Whose Truth is the “Ground Truth”? College Admissions Essays and Bias in Word Vector Evaluation Methods

Noah Arthurs
Stanford University
narthurs@cs.stanford.edu

AJ Alvero
Stanford University
ajalvero@stanford.edu

ABSTRACT

Word vectors are widely used as input features in natural language processing (NLP) tasks. Researchers have found that word vectors often encode the biases of society, and steps have been taken towards debiasing the vectors themselves. However, little has been said about the fairness of the methods used to evaluate the quality of vectors. Analogical and word similarity tasks are commonplace, but both rely on purportedly ground truth statements about the semantic relationships between words (e.g. “man is to woman as king is to queen”). These analogies look reasonable when only taking into account the literal meanings of words, but two issues arise: (1) people don’t always use words in a literal sense, and (2) the same word may be used differently by different groups of people. In this paper, we split a dataset of over 800,000 college admissions essays into quartiles based on reported household income (RHI) and train sets of word vectors on each quartile. We then test these sets of vectors on common intrinsic evaluation tasks. We find that vectors trained on the essays of higher income students encode more of each task’s target semantic relationships than vectors trained on the essays of lower income students. These results hold even when controlling for word frequency. We conclude that the tasks themselves are biased towards the writing of higher income students, and we challenge the notion that there exist ground truth semantic relationships that word vectors must encode in order to be useful.

1. INTRODUCTION

Text analysis has grown into an important topic, with researchers from education, industry, social sciences, humanities, and traditional STEM programs harnessing large amounts of textual data that is widely available and relatively easy to access. Text data is usually very sparse (as most words do not appear in most documents) and difficult to use as input for mathematical models. This has given rise to a variety of vectorization methods that include simple word counting, statistical methods like TF-IDF, and neural methods used to generate dense representations called word vectors. Word vectors have been shown to produce high quality results in a variety of machine learning (ML) and natural language processing (NLP) tasks, but this potentially comes with a social cost. Research has shown that word vectors propagate the gender and racial biases found in society [19, 7, 10]. However, little has been said about the fairness of the methods that we use to evaluate vectors.

After a set of word vectors has been trained on a corpus, researchers and engineers want to evaluate the quality of the vectors. As a result, a standard set of word vector evaluation tasks [41] has been developed in order to measure how useful and generalizable a given set of vectors is. When researchers propose new methods for training word vectors, they demonstrate the performance of their methods by evaluating the resulting vectors on these tasks. Furthermore, when NLP systems are built, vectors that perform well on these tasks are most likely to be chosen.

Word vector evaluation tasks are either intrinsic (performed directly on the vectors) or extrinsic (performed by using the vectors as inputs for a downstream task). Intrinsic evaluation is popular because it is very inexpensive, but it relies on having some secondary notion of what makes vectors useful. The most popular methods assume that there are “ground truth” semantic relationships that a set of word vectors must encode in order to be useful. However, due to sociolinguistic variation, not all language communities share the same semantic relationships [4]. As a result, in order for these tasks to be fair, they need to use semantic relationships that are universal: if semantic relationship R holds in the language patterns of group G but not group H, then the usage of R in an intrinsic evaluation task will bias researchers towards sets of vectors that model group G’s language usage better than group H’s. We use this framework to evaluate the fairness of two popular forms of intrinsic evaluation, analogical tasks and word similarity tasks.

When working with large text corpora, especially in educational contexts, it is important to consider the role of sociolinguistic variation [27]. In particular, students have been punished and targeted for their language practices if they are perceived to be different from the “mainstream” [38, 42]. Understanding how social variation in language affects word vectors is necessary in order to tackle two critical issues. The first is the question, “whose language is being modeled?” Word vectors are meant to capture something about the semantics of each word. If theories of sociolinguistic variation

Noah Arthurs and Aj Alvero "Whose Truth is the "Ground Truth"? College Admissions Essays and Bias in Word Vector Evaluation Methods" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020),* Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 342 - 349

tell us that people from different groups use language in different ways, then we must wonder if standard word vector sets like GloVe [35] are serving everyone equally. Second, we must ask, "how does fairness change across contexts?" In NLP, word vectors that perform well on intrinsic evaluations are used across many different contexts. However, if fairness involves taking sociolinguistic variation into account, it may not be the case that vectors that are unbiased in one context are biased in another.

Educational agencies and institutions are also increasingly relying on algorithms to help with decision-making processes. College admissions offices have been pushed to use AI [3] but have legal and ethical mandates to ensure process fairness for applicants based on their demographics and/or protected statuses, like race, gender, and religion. As the number college applications rise and the need to hire reviewers increases, applicant admissions essays are a likely candidate for some form of automation. Research on the essays encode some degree of applicant gender and social class [2], making careful adoption of AI necessary. If sociolinguistic variation is not taken into account, algorithms have a high chance of reproducing social inequalities.

We address these issues by analyzing a corpus of over 800,000 college admissions essays (CAE) submitted to a selective, multi-campus university system. In addition to the essays, we have a variety of author metadata, including each student's reported household income (RHI). We split the dataset into quartiles by RHI and train one set of word vectors from scratch on each quartile. After training, we find that on both the analogy and similarity tasks, the vectors trained on the writing of higher income students encode more of the target semantic relationships than vectors trained on the writing of lower income students. This indicates that the tasks can be biased against the writing of lower income students.

Our contributions are:

- to challenge the paradigm of "ground truth" labels for intrinsic evaluation by starting with the premise that language distributions vary along demographic characteristics.
- to provide a method for auditing the fairness/bias of an evaluation task, complementing existing methods for auditing the fairness/bias of word vectors themselves.
- to contribute to the educational scholarship of higher education by characterizing sociolinguistic variation in college admissions essays using established AI techniques.

2. BACKGROUND

2.1 Word Vectors

In NLP, word vectors (or word embeddings) are the standard way translate words into input features for machine learning models. Popular word vector training algorithms like word2vec [30] and GloVe [35] are based on the distributional hypothesis, the idea that a word's meaning is encoded in its co-occurrences with other words. In particular, word2vec tries to learn features which can be used to predict a word from its context (or vice versa), and GloVe trains directly from a co-occurrence matrix. Both models take in large corpora of texts and create a dense representation of every word, usually in 100- to 300-dimensional space.

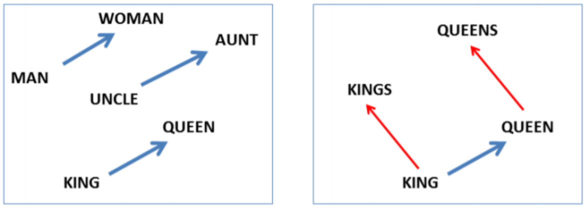

Figure 1: Illustration of simple vector operations modeling semantic (left) and syntactic (right) relationships in vector space from [32]

2.2 Word Vector Evaluation

As described above, vectors can be evaluated intrinsically or extrinsically. Intrinsic evaluation, which is the focus of this study, involves directly examining the relationships between vectors. Intrinsic evaluation has the advantage of being much faster and more lightweight, but it comes with two downsides. The first is that intrinsic tasks do not resemble the use cases of word vectors as much as extrinsic tasks do. The second is that intrinsic tasks rely on "ground truth" human judgments about what the relationships between vectors *should* be.

The word analogy task is based on the idea that analogical relationships between words (e.g. "man is to woman as king is to queen") should be encoded in word vectors as parallelograms (i.e. the vector that connects "man" to "woman" would be the same as the one that connects "king" to "queen"). Mathematically, this means that:

$$v_{\text{queen}} - v_{\text{woman}} \approx v_{\text{king}} - v_{\text{man}}$$

This kind of relationship has been found to hold for both for semantic (meaning-based) and syntactic (grammar-based) relationships (left and right sides of figure 1). The idea for the analogy task dates back to the 1990's [17], but it was not proposed as a word vector evaluation technique until 2013 [31, 30]. Since then, it is common practice to compare sets of vectors on their ability to "solve" word analogies.

Word similarity is based on the idea that similar words (i.e. words that are used in similar contexts) should have similar word vectors. The word similarity task starts with a list of word pairs and involves finding the correlation between ground truth similarities between the words in each pair and the similarities between their corresponding vectors. The similarity between two vectors is in practice measured by taking their cosine similarity. The cosine similarity of two vectors $\vec{a}$ and $\vec{b}$ is defined as:

$$\frac{\vec{a} \cdot \vec{b}}{||\vec{a}||_2 ||\vec{b}||_2}$$

which is equal to 0 if $\vec{a}$ and $\vec{b}$ are orthogonal, 1 if they are in the same direction, or (most of the time) something in between. The ground truth similarities, on the other hand, rely on human judgment. This task remains largely unchanged since its first iteration in 1965, when Rubenstein and Goodenough [39] set out to test the distributional hypothesis. The big difference is that modern datasets for this task starting in 2002 with WS-353 [18] involve larger numbers of word pairs.

Both of these tasks require "ground truth" labels of some sort. The analogy task requires a list of analogies that the vectors are being tested for, while the similarity task requires ratings of the similarities between many pairs of words. These labels are problematic for two reasons. First, it has been pointed out that the labels for these tasks do not take into account the fact that words can take on many different meanings depending on context (polysemy). Second, word use and semantic intent vary along social dimensions, meaning that labels may reflect the language use of some groups better than others, thus creating bias. This second issue is the focus of our study.

2.3 Word Vector Critiques

It has been found that word vectors encode the biases present in their training data [7], and word vectors have been used to quantify the biases that exists in society [19]. Two methods have emerged for reducing bias in word embeddings: we can change our training process in order to penalize biased vectors [7], or we can identify and remove the training documents that are the source of the most bias [8].

Intrinsic evaluation methods have also fallen under scrutiny. Both the analogy task [14, 37] and the similarity task [16] have been criticized for relying on the fuzzy relationship between word similarity and vector similarity, and for not taking polysemy into account. Lastly, both tasks have been found to be poor predictors of extrinsic performance [11].

Although there have been numerous critiques of the bias encoded in word vectors and numerous critiques of intrinsic evaluation tasks, little has been said about whether or not intrinsic evaluation is biased in theory or practice. This study answers that question by identifying whether the "ground truth" semantic relationships prescribed by intrinsic evaluation tasks are shared by students of all income levels.

2.4 Sociolinguistic Variation

Language variation across spatial, demographic, and temporal dimensions is the bedrock theory behind sociolinguistics. Applied research in sociolinguistics often seeks to ameliorate systems and processes mediated through language, especially law [23] and education [36]. Relevant to this study, Bamman et al. showed significant regional variation in cosine similarity of word vectors for common words, such as "wicked" and "city" [4]. Sociolinguists are using computational methods to investigate language variation [33], but a general integration of sociolinguistics into NLP could help researchers identify and address biases.

A more equitable educational data science using text should therefore consider linguistic variation at the forefront of analysis. ML models and systems that do not account for this risk classifying everyday language practices as hate speech, as was found to be the case with tweets written by AAVE users [40, 12]. Large datasets with student level metadata, like the data analyzed in this paper, will become increasingly common in education. Even basic sociolinguistic principles could help researchers address linguistic variation, use variation as a dependent variable, or explain how and why certain data correlates along various social dimensions to address the complicated relationship with student characteristics and language.

2.5 Household Income and College Admissions

Research on college admissions consistently shows that the college admissions process is easier for students from high income households. Studies have shown that standardized testing is strongly correlated with household income and other proxies for wealth [13], especially for black and white students. Other elements of the college application, such as financial aid forms [6] and the steps of the entire application process [26] are also more easily navigated by wealthy families than students from lower socioeconomic backgrounds. Family wealth is itself reflective of many racial and gender inequalities in the US [24].

The college admissions essay (CAE), has faced less scrutiny than standardized testing but some research has shown relationships to student identity and essay content. Using a corpus of CAE written by applicants in Britain, Jones [22] found that students from higher social classes wrote longer essays, had fewer spelling and grammatical errors, and tended to invoke markers of their higher social standing, such as the name of their elite school. Research by Kirkland & Hansen [25] found similar differences along income in diversity statement essays. They found that students from different racial backgrounds but similar socioeconomic levels wrote similar essays. Other studies have tested writing interventions with lower income students to teach them the genre of the CAE [15]. They found when students from a low income high school were explicitly trained on what they should include in their CAE, the average score of their essays on a rubric-based rating was higher than students that did not receive the intervention.

As universities move towards test optional admissions, fair analysis of CAE will become even more critical. If student backgrounds are not explicitly considered when using ML on CAE, new forms and abstractions of bias could be introduced into college admissions. However, computational methods can also shed light on potential issues of fairness in the essays. For example, Pennebaker et al. found that increased usage of function words (eg. pronouns and articles) and less personal narrative writing was positively correlated with college GPA [34].

3. DATA

The data for this study were 826,624 CAE submitted by applicants to a multi-campus, US public university system. The CAE were written across three academic years: 2015-2016, 2016-2017, and 2017-2018. These CAE were required components of the application, not additional essays submitted for honors programs, scholarships, or anything else peripheral to the main application. For this study, we removed essays that were under 100 characters and/or were

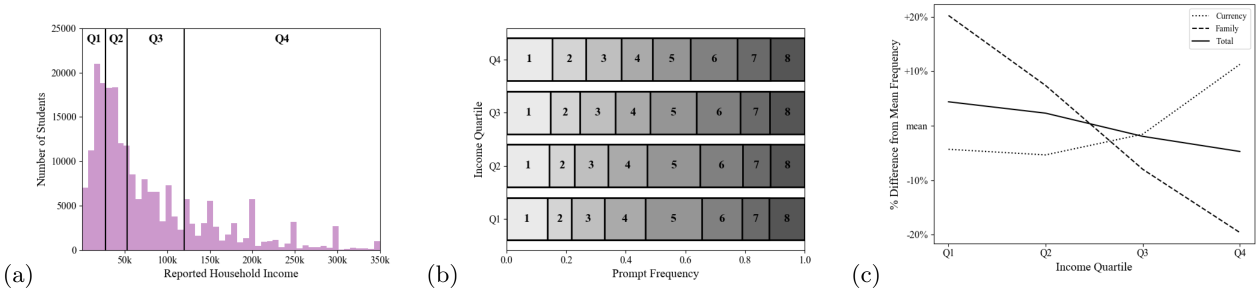

Figure 2: (a) Histogram of student RHI's with quartile boundaries marked. 8420 students with RHI >$350k not included in plot (but included in this study). (b) Proportion of essays written for each prompt by income quartile. (c) Frequency of word usages from different subsets of the Google Analogy Test Set by income quartile. Each frequency is compared to the global mean. Total refers to all of the words in the dataset.

written by students who did not report their household income. After this filtering step, 812,020 essays remained.

3.1 Reported Household Income

A variety of metadata about each applicant and document were included as part of the dataset, but this study focuses on the reported household income (RHI) provided by each applicant. It is important to note that RHI is not an objective measure of a family's household income. When students are accepted to a university, they provide any pertinent information and documentation (such as tax return forms, W-2, etc.). However, when the application is under review before any official admission decision is made, the only income and wealth information available to a reader is the RHI.

RHI was chosen as the variable of interest for several reasons. First, language variation along class and income lines has been well established in sociolinguistic literature [5, 29]. Splitting by quartile is also a relatively crude metric, and if qualitatively and quantitatively different results emerge in the vectors across quartiles then the problem could be both fundamental and grave. For example, we might expect that the top and bottom quartile have noticeable, measurable differences, but we would not expect the second and third quartiles to be substantively different. Finally, if there are correlations between CAE and income similar to other components of the application and income, new approaches and understanding of fairness and college admissions should be considered, as well as the role of CAE in decision-making. This would push ML fairness research in college admissions to think carefully and critically about data and outcomes, as language variation is not as neat as racial or gender parity but almost always arises.

In the dataset, the average RHI is $96,746, the median is $53,000, and the standard deviation is $125,000. Figure 2(a) shows distribution of income levels as well as the boundaries between the quartiles.

3.2 Prompt Choice

In 2015-2016, students had to write two personal statements to the same two prompts, meaning every applicant wrote two essays. In 2016-2017 and 2017-2018, students selected four prompts to write for from eight possible choices (70 possible combinations of prompts). The eight prompts were distinctive in theme, and if students from a certain quartile were responding to a prompt or group of prompts at significantly higher rates than students from other quartiles, our analysis could be skewed. However, figure 2(b) demonstrates that there are only mild differences in prompt choice across the income quartiles.

3.3 Word Distribution Variation

One possible source of error in this dataset is the difference in word usage between students of different backgrounds. This is a source of error because word vector training algorithms rely on large sample sizes in order to properly learn the contexts in which a given word appears. Our quartiles contain about 70 million tokens each, which is on the low end for word2vec datasets. This means that the quality of a given vector is very sensitive to that word's frequency within the data, a well-known issue that is an active topic of NLP research [20]. Practically speaking, if the word vectors trained on one quartile are able to solve an analogy that the vectors trained on another quartile fail to solve, then this could be due to the relevant words appearing more often in the first quartile.

Figure 2(c) shows the difference in word frequencies by quartile for three different subsets of the Google Analogy Test Set (GATS) [30]. We find that low income students use words from the analogy task more often overall, but this does not tell the whole story. We find that low income students use "family" words more often than high income students by a large margin, and we find (not too surprisingly) that high income students name foreign currencies more often than low income students by an even larger margin. This means that the vectors trained on the essays of low income students have an advantage on the "family" analogies, while the vectors trained on the high income students have an advantage on the "currency" analogies. We will take these word distribution-based advantages into account when analyzing the results of the analogy task.

4. METHODOLOGY

4.1 Vector Training

As mentioned above, we separately trained one set of word vectors on the writing from each income quartile. We chose to train our vectors using a word2vec Skip-Gram model in order to stay in line with Allen [1] who showed mathemati-

	“Viable” Analogies					“Q1 Advantage” Analogies		
GATS Subset	n	Q1	Q2	Q3	Q4	n	Q1	Q4
Family	420	0.629	**0.733**	0.702	0.681	348	0.672	**0.710**
Semantic	2446	0.191	0.222	0.233	**0.242**	391	0.609	**0.645**
Syntactic	9553	0.382	0.381	0.407	**0.451**	3307	0.433	**0.488**
Total	11999	0.343	0.349	0.372	**0.408**	3698	0.451	**0.505**

Table 1: Accuracy of each income quartile’s vectors on different subsets of the Google Analogy Test Set. “Viable” refers to analogies whose words appeared at least once in each training set. “Q1 Advantage” refers to viable analogies whose words appeared more often in Q1 (the essays of the lowest income students) than in Q4 (the essays of the highest income students).

cally that vectors trained in this manner would find analogies that exist in the training data.

We trained vectors of size 100 for 20 epochs using a window size of 5. We made all letters lowercase before training, but did not filter stopwords or punctuation. It is possible that changes to these hyperparameters would change the results of the study, but we feel that these are all reasonable choices given the dataset that we started with.

4.2 Vector Evaluation

For the analogy task, we use the Google Analogy Test Set (GATS) [30], which contains 19544 analogies, 8,869 of which are semantic, and 10,675 of which are syntactic. We consider a set of vectors to have “solved” the analogy “A is to B as C is to D” if the closest vector by cosine similarity to $C-A+B$ is D.

We evaluate our vectors on three similarity datasets, all of which are standard intrinsic evaluation tasks:

1. WS-353 [18] consists of 353 pairs of words along with their similarities rated on a scale from 0 to 10 by 13-16 subjects.
2. MEN [9] consists of 3000 word pairs whose similarities were determined by having subjects make binary comparisons between pairs of pairs of words rather than rating similarity directly.
3. SimLex-999 [21] consists of 999 pairs of words whose similarity was rated on a scale from 1 to 7 by 500 subjects. As opposed to the first two sets, SimLex-999 explicitly tries to avoid assigning high similarity scores to pairs of words that are associated but not similar (e.g. “coffee” and “cup”).

We measure similarity task performanace using Spearman correlation.

5. RESULTS

5.1 Analogy Results

Table 1 shows the accuracy of each income quartile’s vectors on different subsets of GATS. The “Family” subset (as it is called in the original dataset) contains analogies between pairs of words that differ according to gender (e.g. “husband is to wife as grandpa is to grandma”). We chose to look at this subset in particular because it is the only semantic section of GATS whose words were used frequently by students of all four quartiles. The other semantic sections of GATS (e.g. identifying currencies and world capitals) contained words used very infrequently by lower RHI students. We also split the entire dataset into semantic relations and syntactic relations. Semantic relations rely on word meaning (including the “Family” subset), while syntactic relations rely on morphological/grammatical differences (e.g. “bad is to worse as big is to bigger”). Finally, “Total” refers to the use of GATS in its entirety.

The first time we performed this experiment, we included all “viable” analogies (presented on the left side of Table 1). A viable analogy is one where all four words appear in each of the four sets of word vectors. With this setup, the Q1 vectors performed worst on all subsets, while the Q4 vectors performed best on all subsets except for “Family.” The difference in performance between Q1 and Q4 is very similar (5-7% of all analogies) between the semantic and syntactic subsets. This indicates that the differences we are observing are not only limited to word meaning, but to word usage as well.

Figure 2(c) shows that low RHI students use the words from GATS more frequently than high RHI students. This indicates that word distribution variation generally favors the lower RHI vectors, meaning that the higher RHI vectors performed better *despite* these variations. However, overall average word usage does not necessarily tell the whole story. It might still be the case, for example, that high RHI students use more of the words in the dataset more frequently than low RHI students. In order to more convincingly deal with the word distribution variation problem, we ran this experiment again, including what we call “Q1 advantage” analogies (presented on the right side of Table 1). An analogy has “Q1 advantage” if it is viable and its words appear more frequently in Q1 than in Q4 (i.e. the words are used more often by low RHI students).

Even when restricting ourselves to “Q1 advantage” analogies, the Q4 vectors outperform the Q1 vectors on each subset of the data, and by margins only slightly smaller than in the first experiment. This convincingly shows that word distribution variation is not to blame for the difference in performance we originally observed, as even when we *only* tested on analogies where Q1 has a word frequency advantage, the Q4 vectors solved far more of the analogies in GATS. This indicates that the observed differences in performance are due to the relationship between the analogies in GATS and the ways in which students of different quartiles use words differently.

Dataset	Q1	Q2	Q3	Q4
WS-353	0.594	0.615	**0.619**	0.583
MEN	0.592	0.625	0.650	**0.666**
SimLex-999	0.336	0.344	0.346	**0.352**

Table 2: Spearman correlation of each income quartile's vectors on three word similarity tasks. Agreement with "ground truth" scores rises as income rises.

5.2 Similarity Results

Table 2 shows the results of each quartile's vectors on each of our three word similarity tasks. Performance is reported using Spearman Correlation, although the results looked largely the same using Pearson Correlation. Note that with the exception of WS-353 (the smallest dataset), similarity task performance increased monotonically with income. This indicates that the similarity scores generated for these evaluation tasks are more in line with the way that high income students use language than the way that low income students use language. We did not filter the similarity tasks according to word frequencies, as the overall frequencies of the words in each task were very similar across the four training sets.

5.3 Qualitative Results

Word vectors by their nature pick up on semantic relationships between words [1]. It then follows that the underlying cause of these differences in intrinsic evaluation performance is a difference in word meaning between the RHI quartiles. Word vectors allow us to measure the similarity in meaning between two words in a dataset by using the cosine similarity of those two words' vectors.

Table 3 shows the words most similar to "money" according to the Q1 and Q4 vectors. We find that while low RHI students are talking about "rent", "expenses", and "bills" when they talk about money, the high RHI students are talking about "savings", "donations", and their "allowance."[1] This shows how a student's experience influences the way they use language. There are probably many other words that demonstrate similar qualitative difference and variation across quartiles, but an exhaustive search through them will be considered for future study. Importantly, human readers would be able to detect the differences between the most similar vectors between Q1 and Q4, even if those differences might be subtle. For both vectors, there are clear connections to money, but the differences in how a high income student writes about money and a low income student writes about money is also clear from our qualitative assessment.

For many scholars, especially sociolinguists, the differences seen in the qualitative results alone would be firm evidence of socio-semantic variation in CAE. Research in education have consistently found that students from different social classes experience and navigate schools differently and therefore rely on different language practices to negotiate their pathways in school [28]. Sociolinguistic variation in education has therefore been widely used to study and understand larger processes of social stratification and inequality. Though our qualitative analysis might not possess the depth of ethnographic research, it could still provide useful insights into how student background and experiences shape their language practices.

	Q1		Q4	
Rank	Word	Similarity	Word	Similarity
1	cash	0.768	funds	0.811
2	funds	0.754	fund	0.771
3	savings	0.724	monies	0.755
4	earnings	0.710	profits	0.744
5	rent	0.709	dollars	0.738
6	payment	0.705	savings	0.724
7	groceries	0.672	donations	0.701
8	expenses	0.669	donate	0.698
9	bills	0.6659	allowance	0.689
10	pay	0.659	goods	0.687

Table 3: Bills vs. Allowance: the words most (cosine) similar to "money" according to the Q1 (low income) and Q2 (high income) word vectors.

6. CONCLUSION

We have found that two standard intrinsic evaluation tasks (similarity and analogy) are biased against the writing of lower income students. Word vectors trained on the writing of lower income students systematically perform worse on similarity and analogy tasks than the vectors trained on the writing of higher income students. These findings do *not* indicate anything about writing quality. Rather, our results indicate that the "ground truth" semantic relationships included in these tasks are not the ground truth for everyone.

If analogies arise naturally in word vectors, then we could view the analogy task as a way of measuring what analogies exist for a given training set. If an analogy does not exist in the vectors of a given quartile, then we might say that the students who wrote those essays do not see those words as analogous. Under this perspective, our results can also serve as a way of quantifying the patterns in word usage between students of different income levels. If these patterns are not considered in large scale text analyses in education, word vectors and the many downstream tasks that use them as input could systematically bias the language practices of students based on their social class.

7. FUTURE WORK

We hope that these techniques will be used to audit other word vector evaluation tasks, both intrinsic and extrinsic. We also hope that there will be more discourse surrounding the fairness of evaluation tasks, especially given the increased use of word vectors in educational contexts. However, more work needs to be done in order to determine whether and how it is possible to debias intrinsic evaluation of word vectors with respect to various social dimensions.

Acknowledgements

We would like to thank our data providers and the Student Narrative Lab at the Stanford Graduate School of Education.

[1] Though not included in the table, we found that Q2's words were very similar to Q1's and Q4's words were very similar to Q3's.

8. REFERENCES

[1] C. Allen and T. Hospedales. Analogies explained: Towards understanding word embeddings. *arXiv preprint arXiv:1901.09813*, 2019.

[2] A. Alvero, N. Arthurs, a. l. antonio, B. W. Domingue, B. Gebre-Medhin, S. Giebel, and M. L. Stevens. AI and holistic review: Informing human reading in college admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 200–206, 2020.

[3] E. C. Baig. Who's going to review your college applications – a committee or a computer?, Dec 2018.

[4] D. Bamman, C. Dyer, and N. A. Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, 2014.

[5] B. B. Bernstein. *Class, codes and control: Applied studies towards a sociology of language*, volume 2. Psychology Press, 2003.

[6] E. P. Bettinger, B. T. Long, P. Oreopoulos, and L. Sanbonmatsu. The role of application assistance and information in college decisions: Results from the h&r block fafsa experiment. *The Quarterly Journal of Economics*, 127(3):1205–1242, 2012.

[7] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[8] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel. Understanding the origins of bias in word embeddings. *arXiv preprint arXiv:1810.03611*, 2018.

[9] E. Bruni, N.-K. Tran, and M. Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

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

[11] B. Chiu, A. Korhonen, and S. Pyysalo. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP*, pages 1–6, 2016.

[12] T. Davidson, D. Bhattacharya, and I. Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019.

[13] E. J. Dixon-Román, H. T. Everson, and J. J. McArdle. Race, poverty and sat scores: Modeling the influences of family income on black and white high school students' sat performance. *Teachers College Record*, 115(4):1–33, 2013.

[14] A. Drozd, A. Gladkova, and S. Matsuoka. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, 2016.

[15] J. S. Early and M. DeCosta-Smith. Making a case for college: A genre-based college admission essay intervention for underserved high school students. *Journal of Writing Research*, 2(3), 2010.

[16] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*, 2016.

[17] S. Federici, S. Montemagni, and V. Pirrelli. Inferring semantic similarity from distributional evidence: an analogy-based approach to word sense disambiguation. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.

[18] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131, 2002.

[19] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[20] C. Gong, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu. Frage: Frequency-agnostic word representation. In *Advances in neural information processing systems*, pages 1334–1345, 2018.

[21] F. Hill, R. Reichart, and A. Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

[22] S. Jones. "ensure that you stand out from the crowd": A corpus-based analysis of personal statements according to applicants' school type. *Comparative Education Review*, 57(3):397–423, 2013.

[23] T. Jones, J. R. Kalbfeld, R. Hancock, and R. Clark. Testifying while black: An experimental study of court reporter accuracy in transcription of african american english. *Language*, 95(2):e216–e252, 2019.

[24] A. Killewald. Return to being black, living in the red: A race gap in wealth that goes beyond social origins. *Demography*, 50(4):1177–1195, 2013.

[25] A. Kirkland and B. B. Hansen. "how do i bring diversity?" race and class in the college admissions essay. *Law & Society Review*, 45(1):103–138, 2011.

[26] D. Klasik. The college application gauntlet: A systematic analysis of the steps to four-year college enrollment. *Research in Higher Education*, 53(5):506–549, 2012.

[27] W. Labov. *Language in the inner city: Studies in the Black English vernacular*, volume 3. University of Pennsylvania Press, 1972.

[28] A. Lareau. *Unequal childhoods: Class, race, and family life*. Univ of California Press, 2011.

[29] D. Lawton. *Social class language and education.* Routledge, 2006.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[32] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic

regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[33] D. Nguyen, A. S. Doğruöz, C. P. Rosé, and F. de Jong. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.

[34] J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844, 2014.

[35] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[36] J. R. Rickford. Linguistics, education, and the ebonics firestorm. *Dialects, Englishes, creoles, and education*, pages 71–92, 2006.

[37] A. Rogers, A. Drozd, and B. Li. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 135–148, 2017.

[38] J. D. Rosa. Standardization, racialization, languagelessness: Raciolinguistic ideologies across communicative contexts. *Journal of Linguistic Anthropology*, 26(2):162–183, 2016.

[39] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[40] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1668–1678, 2019.

[41] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, 2015.

[42] T. Skutnabb-Kangas and R. Dunbar. *Indigenous children's education as linguistic genocide and a crime against humanity?: a global view*. Gáldu Kautokeino, Norway, 2010.