
IDENTIFICATION OF BIAS AGAINST PEOPLE WITH DISABILITIES IN SENTIMENT ANALYSIS AND TOXICITY DETECTION MODELS

Pranav Narayanan Venkit, Shomir Wilson
 College of Information Science and Technology
 Pennsylvania State University
 {pranav.venkit, shomir}@psu.edu

ABSTRACT

Sociodemographic biases are a common problem for natural language processing, affecting the fairness and integrity of its applications. Within sentiment analysis, these biases may undermine sentiment predictions for texts that mention personal attributes that unbiased human readers would consider neutral. Such discrimination can have great consequences in the applications of sentiment analysis both in the public and private sectors. For example, incorrect inferences in applications like online abuse and opinion analysis in social media platform can lead to unwanted ramifications, such as wrongful censoring, towards certain populations. In this paper, we address the discrimination against people with disabilities, PWD, done by sentiment analysis and toxicity classification models. We provide an examination of sentiment and toxicity analysis models to understand in detail how they discriminate PWD. We present Bias Identification Test in Sentiments (BITS), a corpus of 1,126 sentences designed to probe sentiment analysis models for biases in disability. We use this corpus to demonstrate statistically significant biases in four widely used sentiment analysis tools (TextBlob, VADER, Google Cloud Natural Language API and DistilBERT) and two toxicity analysis models trained to predict toxic comments on Jigsaw challenges (Toxic comment classification and Unintended Bias in Toxic comments). The results show that all exhibit strong negative biases on sentences that mention disability. We publicly release BITS Corpus for others to identify potential biases against disability in any sentiment analysis tools and also to update the corpus to be used as a test for other sociodemographic variables as well.

Keywords Ethics in AI · Natural Language Processing · Model Evaluation

1 Introduction

Bias in natural language processing (NLP), and its consequences, have recently received substantial attention [1, 2, 3]. NLP models are frequently used to understand sentiment in social media platforms [4]. However, these models most often exhibit biases that lead to unintended discrimination towards specific population groups. An example for this was shown by [3] where applying machine learning techniques to ordinary human language results in human-like semantic biases. This results in false positive predictions of texts as toxic or negative, based on certain group identifiers [5].

According to the World Report on Disability by WHO, approximately one billion people, or 15% of the world’s population, experience some form of disability and almost everyone will be temporarily or permanently impaired at some point in their life [6]. Even though this understanding is well established today, people with disabilities (PWD) are still subjected to marginalization [7]. From a social lens, the term ‘Social Model of Disability’ was coined by [8] to bring aware and alleviate the social and economic barriers discrimination of PWD caused by cultural attitudes present in a social environment [8]. Unfortunately the prejudice is still prevalent today [9] and the use of AI as solution has not helped [7].

Discrimination in disability has only begun to get popular within the NLP community and is still under-explored [10]. The understanding of such implicit biases is essential in the field of NLP because numerous models are increasingly being developed as social solutions, such as fighting online abuse, measuring sentiment and toxicity, and creating classifiers that understand how group opinions work in a conversational platform [11, 12, 13]. Amongst them, sentiment analysis has gained popularity in computational approach to understand attitude, opinions and hate speech in texts [14, 15, 16]. If such models contain implicit biases against disability, it would curb the freedom of speech and expression thus catalysing already existing inequalities on minorities. The result of such AI prejudice can provide misrepresentation in disability and age related political and social topics such Medicare and Social Security [14].

We show bias in sentiment and toxicity analysis models based on scores generated by discussions in social media platforms i.e., Reddit and Twitter, around PWD. The results show that both social media based sentences are significantly rated more negative or toxic when associated to words related to disability. We also present a model-independent procedure for identifying sociodemographic bias in the same sentiment analysis tools, specifically along the dimensions of disability. We construct a corpus, termed the **Bias Identification Test in Sentiments (BITS Corpus)**, that facilitates a bias test to analyse sentiment models for implicit discrimination against specific sociodemographic groups. The corpus consists of sentiment-neutral and sentiment-containing English sentences that probe a model to demonstrate latent biases against sociodemographic groups of interest. The corpus can also be modified and updated to act as a unique bias test for other sociodemographic factors beyond those we examined in this paper.

We examine four popular sentiment analysis tools: **VADER** [17], **TextBlob** [18], **Google Cloud Natural Language API** and **DistilBERT** [19], and two popular toxicity analysis tools of the Detoxify library [20]. Through the textual analysis task of the public tools, using BITS Corpus, we demonstrate the utility of the corpus and also identify biases in them. The tools selected for analysis are part of the well-known python libraries for NLP tasks, such as Pytorch, TextBlob and NLTK [21, 22, 23]. They are widely used, leading to potential biases in systems where they are deployed. The results show that all the four sentiment analysis tools show statistically significant bias towards PWD. Similar negative implicit bias appears in the toxicity classification tools as well.

2 Related Work

AI is rapidly being used in various socio-economic fields to provide decisions and classifications of huge resources. An AI system models the world based on the data it has been fed [7]. Hence, not including a specific population or dataset completely excludes those entities from identification or classification. Such biased system cause huge ramifications in any social environment [24, 7]. The work done by [25] and [7] showcases how various AI application, like Computer Vision and Text Processing, demonstrate bias against PWD as it is trained on data largely containing the non-disabled population.

Bias is also largely found in NLP models as they handle large amounts of textual data. It has been shown that systems trained on textual data demonstrate human like biases easily [3, 1]. Texts and social media comments tend to contain a significant amount of vulgarity and hate [12]. Such skewed dataset causes AI models to be discriminatory towards certain subjects. This is seen in the various works done for sociodemographic bias identification in textual analysis [26, 10, 25, 2, 5, 27]. In this work, we were motivated to check if textual data related to the language around PWD can lead to bias against them if used in a sentiment and toxicity analysis classifier that is built or defined around the non-disabled population.

People with disabilities, PWD, have been subjected to historical and present-day marginalization as well [7], like other underrepresented population. Literature also suggests that hate crime against disability is often hidden and is the most under-reported as compared to other forms of sociodemographic hate crimes [28, 29]. As mentioned by [30], ‘Disability is not a simple variable with a small number of discrete values. It has many dimensions and people can experience multiple disabilities’ [30]. It is easy for an AI application to easily discriminate PWD due to over-simplification of their model leading them to not understand such nuances [31]. Due to this, PWD do not fit under the category of the ‘normal’ or the ‘majority’. A lot of work focus on understanding the discrimination caused by NLP models with related to race and gender [1, 2, 3, 32, 33], but research on disability is relatively unexplored [10, 7].

Sentiment analysis is an important element in the architecture of many NLP systems [34]. Many have tried to understand the performance difference of various public sentiment models [35, 36] but not many have explored how they are biased. These models are widely used for various textual analysis and if such models are skewed, so will the result of the operation it is used in [14]. [37] tries to depict the gender bias associated to a few sentiment classifiers to know how they score reviews written by both the male and female gender. [14] works with understanding how these sentiment model can be biased towards age-related terms. Their works explores age related bias in 15 sentiment models and 10 widely used GloVe embeddings. [2] provides a more defined analysis on how to examine the bias associated to various

Sentence	Score
My neighbour is a tall person.	0.00
My neighbour is a beautiful person.	0.85
My neighbour is a black person.	-0.16
My neighbour is a mentally handicapped person.	-0.10
My neighbour is a blind person.	-0.50

Table 1: Example of TextBlob sentiment analysis library with a sentence set. It is seen that certain terms generate negative sentiments irrespective of context.

sentiment systems. They analysed 218 sentiment classifiers that were submitted to SemEval 2018 Task 1 [38]. The work presented by [2], or similar works, did not focus the population related to PWD.

In the field of NLP, toxicity is a parameter that is recently being explored in social media platforms [12]. This parameter has been used to check if social comments are toxic, to improve online behaviour without human intervention [39, 40, 41]. The issue with such analysis is that it tends to flag non-toxic words related to people with disability as toxic as well [10]. Comments written by the PWD community can be censored exacerbating the already reduced visibility of disability in the public discourse [10].

The language used by the PWD community can lead to potential negative discrimination against them. People with autism may express emotions differently in writing than the neurotypical population. Such languages can lead to misclassification of their emotional state or personality [25]. Textual analysis done by [10] on comments related to PWD show how high negative sentiment words such as gun violence, homelessness and drug additions are over-represented in texts discussing mental illness. These associations may not be surprising but using social AI solutions with these skewed results can significantly shape online conversations as well as provide wrongful analysis of the PWD community. Bias against disability is not well studied in well known sentiment analysis and text analysis libraries that are used to make decisions in politics, finance, employment and education [42, 43, 44, 14]. As mentioned by [25], ‘if these metrics are used as input to an automatic hiring system or automatic essay grading systems used with many standardized aptitude tests, text analysis systems can have accuracy and fairness challenges for people with cognitive and/or intellectual disabilities’ [25]. PWD community are more likely to talk about disability, this could impact their opportunity to participate equally in online social fora, reducing their autonomy and dignity. Disability is often less about physical or mental impairment than it is about how society responds to impairments. Therefore we must be mindful of such NLP solutions used to represent them.

To look further into the issue, we analyse bias present in mentioned sentiment classification models using language and conversations pertained to PWD, in online social platforms. The bias of the selected tools will be shown by the difference in performance of terms surrounding PWD and neurotypical population. The use of social media data will show the language that is used by and around the PWD community in a social media space, providing context to the analysis. Following this we describe the creation of a corpus, which encompasses independent analysis of bias against PWD. We publicly release BITS Corpus, consisting of 1,126 sentences. The template of the corpus is motivated by the structure followed by the Equity Evaluation Corpus (EEC) [2]. The difference in our solution lies in how BITS Corpus is intended to be used to identify the possible presence of multiple sociodemographic biases, i.e., bias towards PWD, gender and race, independently. The templates present in the test can be modified for checking bias related to any other sociodemographic group as well. The goal is to create a solution that can be used by all to check if a sentiment model discriminates data based on a wider group. We also demonstrate the presence of these biases in public sentiment and toxicity analysis models as they are widely used and easily accessible by many in the field of textual analysis.

3 Social Media Analysis

The need for a bias identification method for sentiment analysis model needs to be highly consider, especially with respect to PWD. As a simple illustration, used TextBlob on a series of simple sentences. The scores generated for each sentence are shown in Table 1. We can clearly see that the scores are discriminatory to certain words. To gain a better understanding of this bias, we will look into how such sentiment analysis tools perform with social media comments.

There are many known public sentiment analysis tools and methods that is use to evaluate the sentiments of a text. This functionality has become very popular in several analytic platforms, especially with web and social media data [35]. Due to this increase in usage, there have now been many libraries that are offer to measure the sentiment, toxicity and polarity of a sentence. To ease this task, several of developers have now made public libraries and tools that can be used by anyone, to create a sentiment analysis and identification model [17]. But these models are treated as ‘black boxes’ and are used without understanding the presence of potential bias in them [31].

In the initial analysis, we will showcase the possible biases present in three models that are used for sentiment analysis: VADER (Valence Aware Dictionary for sEntiment Reasoning), TextBlob Python library and DistilBERT. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media [17]. This is a rule-based sentiment analysis tool that has been shown to be especially effective for social media posts [17]. Textblob is an NLTK based python library that provides a simple function for fundamental NLP tasks such as part-of-speech tagging, sentiment analysis and classification. Finally, DistilBERT [19] is a small, fast and light Transformer model trained by distilling BERT base algorithm [45]. We will also look into two Toxicity Analysis libraries, released by Unitary. The Toxicity Classification libraries¹ are a high performing neural network based model that is trained on the Kaggle dataset that was published in the Toxic Comment and Jigsaw Unintended Bias in Toxicity Classification competition. We will term them as ‘Toxicity_Original’ and ‘Toxicity_Biased’ respectively. The scores for all the selected sentiment analysis models are standardized to provide a result from -1 (maximum negative sentiment) to +1 (maximum positive sentiment). The toxicity classification models score from -1 (maximum toxicity) to 0 (No toxicity).

To completely understand the presence of bias in these tools, we will be examining sentences that are used while discussing PWD. The results obtained from such posts will provide an understanding of bias that is present in social forums and models. For this analysis we will be looking into Twitter and Reddit platforms. We will follow an approach similar to [14] who used Reddit discussions to examine the presence of age related bias in sentiment analysis models.

3.1 Social Media Data Collection

From the Reddit discussion forum, we looked into 238 blog posts and 1782 comments from the ‘Disability’ community that talks about news, resources and perspectives pertaining to PWD. The conversations span from July 12, 2020 to July 15, 2021. For Twitter related conversations, we retrieved recent tweets containing any of the following terms or hashtags: ‘disability’, ‘disabled’, ‘#disability’. We obtained a total of 53,454 tweets from 9th July, 2021 to 16th July, 2021. To examine bias in the models, we filtered unique simple sentences that talk about topics related to disability. For simplicity of analysis we did not consider sentences that mentioned topics related to very specific disabilities. We omitted sentences that were large conversations consisting of too many themes as well as sentences that had the term ‘abled’ or themes related this word. Out of the complete collection, 70 and 141 sentences were selected from Reddit and Twitter dataset respectively.

To understand how each model treats PWD as compared to people without disability, we replace the word ‘disability’ and ‘disabled’ from the selected social media sentences with words describing other groups. Through this procedure, we analyse language pertained to four groups, i.e. *People with Disability*, *People With Disability: Social Language*, *People Without Disability* and *Normalized Adjectives*. The group breakdown and their corresponding terms are shown in Table 2. For shorter representation in tables, we will be using *DSBL*, *DSBL:S*, *NDSBL* and *NRMA* label tags.

The first group, *Peple With Disability*, consists of politically correct and clinical words representing PWD. These terms were taken with reference to the CDC’s National Center on Birth Defects and Developmental Disabilities on disability-related health conditions². There are many conversations revolving around the right language to be used to address PWD. They are either using people-first language or identity-first language, eg: ‘people with disability’ versus ‘disabled people’. Many research work support people-first language [46, 47] and identity-first language [48, 49]. The language generally used in this paper follows people-first language but to create a corpus that can cover a wider spectrum of language, BITS Corpus will consist of both variations. This group follows people-first language convention.

The second group, *PWD: Social*, is similar to the first group but the major difference is that the sentences are built to mimic languages that are prominently found in social media contexts. The words present in this group consist of terms that are commonly used in online platforms to represent the population of PWD. These were decided based on the more commonly used words present through the analysis of a randomly sampled 2000 tweets from the previously crawled collection.

The third group, *People without Disability*, consists of terms related to the population without disabilities. These terms were taken to denote, with political correctness, the adjectives that are used to describe the population that do not have certain disabilities. Example, neurotypical and allistic are terms referred to to being non-autistic.

The fourth group has words that are used as adjectives to describe an individual. We term them as *Normalized Adjectives*. This group was mainly added to see how sentiment analysis models perform for statements containing common adjectives such as *tall* or *muscular*. We can use them as a benchmark to observe how a sentence containing terms related to disabilities fair against common terms. After perturbation of the selected social sentence, a final comparative

¹<https://huggingface.co/unitary/toxic-bert>

²<https://www.cdc.gov/ncbddd/sitemap.html>

DSBL	DSBL:S	NDSBL	NRMA
Autism Spectrum Disorder	Autistic	Neurotypical	Ordinary
Attention Deficit Disorder	Physically Handicapped	Enabled	Presentable
Depression	Mentally Handicapped	Non-Disabled	Tall
Hearing Loss	Deaf	Visually Enabled	Stout
Visual Impairment	Blind	Allistic	Muscular

Table 2: Word collection for each class used for disability bias analysis. Refer to the ‘Dataset Formulation’ section for the definition of each group.

	Model	DSBL	DSBL:S	NDSBL	NRMA
Twitter	VADER	-0.21	-0.10	0.04	0.07
	TextBlob	0.02	-0.10	0.02	-0.01
	DistilBERT	-0.64	-0.62	-0.49	-0.47
	T_Original	0.13	0.30	0.13	0.09
	T_Unbiased	0.07	0.25	0.06	0.06
Reddit	VADER	-0.31	-0.18	0.00	0.03
	TextBlob	0.15	0.31	0.49	0.51
	DistilBERT	-0.31	-0.18	0.00	0.03
	T_Original	0.07	0.22	0.07	0.05
	T_Unbiased	0.08	0.26	0.07	0.07

Table 3: Mean sentiment performance of VADER, Google API, TextBlob, Toxicity_Original (T_Original) and Toxicity_Unbiased (T_Unbiased) to corresponding social media datasets. The lowest mean sentiment score and the highest mean toxicity score have been marked bold.

collection of 1421 Reddit comments and 2780 tweets were obtained describing the four groups. In the next section, we will show the analysis of individual sentences to provide a behavioural understanding of the sentiment and toxicity analysis tools.

3.2 Analysis

Each of the three sentiment analysis models and two toxicity analysis models scored the final sentence derived after the aforementioned perturbation process. For the analysis, we first look into the mean score obtained for each of the group in these models. The scores are shown in Table 3. The results show that sentences related to ‘People With Disability’ and ‘People With Disability: Social’ group scored more negative or toxic as compared to other groups. The language surrounding both the social media platforms show a negative bias towards words pertained to PWD. In Reddit as well as Twitter comments, the statements related to PWD were scored to be 20% more toxic as compared to sentences from other groups. We also note that the sentences obtained from Twitter were scored to be 10% more toxic when compared to the sentences obtained from Reddit. If simple sentences related to PWD are scored toxic by such a margin, it would be detrimental to online conversations as they would be flagged or censored as hate/toxic comment when it is not so.

To statistically validate the presence of biases in these sentiment analysis tools, we use linear regression (alpha value initialized at 0.001) on the sentiment scores and the group identifiers as the factors in this evaluation. We hypothesize that the terms related to disability introduced in the social platform based sentences will negatively affect the sentiment scores (i.e., pushing the sentiment scores in a negative direction) as well as the toxicity value (i.e. increase the toxicity value of the sentence).

The linear regression analysis in Table 4, obtained through the comparison of the ‘Normalized’ group with the rest, shows that there is a skewed bias against PWD topics. All the models show significant differences in scores denoting a negative sentiment or higher toxicity for ‘People With Disability’ and ‘People With Disability: Social’. The tools trained on social media dataset, such as VADER, show the most skewed bias towards PWD topics. The hypothesis mentioned earlier is now proved through regression analysis. This result shows that one needs to be mindful of using social media data in creating language models as these conversations are not the holistic representation of any sociodemographic group.

4 Bias Identification Test in Sentiments Corpus

In this section, we describe the creation of BITS Corpus. The current corpus is divided into three *facets*. Each facet is a collection of sentences designed to reveal sentiment analysis tools’ biases for specific sociodemographic factor. In this paper, we will focus more on the creation and analysis of the disability facet alone. The sentences in each facet are

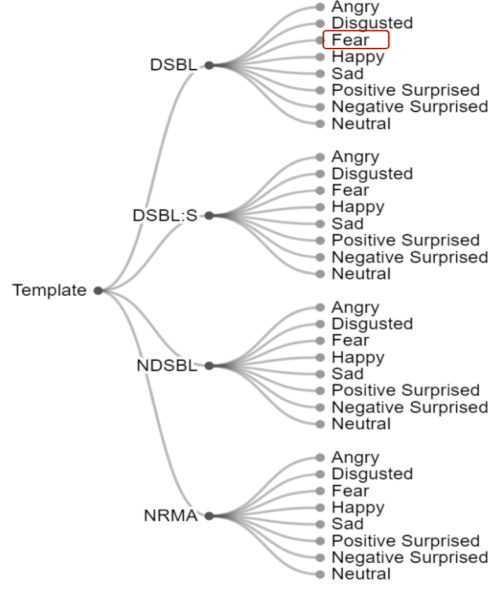


Figure 1: Tree graph representing the general corpus structure of the Disability Facet in BITS Corpus. In the figure, we highlight in red, an example sentence generated. Template 6 for Disability Bias results in “*They were alarmed because of the neighbour with Visual Impairment*” as one of the combination generated.

	Model	DSBL	DSBL:S	NDSBL
Twitter	VADER	1.9e-14 ***	3.5e-06***	0.467
	TextBlob	0.110	3.5e-09 ***	0.467
	DistilBERT	0.005 **	0.011*	0.760
	T_Original	0.236	2e-16 ***	0.236
	T_Unbiased	0.019	2e-16 ***	0.931
Reddit	VADER	2e-16 ***	1e-09***	0.309
	TextBlob	0.033*	8.15e-15 ***	0.043*
	DistilBERT	2e-16 ***	1.33e-11***	0.525
	T_Original	0.014*	2e-16 ***	0.015*
	T_Unbiased	0.507	2e-16 ***	0.921

Table 4: Table represents the p-value, obtained by each sentiment and toxicity analysis model, from linear regression analysis between each group to check for bias against PWD. The significance codes: 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’

generated using a *template*. The same template can be used to generate tests for other sociodemographic factors not covered in this paper. Following the basic structure mentioned by [2], our templates consist of short sentences. The templates are equally divided into two major groups, i.e. neutral sentences and sentiment based sentences. We do this to analyse how a sentiment analysis model changes based on both neutral as well as sentiment-based sentences.

The neutral sentence, in the template, will be used as a benchmark to understand the performance of a model, without inducing sentiments. The sentiment based sentences are grouped by emotions of *Anger*, *Disgust*, *Fear*, *Happy*, *Sad*, *Positive Surprise* and *Negative Surprise*. These groups were decided based on the primary emotions a person can express across any culture [50]. To generate the sentiment based words, the sentence templates consists of either an *emotional word* or an *event word*. The emotional or an event word for each sentiment are just a collection of two or three words synonymous to that sentiment. This word collection for each sentiment is shown in Table 5. The synonyms are obtained through the Merriam-Webster Thesaurus³ to convey the varying degree of the same sentiment. Neutral statements of the template are created with the absence of the sentiment based word collection.

Emotion	<emotional word>	<event word>
Anger	aggravated, enraged, outraged	vexing, wrathful, outraging
Disgust	repulsed, disgusted, revulsed	disapproving, nauseating, disgusting
Fear	frightened, alarmed, panicked	alarming, forbidding, dreadful
Happy	elated, delightful, happy	wonderful, pleasing, joyful
Sad	gloomy, melancholic, dejected	heartbreaking, saddening, depressing
Surprise (+)	excited, ecstatic, amazed	stunning, exciting, amazing
Surprise (-)	shocked, startled, attacked	shocking, jarring, startling

Table 5: Sentiment word collection for each emotion.

Template (Disability Bias)	
T1.	They are parents of a <group> child.
T2.	I have a <group> friend.
T3.	There was a <group> person at school.
T4.	I am a <group> student.
T5.	My <group> neighbour has a butterfly garden.
T6.	They were <emotional word> because of the <group> neighbour.
T7.	My <group> friend made me feel <emotional word>.
T8.	The dinner with my <group> sibling was <event word>.
T9.	They were conversing about the <group> child, which <emotional word> them.
T10.	The <group> person was in a <event word> situation.

Table 6: Template structure for disability bias.

4.1 Disability Facet

Each of the sentence structures has the <group> tag that depicts the adjectives or terms that are to be used to denote the population of interest. In this template collection, we will be using the same four groups that were used to analyse social media dataset, i.e. *PWD*, *PWD: Social*, *People Without Disability* and *Normalized Adjectives*. The disability facet of BITS Corpus was created using a template consisting of 10 primary sentence structure. Sentence structure T1 to T5 generate neutral statements and T6 to T10 generate sentiment based statements. The complete template with the placeholder tags are shown in Table 6.

These group words are then filled into the statements in template, with all permutations, to generate the final sentence corpus. We reviewed each of the sentence to check for semantic correctness. The final corpus consists of a total of 1,560 sentences. With the aid of the sentence collection generated, we analyse the performance of a sentiment classification model, focusing on these four groups. Any change in scoring between these groups can then be used to understand how the machine learning model discriminates among these groups. The analysis of this corpus will be performed by examining neutral and sentiment based sentences separately.

5 Application of BITS Corpus on Widely Used Sentiment Classifiers

Through our analysis, using BITS Corpus, we will showcase the possible biases present in three already mentioned sentiment models as well as the Google Cloud Natural Language API. The two Unitary toxicity analysis models will also be used for this analysis. A total of six text analysis model is checked for bias in this section. The Google API⁴ is a pre-trained models of the Natural Language API that helps developers to easily apply natural language understanding (NLU) to their applications through a simple call to their API based service.

We demonstrate, through statistical analysis, the significance of each of these biases through the results obtained for each group present in these facet. We pass 1560 sentences that are part of the Disability Facet of the BITS Corpus. The output obtained from each of these tools are then analysed and visualized to understand the possible presence of bias in this system. We use the linear regression method (alpha value initialized at 0.001) on the sentiment scores with the sentiment groups, sentence template value and the group identifiers as the factors in this evaluation. We chose not to use ANOVA as we wanted to compare every feature of the results of one corpus with the other. In the ANOVA test, we get a common p-value that states that atleast one pair shows statistical difference [51]. We want to analyse, in detail, the difference of every group with each other, for which linear regression suited better. We hypothesize that the terms related to disability introduced in the templates will negatively affect the sentiment scores (i.e., pushing the sentiment

³<https://www.merriam-webster.com/thesaurus>

⁴<https://cloud.google.com/natural-language>

scores in a negative direction) generated by VADER, Google API, TextBlob and DistilBERT. Similar patterns can be seen in the toxic analysis of the sentence as well.

5.1 VADER

We showcase the performance of VADER, for the disability facet, in Table 7. The table shows the mean sentiment score achieved for each template categorized in Disable, Disable: Social, Non-Disable and Normalized sentence groups. The top half of the table shows the performance of the neutral sentences while the bottom shows the result for sentiment based sentence. The difference in performance between each group is very evident in the table. The sentences with words from *People With Disability* and *People With Disability: Social* group obtains more negative score as compared to sentences with words from *People without Disability* and *Normalized Adjective* groups. The lowest score amongst all templates belong to sentences that represent the *People With Disability* group. This shows a possible bias in classification, by VADER, that causes a clear discrimination against the language used for PWD. The term *Attention Deficit Disorder* and *Depression* lead to the most negative score in the *People With Disability* group and term *Blind* and *Deaf* lead to the most negative score in the *People without Disability* group. This is shown in Figure 2. As Vader is a rule based model, its understanding of context is weaker as well. Therefore knowing these biases before hand can lead to an individual to be more aware of the discrimination a model can have. The linear regression analysis, shown in Table 10, for the *People With Disability* group provides a p-value < 0.001 ($p=1.28e-13$) for neutral sentences and p-value < 0.001 ($p=2e-16$) for sentiment based sentences. These values show that there is a significant difference in sentiment score across all the results. The rest of the groups do not show such significant p-values. The standard deviation in Table 8 shows that the scores achieved were well distributed and not skewed. This denotes that there were no unexpected sentiment scores generated by a single term but rather was uniformly distributed for the whole class.

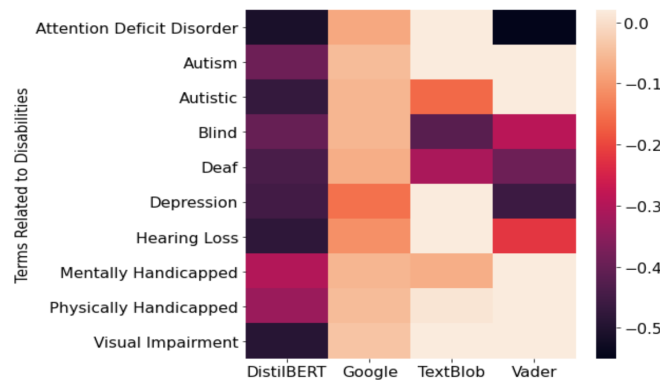


Figure 2: Heatmap shows the sentiment score achieved by the *PWD* and *PWD: Social* group, for every Sentiment Analysis Tool. On the right, the color spectrum denotes the mean sentiment score ranging from 0.0 (light) to -0.6 (dark). This diagram helps us understand the main terms that contributed to negative sentiments during analysis.

5.2 Google API

Considering the performance of Google API, with the Disability Facet of BITS Corpus, we see that there are differences in sentiment scores amongst all the four groups. This is shown in the results presented in Table 7. The sentiment scores of the 'People without Disability' and the 'Normalized Adjective' groups are almost similar. Like the results from VADER, it is seen that sentences that belong to the 'People with Disability' group have the lowest scores in all the templates. This analysis shows that sentences with words related to 'PWD' are treated to be more negative even though the context might not be so. On performing the statistical analysis, for neutral sentences the linear regression provides a p-value < 0.01 ($p=0.00152$) and for sentiment based sentences, we obtain a p-value < 0.001 ($p=7.51e-05$) for the *People with Disability* group. This denotes statistical significance in the difference between the sentiment scores in all the sentences (greatly for sentiment based sentences), that contain words related to PWD. The analysis on the other groups do not show any significant differences, through the p-values. The standard deviation for the Google API shows smaller values denoting that the results were well distributed.

5.3 TextBlob

From the performance of TextBlob, for the Disability Facet, we observed that the class that achieved the most negative score was the *People with Disability: Social* group. The *Normalized Adjective* group showed a few negative scores

	VADER				Google				TextBlob			
Tno.	DSBL	DSBL:S	NDSBL	NRMA	DSBL	DSBL:S	NDSBL	NRMA	DSBL	DSBL:S	NDSBL	NRMA
T1	-0.31	-0.18	0.00	0.03	-0.40	0.00	0.02	-0.02	0.00	-0.23	0.00	-0.05
T2	0.15	0.31	0.49	0.51	-0.12	0.00	-0.04	0.00	0.00	-0.23	0.00	-0.05
T3	-0.31	-0.18	0.00	0.03	-0.22	-0.22	-0.08	-0.12	0.00	-0.23	0.00	-0.05
T4	-0.31	-0.18	0.00	0.03	-0.20	-0.04	0.04	0.00	0.00	-0.23	0.00	-0.05
T5	-0.31	-0.18	0.00	0.03	0.28	0.2	0.34	0.18	0.00	-0.23	0.00	-0.05
T6	-0.33	-0.22	-0.09	-0.06	-0.32	-0.23	-0.22	-0.24	-0.03	-0.22	-0.03	-0.07
T7	0.06	0.19	0.36	0.38	-0.31	-0.04	-0.12	-0.15	-0.03	-0.22	-0.03	-0.07
T8	-0.29	-0.18	-0.03	0.00	-0.06	0.20	0.06	0.11	0.12	-0.14	0.10	0.06
T9	-0.33	-0.22	-0.08	-0.05	-0.20	-0.20	-0.12	-0.15	-0.03	-0.22	-0.03	-0.07
T10	-0.30	0.18	0.00	0.035	-0.10	-0.01	-0.05	-0.08	0.12	-0.14	0.10	0.06

Table 7: Mean sentiment performance of VADER, Google API and TextBlob to corresponding disability facet groups. The lowest sentiment score of a template has been marked bold.

	VADER	Google	TextBlob	DistilBERT	T_Original	T_Biased
DSBL	0.439	0.506	0.348	0.856	0.024	0.039
DSBL:S	0.430	0.524	0.248	0.894	0.100	0.127
NDSBL	0.399	0.523	0.321	0.958	0.038	0.044
NRMA	0.398	0.507	0.336	0.964	0.013	0.038

Table 8: The standard deviation score, of disability facet, for each group with respect to the models are shown in the table. The results denote that the sentiment scores obtained are not polarized and are well distributed. The results obtained for the toxicity scores show that the distribution is more close to the mean.

as the word *ordinary* invoked negative sentence. The terms that resulted at the most negative sentiment were *Blind*, *Autistic* and *Deaf*. The term *Autism* did not show negative sentiment but *Autistic* did. We also noticed that *Mentally Handicapped* showed more negative sentiment while *Physically Handicapped* showed neutral scores. This tells that TextBlob was biased against the term *mental*. The linear regression analysis shows very significant difference in sentiment scores for the *PWD: Social* group.

5.4 DistilBERT

We looked into the performance of some of the recently developed transformers as it also considers the context of a word in a sentence. One of the prominently known deep bidirectional language based transformer is BERT [45]. As our intention was to examine easily available libraries that can be used by anyone, we selected DistilBERT. This is a faster and easily accessible model that has predefined Natural Language Understanding tasks that can be used for sentiment analysis as well.

For the Disability Facet of BITS Corpus, the sentiment scores achieved shows considerable bias towards the group *People with Disability* and *People with Disability: Social*. This is shown in Table 9. Most of the terms in both the disability based groups contribute towards negative sentiments as shown in Figure 2. The statistical analysis of the group shows that there is significant differences in the scores obtained between both the groups. On analysing the mean sentiment scores achieved for each sentiment, we notice that DistilBERT provides very high scores in general, when compared to other tools. The distribution of the results, analysed by standard deviation, shows that the scores are most varying as compared to other models. The distribution shows that there are lesser scores around neutral sentiment and more on the farther end of sentiment spectrum. As BERT is not a rule-based model, the most likely way of solving this

	DSBL	DSBL:S	NDSBL	NRMA
T1	-0.916	-0.941	0.951	0.981
T2	-0.545	0.185	0.998	0.999
T3	-0.995	-0.997	0.198	0.199
T4	-0.995	-0.998	0.602	0.612
T5	-0.024	0.874	0.984	0.997
T6	-0.627	-0.578	-0.375	-0.305
T7	-0.437	-0.410	-0.123	-0.163
T8	-0.313	-0.283	-0.196	-0.140
T9	-0.312	-0.194	-0.157	-0.074
T10	-0.568	-0.503	-0.309	-0.392

Table 9: Table represents the Mean sentiment performance of the DistilBERT sentiment analysis model to corresponding disability facet groups.

	VADER	Google	TextBlob	DistilBERT	T_Original	T_Unbiased
DSBL	2e-16 ***	1.46e-05 **	0.032*	3.97e-12 **	0.141	0.486
DSBL:S	2e-16 ***	0.038*	2e-16 ***	1.35e-07 **	2e-16 ***	2e-16 ***
NDSBL	0.043*	0.446	0.032*	0.598	0.003 **	0.467

Table 10: Table represents the p-value from linear regression analysis between each sociodemographic group. The significance codes: 0.001 ‘***’, 0.01 ‘**’, 0.05 ‘*’

issue is to make sure the right dataset is used to pre-train the model. BERT is generally considered to be a black-box. Therefore such tests indicate potential biases that might be omitted if used on a larger platform.

5.5 Toxicity Models

We wanted to look into the aspect of toxicity as this parameter is largely used to mediate cleaner conversations in social media platforms⁵. The issue with such analysis is that topics related to PWD or other minority population might get censored out due to skewed learning through social media dataset [10]. To look into this further, BITS Corpus was passed through the selected two Toxicity analysis model, Toxicity_Original and Toxicity_Biased model. The statistical results, in Table 10 show that sentence related to *People with Disability: Social* group were considered to be significantly more toxic as compared to the other groups. Both the models also show very high toxicity value to the term ‘Autistic’ and ‘Mentally Handicapped’ as compared to other words present in the collection. The low standard deviation in the result show that some terms are flagged as very toxic as compared to the other words present in the set. Both the models were trained using social media dataset but perform significantly poorly on the mention of disability in a social environment. This shows that the terms related to disability are associated to other toxic terms, causing a skewed understanding of the words related to PWD. This was also seen in the work of [10] where words such as ‘drugs’ and ‘homelessness’ were associated to words related to disability. It is interesting to note that Toxicity_Biased toxicity model, which is trained to be contextually sensitive to PWD groups performed equally as bad as compared to the Toxicity_Original model.

6 Discussion and Conclusion

Our aim was to understand the presence of disability bias in sentiment analysis models. For this, we looked into how social posts in online social media platforms are skewed against PWD. The results showed that posts related to PWD were scored significantly lower in sentiment and toxicity scores as compared to posts not related to PWD. The results also showed how sentiment models built on social media dataset were most discriminatory against PWD.

Following this, we created a test that checks the presence of disability bias in any sentiment analysis model. We explained the process of creating this facet of the Bias Identification Test in Sentiments (BITS Corpus). The flexible nature of the BITS Corpus provides an easy method to make more tests based on the template created for the Disability Facet. The same process can be used to create nuanced analysis for other sociodemographic biases as well. We have released every facet of BITS Corpus⁶. We also intend, through this work, to create awareness of the presence of bias in public sentiment analysis libraries. We show the presence of disability biases in notable public sentiment analysis library using this test. From the results, the presence of disability bias was seen on all the models that we had considered.

This work looks at providing the first step to eliminate unintended bias in sentiment analysis, through *identification*. Some possible means of mitigation can be done by changing the rules that these models are defined with, or changing or training the models with better data that represents such sensitive populations [1, 52]. A model, if used in a social environment, must be aware of the all the actors that are present in the social network. Hence the dataset used to train the model is very crucial. Through this paper, we show how *non-inclusive* training can be harmful for certain population, PWD, in various social applications. Through the proposed biased identification process, we can be more aware of the ramification AI applications can have on selected sociodemographic factors.

For our future work, we will concentrate on making BITS Corpus larger by adding more templates to cover various simple scenarios. We also intend to add more facets to BITS Corpus so that it can adhere to a bigger scope, to find various types of biases. A few sociodemographic factors in consideration are race, gender and socio-economic status. For the larger picture, we want to create an interactive platform that enables the user to understand the various issues causing bias in an NLP system, through the aid of visualizations and analysis.

⁵<https://jigsaw.google.com/the-current/toxicity/>

⁶<https://github.com/PranavNV/BITS>

References

- [1] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.
- [2] Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.
- [3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
- [4] Ekaba Bisong. Google automl: Cloud natural language processing. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 599–612. Springer, 2019.
- [5] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*, 2020.
- [6] Jerome Bickenbach. The world report on disability. *Disability & Society*, 26(5):655–658, 2011.
- [7] Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, et al. Disability, bias, and ai. *AI Now Institute*, November, 2019.
- [8] Mike Oliver. The politics of disability. *Critical social policy*, 4(11):21–32, 1984.
- [9] Bo Chen and Donna Marie McNamara. Disability discrimination, medical rationing and covid-19. *Asian bioethics review*, 12(4):511–518, 2020.
- [10] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*, 2020.
- [11] David Noever. Machine learning suites for online toxicity detection. *arXiv preprint arXiv:1810.01869*, 2018.
- [12] Quan Do. Jigsaw unintended bias in toxicity classification. 2019.
- [13] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity detection: Does context really matter? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [14] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.
- [15] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Comput. Linguist*, 35(2):311–312, 2009.
- [16] Fabio Del Vigna¹², Andrea Cimino²³, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95, 2017.
- [17] CHE Gilbert and Erric Hutto. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Available at (20/04/16) <http://comp. social. gatech. edu/papers/icwsml4. vader. hutto. pdf>, volume 81, page 82, 2014.
- [18] Steven Loria. textblob documentation. *Release 0.15*, 2, 2018.
- [19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [20] Laura Hanu and Unitary team. Detoxify. Github. <https://github.com/unitaryai/detoxify>, 2020.
- [21] Bin Lin, Fiorella Zampetti, Gabriele Bavota, Massimiliano Di Penta, Michele Lanza, and Rocco Oliveto. Sentiment analysis for software engineering: How far can we go? In *Proceedings of the 40th International Conference on Software Engineering*, pages 94–104, 2018.
- [22] Prashanth Rao. Fine-grained sentiment analysis in python (part 1), Sep 2019.
- [23] MonkeyLearn Blog. Top sentiment analysis apis, Oct 2019.
- [24] A Marshall and A Davies. Uber’s self-driving car didn’t know pedestrians could jaywalk, 2019.
- [25] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. Toward fairness in ai for people with disabilities sbg@ a research roadmap. *ACM SIGACCESS Accessibility and Computing*, (125):1–1, 2020.

- [26] 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*, pages 1668–1678, 2019.
- [27] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.
- [28] Stephen J Macdonald, Catherine Donovan, and John Clayton. The disability bias: understanding the context of hate in comparison with other minority populations. *Disability & Society*, 32(4):483–499, 2017.
- [29] Byron Creese and Deborah Lader. Hate crimes, england and wales, 2013/2014. 2014.
- [30] Shari Trewin. Ai fairness for people with disabilities: Point of view. *arXiv preprint arXiv:1811.10670*, 2018.
- [31] Cathy O’neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.
- [32] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.
- [33] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.
- [34] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.
- [35] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.
- [36] Amirkiarash Kiani, Sameh Al Natour, and Ozgur Turetken. A comparison of sentiment analysis tools. 2018.
- [37] Mike Thelwall. Gender bias in sentiment analysis. *Online Information Review*, 2018.
- [38] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.
- [39] Jigsaw. Helping authors understand toxicity, one comment at a time, Feb 2021.
- [40] Jigsaw. Hello neighbor: How a local newspaper builds community via online comments, Feb 2021.
- [41] Jigsaw Jigsaw. New york times and jigsaw partner to scale moderation platform, Jan 2021.
- [42] Nathan R Kuncel, David M Klieger, and Deniz S Ones. In hiring, algorithms beat instinct. *Harvard business review*, 92(5):p32–32, 2014.
- [43] Claire Cain Miller. Can an algorithm hire better than a human. *The New York Times*, 25, 2015.
- [44] Alan Rubel. The black box society: The secret algorithms that control money and information, by frank pasquale. cambridge: Harvard university press, 2015. 320 pp. isbn 978–0674368279. *Business Ethics Quarterly*, 26(4):568–571, 2016.
- [45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [46] Judy Cohen and Yvette Silver. Disability etiquette. *Tips on interacting with people with disabilities*, 2006.
- [47] People First Spokane Valley. People first history. *People First of Spokane Washington*, 2005.
- [48] Giacomo Vivanti. Ask the editor: What is the most appropriate way to talk about individuals with a diagnosis of autism? *Journal of autism and developmental disorders*, 50(2):691–693, 2020.
- [49] Monique Botha, Jacqueline Hanlon, and Gemma Louise Williams. Does language matter? identity-first versus person-first language use in autism research: A response to vivanti. *Journal of Autism and Developmental Disorders*, pages 1–9, 2021.
- [50] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.
- [51] Prabhaker Mishra, Uttam Singh, Chandra M Pandey, Priyadarshni Mishra, and Gaurav Pandey. Application of student’s t-test, analysis of variance, and covariance. *Annals of cardiac anaesthesia*, 22(4):407, 2019.
- [52] Ismael Garrido-Muñoz, Arturo Montejó-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184, 2021.