International Journal of Computational and Biological Intelligent Systems (IJCBIS)
Volume 04, Issue 01, October 2022

Automatic Gender Bias Detection in Digital Content

Ashley Hua[#1], Nicholas Lu[#2], Mingchuan Cheng[#3], Andrew Wang[#4], Jason Minghan Dong[#5]

[#1] *Basis Independent Silicon Valley*
1290 Parkmoor Avenue, San Jose, CA, USA 95126
ashleyhua100@gmail.com

[#2] *The Haverford School*
450 Lancaster Avenue, Haverford, PA, USA 19041
nichlu@haverford.org

[#3] *West High School*
6976 S Pine Mountain Dr. Salt Lake City, UT, USA 84121
heavytoothpaste@gmail.com

[#4] *Lower Moreland High School*
555 Red Lion Rd, Huntingdon Valley, PA, USA 19006
andrewxwg@gmail.com

[#5] *College Jean de Brebeuf*
9723 Rue Jeannette, Montreal, Quebec, Canada, H8R 1S6
jasonminghandong@gmail.com

***Abstract*—**The Internet and social media play an essential role in people's daily lives. There is tremendous growth in user generated content which expedites biased views online. Gender bias is a prominent type of bias. The stereotypes of each gender today can be traced back to ancient times. Artificial Intelligence (AI) can be used to mitigate the risk of gender bias when used with caution. Our research aims to use Natural Language Processing (NLP) algorithms to identify gender biased words and their linguistic context to avoid using these biased words when referring to groups of people. Also, our solution can be used to promote unbiased business brands and product descriptions.

Keywords—artificial intelligence, diversity, equity, and inclusion, gender bias, natural language processing, social media

I. INTRODUCTION

Over the past decade, internet and social media usage have become extremely common. In 2022, 4.48 billion people are using social media worldwide with an average of 147 minutes daily [1], more than double that of 2.07 billion in 2015 [2]. The user generated content (UGC) on social media expedites bias. While print and broadcasting companies follow certain rules and regulations to check the credibility and accuracy of their content, these regulations are not applied to UGC [3]. Discrimination and bias are the two terms that often get mixed. Discrimination involves obstacles to full participation in human activity; bias involves the cultural identity of the activity [4]. Bias involves a strong feeling in favor of or against one group of people due to their cultural identity, or one side in an argument, often not based on fair judgment [5], and often causes negative impacts like spreading fake news and causing polarization through biased viewpoints [6]. Furthermore, algorithms used to promote content on social media often aid the spread of biased content, due to the tendency of fake news to generate more user interaction [7]. On Twitter, it takes factual stories 6 times as long to reach 1,500 people as it does for false stories, with false news stories being 70% more likely to be retweeted than true stories [8]. In extreme cases, false tweets can lead to dangerous situations, like when a shooter attacked a pizza shop believing that it was a base for a pedophile ring connected to Hilary Clinton during the 2016 election [9].

Gender bias is a prominent type of bias on social media. In the United States, at least 57% of people hold a gender bias attitude [10]. According to the

Cambridge English Dictionary, gender bias is an "unfair difference in the way women and men are treated" [11].

Gender bias has had a very long history, dating back to ancient times. A study shows that it arose around 8000 years ago, when males were drawn more frequently in cave walls than women [12]. There have been stereotypes that dominate our minds. Females were seen as inferior to males because of the belief that women were intellectually and physically inferior to men. The logic behind this was that men more often fought in wars and often competed for women. As a result, males have become "more evolved" than women [13]. Fueled by these ideas, society has passed down these concepts as facts, thus creating the stereotypes we have today. In a recent study conducted in 2020 by the United Nations, almost 50% of people worldwide say that men make better political leaders than women and 40% think men are better business executives. Also, around 25% of people think that education is more important for men [10]. Not only damaging in real life, gender bias can subconsciously dictate the biased point of views for large online groups, causing them to further weaponize the bias and spread it [14]. Artificial Intelligence (AI) can potentially mitigate the risk of biases in digital content. Being fed the representative dataset, AI algorithms do not have the unconscious assumptions from humans, resulting in less discrimination [15].

However, there is a dilemma. If not used cautiously, AI can also introduce bias [16]. Machine learning models are trained on data, which causes the fairness of the models to be influenced by the data. When training data disproportionately represents certain groups or has flawed information reflecting historical inequalities, algorithmic bias can appear [17]. Furthermore, the people behind an AI solution can also cause bias. In the AI professional community, only 22% are women and thus are under-represented [18]. An example of an AI program that had gender bias was the Amazon recruiting tool. The data that the program was trained on was from resumes 10 years ago mainly from white men, causing the program to recognize word patterns and penalize the resumes of women who attended women's colleges because of the term "women" in their resumes. Another example of bias is that in a Princeton University study analyzing word associations of an AI program, it was found that European names were perceived to be more pleasant than African-American names and that female meaning words were more associated with the arts than the sciences, due to the training data reflecting human biases [19]. These studies showed gender and racial bias, which has harmful impacts.

There has been prior research done on the topic of AI solving gender bias. In 2021, there was research done to reduce gender bias in news articles by historically finding the ratio of quotes from women to men and working to make that ratio equal so that women are not underrepresented. The Gender Gap Tracker intends to motivate news outlets to diversify their sources and can also be applied to other forms of diversity as well [20]. In 2022, there was research done to detect hate speech, including gender biased content. *Nascimento et al.* find bias-sensitive words in the training dataset as well as in hateful tweets [21]. They use Linear Regression, Support Vector Machines, and Decision Trees with ensemble learning.

Our research aims to provide a solution using AI to detect gender bias in digital content by sorting every relevant word whether it is a stronger female indicator, or it is a stronger male indicator. We utilize Word2Vec, focus on contextual content, and use user generated content in the top ranked bias categories.

II. RELATED WORK

As social media use continues to expand, gender bias detection is necessary to prevent gender bias and abusive language in media.

Natural Language Processing (NLP) is an AI technology that aims to understand the human language and detect features in text and speech. NLP can be used to detect gender bias in online content [22]. In recent years, there is an increasing volume of research papers on this topic.

However, the dilemma is that NLP can be a two-edged sword if not used with caution. The biased training dataset can produce biased NLP models [23]. The following table summarizes recent gender bias detection models and studies about implicit bias in AI in reverse chronological order.

TABLE 1
REVIEW OF EXISTING GENDER BIAS DETECTION AI RESEARCH

Paper Title And Researchers	Main Purpose	Models used and data	Results	Limitations	Year
Gender Bias in IT Entrepreneurship: The Self-Referential Role of Male Overrepresentation in Digital Businesses [24]	Pushes the discourse on gender bias in IT forward by theorizing on the self-fulfilling effect of male overrepresentation in IT entrepreneurship	Used Crunchbase to identify start-ups built around blockchain technologies. Crunchbase contains information about start-ups, investors, founders, trends, milestones, and other related information. To test the data, Conditional Process Analysis was applied using standardized Ordinary Least Squares regression as well as logistic regression.	Males were overrepresented. Female founders are more disadvantaged because they are perceived as too distinct from the prototype before they can demonstrate their potential.	Research has not done justice to the overall group of female founders that were treated as a homogeneous group. Also, the binary concept of gender was adopted and plural gender identities were not accounted for.	2022
Automatic Misogyny Detection in Social Media Platforms using Attention-based Bidirectional-LSTM [25]	Proposes an approach based on an Attention-Based Bidirectional LSTM model used to identify misogyny on social media.	Model used: Attention-based Bidirectional L-STM Dataset: IberEval2018, EVALITA2018, EVALITA2020, HatEval2019, TRAC2020	TF-IDF char ngrams (2,5): F1: 0.896 ACC: 0.897 Lexical features: F1: 0.837 ACC: 0.838 Word2Vec: F1: 0.873 ACC: 0.873 All features: F1: 0.909 ACC: 0.909	Multi-language datasets were unbalanced with limited number of samples	2021
Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms [26]	Examines the prevalence of gender stereotypes on digital media platforms and considers how human efforts to create and curate messages may impact these stereotypes.	Data was collected using Microsoft Bing search engines application programming interface. A number of images of men and women associated with four different occupations (librarian, nurse, computer programmer, civil engineer) on four digital platforms (Twitter, New York Times, Wikipedia, and Shutterstock).	Twitter and Wikipedia showed fewer female subjects for all occupations. The New York Times generally has fewer female subjects than actual labor statistics except for civil engineers. Shutterstock generally shows more female subjects than actual labor statistics except for librarians.	This study was only conducted on four websites: Twitter, The New York Times, Shutterstock, and Wikipedia.	2020
Evaluating Gender Bias in Machine Translation [27]	Analysis of gender bias in Machine Translation	The two datasets used were two corefront resolution datasets composed of English sentences that cast participants into non-stereotypical gender jobs. An automatic gender bias evaluation method for eight target languages with grammatical gender, based on morphological analysis was developed. Some Machine Translation models	All MT systems and target languages were tested and analysis yields that all the MT systems tested are indeed biased.	This study was only conducted for eight target languages and four Machine Translation models.	2019

		used were Google Translate, Microsoft Translator, Amazon Translate, and SYSTRAN.			
Racial Bias in Hate Speech and Abusive Language Detection Datasets [23]	Analysis on how AI may unintentionally introduce implicit bias and negative impact due to the biased training dataset	Twitter tweets labeled with “racism” and “harassment” and “sexism”, a section of this paper specifically focuses on “sexism”	W. & H.: Precision: 0.69 Recall: 0.73 F1: 0.71 W.: Precision: 0.62 Recall: 0.73 F1: 0.67	No ground truth labels for racial identities of authors for Blodgett et al. (2016) dataset. Uncertainty of Experiment 2's results due to the fact of the possibility that words associated with negative categories are used to predict race. Focus only on one dimension of racial bias	2019
Reducing Gender Bias in Abusive Language Detection [22]	To detect gender bias using models trained with abusive language	Three neural models used to garner results: Convolutional Neural Network, Gated Recurrent Unit, and a Bidirectional Gated Recurrent Unit. “abusive” tweets from twitter that are labeled “racist” and “sexist.”	AUC CNN: Random: .881 Fasttext: .906 Word2vec: .906 GRU: Random: .854 Fasttext: .887 Word2vec: .887	Performance loss during fine-tuning due to mitigation methods modifying the data or model	2018

III. OUR PROPOSAL

We propose to apply a Natural Language Processing (NLP) based gender bias framework on digital content. NLP is a field of AI and is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications [28].

Some fundamental approaches in NLP are Bag of Words (BoW) and TF-IDF. BoW keeps track of word occurrences and frequencies. However, it loses information about word order, keeps no semantic meaning of underlying words, and doesn’t properly weight some words.

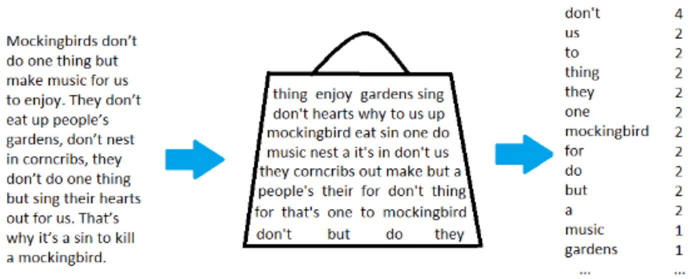
Mockingbirds don’t do one thing but make music for us to enjoy. They don’t eat up people’s gardens, don’t nest in corncribs, they don’t do one thing but sing their hearts out for us. That’s why it’s a sin to kill a mockingbird.
thing enjoy gardens sing don't hearts why to us up mockingbird eat sin one do music nest a it's in don't us they corncribs out make but a people's their for don't thing for that's one to mockingbird don't but do they
don't 4
us 2
to 2
thing 2
they 2
one 2
mockingbird 2
for 2
do 2
but 2
a 2
music 1
gardens 1
... ...

Fig. 1: An example of the bag of words method using “*To Kill a Mockingbird”*

Another popular NLP approach is TF-IDF, which stands for Term Frequency - Inverse Document Frequency. It improves the BoW model by adding weights to the words. It rewards frequent words in a given document and more importantly penalizes those frequent words that are common across documents, which leads to stop words like “the” and non-essential words being eliminated.

Term Frequency (TF) is given by this formula [29]:

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

Whereas the Inverse Document Frequency has:

$$IDF = \log\left(\frac{\text{\# of the documents in the corpus}}{\text{\# of documents in the corpus containing the term}}\right)$$

The TF-IDF of a term is calculated by multiplying TF and IDF scores.

$$TF - IDF = TF \times IDF$$

How do we score the given text? Let’s give an example: say, the word “computing” appears 5 times in a 500-word document. The Term Frequency would be:

$$TF = \frac{5}{500} = 0.01$$

Suppose that there are a total of 500,000 documents in the corpus, and 500 of them contain the word “computing”. The IDF would be:

$$IDF = \log\left(\frac{500000}{500}\right) = 3$$

The TF-IDF score would be:

$$TF - IDF = 0.01 \times 3 = 0.03$$

Compared to BoW, TF-IDF provides a better understanding of text by assigning weights to the words. However, there are a few disadvantages and issues. First, BoW and TF-IDF lose information about word order. Second, they fail to understand context, which is vital to understanding and analysis. Third, words with little meaning like "the", or "a" add noise to the analysis. Several solutions are proposed to solve the above issues.

The first one would be to remove the "stop" words. AI will filter out the common language articles, pronouns, and prepositions that do not impact the general meaning of the sentence, such as "and", "the", "to", "is", etc. Precompiled list of stop words is used for this purpose.

A second way of improving the algorithm of NLP would be to reduce the different forms of a word to its core root. AI can achieve it via stemming and lemmatization. Stemming algorithms are typically heuristic rule-based, whereas lemmatization resolves word to its dictionary form which requires more knowledge about the structure of a sentence and hence demands more computing power.

$$\text{Making} \Longrightarrow \text{Make}$$

$$\text{Laptops} \Longrightarrow \text{Laptop}$$

However, this process leads to more problems for certain words. For example, these ones change the meaning of the word partially or entirely.

$$\text{Modeling} \Longrightarrow \text{Model}$$

These issues can be solved using a model called Word2Vec. Word2Vec is a technique that utilizes layered neural networks to understand the context and similarities between words using large amounts of data, allowing it to categorize words in vector space, and encoding words as vectors, of both semantic and syntactic meaning [30]. Unlike prior mentioned NLP models, which tracked frequency for words in documents instead of grammar or word order, the Word2Vec model is able to understand relationships and semantic meaning of words. Specifically, the Word2Vec model is able to use processes such as stemming, removing affixes to determine the stem of the word, or lemmatization, determining the context in which the word is being used, to give the model greater insight into meaning and similarity. As a result, the Word2Vec model was able to use 300 features in vector form to describe each word, with the distance between word vectors representing similarities. More specifically, words that have greater similarity would share a shorter distance, and as each word is represented as a vector, arithmetic operations, such as subtraction or addition, can be used for a variety of effects.

Some notable examples of Word2Vec are vector subtraction, vector addition, finding similar words, and finding words that do not belong in a group. For example:

$$\text{France} + \text{Berlin} - \text{Germany} = \text{Paris}$$

This is because the capital of France is Paris, and the capital of Germany is Berlin. Because of the vector categories of the words, Word2Vec is able to use vector subtraction to come up with other words. Some other examples can be:

$$\text{vec("king")-vec("man")+vec("woman")= vec("queen")}$$

$$\text{vec("Montreal Canadiens") - vec("Montreal") + vec("Toronto")} \\ = \text{vec("Toronto Maple Leafs")}$$

Besides the above capabilities, Word2Vec can find words that have similar vectors, and can compare vectors to see which word does not fit in the group of words. The Word2Vec model was trained on a Google News dataset consisting of approximately 3 million high occurring words and phrases.

Another important aspect of this study is the usage of Principal Component Analysis (PCA) to identify gender bias. Principal Component Analysis (PCA) is a technique for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss [31]. Using PCA, it is possible to find the principal components, which are variables that are new combinations or mixtures of the initial variables, thus leading to loss of dimensionality but not loss of information [32]. Thus, specific components can be kept, and used to make a feature vector matrix. Using it, the original data can be reoriented to find the principal component axis, or polarities. Given two polarities, PCA can be used to identify opposite words, and orient the collection of three-million-word vectors, collected through Word2Vec, upon the axises. In this study, the opposite polarities were trained with a predetermined list of word pairs, such as ‘Male’ and ‘Female’, in order to determine the magnitude of gender indication of each word or phrase. Thus, using the Word2Vec model, word vectors that are more similar to each of the polarities will be closer to the ends of the axis, while more neutral words will be towards the center. Given the distance between the center of the polarity axis and the word vector, the magnitude and polarity of gender bias can be determined. Thus, the previously determined word vectors are able to be interpreted as a single vector instead of an array of 300 vectors, and PCA can be used to find the gender bias vector and determine bias based on the position of the vector relative to the center, at 0. At positions farther from the center, the word or phrase is more biased. with the sign of the vector indicating what polarity the word is biased toward.

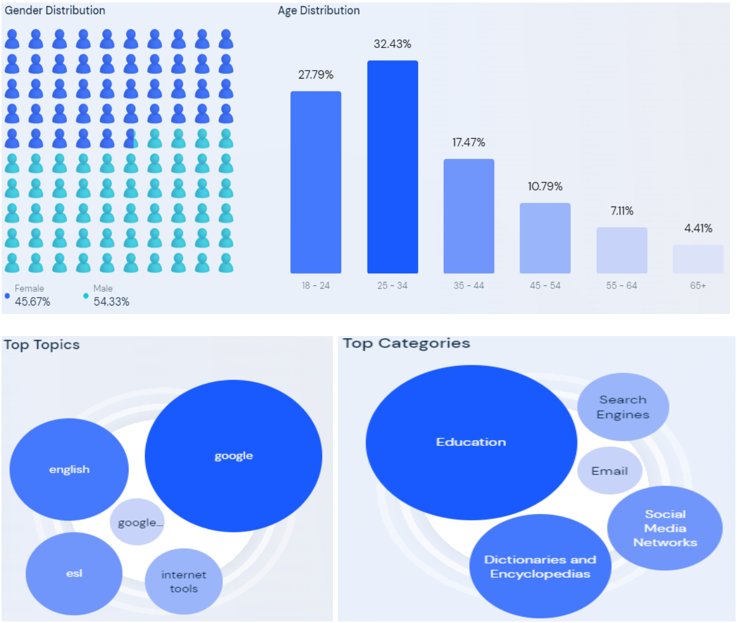

Fig. 2 Statistics regarding visits on teflpedia.com [35]

To identify the predetermined list of word pairs and list of neutral words to train the PCA model, we used search keywords online and picked from top-ranked lists. We used a list from ProofreadingServices.com [33]. Also, we adopted gender word pairs from Teflpedia. According to Teflpedia, "A masculine–feminine pair or gender pair is a pair of words having related meanings but where one is typically masculine and the other typically feminine" [34]. Before adopting the word pairs from Teflpedia, we confirmed that this website is well-visited, safe and suitable for all ages. Also, we checked that its audience has relatively fair representation. Its audience consists of 54.33% male and 45.67% female. The largest age group of visitors is 25 - 34-year-olds [35].

During this study, we discovered that the bias is contextual dependent. The index, in fact, represents the bias of the linguistic context surrounding the word. This is more powerful than the bias index of the word itself.. For example,

bachelor = -0.22185655
daddy = -0.1742111

These two words are clearly male oriented words, but their indexes are neutral and slightly leaning towards feminine side. This means that these words are often used in the context where neutral or slight feminine perspectives or topics are discussed.

Our model implementation can unlock several use cases. First use case is to avoid using gender biased nouns to refer to groups of people.

For example:
policeman = 0.24587852
policewoman = -1.000542
The recommended word should be police officer.

Another example:
salesman = 0.39882642
saleswoman = -1.4926232
The recommended word should be salesperson.

Second use case is to avoid gender bias in brand and product descriptions. 48% of Gen Zers value brands that don’t classify items by gender. In Fact, Hasbro CEO Brian Goldner was ahead of the curve several years ago when he announced that he “eliminated the old delineation of gender” across the company’s brands after learning that 30% of My Little Pony consumers worldwide were boys [36]. Toys”R”Us stopped marketing gender stereotypes to children [37]. Our AI model will help other companies to avoid using gender biased words in their product descriptions and marketing messages to establish a more inclusive brand and boost up the business revenue.

The following table details the gender biased words, and their gender bias indexes.

TABLE 2
GENDER BIASED WORDS AND INDEXES

Type	masculine	feminine	Masculine index	Feminine index
Pronouns	he	she	0.40889728	-1.2777667
	him	her	0.3543523	-1.2540389
	his	hers	0.42674133	-1.1985073
	his	her	0.42674133	-1.2540389
	himself	herself	0.6753487	-1.5069519
Personal titles	Mr.	Miss.	0.2257672	-0.23339172
	Mr.	Mrs.	0.2257672	-0.7876961
	Mr.	Ms.	0.2257672	-1.1576045
	sir	madam	0.31474772	-0.912137
Family-related nouns	bachelor	spinster	-0.22185655	-1.3673965
	brother	sister	0.47126427	-0.97586375
	bridegroom	bride	-0.26767927	-0.7489712
	father	mother	0.22956628	-0.940051
	father	ma	0.22956628	-0.28436863
	father	mam	0.22956628	-0.7481121

	father	mammy	0.22956628	-0.6265792
	father	mom	0.22956628	-0.9909192
	dad	mom	0.13571991	-0.9909192
	daddy	mommy	-0.1742111	-1.0883372
	father	mommy	0.22956628	-1.0883372
	father	mum	0.22956628	-0.83087236
	father	mummy	0.22956628	-0.6182746
	fiancé	fiancée	-0.68740106	-0.64017904
	nephew	niece	0.40459052	-0.9748408
	widower	widow	-0.18767996	-0.63160986
	husband	wife	-0.6734396	-0.5360513
	son	daughter	0.19723879	-1.1382921
	uncle	aunt	0.41697243	-0.8584151
	divorcé	divorcée	-0.29957724	-1.1077904
	grandfather	grandmother	0.33212122	-0.8697305
Gendered job titles	waiter	waitress	-0.07113879	-1.312748
	tailor	seamstress	0.13603967	-1.309859
	actor	actress	0.13313934	-1.4483037
	schoolmaster	schoolmistress	0.25634867	-0.87491935
	policeman	policewoman	0.24587852	-1.000542
	salesman	saleswoman	0.39882642	-1.4926232
	author	authoress	-0.18054138	-0.58150774
	benefactor	benefactress	0.16700771	-0.5744808
	conductor	conductress	0.17292604	-0.4694054
	wizard	witch	0.563896	-1.0102386
	tempter	temptress	0.12943293	-0.91990674
	steward	stewardess	0.05772459	-0.97853327
	poet	poetess	-0.016623972	-1.1690526
	songster	songstress	0.022120306	-1.4416664
	shepherd	shepherdess	0.092920944	-1.0022215
	milkman	milkmaid	0.24349521	-0.7819383
Gendered nouns	boy	girl	0.0118539175	-1.1462415
	gentleman	lady	0.23812568	-1.2437552
	hero	heroine	0.65102285	-1.3479661
	heir	heiress	0.29772472	-1.139049
Religious titles	priest	priestess	0.06698851	-1.2333046
	prophet	prophetess	0.49816358	-0.5623318
	monk	nun	0.1355315	-1.0711473
	god	goddess	0.15256245	-1.1462165

IV. NEXT STEPS

In the process of this research, we also identified the following three areas as the next steps:

1. Expand the study to include other genders. More than 12% of U.S. millennials identify as transgender or gender non-conforming, and a majority believe that gender is a spectrum rather than a man/woman binary. Compared to millennials, Gen Z's views on gender are even more advanced. In the U.S., 56% know someone who uses a gender neutral pronoun and 59% believe forms should include options other than "man" and "woman." Globally, 25% of Gen Zers expect to change their gender identity *at least once* during their lifetime [36]. In this study, we have only considered males and females. As a next step, we suggest to also consider non-binary genders.
2. Focus on slang vocabulary. Slang words, terms, or even spelling, are often used to convey meanings different from their technical definitions, which may confuse or disorient the NLP models that we developed. A training dataset specialized on slang words should be adopted.
3. In the future, we also would like to incorporate other languages. Currently all of the word lists are English based. If we can extend this algorithm to other languages like French, Chinese, Spanish, etc. it would expand its contribution around the world.

V. CONCLUSION

Using NLP algorithms, specifically PCA algorithm trained with gender polarity word pairs, we can detect the gender bias among the top ranked 3 million English words and phrases. We discovered that the index represents the bias of the linguistic context surrounding the word. This is more powerful than the bias index of the word itself. We used this AI model on pronouns, personal titles, family-related nouns, gendered job titles, gendered nouns, and religious titles. The result can be used to raise awareness of the gender bias in online content and call for action to reduce inappropriate biased words. Also, this AI model can be used to promote unbiased business brands and inclusive product descriptions.

REFERENCES

[1] "Global daily social media usage 2022 | Statista." https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/ (accessed Oct. 23, 2022).

[2] "How Many People Use Social Media in 2022? (65+ Statistics)," *Backlinko*, Oct. 10, 2021. https://backlinko.com/social-media-users (accessed Oct. 23, 2022).

[3] M. Vorhaus, "People Increasingly Turn To Social Media For News," *Forbes*. https://www.forbes.com/sites/mikevorhaus/2020/06/24/people-increasingly-turn-to-social-media-for-news/ (accessed Oct. 23, 2022).

[4] "Gender Discrimination and Gender Bias:Different Sides of the Same Coin." http://web.mit.edu/fnl/vol/144/williams.htm (accessed Oct. 23, 2022).

[5] "bias_1 noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com." https://www.oxfordlearnersdictionaries.com/definition/english/bias_1 (accessed Oct. 23, 2022).

[6] "Misinformation and biases infect social media, both intentionally and accidentally." https://theconversation.com/misinformation-and-biases-infect-social-media-both-intentionally-and-accidentally-97148 (accessed Oct. 23, 2022).

[7] A. Hua, "Social Media Fake News: A Review of Case Studies and Psychology," *Int. J. Comput. Biol. Intell. Syst.*, vol. 3, no. 2, Art. no. 2, Oct. 2021, Accessed: Oct. 23, 2022. [Online]. Available: https://ijcbis.org/index.php/ijcbis/article/view/1643

[8] "Study: On Twitter, false news travels faster than true stories," *MIT News | Massachusetts Institute of Technology*. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308 (accessed Oct. 23, 2022).

[9] "19 Real Events Caused by Fake News in the US | Marubeni Corporation." https://www.marubeni.com/en/research/potomac/backnumber/19.html (accessed Oct. 23, 2022).

[10] "6 Dismal Findings From U.N. Report On Gender Bias." https://www.forbes.com/sites/kimelsesser/2020/03/09/6-dismal-findings-from-un-report-on-gender-bias/?sh=3707783d5d1c (accessed Oct. 23, 2022).

[11] "GENDER BIAS | definition in the Cambridge English Dictionary." https://dictionary.cambridge.org/us/dictionary/english/gender-bias?q=gender-bias%29 (accessed Oct. 23, 2022).

[12] "Gender inequality arose 8000 years ago." https://cosmosmagazine.com/history/gender-inequality-arose-8000-years-ago/ (accessed Oct. 23, 2022).

[13] "The history of the human female inferiority ideas in evolutionary biology - PubMed." https://pubmed.ncbi.nlm.nih.gov/12680306/ (accessed Oct. 23, 2022).

[14] "Gender bias online is as harmful as ever. Here are a few ways to fight back. | Mashable." https://mashable.com/article/how-to-challenge-gender-bias-online (accessed Oct. 23, 2022).

[15] "How AI Can End Bias | SAP Insights." https://insights.sap.com/how-ai-can-end-bias/ (accessed Oct. 23, 2022).

[16] "How Artificial Intelligence Perpetuates Gender Imbalance." https://www.oliverwyman.com/our-

expertise/insights/2020/mar/gender-bias-in-artificial-intelligence.html (accessed Oct. 23, 2022).

[17] N. T. L. Barton Paul Resnick, and Genie, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms," *Brookings*, May 22, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ (accessed Oct. 23, 2022).

[18] "When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity." https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity (accessed Oct. 23, 2022).

[19] "Biased bots: Artificial-intelligence systems echo human prejudices," *Princeton University*. https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices (accessed Oct. 23, 2022).

[20] "The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media | PLOS ONE." https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245533#sec001 (accessed Oct. 23, 2022).

[21] "Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/S095741742200447X (accessed Oct. 23, 2022).

[22] J. H. Park, J. Shin, and P. Fung, "Reducing Gender Bias in Abusive Language Detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct. 2018, pp. 2799–2804. doi: 10.18653/v1/D18-1302.

[23] T. Davidson, D. Bhattacharya, and I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets." arXiv, May 29, 2019. Accessed: Oct. 23, 2022. [Online]. Available: http://arxiv.org/abs/1905.12516

[24] "Full article: Gender Bias in IT Entrepreneurship: The Self-Referential Role of Male Overrepresentation in Digital Businesses." https://www.tandfonline.com/doi/full/10.1080/0960085X.2022.2075801 (accessed Oct. 23, 2022).

[25] "Automatic Misogyny Detection in Social Media Platforms using Attention-based Bidirectional-LSTM | IEEE Conference Publication | IEEE Xplore." https://ieeexplore.ieee.org/document/9659158 (accessed Oct. 23, 2022).

[26] V. K. Singh, M. Chayko, R. Inamdar, and D. Floegel, "Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms," *J. Assoc. Inf. Sci. Technol.*, vol. 71, no. 11, pp. 1281–1294, 2020, doi: 10.1002/asi.24335.

[27] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating Gender Bias in Machine Translation." arXiv, Jun. 03, 2019. doi: 10.48550/arXiv.1906.00591.

[28] "Natural language processing - Retresco (EN)." https://www.retresco.com/encyclopedia-article/what-is-natural-language-processing (accessed Oct. 23, 2022).

[29] "TF-IDF — Term Frequency-Inverse Document Frequency – LearnDataSci." https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/ (accessed Oct. 23, 2022).

[30] "Glossary," *NLP-guidance*. https://moj-analytical-services.github.io/NLP-guidance/Glossary.html (accessed Oct. 23, 2022).

[31] "Principal component analysis: a review and recent developments | Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences." https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 (accessed Oct. 23, 2022).

[32] "Principal Component Analysis (PCA) Explained | Built In." https://builtin.com/data-science/step-step-explanation-principal-component-analysis (accessed Oct. 23, 2022).

[33] "73 Pairs of Nouns of Opposite Genders." https://www.proofreadingservices.com/pages/nouns-of-opposite-gender (accessed Oct. 23, 2022).

[34] "Masculine–feminine pair - Teflpedia." https://teflpedia.com/Masculine%E2%80%93feminine_pair (accessed Oct. 23, 2022).

[35] "teflpedia.com Traffic Analytics & Market Share | Similarweb." https://www.similarweb.com/website/teflpedia.com/#demographics (accessed Oct. 23, 2022).

[36] "Companies Can't Ignore Shifting Gender Norms." https://hbr.org/2020/04/companies-cant-ignore-shifting-gender-norms (accessed Oct. 23, 2022).

[37] "Petition · Toys 'R' Us: Stop Marketing Gender Stereotypes to Children · Change.org." https://www.change.org/p/toys-r-us-stop-marketing-gender-stereotypes-to-children (accessed Oct. 23, 2022).