Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories

Kaytlin Chaloner
ADAPT Centre, SCSS
Trinity College Dublin
Ireland
chalonek@tcd.ie

Alfredo Maldonado
ADAPT Centre, SCSS
Trinity College Dublin
Ireland
alfredo.maldonado@adaptcentre.ie

Abstract

Prior work has shown that word embeddings capture human stereotypes, including gender bias. However, there is a lack of studies testing the presence of specific gender bias categories in word embeddings across diverse domains. This paper aims to fill this gap by applying the WEAT bias detection method to four sets of word embeddings trained on corpora from four different domains: news, social networking, biomedical and a gender-balanced corpus extracted from Wikipedia (GAP). We find that some domains are definitely more prone to gender bias than others, and that the categories of gender bias present also vary for each set of word embeddings. We detect some gender bias in GAP. We also propose a simple but novel method for discovering new bias categories by clustering word embeddings. We validate this method through WEAT's hypothesis testing mechanism and find it useful for expanding the relatively small set of well-known gender bias word categories commonly used in the literature.

1 Introduction

Artificial intelligence (AI) acquired from machine learning is becoming more prominent in decision-making tasks in areas as diverse as industry, healthcare and education. AI-informed decisions depend on AI systems' input training data which, unfortunately, can contain implicit racial, gender or ideological biases. Such AI-informed decisions can thus lead to unfair treatment of certain groups. For example, in Natural Language Processing (NLP), résumé search engines can produce rankings that disadvantage some candidates, when these ranking algorithms take demographic features into account (directly or indirectly) (Chen et al., 2018), while abusive online language detection systems have been observed to produce false positives on terms associated with minorities and women (Dixon et al., 2018; Park et al., 2018). Another example where bias (specifically gender bias) can be harmful is in personal pronoun coreference resolution, where systems carry the risk of relying on societal stereotypes present in the training data (Webster et al., 2018).

Whilst gender bias in the form of concepts of masculinity and femininity has been found inscribed in implicit ways in AI systems more broadly (Adam, 2006), this paper focuses on gender bias on word embeddings.

Word embeddings are one of the most common techniques for giving semantic meaning to words in text and are used as input in virtually every neural NLP system (Goldberg, 2017). It has been shown that word embeddings capture human biases (such as gender bias) present in these corpora in how they relate words to each other (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018). For the purposes of this paper, gender bias is understood as the inclination towards or prejudice against one gender.

Several methods have been proposed to test for the presence of gender bias in word embeddings; an example being the Word Embedding Association Test (WEAT) (Caliskan et al., 2017). WEAT is a statistical test that detects bias in word embeddings using cosine similarity and averaging methods, paired with hypothesis testing. WEAT's authors applied these tests to the publicly-available GloVe embeddings trained on the English-language "Common Crawl" corpus (Pennington et al., 2014) as well as the Skip-Gram (word2vec) embeddings trained on the Google News corpus (Mikolov et al., 2013). However, there is a diverse range of publicly-available word embeddings trained on corpora of different domains. To address this, we applied the WEAT test on four sets of word embeddings trained on corpora from four domains: social media (Twit-

Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pages 25–32
Florence, Italy, August 2, 2019.

ter), a Wikipedia-based gender-balanced corpus (GAP) and a biomedical corpus (PubMed) and news (Google News, in order to reproduce and validate our results against those of Caliskan et al. (2017)) (see Section 3).

Caliskan et al. (2017) confirmed the presence of gender bias using three categories of words well-known to be prone to exhibit gender bias: (B1) **career vs. family** activities, (B2) **Maths vs. Arts** and (B3) **Science vs. Arts**. Garg et al. (2018) expanded on this work and tested additional gender bias word categories: (B4) differences on personal descriptions based on **intelligence vs. appearance** and on (B5) physical or emotional **strength vs. weakness**. In this paper, we use these five categories to test for the presence of gender bias in the aforementioned domain corpora. Notice that one of the tested corpora is the gender-balanced GAP corpus (Webster et al., 2018). We specifically chose this corpus in order to test whether the automatic method used to compile it (based on sampling an equal number of male and female pronouns from Wikipedia) yielded a set that was balanced according to these five well-known gender bias word categories. GAP's authors acknowledge that Wikipedia has been found to contain gender biased content (Reagle and Rhue, 2011).

We confirmed bias in all five categories on the Google News embeddings but far less bias on the rest of the embeddings, with the biomedical PubMed embeddings showing the least bias. We did find some bias on GAP. However, given the small size of this corpus, many test words were not present (see Section 4).

The six word categories studied here are word lists manually curated by Psychology researchers based on their studies (e.g. Greenwald et al., 1998). However, it is difficult to establish whether they are exhaustive as there could be other word categories presenting bias, which may well be domain-dependant. In response, we developed a simple method to automatically discover new categories of gender bias words based on word clustering, and measuring statistical associations of the words in each cluster to known female and male attribute words. Assuming that each cluster roughly represents a topic in the corpus, the set of gender bias words in each cluster/topic in the corpus corresponds to a potentially new category of gender-biased words. As far as we are aware, this is the first time a method to discover new gender bias word categories is proposed. We used WEAT's hypothesis testing mechanism to automatically validate the induced gender bias word categories produced by our system. A visual inspection on a sample of these induced categories is consistent with the authors' intuitions of gender bias. We make these induced categories available to other researchers to study.[1] An advantage of this discovery method is that it allows us to detect bias based on a corpus' own vocabulary, even if it is small, as is the case in the GAP corpus embeddings.

2 Previous Work

In word embeddings, words are represented in a continuous vector space where semantically similar words are mapped to nearby points (Goldberg, 2017, ch. 10). The underlying assumption is that words that appear in similar contexts share similar meaning (Harris, 1954; Miller and Charles, 1991). This context-based similarity is operationalised through cosine similarity, a well-established method for measuring the semantic similarity of words in vector space (Schütze, 1998). Recently, however, researchers noticed that cosine similarity was able to exhibit gender biases captured through training on corpora and started developing methods for mitigating this bias (Bolukbasi et al., 2016). Caliskan et al. (2017) then developed the Word Embedding Association Test (WEAT), which is an adaptation of the Implicit Association Test (IAT) from Psychology (Greenwald et al., 1998) to measure biases in word embeddings. The IAT measures a person's automatic association between mental representations of concepts, based on their reaction times. Instead of relying on reaction times, WEAT relies on cosine similarity. WEAT is based on two statistical measures: (1) the effect size in terms of Cohen's d, which measures the association between suspected gender biased words and two sets of reference words (attribute words in WEAT's terminology) known to be intrinsically male and female, respectively; and (2) a statistical hypothesis test that confirms this association. We borrow these statistical measures in this paper. Garg et al. (2018) measured gender bias synchronically across historical data covering 100 years of English language use.

Most work however has concentrated in meth-

[1] Code, generated embeddings and data available at https://github.com/alfredomg/GeBNLP2019

ods for mitigating gender bias in word embeddings. One approach is debiasing learnt corpora (Bolukbasi et al., 2016), which is achieved using algorithms that modify word embeddings in such a way that neutralises stereotypical cosine similarities. Another approach is creating gender-balanced corpora, such as the GAP corpus (balanced corpus of Gendered Ambiguous Pronouns) (Webster et al., 2018). Roughly speaking, GAP was developed by sampling sentences from Wikipedia in such a way that an equal number of male and female personal pronouns was obtained. Its main use is in the evaluatiation of systems that resolve the coreference of gendered ambiguous pronouns in English. In a similar vain, Dixon et al. (2018) builds a balanced corpora that seeks to neutralise toxic mentions of identity terms.

To the best of our knowledge there has not been work testing for bias on corpora from different domains. Also, we believe this is the first time an unsupervised method for discovering new gender bias word categories from word embeddings is proposed.

3 Choice of Word Embeddings

English-language word embeddings were selected with the intention of giving an insight into gender bias over a range of domains and with the expectation that some word embeddings would demonstrate much more bias than others. The word embeddings selected were: (a) Skip-Gram embeddings trained on the Google News corpus[2], with a vocabulary of 3M word types (Mikolov et al., 2013); (b) Skip-Gram embeddings trained on 400 million Twitter micro-posts[3], with a vocabulary of slightly more than 3M word types (Godin et al., 2015); (c) Skip-Gram embeddings trained on the PubMed Central Open Access subset (PMC) and PubMed[4], with a vocabulary of about 2.2M word types (Chiu et al., 2016) and trained using two different sliding window sizes: 2 and 30 words; (d) FastText embeddings trained on the GAP corpus (Webster et al., 2018) by us[5], with a vocabulary of 7,400 word types.

[2] https://tinyurl.com/mpzqe5o

[3] https://github.com/loretoparisi/word2vec-twitter

[4] https://github.com/cambridgeltl/BioNLP-2016

[5] See footnote 1.

4 WEAT Hypothesis Testing

4.1 Experimental Protocol

We largely follow the WEAT Hypothesis testing protocol introduced by Caliskan et al. (2017). The input is a suspected gender bias word category represented by two lists, X and Y, of **target words**, i.e. words which are suspected to be biased to one or another gender. E.g. $X = \{$programmer, engineer, scientist$\}$, $Y = \{$nurse, teacher, librarian$\}$. We wish to test whether X or Y is more biased to one gender or the other, or whether there is not difference in bias between the two lists. Bias is compared in relation to two reference lists of words that represent unequivocally male and female concepts. E.g. $M = \{$man, male, he$\}$, $F = \{$woman, female, she$\}$. In WEAT's terminology these reference lists are called the **attribute words**. Table 1 shows the target and attribute word sets used in our experiments.

The null hypothesis H_o is that there is no difference between X and Y in terms of their relative (cosine) similarity to M and F. Assuming that there is a word embedding vector $\vec{w}$ (trained on some corpus from some domain) for each word w in X, Y, M and F, we compute the following **test statistic**:

$$s(X, Y, M, F) = \sum_{x \in X} s(x, M, F) - \sum_{y \in Y} s(y, M, F) \quad (1)$$

where $s(w, M, F)$ is the **measure of association** between target word w and the attribute words in M and F:

$$s(w, M, F) = \frac{1}{|M|} \sum_{m \in M} \cos(\vec{w}, \vec{m}) - \frac{1}{|F|} \sum_{f \in F} \cos(\vec{w}, \vec{f}) \quad (2)$$

In Caliskan et al. (2017) H_o is tested through a permutation test, in which $X \cup Y$ is partitioned into alternative target lists $\hat{X}$ and $\hat{Y}$ exhaustively and computing the one-sided p-value $p[s(\hat{X}, \hat{Y}, M, F) > s(X, Y, M, F)]$, i.e. the proportion of partition permutations $\hat{X}$, $\hat{Y}$ in which the test statistic $s(\hat{X}, \hat{Y}, M, F)$ is greater than the observed test statistic $s(X, Y, M, F)$. This p-value is the probability that H_o is true. In other words, it is the probability that there is no difference between X and Y (in relation to M and F) and therefore that the word category is *not* biased. The

	Attribute words	M	male, man, boy, brother, he, him, his, son, father, uncle, grandfather
		F	female, woman, girl, sister, she, her, hers, daughter, mother, aunt, grandmother
Target words	B1: career vs family	X	executive, management, professional, corporation, salary, office, business, career
		Y	home, parents, children, family, cousins, marriage, wedding, relatives
	B2: maths vs arts	X	math, algebra, geometry, calculus, equations, computation, numbers, addition
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	B3: science vs arts	X	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	B4: intelligence vs appearance	X	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
		Y	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
	B5: strength vs weakness	X	power, strong, confident, dominant, potent, command, assert, loud, bold, succeed, triumph, leader, shout, dynamic, winner
		Y	weak, surrender, timid, vulnerable, weakness, wispy, withdraw, yield, failure, shy, follow, lose, fragile, afraid, loser

Table 1: Target words used for each gender-bias word category and attribute words used as gender reference

higher this p-value is the less bias there is. Following Caliskan et al. (2017), in this work we consider a word category to have statistically significant gender bias if its p-value is below the 0.05 threshold. Given that a full permutation test can quickly become computationally intractable, in this paper we instead use randomisation tests (Hoeffding, 1952; Noreen, 1989) with a maximum of 100,000 iterations in each test.

4.2 WEAT Results

Before experimentation we expected to find a great deal of gender bias across the Google News and Twitter embedding sets and far less in the PubMed and GAP sets. However, results in Table 2 are somewhat different to our expectations:

Google News We detect statistically significant (p-values in bold) gender bias in all 5 categories (B1-B5) on this corpus. Although one would hope to find little gender bias in a news corpus, given that its authors are professional journalists, bias had already been detected by Caliskan et al. (2017) and Garg et al. (2018) using methods similar to ours. This is not surprising given that women represent only a third (33.3%) of the full-time journalism workforce (Byerly, 2011). In addition, it has been found that news coverage of female personalities more frequently mentions family situations and is more likely to invoke matters of superficial nature, such as personality, appearance and fashion decisions, whereas the focus on men in news coverage tends to be be given to their experience and accomplishments (Armstrong et al., 2006).

Twitter On this social media set, we surprisingly only detected bias on the career vs. family (B1) category, although science vs. maths (B2) is a borderline case with a p-value of just 0.0715, and the rest of the values are not particularly high. We also observe that most effect sizes (Cohen's d) are under 1, indicating relatively weaker associations with the gender-specific attribute words from Table 1. We leave for future work further analysis on this set, however we hypothesise that the idiosyncratic language use common in micro-blogging, such as non-standard spelling and hashtags, divide up the semantic signal of word embeddings, perhaps diluting their association bias. Indeed, the word categories showing most gender bias in the discovery experiments (Section 5) include many hashtags, punctuation marks and words with non-standard spellings such as "alwaaaaays", which will not be tested for bias using standard-spelling target words.

PubMed This biomedical set showed the least gender bias, which was expected given its scientific nature. However, it has been documented that gender bias exists in biomedical studies given that more clinical studies involve males than females, and also based on the differences in which male and female patients report pain and other medical complaints and, in turn, the differences in which male and female health practitioners record and understand these complaints (Fillingim et al., 2009). It is possible that gender bias is still present in these texts but it is manifested differently and perhaps cannot be detected through word embed-

	Google News		Twitter		PubMed w2		PubMed w30		GAP	
Categories	p	d	p	d	p	d	p	d	p	d
B1: career vs family	**0.0012**	1.37	**0.0029**	1.31	0.7947	-0.42	0.0962	0.67	**0.0015**	1.44
B2: maths vs arts	**0.0173**	1.02	0.1035	0.65	0.9996	-1.40	0.9966	-1.20	0.0957	1.04
B3: science vs arts	**0.0044**	1.25	0.0715	0.74	0.9797	-0.98	0.7670	-0.37	0.1434	0.71
B4: intelligence vs appearance	**0.0001**	0.98	0.1003	0.37	0.2653	0.18	0.0848	0.36	0.9988	-0.64
B5: strength vs weakness	**0.0059**	0.89	0.2971	0.20	0.0968	0.48	**0.0237**	0.72	**0.0018**	0.77

Table 2: WEAT hypothesis test results for corpora tested for five well-known gender-biased word categories. p-values in bold indicate statistically significant gender bias ($p < 0.05$).

dings. Also of note is that across all five categories, bias is greater (smaller p-values) on the 30-word window set than on the 2-word window set. It is known that window size affects semantic similarity: larger window sizes tend to capture broader, more topical similarities between words whilst smaller windows capture more linguistic or even syntactic similarities (Goldberg, 2017, Sec. 10.5). We leave for future work further analysis on the bias effects of window sizes.

GAP Whilst GAP was specifically developed with gender balance in mind, we did find some degree of gender bias. In fact, given that it is derived from a gender-biased source text (Wikipedia), we actually expected to measure a higher degree of gender bias. This relatively low bias measurement could be due in part to the fact that GAP's vocabulary lacks many of the attribute and target word lists used in the tests. Table 3 shows the number of out-of-vocabulary words from these lists in PubMed and GAP (Google News and Twitter did not have any out-of-vocabulary words). Notice that the category missing most target words (intelligence vs. appearance category, B4) shows the least bias. However, the second category that misses most words (strength vs weakness, B5) does indeed show bias to a medium-high effect size of 0.77. This difficulty in assessing the reliability of these tests, in the face of a relatively high number of out-of-vocabulary attribute and target words, is one of the reasons that inspired us to develop a method for discovering new categories of biased words from an embedding set's own vocabulary. Section 5 covers this method.

5 Discovering New Gender Bias Word Categories

We propose a method for automatically detecting new categories of gender-biased words from a word embedding set. The simplest method in-

	Attrs.		Target Words									
			B1		B2		B3		B4		B5	
	M	F	X	Y	X	Y	X	Y	X	Y	X	Y
TOTAL	11	11	8	8	8	8	8	8	25	25	15	15
PubMed	0	0	0	0	0	0	0	0	0	1	0	0
GAP	0	1	1	1	6	1	4	1	21	18	7	9

Table 3: Number of out-of-vocabulary target and attribute words in the PubMed and GAP embeddings. Google and Twitter embeddings contain all words.

volves constructing a list of male- and female-biased words from a word embedding vocabulary through eq. (2). However, the resulting list would not have a topical or semantic cohesion as the categories B1-B5 have. We propose instead to first cluster the word vectors in an embedding set and then return a list of male- and female-associated word lists per cluster. We expect these cluster-based biased word lists to be more topically cohesive. By controlling for the number and size of clusters it should be possible to find more or less fine-grained categories.

We cluster embedings using K-Means++ (Arthur and Vassilvitskii, 2007), as implemented by scikit-learn (Pedregosa et al., 2011), using 100 clusters for GAP and 3,000 for Google News, Twitter and PubMed (window size 30 only). This algorithm was chosen as it is fast and produces clusters of comparable sizes. For each cluster we then return the list of n most male- and female-associated words (as per eq. 2): these are the discovered gender bias word categories candidates. Table 4 shows a selection of these candidates.[6]

Upon visual inspection, most of these candidates seem to be somewhat cohesive. We notice that on Google News and GAP many of the clusters relate to people's names (Google News cluster 2369) whilst others mix people's names with

[6] All candidates in paper repo. See footnote 1.

	Clus.	Gender Biased words: Male	Gender Biased words: Female	d
Google News	2763	eating_cheeseburgers, Tuna_Helper, Kielbasa, Turtle_Soup, beef_patty_topped, noodle_stir_fry, fried_broiled, trencherman, magnate_Herman_Cain, knockwurst, cracklins, hearty, juicy_steaks, Philly_Cheesesteak, duck_goose, PBJ, loafs, Eat_MREs, Cheddar_cheeses, pizzas_salads	Gingersnap, Blueberry_Pie, champagne_truffles, bake_brownies, Bon_Bons, Seasonal_Fruit, bakes_cakes, Sinfully, Lemon_Curd, Tagalong, Godiva_Chocolate, brownie_bites, Adoree, apple_crisps, Elinor_Klivans, Mud_Pie, decorate_cupcakes, granola_cereal, baked_apple_pie, cakes_cupcakes	1.97
	2369	Luke_Schenscher, Stetson_Hairston, Jake_Odum, Maureece_Rice, Errick_Craven, Marcus_Hatten, Jeremiah_Rivers, JR_Pinnock, Tom_Coverdale, Isacc_Miles, Brian_Wethers, Jeff_Varem, Matt_Pressey, Tyrone_Barley, Tavarus_Alston, Kojo_Mensah, Marcellus_Sommerville, Lathen_Wallace, Jordan_Cornette, Willie_Deane	Jayne_Appel, Cappie_Pondexter, Betty_Lennox, Kara_Lawson, Janel_McCarville, Lisa_Leslie, Deanna_Nolan, Sancho_Lyttle, Seimone_Augustus, Candice_Wiggins, Gearlds, Jessica_Davenport, Plenette_Pierson, Wisdom_Hylton, Lindsey_Harding, Yolanda_Griffith, Elena_Baranova, Loree_Moore, Taurasi, Noelle_Quinn	1.97
	2995	vetran, defens, ennis, 3AW_Debate, efore, carrer, redknapp, exellent, shanny, slater, shanahan, afridi, brees, westbrook, Thudd, dirk, feild, righ, duhhh, arsene_wenger	lolita, shiloh, beverly_hills, middleton, extr, leah, dwts, sophie, aniston, kathryn, liza, kristen_stewart, celine, kristin, tess, elena, alexandra, versace, alison, michelle_obama	1.97
	2424	brewing_vats, Refrigeration_Segment, Sealy_mattress, anesthesia_workstations, outdoor_Jacuzzis, Otis_elevators_Carrier, panels_moldings, Van, CPVC_pipes, refurbed, dome_coverings, covered_amphitheater_Astroturf, JES_Restaurant, beakers_flasks_wiring_Hazmat, Home_3bdrm_2ba, mold_peeling_paint, Zanussi, hoovers, Brendan_Burford, grills_picnic_tables	Corningware, Lowenborg_Frick, Aveda_bath, upholstery_carpets, bedside_commodes, Janneke_Verheggen_spokeswoman, hipster_tastemaker_Kelly_Wearstler, vacuuming_robot, dinettes, comforters_sheets, robes_slippers, Wolfgang_Puck_Bistro, neck_massager, breakrooms, china_cutlery, jewelers_florists, linen_towels, Frette_sheets, holding_freshly_diapered, china_flatware	1.97
Twitter	7	#HEG_NUMBER_, #zimdecides, #goingconcern, #twobirdswithonestone, #BudSelig, #LeedsUnitedAFC, #Buyout, #Batting, #rickyhatton, #baddeal, #ChaudhryAslam, #radio_NUMBER_today, DETROYT, #findtim, #houndsleadthepack, #Warcriminal, #spil, #kenyakwanza, #commonpurpose, #patriotway	#Charade, #lcchat, #charlizetheron, #horoscopo, #DiamondJubilee, #paternity, #tombola, #singlepeople, #Fabian, #Flipper, #toilettraining, #eca_NUMBER_, #financialadvisers, #tn_NUMBER_, Swift-boated, #dailytips, #Aramex, #MBCAware, #Glaciers, #RiverValley	1.97
	991	hewy, Suchecki, furgie, Huebert, bseball, jump-pass, gaudreau, #Thakur, lookalikey, lavalle, _NUMBER_verse, Timonen's, #Kipper, Kouleas, Mannjng, #wetpanda, oriels, Drowney, Brucato, sczesney	perrrrlease, VoteVictorious, olone, Chick-fi-la, InVasion, #DinahTo_NUMBER_K, Chika's, relaxxxxxxxxzzzxxzxxzxzxzzzzxxssssseeeeeeezzzxxx, Heliodore, Kandia, Shakeita, flowers/plants, dress/heels, sexy-times, #sobersunday, Teonia, shrinkies, Kokiri's, solutely, gayke	1.97
	2998	Swishhhhhh, Johnsonnn, #dadadadadada, Coyh, #HappyBdaySpezz, #RawRivals, #Fastandthefurious, #HedoIsTheBestOnTwitter, LeJames, #spursheat, #sidelined, #BackInYourBox, #fuckmanutd, #FellainiFacts, #LosBravos, #GreatestPlayerIHaveEverSeen, TouchDwn, #celticsvsknicks, Irte, #MSQuotes	#tiffanymynx, #solvetheriddle, #mandybright, #getthehelloutofhere, #RedSoles, #that'lldopig, #MeetVirginia, #freenudes, #KeepitClassy, #lushlive, #PeaceMessageIn_NUMBER_FromCarolis hFamily, #GotOne, #itsnotfine, #thursdayhurryup, #gentletweets, #InDesperateNeed, #PackInMore, masterpieceeee, somebodytellmewhy, #biggestflaw	1.97
PubMed	2993	JSK, 1938-1952, Johann-Wolfgang-Goethe, Winstein, Argenteuil, Alfried, Critica, NUST, traumatologie, Saarow, Urologische, carDiac, Neustadt/Saale, Massy, Umgebung, 1925-1927, Eli-Lilly, Commented, Senden, Maisons	gynecology-important, non-CEE, Breast-, Oketani, RiSk, MIREC, NASPAG, Cervitula, PDCU, cervical-ripening, Step-2, Kimia, Skin-to-Skin, Eeva, AdHOC, NMDSP, lipid-management, CEPAM, NCT00397150, Mass-screening	1.97
	1092	3beta, 19-NA, Leyding, u-PAI, nonpinealectomized, DHA-s, hydrocortisone-supplemented, misulban, Burd, CHH/KS, NEP28, adrenostasis, JNF, dihydrotestosterone-, appetite-stimulatory, P-hGH, d-Leu6, GIP-treated, alpha-methyl-DHT, pineal-gland, DESPP	embryo/foetus, hormonally-dependent, gland-stroma, 16α-OHE1, conceptus-produced, Lactogenesis, nERalpha, mid-reproductive, relaxin-deficient, 64.0-kDa, pre-synchronized, 17alpha, 20alpha-dihydroxypregn-4-en-3-one, pseudogestation, foetectomy, E2-dominated, Pinopodes, midpseudopregnancy, estrous-cycle, LH-only, blood-mammary	1.97
	559	examiniations, cholangiocarcinoma., phlebolithiasis, Celiacography, 79-year-old-woman, 6 cases, spermatocystitis, microbladder, otorhino-laryngological, FACD, neurogen, Cryodestruction, bladder-, Diverticuli, pseudo-angina, epididymo-testicular, Rendu-Weber-Osler, Dystopic, Lazorthes, AISO	Conisation, TV-US, sonohysterogram, peri, tracheloscopy, Salpingo-oophorectomy, hysteroscopic-guided, endosalpingitis, perifimbrial, previa-percreta, hystero-salpingography, pudendum, auto-amputated, hysterometry, Peritubal, isthmocervical, ovserved, hemosalpinx, Hysterosalpingo, necrosis/dehiscence	1.97
GAP	84	critical, artistic, commercial, vocal, era, pop, article, project, comic, projects, commercials, science, artists, editions, critic, popular, sports, introduction, articles, vocals	roles, films, television, two, drama, producer, worker, musicians, producers, magazine, produce, programmes, version, products, credits, music, opera, portrayal, features, direct	1.74
	93	captures, tribe, struck, brain, Asgard, capacity, coin, reinforcements, favour, corpse, assault, license, referee, system, aide, proceedings, strigoi, loyalty, Yu, energy	Owen, green, parole, rapper, telephone, together, personally, shoe, heroine, chosen, between, storyline, clothes, ghost, daily, Pink, spell, neighborhood, adult, Ramona	1.95
	73	resulted, responded, considered, constructed, used, respected, accused, committed, ordered, recognized, participated, charged, recommended, focused, devoted, instructed, captured, regarded, demonstrated, controlled	played, disappeared, stayed, named, arranged, betrayed, hatred, displayed, Damaged, danced, shared, Jared, Named, hosted, abandoned, teamed, separated, Voiced, appealed, welcomed	1.91
	18	treat, inside, demands, capable, proceeds, crash, skills, buy, far, unable, cash, struggle, promises, guilty, threat, fun, engage, bail, boat, toward	stolen, actually, friend, even, stays, fallen, Tina, sit, sex, doll, alive, sick, night, totally, boy, sheet, step, knew, still, Esme	1.91

Table 4: Selection of induced gender bias word categories per cluster.

more obviously biased words (Google News cluster 2995 and most GAP clusters). It is clear that this method detects thematically-cohesive groups of gender-associated words. However, not all words seem to be genuinely gender biased in a harmful way. We leave for future work the development of a filtering or classification step capable of making this distinction.

In order to test whether the candidates' bias is statistically significant, we applied the full WEAT hypothesis testing protocol, using randomised tests of 1,000 iterations per cluster to make the computation tractable. All clusters across all embedding sets returned a p-value < 0.001. The effect size (Cohen's d) was quite high across all clusters, averaging 1.89 for Google News, 1.87 for Twitter, 1.88 for PubMed and 1.67 for GAP. We leave for future work to conduct a human-based experiment involving experts on gender bias on different domains and languages other than English to further validate our outputs. Emphasis will be placed on assessing the usefulness of this tool for domains and languages lacking or seeking to develop lists of gender bias word categories.

6 Conclusions and Future Work

We have shown that there are varying levels of bias for word embeddings trained on corpora of different domains and that within the embeddings, there are different categories of gender bias present. We have also developed a method to discover potential new word categories of gender bias. Whilst our clustering method discovers new gender-associated word categories, the induced topics seem to mix harmless gender-associated words (like people names) with more obviously harmful gender-biased words. So as a future development, we would like to develop a classifier to distinguish between harmless gender-associated words and harmful gender-biased words. We wish to involve judgements by experts on gender bias in this effort, as well as exploiting existing thematic word categories from lexical databases like WordNet (Fellbaum, 1998), ontologies and terminologies. At the same time, we will also seek to measure the negative impact of discovered categories in NLP systems' performance. We also wish to more closely investigate the relationships between different word embedding hyperparameters, such sliding window size in the PubMed set, and their learned bias.

Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. We wish to thank our anonymous reviewers for their invaluable feedback.

References

Alison Adam. 2006. *Artificial knowing: Gender and the thinking machine*. Routledge.

Cory L Armstrong, Michelle LM Wood, and Michelle R Nelson. 2006. Female news professionals in local and national broadcast news during the buildup to the Iraq War. *Journal of Broadcasting & Electronic Media*, 50(1):78–94.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The Advantages of Careful Seeding. In *Proceedings of the 2007 ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, New Orleans, LA.

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.

Carolyn M Byerly. 2011. *Global Report on the Status Women in the News Media*. Washington, DC: International Women's Media Foundation [IWMF].

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

Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 651. ACM.

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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans, LA.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Roger B. Fillingim, Christopher D. King, Margarete C. Ribeiro-Dasilva, Bridgett Rahim-Williams, and Joseph L. Riley. 2009. Sex, Gender, and Pain: A Review of Recent Clinical and Experimental Findings. *Journal of Pain*, 10(5):447–485.

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.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153.

Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Wassily Hoeffding. 1952. The large sample power of tests based on permutations. *The Annals of Mathematical Statistics*, 23(2):169–192.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Eric W. Noreen. 1989. *Computer-intensive methods for testing hypotheses : an introduction*. Wiley.

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, Brus.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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*, pages 1532–1543, Doha.

Joseph Reagle and Lauren Rhue. 2011. Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5:21.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.