizona State University*.

[12] Adam Pazke and Gregory Chanan Edward Yang Zachary DeVito Zeming Lin Alban Desmaison Luca Antiga Adam Lerer Sam Gross, Soumith Chintala. Automatic differentiation in pytorch. In *NIPS 2017 Workshop Autodiff*, 2017.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations, San Diego*, 2015.

[14] Jacob Devlin, Chang, Kenton Lee, and Kristina Toutanova. 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*, 2019.

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. In *Journal of Machine Learning Research*, 2014.

[16] Stephen Mussman Percy Liang Chaganty, Arun Tejasvi. The price of debiasing automatic metrics in natural language evaluation. In *Association for Computational Linguistics*, 208.

A Appendix

B Key Information to include

Mentors: We would like to thank our mentor Shikhar Murty for guidance on the project. We received advice on model training from Reid Pryzant, who constructed the model we use for the project. We also consulted Dr. Lindsay Felt, who teaches PWR 1: The Rhetoric of Disability at Stanford, on best practices for constructing our corpus.