Machine Unlearning Human Biases: Inclusive Word Embeddings by Excluding Biased Texts

Shikhar Singla
London Business School
ssingla@london.edu

Abstract

Word embeddings exhibit biases such as racial and gender biases due to the presence of these biases in the training corpus. Usage of these algorithms can increase the stereotypes in various contexts. We present a simple and generalizable approach of detecting the parts of a corpus that affect the bias and show how removing those parts can debias the word embeddings. The approach finds words that link the target words for a group and biased or attribute words (indirect bias). Unlike prior work, our approach a) removes the biases completely, b) removes indirect bias, and c) can be generalized to any type of bias, downstream task or word embedding model. We apply our methodology on Wikipedia and American National Corpus (ANC) for Word2Vec and GloVe models on the racial and gender biases. It is highly accurate in removing the biases without affecting the performance of the models in capturing semantic information.

1 Introduction

Word embeddings capture word semantics and are used in many NLP models and downstream tasks. However, they contain human biases such as racial, gender or religious since these biases are present in the training data. (Caliskan et al. (2017), Bolukbasi et al. (2016), Garg et al. (2018)).

Many debiasing solutions have been proposed (Bolukbasi et al. (2016), Zhao et al. (2018b), Kaneko and Bollegala (2019), Kaneko and Bollegala (2021)) in the literature. The existing methods do not remove the bias completely (Gonen and Goldberg (2019), Kaneko and Bollegala (2021)).

The reason pointed out by Gonen and Goldberg (2019) for this is twofold. First, the proposed methods reduce the direct bias but not the indirect bias. Direct bias originates from the relationship between target words of a group (she, her, woman etc., for women) and the attribute words (words related to family or professions). Indirect bias is the bias that stems from the words that are closely associated with both the target and attribute words. The bias remains since the association of these words with target words is not changed much. Second, many of these solutions are post-processing methods and do not alter the training data, which is the source of the biases.

Finally, many biases have been recognized in word embeddings - for example, a disproportionate association of women with family than career and stereotyped association of certain professions with women. No generalized method exists which can debias such different kinds of biases.

To overcome these challenges, we develop a simple and generalizable approach to debias word embeddings. Our methodology has three steps. We will explain the method with the example of the bias pointed out in Caliskan et al. (2017) - women have a higher association with family terms than career terms compared to men. The first step is to train a word embedding model which has the bias we want to remove. Second, find the words (link words) close to both the target words for a group (common female names) and the attribute words (family terms). We use cosine similarity to get the association with target words and a polarity score assigned by the approach devised by Rothe et al. (2016) to get the association with family (vs career) terms. The final step is to find the sentences/articles that have both the target words and the link words and re-train the model without these sentences/articles. The number of words is the only parameter that needs to be changed until the bias is zero. This approach on choosing the right N removes the bias entirely by focusing on the indirect bias. Since it works on the association between words, it can be generalised to any bias or downstream task.

We apply our approach on Wikipedia and ANC corpora for Word2Vec (Mikolov et al. (2013a) and GloVe (Pennington et al. (2014) word embeddings for racial and gender biases measured with Word Embedding Association Test (WEAT) (Caliskan et al. (2017)). Our experimental results show that the method accurately removes the biases completely without affecting the performance on word embedding benchmarks.

2 Related Work

Word embeddings are vector representations of words learned from training documents that capture syntactic and semantic relationships between words such Man is to King as Woman is to Queen (Mikolov et al. (2013b). They have also been shown to maintain discriminatory biases present in the corpus, such as Man is to Computer Programmer as Woman is to Homemaker (Bolukbasi et al. (2016)).

Caliskan et al. (2017) introduced the WEAT and showed that word embeddings contain human biases. For example, African-American names are more closely associated with unpleasant terms than pleasant terms compared to European-American names.

Bolukbasi et al. (2016), Zhao et al. (2018b) and Kaneko and Bollegala (2019) propose methods to debias embeddings, reducing the bias according to their definitions. Zhang et al. (2018) use adversarial learning to debias word embeddings. Brunet et al. (2019) devise a method to identify resulting WEAT bias from each document in the training corpus. Zhao et al. (2018a) show that biases remain in coreference resolution, a downstream task. Zhao et al. (2019), Bordia and Bowman (2019) and May et al. (2019) show biases exist in contextualized word embeddings as well.

3 Background

3.1 The Word Embedding Association Test

The Word Embedding Association Test (WEAT) is the word embedding analog of Implicit Association Test (Greenwald et al. (1998)). It measures biases in word embeddings.

The measure only depends on two sets of target words X, Y of equal size (e.g. male and female names) and two sets of attribute words A, B (e.g. career vs family).

The similarity between two words a and b is measured by the cosine similarity of their vectors, cos(a, b). The difference in association of word w with the attribute sets A and B is defined as follows:

$$\begin{aligned} s(w, A, B) = mean_{a \in A} cos(w, a) - \\ mean_{b \in B} cos(w, b) \end{aligned} \tag{1}$$

The effect size of the bias is defined by the normalized measure as follows:

$$\frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std-dev_{w \in X \cup Y} s(w, A, B)} \tag{2}$$

Where $mean$ and $std - dev$ refer to mean and standard deviation, respectively.

4 Methodology

The first step is to train a word embedding model which has the bias we want to remove. The second step is to assign each word in the corpus a score on the attribute dimension (e.g., pleasant vs unpleasant or career vs family). We use DENSIFIER (introduced in Rothe et al. (2016)) for this.[1] It uses an orthogonal transformation of the embedding space to generate an ultradense embedding for each word. This method achieves state-of-the-art performance on assigning polarity scores to each word in the model on any dimension (e.g., positive vs negative, career vs family, extreme female vs extreme male occupations) using few seed words ((Hamilton et al., 2016)). Seed words, in our case, are two sets of attribute words (e.g., for the career vs family bias; our seed words are - Family words: home, parents, children, family, cousins, marriage, wedding, relatives. Career words: executive, management, professional, corporation, salary, office, business, career). We standardize the polarity scores to have zero mean and unit variance. Table 1 provides an example from one of the Word2Vec models using Wiki corpus of the 20 words with the highest polarity in the two directions of the career vs family dimension. A simple technique to assign polarity, like the mean of cosine similarities with the seed words, also works.

We then consider the N nearest neighbors (N) using cosine similarity of target words (e.g., for the career vs family bias; our target words are - Female names: amy, joan, lisa, sarah, diana, kate, ann, donna. Male names: john, paul, mike, kevin, steve, greg, jeff, bill) that have $polarity \notin (-1, 1)$.[2]

[1] We use code from Hamilton et al. (2016).

[2] This is to focus on words that have high polarity in either direction, other cutoffs also work well.

Highest family assn. words	**Polarity**	**Highest career assn. words**	**Polarity**
parents	-4.50	management	3.70
cousins	-4.38	business	3.64
relatives	-4.27	savills	3.64
marriage	-4.24	ict	3.63
wedding	-4.21	nsf	3.59
family	-4.17	executive	3.58
nieces	-4.15	nse	3.57
children	-4.11	consultancy	3.56
great-grandparents	-4.05	audit	3.52
aunts	-4.04	office	3.51
daughters	-3.93	salary	3.48
uncles	-3.91	dikman	3.45
sister-in-law	-3.86	nec	3.44
mother	-3.76	marketing	3.44
marrying	-3.69	professional	3.43
mothers	-3.68	rds	3.38
marry	-3.64	industry	3.36
daughter	-3.63	corporation	3.33
great-grandmother	-3.61	easa	3.31
stepchildren	-3.54	accountants	3.29

Table 1: Words with the highest polarity in the two directions of the career vs family dimension using DENSIFIER

Table 2 shows the top 20 words, which are the highest on the family (career) direction and are nearest neighbors of female (male) target words but not present in the attribute words from the same model as in Table 1.

The final step is to remove (duplicate) the sentences which increase (decrease) the unwanted (wanted) associations. For example, in the career vs family bias, we want to reduce the association between men and science and women and family. On the other hand, we want to increase the association between women and science and men and family. Duplicate is to have the same sentence twice in the corpus. For each target word, we remove (duplicate) the sentences where the target word and any of its nearest neighbours (among N) with high polarity and unwanted (wanted) associations occur together. This step is crucial as it removes the indirect bias (Gonen and Goldberg (2019)) and removes the bias entirely, which other methods are not able to do. We also remove/duplicate sentences where target words and attribute words appear together (direct bias). Finally, we re-train the model with the new corpus and choose N until the bias is zero.

N is the only parameter that needs to be chosen in the methodology until the bias reduces to the

Family	**Career**
nieces	nsf
aunts	budget
daughters	cfo
uncles	yost
sister-in-law	fiscal
mother	expenditures
marry	co-president
daughter	buyout
great-grandmother	five-year
thumbelina	exxon
tsarina	jackpot
wife	fema
stepmother	chicago-based
kayise	jic
perses	underwriting
half-brothers	american-based
yalti	northrop
princesses	iacocca
clytemnestra	statistician
betrothed	shutdown

Table 2: Words with the highest polarity on the family (career) direction and associated with female (male) target words but not present in the attribute words from the same model as in Table 1.

desired level. The higher the value of N, the more is the reduction in the bias. Only removal of sentences[3] without duplicating works as well, but the value of N is higher.

A sentence similarity methodology (Arora et al. (2017)) could be used to rank the sentences on the bias they create. The similarity between the sentence vector (calculated using the SIF method proposed by Arora et al. (2017), which simply removes the first principal component of the weighted average of the word vectors) and an average of bias creating N nearest neighbor word vectors of the target words will give the value of bias for each sentence. This could be used to exclude (duplicate) only the most bias increasing (decreasing) documents. It could be optimized to make minimum changes to the corpus and further control the reduction in bias along with N. We do not use sentence similarity for this paper to keep the methodology simpler.[4]

We will show in Section 6 that methodology does not affect the performance of word embedding models in capturing semantic information. The method is generalizable to any bias or word embedding.

5 Experimental Setup

5.1 Word Embeddings and Datasets

In our experiments we use Word2Vec (Mikolov et al. (2013a)) and GloVe (Pennington et al. (2014)) word embeddings and Simple Wikipedia (`https://dumps.wikimedia.org/`) and American National Corpus (`https://www.anc.org/data/oanc/`) corpora.[5] We use skip-gram architecture for Word2Vec and 300-dimensional word vectors for both embeddings. We pre-process both corpora to lowercase tokens, remove words with digits or punctuation and keep words that occur a minimum of 10 and 100 times for ANC and Wikipedia, respectively. We train both embeddings for ten iterations.

5.2 Multiple Repetitions

Due to the stochastic nature of the optimization used in training word embeddings, the cosine similarities between small sets of words can differ in two different embeddings trained on the same corpus (Antoniak and Mimno (2018) and Brunet et al. (2019)). Hence, we repeat each experiment 30 times and get a distribution of the WEAT of the original model, WEAT of the debiased model, benchmark performances before and after debiasing. This allows us to perform hypothesis testing.

5.3 WEAT

We use male and female names as target words along with career and family attributes (WEAT 6 from Caliskan et al. (2017)) for gender bias. The target and attribute sets are from Caliskan et al. (2017).

We use European-American and African-American names as target words along with pleasant and unpleasant attributes (WEATs 3-5 from Caliskan et al. (2017)) for racial bias. We randomly select 25 target names and attribute words from the combined sets of WEATs 3-5 from Caliskan et al. (2017) every time a word embedding is trained.[6] Table 3 shows the experiments performed. WEAT effect size for pleasant vs unpleasant (European-American vs African-American names) is statistically lesser than 0 for the ANC corpus; hence we exclude it from the list of experiments.

5.4 Word Embedding Benchmarks

It is essential that a debiasing method removes only biases, but its performance in capturing the semantic information does not decrease. We could test if debiased embeddings' performance is similar to the original embeddings on a variety of similarity, analogy and categorisation tasks (Levy et al. (2015) and Jastrzebski et al. (2017)). But, since we change corpus parts pertaining only to a particular bias, performance on the benchmarks datasets, which cover a wide range of subtasks like animals, gender, cities etc., could be misleading.

For example, in the career vs family WEAT, only the parts of the corpus containing the gender-specific words are altered, so the performance in capturing semantic information regarding gender could have decreased. To capture this, we combine the gender-specific subtask of the analogy

[3] We say just removal in the other sections of the paper for simplicity.

[4] Code and data used in the paper will be made available at our GitHub repository.

[5] Word2Vec and GloVe are trained using codes from `https://github.com/RaRe-Technologies/gensim` and `https://github.com/maciejkula/glove-python`, respectively.

[6] The complete sets of target and attribute words for both biases are provided in the Appendix A.

(1)	(2)	(3)	(4)	(5)
Data	**Model**	**Target Words**	**Attribute Words**	**Experiment**
Wiki	Word2Vec	European-American vs. African-American names	Pleasant vs. unpleasant	E1
Wiki	GloVe	European-American vs. African-American names	Pleasant vs. unpleasant	E2
ANC	Word2Vec	Male vs. female names	Career vs. family	E3
ANC	GloVe	Male vs. female names	Career vs. family	E4
Wiki	Word2Vec	Male vs. female names	Career vs. family	E5
Wiki	GloVe	Male vs. female names	Career vs. family	E6

Table 3: Experiments performed

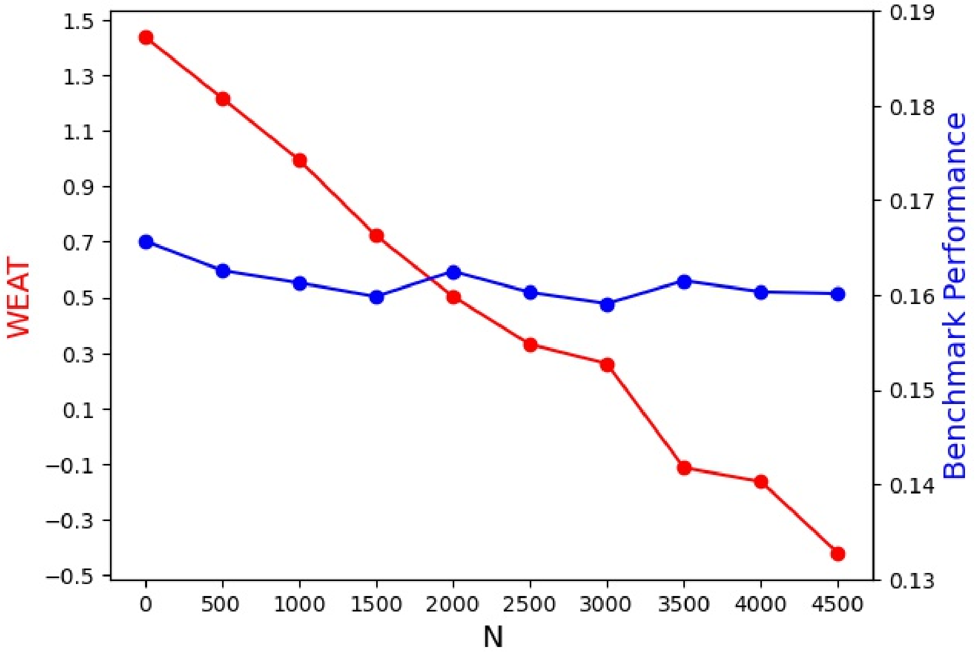

Figure 1: WEAT and benchmark performance plotted with N for Experiment E5 from Table 3.

task from WordRep (Gao et al. (2014)) and Google (Mikolov et al. (2013b)). These are the only two from Jastrzebski et al. (2017) benchmarks that have gender-specific subtasks. None of the tests has race-specific subtasks; hence no benchmark comparison is made for Experiments E1 and E2 (Table 3).

6 Results

In Figure 1, we plot how the mean WEAT effect size (left Y-axis) and benchmark performance (right Y-axis) change as the nearest neighbors considered (N) for Experiment E5 (Table 3) increases. We choose Experiment E5 for the plot as the original effect size is high, and the N to get it to zero is 3500. The WEAT decreases as N increases on the left Y-axis, whereas benchmark performance does not decrease (right Y-axis). We see that WEAT decreases from 1.84 (Table 4, column 2, E5) to 1.44 when $N = 0$. This is due to the removal of direct bias since only sentences where target words and attribute words appear together are added/removed when $N = 0$. As N increases, indirect bias keeps reducing. For N above 3500, WEAT is substantially lesser than 0.

In Table 4, Column 4, we show that for the reported N (nearest neighbors), debiased WEAT is significantly less than zero (p < 0.01) for all experiments where original WEAT is significantly greater than 0. In comparison, WEAT remains high and statistically significant after the existing debiasing methods (please refer to Section 4 in Gonen and Goldberg (2019) and Table 4 in Kaneko and Bollegala (2021)). We report means of original and debiased WEAT (Columns 2 and 3).

In Table 5 (same N as Table 4), Column 4, we show that debiased benchmark performance is not significantly less than original benchmark performance (p < 0.01). The difference in performance

(1)	(2)	(3)	(4)	(5)
Experiment	**WEAT Org.**	**WEAT Debiased**	**Pr(WEAT Debiased < 0)**	**N**
E1	0.285	-0.684	0.0000*	0
E2	0.468	-0.078	0.0002*	800
E3	1.241	-0.172	0.0000*	1000
E4	0.335	-0.142	0.0000*	70
E5	1.835	-0.114	0.0086*	3500
E6	1.705	-0.329	0.0000*	900

Table 4: WEAT of original and debiased models. * indicates statistical significance at p < 0.01 for p-values reported in Column 4. Experiments are from Table 3.

(1)	(2)	(3)	(4)	(5)	(6)
Experiment	**Prf. Debiased**	**Prf. Org. - Prf. Debiased**	**Pr(Prf. Org. > Prf. Debiased)**	**Corpus Org.**	**Corpus Debiased**
E1	-	-	-	1,430,872	1,430,854
E2	-	-	-	1,430,872	1,424,784
E3	0.145	0.0003	0.4091	1,092,231	1,090,979
E4	0.005	-0.0006	0.8952	1,092,231	1,091,086
E5	0.162	0.0030	0.0189	1,430,872	1,427,811
E6	0.064	-0.0006	0.9015	1,430,872	1,439,929

Table 5: Performance and corpus size of original and debiased models. * indicates statistical significance at p < 0.01 for p-values reported in Column 4. Experiments are from Table 3.

(Column 3) is minimal for all experiments.[7] We report the number of sentences in the original corpus and the mean number of sentences in the debiased corpus (Columns 5 and 6).

7 Conclusion

We proposed a generalizable and straightforward method that removes direct and indirect bias from word embeddings by removing the biased parts of the corpus. Multiple experimental results show that the proposed method removes the discriminatory biases entirely without affecting the performance of the models in capturing semantic information.

[7] Benchmark performance on the complete set of tests (Levy et al. (2015) and Jastrzebski et al. (2017)) is also not significantly different for all experiments; we do not report the numbers due to sake of brevity.

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

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. *Advances in neural information processing systems*, 29:4349–4357.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*, pages 803–811. PMLR.

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

Bin Gao, Jiang Bian, and Tie-Yan Liu. 2014. Wordrep: A benchmark for research on learning word representations. *arXiv preprint arXiv:1407.1640*.

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.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

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.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, volume 2016, page 595. NIH Public Access.

Stanisław Jastrzebski, Damian Leśniak, and Wojciech Marian Czarnecki. 2017. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. *arXiv preprint arXiv:1702.02170*.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv preprint arXiv:1906.00742*.

Masahiro Kaneko and Danushka Bollegala. 2021. Dictionary-based debiasing of pre-trained word embeddings. *arXiv preprint arXiv:2101.09525*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

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

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.

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.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

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, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

A Appendix

Career vs Family WEAT:
Male names: john, paul, mike, kevin, steve, greg, jeff, bill
Female names: amy, joan, lisa, sarah, diana, kate, ann, donna
Career: executive, management, professional, corporation, salary, office, business, career
Family: home, parents, children, family, cousins, marriage, wedding, relatives

Pleasant vs Unpleasant WEAT:
African-American names: alonzo, jamel, lerone, percell, theo, alphonse, jerome, leroy, rasaan, torrance, darnell, lamar, lionel, rashaun, tyree, deion, lamont, malik, terrence, tyrone, everol, lavon, marcellus, terryl, wardell, aiesha, lashelle, nichelle, shereen, temeka, ebony, latisha, shaniqua, tameisha, teretha, jasmine, latonya, shanise, tanisha, tia, lakisha, latoya, sharise, tashika, yolanda, lashandra, malika, shavonn, tawanda, yvette, darnell, hakim, jermaine, kareem, jamal, leroy, rasheed, tremayne, tyrone, aisha, ebony, keisha, kenya, latonya, lakisha, latoya, tamika, tanisha
European-American names: adam, chip, harry, josh, roger, alan, frank, ian, justin, ryan, andrew, fred, jack, matthew, stephen, brad, greg, jed, paul, todd, brandon, hank, jonathan, peter, wilbur, amanda, courtney, heather, melanie, sara, amber, crystal, katie, meredith, shannon, betsy, donna, kristin, nancy, stephanie, bobbie-sue, ellen, lauren, peggy, sue-ellen, colleen, emily, megan, rachel, wendy, brad, brendan, geoffrey, greg, brett, jay, matthew, neil, todd, allison, anne, carrie, emily, jill, laurie, kristen, meredith, sarah
Pleasant: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation, joy, love, peace, wonderful, pleasure, friend, laughter, happy
Unpleasant: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit, agony, terrible, horrible, nasty, evil, war, awful, failure