Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings

Anthony Rios[1], **Reenam Joshi**[2], **and Hejin Shin**[3]
[1]Department of Information Systems and Cyber Security
[2]Department of Computer Science
[3]Library Systems
University of Texas at San Antonio
{Anthony.Rios, Reenam.Joshi, Hejin.Shin}@utsa.edu

Abstract

Gender bias in biomedical research can have an adverse impact on the health of real people. For example, there is evidence that heart disease-related funded research generally focuses on men. Health disparities can form between men and at-risk groups of women (i.e., elderly and low-income) if there is not an equal number of heart disease-related studies for both genders. In this paper, we study temporal bias in biomedical research articles by measuring gender differences in word embeddings. Specifically, we address multiple questions, including, How has gender bias changed over time in biomedical research, and what health-related concepts are the most biased? Overall, we find that traditional gender stereotypes have reduced over time. However, we also find that the embeddings of many medical conditions are as biased today as they were 60 years ago (e.g., concepts related to drug addiction and body dysmorphia).

1 Introduction

It is important to develop gender-specific best-practice guidelines for biomedical research (Holdcroft, 2007). If research is heavily biased towards one gender, then the biased guidance may contribute towards health disparities because the evidence drawn-on may be questionable (i.e., not well studied). For example, there is more research funding for the study of heart disease in men (Weisz et al., 2004). Therefore, the at-risk populations of older women in low economic classes are not as well-investigated. Therefore, this opens up the possibility for an increase in the health disparities between genders.

Among informatics researchers, there has been increased interest in understanding, measuring, and overcoming bias associated with machine learning methods. Researchers have studied many application areas to understand the effect of bias. For example, Kay et al. (2015) found that the Google image search application is biased (Kay et al., 2015). Specifically, they found an unequal representation of gender stereotypes in image search results for different occupations (e.g., all police images are of men). Likewise, ad-targeting algorithms may include characteristics of sexism and racism (Datta et al., 2015; Sweeney, 2013). Sweeney (2013) found that the names of black men and women are likely to generate ads related to arrest records. In healthcare, much of the prior work has studied the bias in the diagnosis process made by doctors (Young et al., 1996; Hartung and Widiger, 1998). There have also been studies about ethical considerations about the use of machine learning in healthcare (Cohen et al., 2014).

It is possible to analyze and measure the presence of gender bias in text. Garg et al. (2018) analyzed the presence of well-known gender stereotypes over the last 100 years. Hamberg (2008) shown that gender blindness and stereotyped preconceptions are the key cause for gender bias in medicine. Heath et al. (2019) studied the gender-based linguistic differences in physician trainee evaluations of medical faculty. Salles et al. (2019) measured the implicit and explicit gender bias among health care professionals and surgeons. Feldman et al. (2019) quantified the exclusion of females in clinical studies at scale with automated data extraction. Recently, researchers have studied methods to quantify gender bias using word embeddings trained on biomedical research articles (Kurita et al., 2019). Kurita et al. (2019) shown that the resulting embeddings capture some well-known gender stereotypes. Moreover, the embeddings exhibit the stereotypes at a lower rate than embeddings trained on other corpora (e.g., Wikipedia). However, to the best of our knowledge, there has not been an automated temporal study in the change

of gender bias.

In this paper, we look at the temporal change of gender bias in biomedical research. To study social biases, we make use of word embeddings trained on different decades of biomedical research articles. The two main question driving this work are, In what ways has bias changed over time, and Are there certain illnesses associated with a specific gender? We leverage three computational techniques to answer these questions, the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), the Embedding Coherence Test (ECT) (Dev and Phillips, 2019), and Relational Inner Product Association (RIPA) (Ethayarajh et al., 2019). To the best of our knowledge, this will be the first temporal analysis of bias of word embeddings trained on biomedical research articles. Moreover, to the best of our knowledge, this is the first analysis that measures the gender bias associated with individual biomedical words.

Our work is most similar to Garg et al. (2018). Garg et al. (2018) study the temporal change of both gender and racial biases using word embeddings. Our work substantially differs in three ways. First, this paper is focused on biomedical literature, not general text corpora. Second, we analyze gender stereotypes using three distinct methods to see if the bias is robust to various measurement techniques. Third, we extend the study beyond gender stereotypes. Specifically, we look at bias in sets of occupation words, as well as bias in mental health-related word sets. Moreover, we quantify the bias of individual occupational and mental health-related words.

In summary, the paper makes the following contributions:

- We answer the question; How has the usage of gender stereotypes changed in the last 60 years of biomedical research? Specifically, we look at the change in well-known gender stereotypes (e.g., *Math vs Arts*, *Career vs Family*, *Intelligence vs Appearance*, and occupations) in biomedical literature from 1960 to 2020.

- The second contribution answers the question; What are the most gender-stereotyped words for each decade during the last 60 years, and have they changed over time? This contribution is more focused than simply looking at traditional gender stereotypes. Specifically, we analyze two groups of words: occupations and mental health disorders. For each group, we measure the overall change in bias over time. Moreover, we measure the individual bias associated with each occupation and mental health disorder.

2 Related Work

In this section, we discuss research related to the three major themes of this paper: gender disparities in healthcare, biomedical word embeddings, and bias in natural language processing (NLP).

2.1 Gender Disparities in Healthcare.

There is evidence of gender disparities in the healthcare system, from the diagnosis of mental health disorders to differences in substance abuse. An important question is, Do similar biases appear in biomedical research? In this work, while we explore traditional gender stereotypes (e.g., *Intelligence vs Appearance*), we also measure potential bias in the occupations and mental health-related disorders associated with each gender.

With regard to mental health, as an example, affecting more than 17 million adults in the United States (US) alone, major depression is one of the most common mental health illnesses (Pratt and Brody, 2014). Depression can cause people to lose pleasure in daily life, complicate other medical conditions, and possibly lead to suicide (Pratt and Brody, 2014). Moreover, depression can occur to anyone, at any age, and to people of any race or ethnic group. While treatment can help individuals suffering from major depression, or mental illness in general, only about 35% of individuals suffering from severe depression seek treatment from mental health professionals. It is common for people to resist treatment because of the belief that depression is not serious, that they can treat themselves, or that it would be seen as a personal weakness rather than a serious medical illness (Gulliver et al., 2010). Unfortunately, while depression can affect anyone, women are almost twice as likely as men to have had depression (Albert, 2015). Moreover, depression is generally higher among certain demographic groups, including, but not limited to, Hispanic, non-Hispanic black, low income, and low education groups (Bailey et al., 2019). The focus of this paper is to understand the impact of these mental health disparities in word embeddings trained on biomedical corpora.

2.2 Biomedical Word Embeddings.

Word embeddings capture the distributional nature between words (i.e., words that appear in similar contexts will have a similar vector encoding). Over the years, there have been multiple methods of producing word embeddings, including, but not limited to, latent semantic analysis (Deerwester et al., 1990), Word2Vec (Mikolov et al., 2013a,b), and GLOVE (Pennington et al., 2014). Moreover, pretrained word embeddings have been shown to be useful for a wide variety of downstream biomedical NLP tasks (Wang et al., 2018), such as text classification (Rios and Kavuluru, 2015), named entity recognition (Habibi et al., 2017), and relation extraction (He et al., 2019). In Chiu et al. (2016), the authors study a standard methodology to train good biomedical word embeddings. Essentially, they study the impact of the various Word2Vec-specific hyperparameters. In this paper, we use the strategies proposed in Chiu et al. (2016) to train optimal decade-specific biomedical word embeddings.

2.3 Bias and Natural Language Processing.

Unfortunately, because word embeddings are learned using naturally occurring data, implicit biases expressed in text will be transferred to the vectors. Bias (and fairness) is an important topic among natural language processing researchers. Bias has been found in word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018, 2019), text classification models (Dixon et al., 2018; Park et al., 2018; Badjatiya et al., 2019; Rios, 2020), and in machine translation systems (Font and Costa-jussà, 2019; Escudé Font, 2019). In general, each paper generally focuses on either testing whether bias exists in various models, or on removing bias from classification models for specific applications.

Much of the work on measuring (gender) bias using word embeddings neither studies the temporal aspect (i.e., how bias changes over time) nor focuses on biomedical research (Chaloner and Maldonado, 2019). For example, Caliskan et al. (2017) studied the bias in groups of words—focusing on traditional gender stereotypes. Kurita et al. (2019) expanded on Caliskan et al. (2017) to generalize to contextual word embeddings. Garg et al. (2018) developed a technique to study 100 years of gender and racial bias using word embeddings. They evaluated the bias over time using the US Census as a baseline to compare embedding bias to demographic and occupation shifts. There has also been work on measuring bias in sentence embeddings (May et al., 2019). Furthermore, there has been a significant amount of research that explores different ways to measure bias in word embeddings (Caliskan et al., 2017; Dev and Phillips, 2019; Ethayarajh et al., 2019). In this work, we make use of many of the bias measurement techniques (Caliskan et al., 2017; Dev and Phillips, 2019; Ethayarajh et al., 2019) to apply them to the biomedical domain.

Year	# Articles
1960-1969	1,479,370
1970-1979	2,305,257
1980-1989	3,322,556
1990-1999	4,109,739
2000-2010	6,134,431
2010-2020	8,686,620
Total	26,037,973

Table 1: The total number of articles in each decade.

3 Dataset

We analyze PubMed-indexed titles and abstracts published anytime between 1960 and 2020. The total number of articles per decade are shown in Table 1. The text is lower-cased and tokenized using the SimpleTokenizer available in GenSim (Khosrovian et al., 2008). We find that the total number of papers have grown substantially each decade, from 1.4 million indexed articles in the 1960s to 8.6 million in the 2010s. Yet, the rate of growth stayed relatively stable each decade.

4 Method

We train the Skip-Gram model on PubMed-indexed titles and abstracts from 1960 to 2020. The hyperparameters of the Skip-Gram model are optimized independently for each decade. Next, given the best set of embeddings for each decade, we explore three different techniques to measure bias: the Word Embedding Association Test (WEAT), the Embedding Coherence Test (ECT), and the Relational Inner Product Association (RIPA). Each method allows us to quantify bias in different ways, such as comparing multiple sets of words (e.g., comparing the bias with respect to *Career vs Family*), comparing a single set of words (e.g., occupations), and measuring the bias of individual words (e.g., nurse). In this section, we briefly discuss the procedure we used to train the word embeddings,

Attribute Words	Male vs Female	X	male, man, boy, brother, he, him, his, son, father, uncle, grandfather
		Y	female, woman, girl, sister, she, her, hers, daughter, mother, aunt, grandmother
Target Words	Career vs Family	A	executive, management, professional, corporation, salary, office, business, career
		B	home, parents, children, family, cousins, marriage, wedding, relatives
	Math vs Art	A	math, algebra, geometry, calculus, equations, computation, numbers, addition
		B	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	Science vs Art	A	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy
		B	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	Intelligence vs Appearance	A	precocious, resourceful, inquisitive, genius, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, apt, venerable, imaginative, shrewd, thoughtful, wise, smart, ingenious, clever, brilliant, logical, intelligent
		B	alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong
	Weak vs Strong	A	power, strong, confident, dominant, potent, command, assert, loud, bold, succeed, triumph, leader, shout, dynamic, winner
		B	weak, surrender, timid, vulnerable, weakness, wispy, withdraw, yield, failure, shy, follow, lose, fragile, afraid, loser

Table 2: Attribute and Target words words used by WEAT to measure the presence of traditional gender stereotypes in biomedical literature.

as well as provide descriptions of each of the bias measurement techniques.

4.1 Word2Vec Model Training.

We train a Skip-Gram model using GenSim (Khosrovian et al., 2008). Following Chiu et al. (2016), we search over the following key hyper-parameters: Negative sample size, sub-sampling, minimum-count, learning rate, vector dimension, and context window size. See Chiu et al. (2016, Table 2) for more details.

To find the best model, as we search over the various hyper-parameters, we make use of the UMLS-Sim dataset (McInnes et al., 2009). UMLS-Sim consists of 566 medical concept pairs for measuring similarity. The degree of association between terms in UMLS-Sim was rated by four medical residents from the University of Minnesota medical school. All these clinical terms correspond to Unified Medical Language System (UMLS) concepts included in the Metathesaurus (Bodenreider, 2004). Evaulation is performed using Spearman's rho rank correlation between a vector of cosine similarities between each of the 566 pairs of words and their respective medical-resident ratings. Intuitively, the ranking of the pairs using cosine similarity, from most similar pairs to the least, should be similar to the human (medical expert) annotations.

4.2 Word Embedding Association Test

The implicit bias test measures unconscious prejudice (Greenwald et al., 1998). WEAT is a generalization of the implicit bias test for word embeddings, measuring the association between two sets of target concepts and two sets of attributes. We use the same target and attribute sets from Kurita et al. (2019). We list the targets and attributes in Table 2. The attribute sets of words are related to the groups in which the embeddings are biases towards or against, e.g., *Male vs Female*. The words in the target categories—*Career vs Family*, *Math vs Arts*, *Science vs Arts*, *Intelligence vs Appearance*, and *Strength vs Weakness*—represent the specific types of biases. For example, using the attributes and targets, we want to know whether the learned embeddings that represent *men* are more related to *career* than the *female*-related words (i.e., test if female words are more related to family, than male words).

Formally, let X and Y be equal-sized sets of *target* concept embeddings and let A and B be sets of *attribute* embeddings. To measure the bias, we follow Caliskan et al. (2017), which defines the following *test statistic* that is the difference between the sums over the respective target concepts,

$$s(X, Y, A, B) = \left[\sum_{x \in X} s(x, A, B)\right] - \left[\sum_{y \in Y} s(y, A, B)\right]$$

where $s(w, A, B)$ measures the association between a single target word w (e.g., career) with

each of the attribute (gendered) words as

$$s(w, A, B) = \Big[\sum_{a \in A} \cos(\vec{\mathbf{w}}, \vec{\mathbf{a}})\Big] - \Big[\sum_{b \in B} \cos(\vec{\mathbf{w}}, \vec{\mathbf{b}})\Big],$$

such that $\cos()$ represents the cosine similarity between two vectors. $\vec{\mathbf{w}} \in \mathbb{R}^d$, $\vec{\mathbf{a}} \in \mathbb{R}^d$, and $\vec{\mathbf{b}} \in \mathbb{R}^d$ represents the word embedding for x, y, and w, respectively. Similarly, d is the dimension of each word embedding. Instead of using the test statistic directly, to measure bias, we use the *effect size*. Effect size is a normalized measure of the separation of the two distributions, defined as

$$\frac{\mu_{x \in X}\big[s(x, A, B)\big] - \mu_{y \in Y}\big[s(y, A, B)\big]}{\sigma_{w \in X \cup Y} s(w, A, B)}$$

where $\mu_{x \in X}$ and $\mu_{y \in Y}$ represent the mean score over target words for a specific attribute word. Likewise, $\sigma_{w \in X \cup Y}$ is the standard deviation of the scores for the word w in the union of X and Y. Intuitively, a positive score means that the attribute words in X (e.g., male, man, boy) are more similar to the target words A (e.g., strong, power, dominant) than Y (e.g., female, woman, girl). Moreover, larger effects represent more biased embeddings.

As previously stated, the Attribute and Target words are from Kurita et al. (2019). It is important to note that the list is manually curated. Moreover, the bias measurement can change depending on the exact list of words. RIPA is more robust to slight changes to the attribute words than WEAT (Ethayarajh et al., 2019).

4.3 Embedding Coherence Test.

We also explore a second method of measuring bias, the Embedding Coherence Test (ECT) (Dev and Phillips, 2019). Unlike WEAT, it compares the attribute Words (e.g., *Male vs Female*) with a single target set (e.g., *Career*). Thus, we do not need two contrasting target sets (e.g., *Career vs Family*) to measure bias. We take advantage of this to measure bias associated with occupations and mental health-related disorders. Specifically, we use a total of 290 occupation words and 222 mental health-related words. The occupation words come from prior work measuring per-word bias (Dev and Phillips, 2019). To form a list of mental health words, we use the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), a taxonomic and diagnostic tool published by the American Psychiatric Association (Association et al., 2013). For each mental health disorder in DSM-5, which are generally multi-word expressions, we split it into individual words. Next, we manually remove uninformative adjective and function words. For example, the disorder "Specific learning disorder, with impairment in mathematics" is tokenized into the following words: "learning", "disorder", "impairment", and "mathematics". A complete listing of the occupational and mental health words can be found in the appendix.

Year	Sim	Pair Cnt
1960-1969	.6586	101
1970-1979	.6715	207
1980-1989	.7033	277
1990-1999	.7282	265
2000-2010	.7078	272
2010-2020	.6867	306

Table 3: Quality of the embeddings trained for each decade, measured using the UMLS-Sim dataset. Sim represents Spearman's rho ranking correlation. Pair count is the number of UMLS-Sim's word-pairs that were present in that decades embeddings.

Formally, ECT first computes the mean vectors for the attribute word sets X and Y, defined as

$$\vec{\mathbf{v}}_X = \frac{1}{|X|} \sum_{x \in X} \vec{\mathbf{x}}$$

where $\vec{\mathbf{v}}_X \in \mathbb{R}^d$ and $|X|$ represents the number of words in category X. $\vec{\mathbf{v}}_Y$ is calculated similarly.

For both $\vec{\mathbf{v}}_X$ and $\vec{\mathbf{v}}_Y$, ECT computes the (cosine) similarities with all vectors $a \in A$, i.e., the cosine similarity is calculated between each target word a and $\vec{\mathbf{v}}_X$ and stored in $s_X \in \mathbb{R}^{|A|}$. The two resultant vectors of similarity scores, s_X (for X) and s_Y (for Y) are used to obtain the final ECT score. It is the Spearman's rank correlation between the rank orders of s_X and s_Y—the higher the correlation, the lower the bias. Intuitively, if the correlation is high, then the rank of target words based on similarity is correlated when calculated for the both X and Y (i.e., male and female).

4.4 Relational Inner Product Association.

While ECT only requires a single target set, both WEAT and ECT [1] calculate the bias between sets

[1] The cosine similarities from ECT *can* be used to measure scores for individual words, but it is not as robust as RIPA (Ethayarajh et al., 2019).

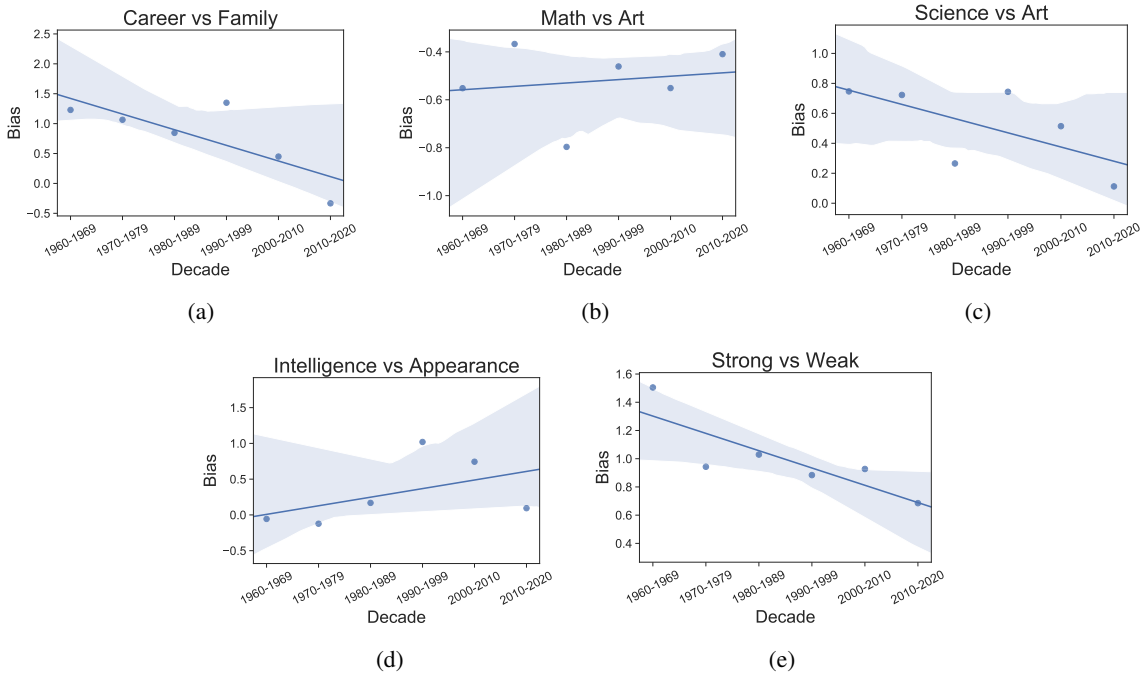


Figure 1: Each subfigure plots the bias measures using WEAT for one of five gender stereotypes: (a) Career vs Family, (b) Math vs Art, (c) Science vs Art, (d) Intelligence vs Appearance, and (e) Strong vs Weak. A bias score of zero represents no bias, i.e., no measurable difference between the two target categories for each gender. The shaded area of each subplot represents the bootstrap estimated 95% confidence interval.

of words. However, neither approach calculates a robust bias score for individual words. To study the most gender biased words over time, we make use of RIPA (Ethayarajh et al., 2019). Intuitively, RIPA uses a single vector to represent gender, then each word is scored by taking the dot product between the gender embedding and its respective embedding. The sign of the score will determine whether the embedding is more male or female-related.

The major aspect of RIPA is creating the gender embedding. Formally, given S, a non-empty set of ordered word pairs (x, y) (e.g., ('man', 'woman'), ('male', 'female')) that defines the gender association, we take the first principal component of all the difference vectors $\{\vec{\mathbf{x}} - \vec{\mathbf{y}} | (x, y) \in S\}$, which we call the relation vector $\vec{\mathbf{g}} \in \mathbb{R}^d$—that would be a one-dimensional bias subspace. Then, for some word vector $\vec{\mathbf{w}} \in \mathbb{R}^d$ the dot product is taken with $\vec{\mathbf{g}}$ to measure bias.

5 Results

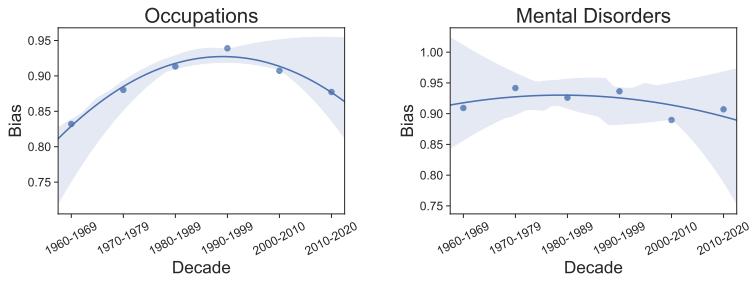
In this section, we present the results of our study in four parts. First, we report the embedding quality using UMLS-sim. Second, we study the temporal bias of traditional gender stereotypes, such as *Career vs Family* and *Strong vs Weak*. Ideally, we want to understand how, and which, stereotypes have changed over time. To understand the biased stereotypes, we make use of the WEAT method. Third, we look at whether occupational and mental health-related words are biased, and how the bias has changed over time. For this result, we only use a single set of target words. Thus, we make use of ECT. Fourth, we use RIPA to find the most biased words for each gender in each decade.

5.1 Embedding Quality.

In Table 3, we report the quality of each decade's embeddings based on the UMLS-sim dataset. Overall, we find that the quality consistently improves until the 1990s, however, we see drops in the 2000s and 2010s. We hypothesize that the reason for the decrease in embedding quality is because of the growth of research articles indexed on PubMed. Intuitively, word embeddings are only able to capture a single sense of a word. However, given the breadth of articles PubMed indexes—from machine learning (e.g., BioNLP) to biomaterials—multiple word meanings are being stored in a single vector. Thus, the overall quality begins to drop.

5.2 Traditional Gender Stereotypes.

In Figure 1, we plot the bias scores reported using WEAT. Remember, a large positive score means



(a) (b)

Figure 2: ECT bias estimates for both the set of occupation and mental disorder words. The shaded area of each subplot represents the bootstrap estimated 95% confidence interval.

that the $male$ words are more similar to the targets A (e.g., career) than the $female$ words. There is no measurable bias with a value of zero. Overall, we find that the results from the WEAT test vary depending on the stereotype. For *Career vs Family*, in Figure 1a, we find a steady linear decrease in bias each decade—with the exception of the 1990s. We also find similar linear decreases in bias for both *Science vs Art* and *Strong vs Weak* (Figures 1c and 1e). In Figure 1b, for *Math vs Art*, however, the bias stays relatively static, i.e., it does not dramatically change over time. Moreover, the WEAT score for *Math vs Art* is negative, meaning that the female words are more similar to math than the male words. Likewise, for *Intelligence vs Appearance* (Figure 1d), we see relatively little bias from 1960 to 1989, however, in the 1990s and 2000s, we had a substantial jump in the bias score.

Our evaluation supports prior work evaluating bias in biomedical word embeddings (e.g., *Strong vs Weak* is the most biased stereotype in biomedical literature) (Chaloner and Maldonado, 2019, Table 2). However, we also find differences when measuring bias over time. For example, we find that from 2010 to 2019 there is not a lot evidence for the *Career vs Family* stereotype in biomedical corpora, matching the results from Chaloner and Maldonado (2019, Table 2). Yet, this is only a recent phenomenon. The embeddings trained on articles published from 1990 to 1999 exhibit a *Career vs Family* bias score greater than 1.5. Overall, comparing to Chaloner and Maldonado (2019, Table 2), this means that the bias in recently published biomedical literature may not be as strong as what is found in general text corpora. But, if we exclude the most recent decade's embeddings, the bias in biomedical literature becomes much stronger. Future work should explore comparing the temporal bias in general text corpora to what is found in biomedical literature.

5.3 Occupational and Mental Health Bias.

In Figure 2, we report the gender bias results from ECT on two categories: occupations (e.g., doctor, nurse, teacher) and mental health disorders (e.g., depression, alcoholism, PTSD). Again, unlike WEAT, ECT calculates bias scores on a single target set of words. Therefore, we do not need two contrasting target word sets (e.g., *Math vs Art*), instead we can focus on bias for a single set (e.g., *Math*). Also, the larger the score, the lower the bias—a score of one would represent no difference between male and female words for that specific target set. Interestingly, we find that the ECT scores follow a similar pattern as found in Table 3, the better the embedding quality, the lower the bias.

Comparing Figures 2a and 2b, we find that the word embeddings for both occupations and mental disorders have relatively little bias in the 1990s. Furthermore, while there was small variation, mental disorders experienced little change in bias decade-by-decade. Yet, occupation-related words had a substantial amount of bias in the 1960s and 1970s. Moreover, we find that the bias related to occupations experienced more change, than mental disorders, starting 0.83 in the 1960s and increase by more than ten points to 0.94 in the 1990s. Whereas, mental disorder-related bias scores only ranged from 0.90 to 0.94.

5.4 Biased Words.

In Figure 2, we analyze the bias of individual occupational and mental health-related words. We found a substantial change in the bias of occupational-related words.

We found little change in the bias of mental health-related words since the 1960s. Yet, while

	Male					Female				
	1970-1979	**1980-1989**	**1990-1999**	**2000-2010**	**2010-2020**	**1970-1979**	**1980-1989**	**1990-1999**	**2000-2010**	**2010-2020**
					Occupations					
1	promoter	conductor	chef	dentist	mediator	teacher	housewife	neurosurgeon	swimmer	priest
2	collector	chef	baker	counselor	promoter	professor	teenager	pediatrician	baker	fisherman
3	investigator	biologist	astronaut	librarian	dentist	counselor	bishop	educator	butcher	teenager
4	principal	collector	swimmer	pharmacist	principal	physician	lawyer	teenager	medic	chef
5	baker	dad	prisoner	teenager	collector	pediatrician	pediatrician	counselor	barber	writer
6	researcher	singer	mechanic	bishop	cop	consultant	athlete	neurologist	physicist	nanny
7	character	chemist	character	acquaintance	conductor	doctor	physician	consultant	soldier	historian
8	mechanic	butler	worker	cardiologist	substitute	student	pathologist	dentist	baron	president
9	analyst	mechanic	soldier	promoter	coach	lawyer	educator	athlete	director	inventor
10	conductor	promoter	analyst	attorney	employee	pathologist	carpenter	doctor	singer	housewife
					Mental Disorders					
1	caffeine	cannabis	separation	lacunar	lacunar	dysmorphic	factitious	binge	dissociative	munchausen
2	restrictive	hypnotic	restrictive	bulimia	circadian	psychogenic	dysmorphic	nervosa	coordination	mutism
3	attachment	caffeine	coordination	erectile	nicotine	anorexia	nervosa	bulimia	separation	factitious
4	separation	coordination	dyskinesia	gambling	gambling	adolescent	mutism	opioid	parasitosis	dysmorphic
5	circadian	hallucinogen	conversion	bereavement	phencyclidine	nervosa	bulimia	hypersomnia	terror	hysteria
6	coordination	dependence	mathematics	binge	ocpd	mutism	tourette	narcolepsy	hysteria	cotard
7	benzodiazepine	attachment	attachment	nervosa	cocaine	infancy	infancy	anorexia	conversion	claustrophobia
8	dependence	mathematics	residual	mood	insomnia	munchausen	episode	panic	malingering	ekbom
9	selective	restrictive	parasitosis	depressive	sleep	factitious	anorexia	korsakoff	tic	diogenes
10	conversion	pdd	developmental	polysubstance	caffeine	disorder	munchausen	factitious	munchausen	encopresis

Table 4: The top ten words with the largest RIPA scores (i.e., the most biased) across each decade. The RIPA scores are reported for both occupations and mental health disorders. While all the listed words are biased, they are ranked starting with the most biased word to the least.

we found little change in mental health bias overall, are there at least a few disorders that changed over time? Moreover, we found a slight bias in mental health terms, therefore, What are the biased terms in each group? We look at the most gender biased occupational and mental health-related terms for each decade in Table 4. Because of space limitations, we only display the gendered words from the 1970s to the 2010s. The words from the 1960s can be found in the appendix. The word-level scores were generated using RIPA. First, for occupations, the words vary between male and female. For example, in the 1970s, male-related words include "mechanic", "principal", and "investigator". The female-related words include "teacher", "counselor", and "pediatrician". Interestingly, the jobs associated with men such as "principal" and "researcher" are positions with power over the jobs associated with woman. For example "principals" (male) have power over "teachers" (female) and "researchers" (male) have power over "students" (female). We also find other well-known occupations appear to be gender-related. For instance, "butler" in the 1980s is associated to male while "nanny" is related to female in the 2010s.

With regard to mental health, we find that disorders associated with well-known gender disparities appear to be biased using RIPA (Organization, 2013). For example, through the last 60 years, words associated with addictions are male-related, e.g., "caffeine", "cannabis", "nicotine", and "gambling". Similarly, disorders related to appearance are more female-related, e.g., "dysmorphic" [2] and "anorexia". We also find that disorders related to emotions are more female-related, such as "munchausen" [3], "hysteria" [4], and "terror". Interestingly, we find that the word "hysteria" is heavily biased in the 2010s. Even though the diagnosis of female hysteria substantially fell in the 1900s (Micale, 1993), it still seems to be a biased term. We want to note that this could simply be caused by research studying mental health diagnosis bias in women, however, the underlying cause of why the term is biased in the 2010s is left for future work.

6 Discussion

In this section, we discuss the impact of the results on two stakeholders of this research: BioNLP researchers and general biomedical researchers. Furthermore, we discuss the limitations of focusing on binary gender (*Male vs Female*).

[2]Dysmorphia is a mental health disorder in which you can't stop thinking about one or more perceived defects or flaws

[3]Munchausen is a mental disorder in which a person repeatedly and deliberately acts as if he or she has a physical or mental illness

[4]Hysteria is a (biased) catch-all for symptoms including, but not limited to, nervousness, hallucinations, and emotional outbursts.

6.1 Impact on BioNLP researchers.

The results in this paper are important for BioNLP research in two ways. First, we have produced decade-specific word embeddings. [5] Therefore, BioNLP research can use the embeddings to study other historical phenomenon in biomedical research articles. Second, the analysis of historical bias in biomedical research in this paper can be applied to other domains, beyond occupations and mental disorders.

6.2 Impact on Biomedical Researchers.

With regard to general biomedical researchers (e.g., medical researchers and biologist), this work can provide a way to measure which demographics current research is leaning towards in an automated fashion. As discussed in Holdcroft (2007), if research is heavily focused on a single gender, then health disparities can increase. Treatments should be explored equally for all at-risk patients. Furthermore,with the use of contextual word embeddings (Scheuerman et al., 2019), implicit bias measurement techniques can be used as part of the writing process to avoid gendered language when it is not necessary (e.g., using singular they vs he/she).

6.3 A Note About Gender.

Similar to prior work measuring gender bias (Chaloner and Maldonado, 2019), we focus on binary gender. However, it is important to note that the results for binary gender do not necessarily generalize to other genders, including, but not limited to, binary trans people, non-binary people, gender non-conforming people (Scheuerman et al., 2019). Therefore, we want to explicitly note that **our research does not necessarily generalize beyond binary gender**. In future work, we recommend that researcher's studies should be performed for other genders, beyond simply studying *Male vs Female*.

How can this study be expanded beyond binary gender? The three bias measurement techniques studied in this paper (i.e., WEAT, ECT, and RIPA) require sets of words representing a single gender (e.g., boy, men, male). Unfortunately, there is not a large number of words to represent every gender of interest. A promising area of research is to explore bias in contextual word embeddings. With the use of contextual word embeddings (Kurita et al., 2019), we can measure the bias of individual words across many contexts. Thus, we can possibly overcome the problem of a limited number of words per gender.

[5] https://github.com/AnthonyMRios/Gender-Bias-PubMed

7 Conclusion

In this paper, we studied the historical bias present in word embeddings from 1960 to 2020. In summary, we found that while some biases have shown a consistently decrease over time (e.g., *Strong vs Weak*), others have stayed relatively static, or worse, increased (e.g., *Intelligence vs Appearance*). Moreover, we found that the gender bias towards occupations has substantially changed over time, showing that in the past, there was more gender bias associated with certain jobs.

There are two major avenues for future work. First, this work quantified various aspects of gender bias over time. However, we do not know why the bias is present in the word embeddings. For example, is the word "hysteria" biased in 2010 because researchers are associating it with women implicitly, or is it that researchers are studying the historical usage of the diagnosis to ensure the diagnosis is not made because of implicit bias in the future? Thus, our future work will focus on causal studies of bias in biomedical literature. Second, we simply independently trained Skip-Gram word embeddings for each decade. However, recent work has shown that dynamic embeddings, rather than static (decade-specific), perform better with regard to analyzing public perception over time (Gillani and Levy, 2019). Future work will focus on developing new techniques to study bias temporally. Moreover, many techniques may depend on the magnitude of the bias, therefore, we plan to analyze the circumstances in which one embedding approach may measure bias (e.g., Skip-Gram) better than another (e.g., dynamic embeddings).

Acknowledgements

We would like to thank the anonymous reviewers for their invaluable help improving this manuscript. This material is based upon work supported by the National Science Foundation under Grant No. 1947697.

References

Paul R Albert. 2015. Why is depression more prevalent in women? *Journal of psychiatry & neuroscience: JPN*, 40(4):219.

American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pages 49–59.

Rahn Kennedy Bailey, Josephine Mokonogho, and Alok Kumar. 2019. Racial and ethnic differences in depression: current perspectives. *Neuropsychiatric disease and treatment*, 15:603.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

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

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

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174.

I Glenn Cohen, Ruben Amarasingham, Anand Shah, Bin Xie, and Bernard Lo. 2014. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health affairs*, 33(7):1139–1147.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Conference on AI, Ethics, and Society*.

Joel Escudé Font. 2019. Determining bias in machine translation with deep learning techniques. Master's thesis, Universitat Politècnica de Catalunya.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705.

Sergey Feldman, Waleed Ammar, Kyle Lo, Elly Trepman, Madeleine van Zuylen, and Oren Etzioni. 2019. Quantifying sex bias in clinical studies at scale with automated data extraction. *JAMA network open*, 2(7):e196700–e196700.

Joel Escudé Font and Marta R Costa-jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

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

Nabeel Gillani and Roger Levy. 2019. Simple dynamic word embeddings for mapping perceptions in the public sphere. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 94–99.

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

Amelia Gulliver, Kathleen M Griffiths, and Helen Christensen. 2010. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC psychiatry*, 10(1):113.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Katarina Hamberg. 2008. Gender bias in medicine. *Women's Health*, 4(3):237–243.

Cynthia M Hartung and Thomas A Widiger. 1998. Gender differences in the diagnosis of mental disorders: Conclusions and controversies of the dsm–iv. *Psychological bulletin*, 123(3):260.

Bin He, Yi Guan, and Rui Dai. 2019. Classifying medical relations in clinical text via convolutional neural networks. *Artificial intelligence in medicine*, 93:43–49.

Janae K Heath, Gary E Weissman, Caitlin B Clancy, Haochang Shou, John T Farrar, and C Jessica Dine. 2019. Assessment of gender-based linguistic differences in physician trainee evaluations of medical faculty using automated text mining. *JAMA network open*, 2(5):e193520–e193520.

Anita Holdcroft. 2007. Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine*, 100(1):2.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM.

Keyvan Khosrovian, Dietmar Pfahl, and Vahid Garousi. 2008. Gensim 2.0: a customizable process simulation model for software process evaluation. In *International conference on software process*, pages 294–306. Springer.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.

Bridget T McInnes, Ted Pedersen, and Serguei VS Pakhomov. 2009. Umls-interface and umls-similarity: open source software for measuring paths and semantic similarity. In *AMIA annual symposium proceedings*, volume 2009, page 431. American Medical Informatics Association.

Mark S Micale. 1993. On the" disappearance" of hysteria: A study in the clinical deconstruction of a diagnosis. *Isis*, 84(3):496–526.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *ICLR Workshop*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

World Health Organization. 2013. *Gender Disparities in Mental Health*. World Health Organization.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

LA Pratt and DJ Brody. 2014. Depression and obesity in the us adult household population, 2005-2010. *NCHS data brief*, (167):1–8.

Anthony Rios. 2020. FuzzE: Fuzzy fairness evaluation of offensive language classifiers on african-american english. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.

Anthony Rios and Ramakanth Kavuluru. 2015. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 258–267.

Arghavan Salles, Michael Awad, Laurel Goldin, Kelsey Krus, Jin Vivian Lee, Maria T Schwabe, and Calvin K Lai. 2019. Estimating implicit and explicit gender bias among health care professionals and surgeons. *JAMA network open*, 2(7):e196545–e196545.

Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue*, 11(3):10.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

Daniel Weisz, Michael K Gusmano, and Victor G Rodwin. 2004. Gender and the treatment of heart disease in older persons in the united states, france, and england: a comparative, population-based view of a clinical phenomenon. *Gender medicine*, 1(1):29–40.

Terry Young, Rebecca Hutton, Laurel Finn, Safwan Badr, and Mari Palta. 1996. The gender bias in sleep apnea diagnosis: are women missed because they have different symptoms? *Archives of internal medicine*, 156(21):2445–2451.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proc. of EMNLP*, pages 4847–4853.

A 1960s Most Biased Words

Male:

- physician
- doctor
- president
- dentist
- psychiatrist
- surgeon
- student
- nurse
- worker
- professor

Female:

- substitute
- principal
- editor
- baker
- character
- author
- pharmacist
- scientist
- therapist
- teacher

B Mental Health-Related Terms

[abuse, acute, adaptation, adjustment, adolescent, adult, affective, agoraphobia, alcohol, alcoholic, alzheimer, amnesia, amnestic, amphetamine, anorexia, anosognosia, anterograde, antisocial, anxiety, anxiolytic, asperger, atelophobia, attachment, attention, atypical, autism, autophagia, avoidant, avoidant, restrictive, barbiturate, behavior, benzodiazepine, bereavement, bibliomania, binge, bipolar, body, borderline, brief, bulimia, caffeine, cannabis, capgras, catalepsy, catatonia, catatonic, childhood, circadian, claustrophobia, cocaine, cognitive, communication, compulsive, condition, conduct, conversion, coordination, cotard, cyclothymia, daydreaming, defiant, deficit, delirium, delusion, delusional, delusions, dependence, depersonalization, depression, depressive, derealization, dermatillomania, desynchronosis, deux, developmental, diogenes, disease, disorder, dissociative, dyscalculia, dyskinesia, dyslexia, dysmorphic, eating, ejaculation, ekbom, encephalitis, encopresis, enuresis, epilepsy, episode, erectile, erotomania, exhibitionism, factitious, fantastica, fetishism, fregoli, fugue, functioning, gambling, ganser, grandiose, hallucinogen, hallucinosis, histrionic, huntington, hyperactivity, hypersomnia, hypnotic, hypochondriasis, hypomanic, hysteria, ideation, identity, impostor, induced, infancy, insomnia, intellectual, intermittent, intoxication, kleptomania, korsakoff, lacunar, lethargica, love, major, maladaptive, malingering, mania, mathematics, megalomania, melancholia, misophonia, mood, munchausen, mutism, narcissistic, narcolepsy, nervosa, neurocysticercosis, neurodevelopmental, nicotine, nightmare, nos, obsessive, obsessive–compulsive, ocd, ocpd, oneirophrenia, opioid, oppositional, orthorexia, pain, panic, paralysis, paranoid, parasitosis, parasomnia, parkinson, partialism, pathological, pdd, perception, persecutory, personality, pervasive, phencyclidine, phobia, phobic, phonological, physical, pica, polysubstance, posttraumatic, pseudologia, psychogenic, psychosis, psychotic, ptsd, pyromania, reactive, residual, retrograde, rumination, schizoaffective, schizoid, schizophrenia, schizophreniform, schizotypal, seasonal, sedative, selective, separation, sexual, sleep, sleepwalking, social, sociopath, somatic, somatization, somatoform, stereotypic, stockholm, stress, stuttering, substance, suicidal, suicide, tardive, terror, tic, tourette, transient, transvestic, tremens, trichotillomania, truman, withdrawal, wonderland]

C Occupations

[detective, ambassador, coach, officer, epidemiologist, rabbi, ballplayer, secretary, actress, manager, scientist, cardiologist, actor, industrialist, welder, biologist, undersecretary, captain, economist, politician, baron, pollster, environmentalist, photographer, mediator, character, housewife, jeweler, physicist, hitman, geologist, painter, employee, stockbroker, footballer, tycoon, dad, patrolman, chancellor, advocate, bureaucrat, strategist, pathologist, psychologist, campaigner, magistrate, judge, illustrator, surgeon, nurse, missionary, stylist, solicitor, scholar, naturalist, artist, mathematician, businesswoman, investigator, curator, soloist, servant, broadcaster, fisherman, land-

lord, housekeeper, crooner, archaeologist, teenager, councilman, attorney, choreographer, principal, parishioner, therapist, administrator, skipper, aide, chef, gangster, astronomer, educator, lawyer, midfielder, evangelist, novelist, senator, collector, goalkeeper, singer, acquaintance, preacher, trumpeter, colonel, trooper, understudy, paralegal, philosopher, councilor, violinist, priest, cellist, hooker, jurist, commentator, gardener, journalist, warrior, cameraman, wrestler, hairdresser, lawmaker, psychiatrist, clerk, writer, handyman, broker, boss, lieutenant, neurosurgeon, protagonist, sculptor, nanny, teacher, homemaker, cop, planner, laborer, programmer, philanthropist, waiter, barrister, trader, swimmer, adventurer, monk, bookkeeper, radiologist, columnist, banker, neurologist, barber, policeman, assassin, marshal, waitress, artiste, playwright, electrician, student, deputy, researcher, caretaker, ranger, lyricist, entrepreneur, sailor, dancer, composer, president, dean, comic, medic, legislator, salesman, observer, pundit, maid, archbishop, firefighter, vocalist, tutor, proprietor, restaurateur, editor, saint, butler, prosecutor, sergeant, realtor, commissioner, narrator, conductor, historian, citizen, worker, pastor, serviceman, filmmaker, sportswriter, poet, dentist, statesman, minister, dermatologist, technician, nun, instructor, alderman, analyst, chaplain, inventor, lifeguard, bodyguard, bartender, surveyor, consultant, athlete, cartoonist, negotiator, promoter, socialite, architect, mechanic, entertainer, counselor, janitor, firebrand, sportsman, anthropologist, performer, crusader, envoy, trucker, publicist, commander, professor, critic, comedian, receptionist, financier, valedictorian, inspector, steward, confesses, bishop, shopkeeper, ballerina, diplomat, parliamentarian, author, sociologist, photojournalist, guitarist, butcher, mobster, drummer, astronaut, protester, custodian, maestro, pianist, pharmacist, chemist, pediatrician, lecturer, foreman, cleric, musician, cabbie, fireman, farmer, headmaster, soldier, carpenter, substitute, director, cinematographer, warden, marksman, congressman, prisoner, librarian, magician, screenwriter, provost, saxophonist, plumber, correspondent, organist, baker, doctor, constable, treasurer, superintendent, boxer, physician, infielder, businessman, protege]