Probabilistic Bias Mitigation in Word Embeddings

Hailey Joren
Harvard College
Cambridge, MA
hjoren@ucsd.edu

David Alvarez-Melis
Massachusetts Institute of Technology
Cambridge, MA
dalvmel@mit.edu

Abstract

It has been shown that word embeddings derived from large corpora tend to incorporate biases present in their training data. Various methods for mitigating these biases have been proposed, but recent work has demonstrated that these methods hide but fail to truly remove the biases, which can still be observed in word nearest-neighbor statistics. In this work we propose a probabilistic view of word embedding bias. We leverage this framework to present a novel method for mitigating bias which relies on probabilistic observations to yield a more robust bias mitigation algorithm. We demonstrate that this method effectively reduces bias according to three separate measures of bias while maintaining embedding quality across various popular benchmark semantic tasks.

1 Introduction

Word embeddings, or vector representations of words, are an important component of Natural Language Processing (NLP) models and necessary for many downstream tasks. However, word embeddings, including embeddings commonly deployed for public use, have been shown to exhibit unwanted societal stereotypes and biases, raising concerns about disparate impact on axes of gender, race, ethnicity, and religion [1, 2]. The impact of this bias has manifested in a range of downstream tasks, ranging from autocomplete suggestions [3] to advertisement delivery [4], increasing the likelihood of amplifying harmful biases through the use of these models.

The most well-established method thus far for mitigating bias[1] relies on projecting target words[2] onto a bias subspace (such as a gender subspace) and subtracting out the difference between the resulting distances [1]. On the other hand, the most popular metric for measuring bias is the WEAT statistic [2], which compares the cosine similarities between groups of words. However, WEAT has been recently shown to overestimate bias as a result of implicitly relying on similar frequencies for the target words [5], and Gonen and Goldberg [6] demonstrated that evidence of bias can still be recovered after geometric bias mitigation by examining the neighborhood of a target word among socially-biased words.

In response to this, we propose an alternative framework for bias mitigation in word embeddings that approaches this problem from a probabilistic perspective. The motivation for this approach is two-fold. First, most popular word embedding algorithms are probabilistic at their core – i.e., they are trained (explicitly or implicitly [7]) to minimize some form of word co-occurrence probabilities. Thus, we argue that a framework for measuring and treating bias in these embeddings should take into account, in addition to their geometric aspect, their probabilistic nature too. On the other hand,

[1] We intentionally do not reference the resulting embeddings as "debiased" or free from all gender bias, and prefer the term "mitigating bias" rather that "debiasing," to guard against the misconception that the resulting embeddings are entirely "safe" and need not be critically evaluated for bias in downstream tasks.

[2] Throughout this paper, we use *word* interchangeably with the vector representing the word in an embedding.

Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. (non-archival)

the issue of bias has also been approached (albeit in different contexts) in the fairness literature, where various intuitive notions of equity such as equalized odds have been formalized through probabilistic criteria. By considering analogous criteria for the word embedding setting, we seek to draw connections between these two bodies of work.

We present experiments on various bias mitigation benchmarks and show that our framework is comparable to state-of-the-art alternatives according to measures of geometric bias mitigation and that it performs far better according to measures of neighborhood bias. For fair comparison, we focus on mitigating a binary gender bias in pre-trained word embeddings using SGNS (skip-gram with negative-sampling), though we note that this framework and methods could be extended to other types of bias and word embedding algorithms.

2 Background

Geometric Bias Mitigation Geometric bias mitigation uses the cosine distances between words to both measure and remove gender bias [1]. This method implicitly defines bias as a geometric asymmetry between words when projected onto a subspace, such as the gender subspace constructed from a set of gender pairs such as $\mathcal{P} = \{(he, she), (man, woman), (king, queen)...\}$. The projection of a vector v onto B (the subspace) is defined by $v_B = \sum_{j=1}^{k}(v \cdot b_j)b_j$ where a subspace B is defined by k orthogonal unit vectors $B = b_1, ..., b_k$.

WEAT The WEAT statistic [2] demonstrates the presence of biases in word embeddings with an effect size defined as the mean test statistic across the two word sets:

$$\frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std\_dev_{w \in X \cup Y} s(w, A, B)} \tag{1}$$

Where s, the test statistic, is defined as: $s(w, A, B) = mean_{a \in A} cos(w, a) - mean_{b \in B} cos(w, a)$, and X,Y,A, and B are groups of words for which the association is measured. Possible values range from -2 to 2 depending on the association of the words groups, and a value of zero indicates X and Y are equally associated with A and B. See Ethayarajh et al. [5] for further details on WEAT.

RIPA The RIPA (relational inner product association) metric was developed as an alternative to WEAT, with the critique that WEAT is likely to overestimate the bias of a target attribute [5]. The RIPA metric formalizes the measure of bias used in geometric bias mitigation as the inner product association of a word vector v with respect to a relation vector b. The relation vector is constructed from the first principal component of the differences between gender word pairs. We report the absolute value of the RIPA metric as the value can be positive or negative according to the direction of the bias. A value of zero indicates a lack of bias, and the value is bound by $[-||w||, ||w||]$.

Neighborhood Metric The neighborhood bias metric proposed by Gonen and Goldberg [6] quantifies bias as the proportion of male socially-biased words among the k nearest socially-biased male and female neighboring words, whereby biased words are obtained by projecting neutral words onto a gender relation vector. As we only examine the target word among the 1000 most socially-biased words in the vocabulary (500 male and 500 female), a word's bias is measured as the ratio of its neighborhood of socially-biased male and socially-biased female words, so that a value of 0.5 in this metric would indicate a perfectly unbiased word, and values closer to 0 and 1 indicate stronger bias.

3 A Probabilistic Framework for Bias Mitigation

Our objective here is to extend and complement the geometric notions of word embedding bias described in the previous section with an alternative, probabilistic, approach. Intuitively, we seek a notion of *equality* akin to that of *demographic parity* in the fairness literature, which requires that a decision or outcome be independent of a protected attribute such as gender. [8]. Similarly, when considering a probabilistic definition of unbiased in word embeddings, we can consider the conditional probabilities of word pairs, ensuring for example that $p(doctor|man) \approx p(doctor|woman)$, and can extend this probabilistic framework to include the neighborhood of a target word, addressing the potential pitfalls of geometric bias mitigation.

Conveniently, most word embedding frameworks allow for immediate computation of the conditional probabilities $P(w|c)$. Here, we focus our attention on the Skip-Gram method with Negative Sampling (SGNS) of Mikolov et al. [9], although our framework can be equivalently instantiated for most other popular embedding methods, owing to their core similarities [7, 10]. Leveraging this probabilistic nature, we construct a bias mitigation method in two steps, and examine each step as an independent method as well as the resulting composite method.

Probabilistic Bias Mitigation This component of our bias mitigation framework seeks to enforce that the probability of prediction or outcome cannot depend on a protected class such as gender. We can formalize this intuitive goal through a loss function that penalizes the discrepancy between the conditional probabilities of a *target* word (i.e., one that should *not* be affected by the protected attribute) conditioned on two words describing the protected attribute (e.g., *man* and *woman* in the case of gender). That is, for every target word we seek to minimize:

$$loss = \sum_{a,b \in \mathcal{P}} p(target|a) - p(target|b) \tag{2}$$

where $\mathcal{P} = \{(he, she), (man, woman), (king, queen), \dots\}$ is a set of word pairs characterizing the protected attribute, akin to that used in previous work [1].

At this point, the specific form of the objective will depend on the type of word embeddings used. For our expample of SGNS, recall that this algorithm models the conditional probability of a target word given a context word as a function of the inner product of their representations. Though an exact method for calculating the conditional probability includes summing over conditional probability of all the words in the vocabulary, we can use the estimation of log conditional probability proposed by Mikolov et al. [9], i.e., $\log p(w_O|w_I) \approx \log \sigma({v'_{wo}}^T v_{wI}) + \sum_{i=1}^{k} [\log \sigma(-{v'_{wi}}^T v_{wI})]$.

Nearest Neighbor Bias Mitigation Based on observations by Gonen and Goldberg [6], we extend our method to consider the composition of the neighborhood of socially-gendered words of a target word. We note that bias in a word embedding depends not only on the relationship between a target word and explicitly gendered words like *man* and *woman*, but also between a target word and socially-biased male or female words. Bolukbasi et al [1] proposed a method for eliminating this kind of indirect bias through geometric bias mitigation, but it is shown to be ineffective by the neighborhood metric [6].

Instead, we extend our method of bias mitigation to account for this neighborhood effect. Specifically, we examine the conditional probabilities of a target word given the $k/2$ nearest neighbors from the male socially-biased words as well as given the $k/2$ female socially-biased words (in sorted order, from smallest to largest). The groups of socially-biased words are constructed as described in the neighborhood metric. If the word is unbiased according to the neighborhood metric, these probabilities should be comparable. We then use the following as our loss function:

$$loss = \sum_{i=0}^{k/2} p(target|m_i) - p(target|f_i), \tag{3}$$

where m and f represent the male and female neighbors sorted by distance to the target word t (we use $L1$ distance).

4 Experiments

We evaluate our framework on fastText embeddings trained on Wikipedia (2017), UMBC webbase corpus and statmt.org news dataset (16B tokens) [12]. For simplicity, only the first 22000 words are used in all embeddings, though preliminary results indicate the findings extend to the full corpus. For our novel methods of mitigating bias, a shallow neural network is used to adjust the embedding. The single layer of the model is an embedding layer with weights initialized to those of the original embedding. For the composite method, these weights are initialized to those of the embedding after probabilistic bias mitigation. A batch of word indices is fed into the model, which are then embedded and for which a loss value is calculated, allowing back-propagation to adjust the embeddings. For each of the models, a fixed number of iterations is used to prevent overfitting, which can eventually

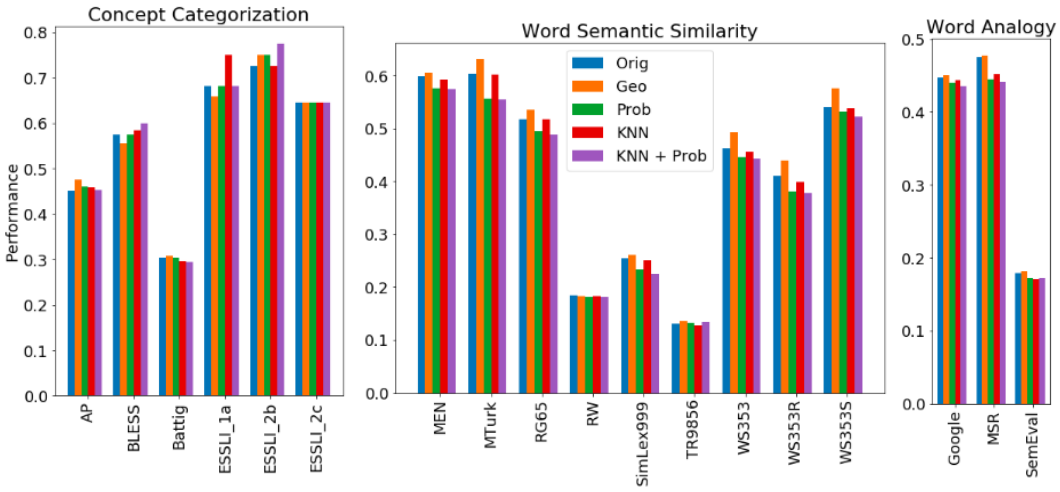

Figure 1: Word embedding semantic quality benchmarks for each bias mitigation method (higher is better). See Jastrzkebski et al. [11] for details of each metric.

	RIPA	Neighborhood
	$\lvert mean \rvert$	$\lvert .5 - mean \rvert$
Original	2.895	0.323
Geometric	0.096	0.328
Simple Probabilistic	0.320	0.250
Nearest Neighbor	1.705	0.083
Composite NN + Prob	0.372	0.034

Table 1: Remaining Bias (as measured by RIPA and Neighborhood metrics) in fastText embeddings for baseline (top two rows) and our (bottom three) methods.

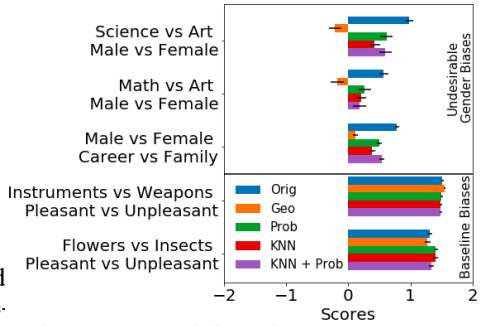

Figure 2: Remaining Bias (WEAT score)

hurt performance on the embedding benchmarks (See Figure 1). We evaluated the embedding after 1000 iterations, and stopped training if performance on a benchmark decreased significantly.

We construct a list of candidate words to debias, taken from the words used in the WEAT gender bias statistics. Words in this list should be gender neutral, and are related to the topics of career, arts, science, math, family and professions (see appendix). We note that this list can easily be expanded to include a greater proportion of words in the corpus. For example, Ethayarajh et al. [5] suggested a method for identifying inappropriately gendered words using unsupervised learning.

We compare this method of bias mitigation with the no bias mitigation ("Orig"), geometric bias mitigation ("Geo"), the two pieces of our method alone ("Prob" and "KNN") and the composite method ("KNN+Prob"). We note that the composite method performs reasonably well according the the RIPA metric, and much better than traditional geometric bias mitigation according to the neighborhood metric, without significant performance loss according to the accepted benchmarks. To our knowledge this is the first bias mitigation method to perform reasonably both on both metrics.

5 Discussion

We proposed a simple method of bias mitigation based on this probabilistic notions of fairness, and showed that it leads to promising results in various benchmark bias mitigation tasks. Future work should include considering a more rigorous definition and non-binary of bias and experimenting with various embedding algorithms and network architectures.

Acknowledgements

The authors would like to thank Tommi Jaakkola for stimulating discussions during the initial stages of this work.

References

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016. URL http://arxiv.org/abs/1607.06520.

[2] Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187, 2016. URL http://arxiv.org/abs/1608.07187.

[3] Issie Lapowsky. Google autocomplete still makes vile suggestions, 2018. URL "https://www.wired.com/story/google-autocomplete-vile-suggestions/".

[4] Catherine E. Tucker Anja Lambrecht. Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. *SSRN*, 2016. doi: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260.

[5] Kawin Ethayarajh, David Kristjanson Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations. In *ACL*, 2019.

[6] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *CoRR*, abs/1903.03862, 2019. URL http://arxiv.org/abs/1903.03862.

[7] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[8] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016. URL http://arxiv.org/abs/1610.02413.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. URL http://arxiv.org/abs/1310.4546.

[10] Tatsunori B Hashimoto, David Alvarez-Melis, and Tommi S Jaakkola. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286, 2016.

[11] Stanislaw Jastrzkebski, Damian Lesniak, and Wojciech Marian Czarnecki. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. *CoRR*, abs/1702.02170, 2017. URL http://arxiv.org/abs/1702.02170.

[12] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. *CoRR*, abs/1712.09405, 2017. URL http://arxiv.org/abs/1712.09405.

A Experiment Notes

For Equation 4, as described in the original work, in regards to the k sample words w_i is drawn from the corpus using the Unigram distribution raised to the 3/4 power.

For reference, the most male socially-biased words include words such as:'john', 'jr', 'mlb', 'dick', 'nfl', 'cfl', 'sgt', 'abbot', 'halfback', 'jock', 'mike', 'joseph',while the most female socially-biased words include words such as:'feminine', 'marital', 'tatiana', 'pregnancy', 'eva', 'pageant', 'distress', 'cristina', 'ida', 'beauty', 'sexuality','fertility'

B Professions

'accountant', 'acquaintance', 'actor', 'actress', 'administrator', 'adventurer', 'advocate', 'aide', 'alderman', 'ambassador', 'analyst', 'anthropologist', 'archaeologist', 'archbishop', 'architect', 'artist', 'assassin', 'astronaut', 'astronomer', 'athlete', 'attorney', 'author', 'baker', 'banker', 'barber', 'baron', 'barrister', 'bartender', 'biologist', 'bishop', 'bodyguard', 'boss', 'boxer', 'broadcaster', 'broker', 'businessman', 'butcher', 'butler', 'captain', 'caretaker', 'carpenter', 'cartoonist', 'cellist', 'chancellor', 'chaplain', 'character', 'chef', 'chemist', 'choreographer', 'cinematographer', 'citizen', 'cleric', 'clerk', 'coach', 'collector', 'colonel', 'columnist', 'comedian', 'comic', 'commander', 'commentator', 'commissioner', 'composer', 'conductor', 'confesses', 'congressman', 'constable', 'consultant', 'cop', 'correspondent', 'counselor', 'critic', 'crusader', 'curator', 'dad', 'dancer', 'dean', 'dentist', 'deputy', 'detective', 'diplomat', 'director', 'doctor', 'drummer', 'economist', 'editor', 'educator', 'employee', 'entertainer', 'entrepreneur', 'envoy', 'evangelist', 'farmer', 'filmmaker', 'financier', 'fisherman', 'footballer', 'foreman', 'gangster', 'gardener', 'geologist', 'goalkeeper', 'guitarist', 'headmaster', 'historian', 'hooker', 'illustrator', 'industrialist', 'inspector', 'instructor', 'inventor', 'investigator', 'journalist', 'judge', 'jurist', 'landlord', 'lawyer', 'lecturer', 'legislator', 'librarian', 'lieutenant', 'lyricist', 'maestro', 'magician', 'magistrate', 'maid', 'manager', 'marshal', 'mathematician', 'mechanic', 'midfielder', 'minister', 'missionary', 'monk', 'musician', 'nanny', 'narrator', 'naturalist', 'novelist', 'nun', 'nurse', 'observer', 'officer', 'organist', 'painter', 'pastor', 'performer', 'philanthropist', 'philosopher', 'photographer', 'physician', 'physicist', 'pianist', 'planner', 'playwright', 'poet', 'policeman', 'politician', 'preacher', 'president', 'priest', 'principal', 'prisoner', 'professor', 'programmer', 'promoter', 'proprietor', 'prosecutor', 'protagonist', 'provost', 'psychiatrist', 'psychologist', 'rabbi', 'ranger', 'researcher', 'sailor', 'saint', 'salesman', 'saxophonist', 'scholar', 'scientist', 'screenwriter', 'sculptor', 'secretary', 'senator', 'sergeant', 'servant', 'singer', 'skipper', 'sociologist', 'soldier', 'solicitor', 'soloist', 'sportsman', 'statesman', 'steward', 'student', 'substitute', 'superintendent', 'surgeon', 'surveyor', 'swimmer', 'teacher', 'technician', 'teenager', 'therapist', 'trader', 'treasurer', 'trooper', 'trumpeter', 'tutor', 'tycoon', 'violinist', 'vocalist', 'waiter', 'waitress', 'warden', 'warrior', 'worker', 'wrestler', 'writer'

C WEAT Word Sets

Words used for WEAT statistic, consisting of baseline bias tests and gender bias tests in the format **X vs Y / A vs B**

Flowers vs Insects / Pleasant vs Unpleasant

X: "aster", "clover", "hyacinth", "marigold", "poppy", "azalea", "crocus", "iris", "orchid", "rose", "bluebell", "daffodil", "lilac", "pansy", "tulip", "buttercup", "daisy", "lily", "peony", "violet", "carnation", "gladiola", "magnolia", "petunia", "zinnia"

Y: "ant", "caterpillar", "flea", "locust", "spider", "bedbug", "centipede", "fly", "maggot", "tarantula", "bee", "cockroach", "gnat", "mosquito", "termite", "beetle", "cricket", "hornet", "moth", "wasp", "blackfly", "dragonfly", "horsefly", "roach", "weevil"

A: "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"

B: "abuse", "crash", "filth", "murder", "sickness", "accident", "death", "grief", "poison", "stink", "assault", "disaster", "hatred", "pollute", "tragedy", "divorce", "jail", "poverty", "ugly", "cancer", "kill", "rotten", "vomit", "agony", "prison"

Instruments vs Weapons / Pleasant vs Unpleasant:

X: "bagpipe", "cello", "guitar", "lute", "trombone", "banjo", "clarinet", "harmonica", "mandolin", "trumpet", "bassoon", "drum", "harp", "oboe", "tuba", "bell", "fiddle", "harpsichord", "piano", "viola", "bongo", "flute", "horn", "saxophone", "violin"

Y: "arrow", "club", "gun", "missile", "spear", "ax", "dagger", "harpoon", "pistol", "sword", "blade", "dynamite", "hatchet", "rifle", "tank", "bomb", "firearm", "knife", "shotgun", "teargas", "cannon", "grenade", "mace", "slingshot", "whip"

A: "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"

B: "abuse", "crash", "filth", "murder", "sickness", "accident", "death", "grief", "poison", "stink", "assault", "disaster", "hatred", "pollute", "tragedy", "divorce", "jail", "poverty", "ugly", "cancer", "kill", "rotten", "vomit", "agony", "prison"

Male vs Female / Career vs Family:

X: "brother", "father", "uncle", "grandfather", "son", "he", "his", "him", "man", "himself", "men", "husband", "boy", "uncle", "nephew", "boyfriend", "king", "actor"

Y: "sister", "mother", "aunt", "grandmother", "daughter", "she", "hers", "her", "woman", "herself", "women", "wife", "aunt", "niece", "girlfriend", "queen", "actress"

A: "executive", "management", "professional", "corporation", "salary", "office", "business", "career", "industry", "company", "promotion", "profession", "CEO", "manager", "coworker", "entrepreneur"

B: "home", "parents", "children", "family", "cousins", "marriage", "wedding", "relatives", "grandparents", "grandchildren", "nurture", "child", "toddler", "infant", "teenager"

Math vs Art / Male vs Female:

X: "math", "algebra", "geometry", "calculus", "equations", "computation", "numbers", "addition", "trigonometry", "arithmetic", "logic", "proofs", "multiplication", "mathematics"

Y: "poetry", "art", "Shakespeare", "dance", "literature", "novel", "symphony", "drama", "orchestra", "music", "ballet", "arts", "creative", "sculpture"

A: "brother", "father", "uncle", "grandfather", "son", "he", "his", "him", "man", "himself", "men", "husband", "boy", "uncle", "nephew", "boyfriend", "king", "actor"

B: "sister", "mother", "aunt", "grandmother", "daughter", "she", "hers", "her", "woman", "herself", "women", "wife", "aunt", "niece", "girlfriend", "queen", "actress"

Science vs Art / Male8 vs Female8:

X:"science", "technology", "physics", "chemistry", "Einstein", "NASA", "experiment", "astronomy", "biology", "aeronautics", "mechanics", "thermodynamics"

Y: "poetry", "art", "Shakespeare", "dance", "literature", "novel", "symphony", "drama", "orchestra", "music", "ballet", "arts", "creative", "sculpture"

A: "brother", "father", "uncle", "grandfather", "son", "he", "his", "him", "man", "himself", "men", "husband", "boy", "uncle", "nephew", "boyfriend"

B: "sister", "mother", "aunt", "grandmother", "daughter", "she", "hers", "her", "woman", "herself", "women", "wife", "aunt", "niece", "girlfriend"