An Analysis of Gender Bias in K-12 Assigned Literature Through Comparison of Non-Contextual Word Embedding Models

Preeti Mohan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2020

Reading Committee:

Emily M. Bender, Chair

Shane Steinert-Threkeld

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

An Analysis of Gender Bias in K-12 Assigned Literature Through Comparison of Non-Contextual Word Embedding Models

Preeti Mohan

Chair of the Supervisory Committee:
Professor Emily M. Bender
Linguistics

Word embeddings are mathematical representations of words computed from a group of texts that a machine learning model is trained on. Generally, words that are similar to each other semantically will be closer together in the vector-space created by the embedding model. The distance between words can be analyzed to understand what words tend to be used in the same contexts in a given group of texts.

In this thesis, I use three different non-contextual methods of training word embedding models, Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), and GloVe (Pennington et al., 2014), on a corpus of literature assigned to students in grades K-12 in the United States to answer three questions:

- It has been shown that children are particularly prone to internalize biases in the content they read and watch (Railsback, 1993; Jacobs, 2003; Slater, 2003). What biases are present in literature assigned to children in grades K-12 in the United States?

- Are different kinds of non-contextual word embeddings sensitive to bias in different ways?

- Is the text from one book enough to detect bias using non-contextual word embedding models?

I find that GloVe embeddings are more sensitive to biases in smaller corpora, while Word2Vec and FastText are more sensitive to biases in large corpora. When looking at the word embeddings from a single book, I see variations in the strength of the words that are the "most gendered" — a book that had stronger gender biases (determined through literary critiques) had words that were more strongly gendered than a book that subverted gender biases (also determined through literary critique).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to express my sincere appreciation to the University of Washington Department of Linguistics for the opportunity to be a part of the Computational Linguistics Masters Program, my adviser, Professor Emily M. Bender for all the advice, guidance, and support provided during the process of the program, and Professor Shane Steinert-Threkeld for his advice and thoughtfulness as a reader of this thesis. I also want to express how thankful I am to my friends and family for their patience and support as I worked through a full-time job, social gatherings, and family visits to complete my degree. Thanks to Alex Stephen, Liz White-Hatton, Karissa Longo, Brandon Waggoner, Emily Lawton, Jake Long, and Eleanor Howell-Shryock for feedback.

DEDICATION

To my parents, Tilak and Vidya Mohan, my brother, Eashwar Mohan, my best friend, Geethika Vemulapalli, and my fiance, Jake Long.

Chapter 1

INTRODUCTION

Word embeddings are vectorized representations of words taken from a corpus of texts that a machine learning model is trained on. Words that are similar to each other semantically (for example, *nice* and *good*) will be closer together in the vector-space created by the embedding model, and words that occur in the same contexts will be close in vector space (Collobert and Weston, 2008; Baroni et al., 2014).

The ability of word embedding models to encode the biases of the texts they are trained on has been well researched and documented, especially on large corpora of data, and between different kinds of non-contextual and contextual word embedding models (Bolukbasi et al., 2016; Caliskan et al., 2017; Rozado, 2020; Basta et al., 2019; Kurita et al., 2019). Caliskan et al. (2017) describe this "encoding of biases" by demonstrating how, when performing tasks equivalent to human psychological studies, computational models that should theoretically be neutral show similar bias as humans do in these psychological studies (see section 2.2.2).

In this thesis, I study the attitudes encoded in non-contextual word embeddings trained on a newly-created corpus with limited available data, literature assigned to K-12 students in the United States. I explore the social attitudes reflected in the data, the sensitivity of different word embedding models to potential bias, and if I can get information about the social attitudes found in a single book.

1.1 Social Biases Reflected in K-12 Literature

Railsback 1993 and Luke et al. 1986, who both approach the topic from the perspective of education and curriculum researchers, discuss how the social attitudes present in books and media affect the thinking and attitudes of those who read them, particularly children.

Railsback (1993) discusses how the models presented in literature reinforce cultural values to children through presenting those values as implied/universal in the text, and how the promotion of gender stereotypes in literature send indirect messages to them about how they should be acting.

Caliskan et al. (2017) present evidence that word embedding models tend to accurately approximate the social attitudes of the texts they are trained on. Having found no previous work on analyzing bias in children's literature using word embeddings, I thought it would be particularly useful to understand if and what biases are present in the text, and see if I can get an indirect understanding of the general messages sent to children in the books they are assigned to read in school, or are recommended to them by libraries and book award systems.

1.2 Sensitivity to Bias in Different Models

As different kinds of word embedding models analyze text differently to compute word vectors, are there differences in how sensitive these models are to bias?

I make the assumption that bias is shown in a set of word embeddings if the distance between the word vector for any target word (such as an adjective like *kind*) and a representative combined vector for a social group (in this case, binary gender vectors for men and women) conforms to social stereotypes, such as in Bolukbasi et al. 2016. A hypothesis relating to this is that I might expect thata given widely held social stereotype to influence how people use language relating to the entities referenced by that stereotype — e.g. if the stereotype held is that "x is more y that z is", people would speak about x and z in different ways in how they relate to y. I then hypothesize that this difference in language usage about x and z relating to y affect distributional patterns enough to influence a word embedding model. Under this hypothesis, the word vector calculated by a model for y will be closer in the mathematical vector space to x than to z.

I create two "group vectors" to represent each binary gender. A group vector is an averaged vector of words that represent that subject — for example, *woman*, *female*, *girl*,

she, etc. would be a part of the word vector representing women. Since the assumption is that semantically related concepts are located near each other in vector space, if the representation for a given word is closer to one of these group vectors than the other, I also assume that it is related more to the words in that group vector than the words in the other. If the word vectors in a given model do not lean towards either of the gendered group vectors I am looking at meaningfully, I assume that that the model does not pick out gender biases — either they are not present in the data, or the model space does not place these words together. If significant distances are found between target words and group vectors that do not conform to the social stereotypes expected to be present in the text, I assume that the model introduces its own, potentially random, biases.

I train three kinds of models, Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), and GloVe (Pennington et al., 2014) (described further in 2.1.1), on the corpus of data, and distances are calculated between target adjectives and gendered group vectors to determine if and what social stereotypes are captured in each model. Does one of these models capture stereotypes more strongly than another?

Contextual word embedding models take word-sense disambiguation and sentence representations into account, but are more computationally expensive than non-contextual word embedding models, which perform simpler operations to get information about words and their representation within a corpus of text.

In this experiment, only non-contextual word embedding models are considered, due to computational limitations and the lack of GPU access through the Hathitrust Research Center Data Capsule computing environments, where the full-text data of books must be accessed from. In section 6.3, I describe potential future experiments using contextual word embedding models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and GPT-2/3 (Radford et al.; Brown et al., 2020).

1.3 Single Text Social Stereotypes

I will also look at word embedding vectors trained on the text of a single book, and compare the closest words to the group vectors from the models trained on these texts to those found in the corpus as a whole, to see if there are differences in the sentiment of the words closest to the group vectors. The books chosen are books, given common literary critique and analysis, I am given reason to believe will deviate from the average biases found in the text corpus. The two books chosen as representative of characterizing demographics in ways outside of the average are *Foundation* by Isaac Asimov, and *Song of Achilles* by Madeleine Miller.

Foundation is widely hailed as one of the most influential speculative fiction novels of all time. The series it is a part of beat *The Lord of the Rings* to win a Hugo award for Best All-Time series in 1966 (The Hugo Awards, 2007). However, Asimov is also well known for his personal misogynistic views towards women, and his characterization of them in his books, largely focusing on their appearance and sexuality over other character traits (Gabler, 2020). In *Foundation*, there are a small few examples of female characters — one being a wife of a planetary ruler who is characterized as a "annoying housewife" stereotype — said to be nagging and belittling. The other appearances of female characters are fleeting, and noted to be side-characters who are given personalities that convey negative stereotypes of women (SL, 2015). Past this, every other character is assumed to be male by the author, with explicit note separating scientists from their "women and children". Asimov's portrayal of men in *Foundation* is also described as "single-minded and flat" (Haines, 2012).

Song of Achilles is hailed for its portrayal as a feminist and LGBTQ+ positive characterization of Homer's *The Iliad*, focusing on the relationship between Patroclus and Achilles. Not only do reviewers claim that the male characters are characterized in nuanced and sensitive ways, but the female characters are non stereotypical, strong, and multifaceted (Flint, 2020; Habit, 2018). *Song of Achilles* has also won several awards — the Over the Rainbow award from the American Library Association for books that express LGBTQ+ experiences, a Stonewall award from the ALA for it's representation of LGBTQ+ characters, and the

Orange Prize for Fiction (now called the Women's Prize for Fiction), a prestigious award given out in the United Kingdom recognizing female authors (American Library Association, 2017).

I train the word embedding models described at the beginning of this chapter on the text of these books to see if the adjectives closest to the group vectors imply the social attitudes reflected in the books based on the literary critiques are identifiable in the distances between the word vectors. Moving forward, I will give an overview of the relevant literature in this area, covering word embedding models, evaluation of attitudes and bias in text, and the importance of understanding bias in literature. Past that, I will be explaining the methodology taken to answer these questions, then I will discuss and analyze the results seen in the data.

Chapter 2
LITERATURE REVIEW

In this chapter I review four bodies of literature that this thesis builds on — different kinds of word embedding models, different ways of evaluating word biases in models, concrete findings from previous model analysis, and the importance of understanding bias in text.

2.1 Differences in Word Embeddings

As a whole, word embedding models are given the task of predicting words, sentences, or other aspects of language given an input. The prediction desired from a model could be the next word in a sentence or the most likely sentence a given word would be used in.

Roughly, word embedding types can be split into two categories, non-contextualized and contextualized. Non-contextualized representations do not necessarily capture polysemy and the usage of words in different contexts, while contextualized word embeddings do capture different representations for different meanings of the same word. A common example of this would be a representation of the word *bank*, which can be used as a money lending institution, or as a geographic feature. Non-contextualized word embeddings would not differentiate between the meanings of the word *bank*, and treat it as one entity, while contextual word embedding models, due to their model architecture, can differentiate between different senses of a word based on it's context.

2.1.1 Non-Contextual Word Embeddings

Word2Vec

Word2Vec (Mikolov et al., 2013) has two model architectures; Contextual Bag of Words (CBOW), which predicts a word given the words surrounding it; and Skip-Gram, which

predicts a surrounding window of context words given a word. Note that the order of the words is not taken into account (this is called a "bag of words"or "windowed" approach).

An example of CBOW would be; given the sentence with an unknown token "Today was really n", the word most likely to be n, given the probabilities of word co-occurances learned from the training corpus, is predicted. So, if "Today was really nice" was seen in the text far more than "Today was really bad", n is much more likely to be predicted to be the word *nice*, given the context "Today was really".

To contrast, Skip-Gram would take a word, such as *marvelous*, and predict the most likely context for that word to appear in. This allows the Skip-Gram model to be more effective for infrequent words — if the word *marvelous* may be used much less in the text than the word *nice*, then "Today was really n" would never predict the word *marvelous* for n, while the Skip-Gram model is likely to predict the context of "today was really" for the word *marvelous*. The words *nice* and *marvelous* are not in competition to be chosen as the most probable, as they would be in the CBOW model.

FastText

FastText (Bojanowski et al., 2017) is a word embedding model that takes character n-grams of a word into account, rather than whole words, with the premise that the morphological structure of a word contains important information about its meaning. So for FastText with n-grams of 3 characters, the word *structure* would be represented as [st, str, tru, ruc, uct, ctu, tur, ure, re]. The vector for the word *structure* would be the combination of the n-gram vectors of its parts.

This allows it to create representations for out-of-vocabulary (OOV) words by summing up the vectors for the n-grams the word contains. FastText, like Word2Vec, can be trained either with Skip-Gram or CBOW (Bojanowski et al., 2017).

	the	dog	sat	on	log
the	0	1	0	1	1
dog	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
log	1	0	0	0	0

Figure 2.1: GloVe Matrix Representation of the Sentence "The dog sat on the log"

GloVe

GloVe (Pennington et al., 2014) creates matrix of size n by n (where n is the number of words in the vocabulary) between words, indicating how many times one word occurs in the context of another.

For example, the sentence "the dog sat on the log" would have a matrix that looks like table 2.1. The word *the* appears in the sentence twice, but only gets one representation in the matrix. If a word occurs next to another word, then the number of times those words occur next to each other (whether it be before or after) is the value of where they intersect in the matrix.

Once this matrix is created, the word vectors for each word in the matrix need to be created. GloVe is trained so that the dot product of the computed word vectors is equal to the logarithm of the probability of those two words co-occuring:

$$w_i \cdot w_j + b_i + b_j = log(X_{ij})$$

where b_i and b_j are scalar biases for words i and j, and X is the co-occurence matrix. For each pair of words i and j in the matrix, the above function should be true, or more specifically,

$$w_i \cdot w_j + b_i + b_j - log(X_{ij}) = 0$$

for each pair of words.

GloVe then defines a cost-function to minimize given the above goal, given each word i and each word j in the vocabulary, V.

$$\sum_{i=1}^{V}\sum_{j=1}^{V} f(X_{ij})(w_i \cdot w_j + b_i + b_j - log(X_{ij}))^2$$

$f(X_{ij})$ is a weighting function to make sure that only common word pairs are not learned from.

$$f(X_{ij}) = \begin{cases} (X_{ij}/X_{MAX})^{\alpha} & \text{if} X_{ij} < X_{MAX} \\ 1 & \text{otherwise} \end{cases}$$

Comparisons

Word2Vec and FastText are purely prediction-based word embedding models. They update (randomly) pre-initialized vectors as they are trained to be able to accurately predict a word given a context, or vice-versa. The word embeddings in prediction based models are learned so that the accuracy of a prediction is maximized over the entire dataset. In these models, the context is important — it is either used for prediction of a specific word, or it itself is predicted.

GloVe is a count-based and prediction-based word embedding model, and it reduces a large, sparse, matrix created by word co-occurrence counts to a determined size. This co-occurance matrix is global, meaning that the local context a word appears in is not taken into account the same way it is in Word2Vec and FastText. Glove aims to predict the logarithm of the co-occurance count between two words, globally. Pennington et al. (2014) also designed GloVe to not overweight frequent co-occurances or rare co-occurances.

Rozado (2020), analyzes Word2Vec (Skip-Gram), GloVe, and FastText, each trained on multiple corpora (Google News, Wikipedia, Twitter, and Common Crawl). He finds that FastText models outperformed Word2Vec and GloVe on semantic similarity and analogy

tasks, potentially due to its usage of word parts, which assume morphological features.

2.1.2 Contextual Word Embeddings

ELMo

An example of a well-known model for contextualized word embeddings is ELMo (Peters et al., 2018). ELMo embeddings are learned from the internal state of a bi-directional long short-term memory model (LSTM) (Hochreiter and Schmidhuber, 1997). An LSTM preserves information from inputs it has already seen — so a traditional unidirectional LSTM would preserve information from the past. A bi-directional LSTM runs information forwards and backwards, so context from the past and the future are preserved internally. Using information from the past and the future, it becomes easier to have more information on what the next word in a sequence should be based on its context. In ELMo, not only are the representations for words dependent on the context in which it is used (as described above), the word representations are also character based instead of word based. This is so that word-part clues (such as roots, prefixes, suffixes, etc.) can be used to try to create context for words that were not seen before (similarly to FastText).

GPT-2 and GPT-3

Another popular model for contextualized word embeddings are OpenAI's GPT-2 and GPT-3 (Radford et al.; Brown et al., 2020). The GPT language models were trained on the 40GB WebText dataset, and are fine-tuned for specific supervised tasks. GPT-2/3 are built using transformer decoder blocks — each of these blocks takes a word embedding and a context vector, and it focuses on deriving information about the word embedding from the context vector.

GPT models are uni-directional (they only look at information going forward, not both forwards and backwards), which means they are only allowed to pay attention to the previous tokens as clues. As an output is generated based on the input and the input's context, that

output is added to the sequence of inputs, and this new sequence becomes the input to the model in the next step. This is called "auto-regression".

The differences between GPT-2 and GPT-3 lie in the number of layers and attention heads, the size of the word embeddings (1600 to 12888 dimensions), context window size (1024 to 2048 tokens), optimiser parameters, and alternating attention patterns (Radford et al.; Brown et al., 2020).

BERT

BERT (Bidirectional Encoder Representations from Transformer; Devlin et al. 2019) is another well-known language model. While GPT-2 uses transformer decoder blocks, BERT uses transformer encoder blocks, and it looks at information forwards and backwards (as its name would suggest). An encoder receives a list of vectors as an input, and processes the list through a self-attention layer. The concept of attention refers to what a model uses as "clues" in the input sequence to create a better encoding for a token. An encoder passes up its input as well as its self-attention to the next encoder in the stack. The vectors produced by encoders can be used as the input for the classifier to be used for a given word-embedding task.

Another differentiation of BERT is that it uses a masked language model. Fifteen percent of the input is masked, and those words are predicted by the model (the output is produced from an incomplete input). This allows the model to better learn the properties of word sequences, by comparing the tokens in those sequences (Devlin et al., 2019).

2.2 Understanding Biases in Models

There have been several ways of trying to understand social biases present in a given corpus of text that have been described in previous work. Most of these methods use pre-determined sets of words that social bias is commonly seen for (adjectives related to appearance and competency, specific careers). If a given model predicts words or demonstrates vector closeness in a specific way that conforms with social stereotypes, it is determined that the model has

picked up bias present in the text, as the model learns its predictions and word contexts based purely on how words are used in the training corpus.

To frame the discussion of previous methods, I first review the work of Blodgett et al. (2020). They perform a survey of 146 papers that analyze bias in NLP, and discuss the problems and inaccuracies that are often seen in the approaches taken by the authors of the papers they survey in both measuring and mitigating bias. They bring up that there are often inconsistent motivations and reasonings for analyzing bias in NLP systems, and that often the reasonings presented are not grounded in in literature outside of the NLP domain, and techniques for analysing bias between different works are not comparable.

In reading the 146 papers chosen through a combination of keywords and manual search, they aimed to categorize the motivations of the papers into five categories — allocational harms (how might an automated system allocate resources unfairly to differnt groups?), representational harms (how might a system represent different groups in different ways?), questionable correlations (how does system behavior correlate features of language with social groups?), vague descriptions of bias (papers that do not explicitly define what the bias is that they are looking at), and surveys/frameworks. In categorizing these papers, they find that 16 percent of papers do state vague motivations (or don't state any motivation at all) for analysing bias, and do not define what it means for a system to be biased or to discriminate. They find that 32 percent of papers do not specify normative reasoning for their motivations, and instead only focus on system performance issues.

They also find that papers do not always describe why the biases found in systems are harmful, in what ways, and to whom. They re-iterate that it is important to understand who is harmed and how by biases found in systems. Relatedly, the conceptualization of bias in different papers is thought about differently, and often allocational and representational harm are mixed up or conflated, where biases in downstream applications are cited as reasoning for analysing the biases found in a system, without the actual measurements known or present for these downstream systems.

They present three recommendations going forward as a framework for researchers ana-

lyzing bias in NLP systems. Firstly, relevant literature outside of NLP should be consulted in understanding the relationship between language and social hierarchies, and representational harms should be treated as harmful in their own right, instead of being conflated with allocational harms. Secondly, it should be stated explicitly why bias in a system is harmful and to whom. Thirdly, members of communities affected by NLP systems should be engaged with.

Going forward in this section, I discuss the methods taken by researchers to understand the biases present in NLP systems, but only note their methodology. In section 2.4, I discuss the literature I reviewed outside of the NLP domain relating to the importance of understanding bias in literature.

2.2.1 Word Vector Embedding Distance

In Bolukbasi et al. 2016, curate lists of occupational words, and create analogy problems to compare the "genderedness" of an occupation, based on crowdsourced data. The format of the analogies used is *he:x = she:y*, where x and y should be words that the embedding model believes to be analogous to *he* and *she* (a discussion of problems with this analogy method is described further in 2.3.2). To retrieve an analogous word y given a predetermined x, the distance between the word vectors for x and y must meet a certain threshold, and then the cosine similarity of the vector created when subtracting the vector for y from the vector for x is computed with the vector created when subtracting the vector for *she* from the vector for *he*. The word y that maximizes this cosine similarity given the predetermined word x is chosen as the analogous word.

Similarly, Garg et al. (2018) create "group vectors" for male and female words (instead of using the vectors for the words *he* and *she*) as well as group vectors associated with race and ethnicity. These group vectors are calculated by taking the average of the word embedding vectors of a set of gendered words (i.e. the "female" group vector could contain the words *she, her, woman*, and both group vectors have twenty total words comprising of nouns and pronouns). The datasets used to choose words for these group vectors, as well

as the occupations and adjectives used to compare bias come from a combination of census data, previous work, and crowdsourcing. The embedding distance between words (adjectives, occupations) and these group vectors (gender, race/ethnicity) are calculated to determine which group vector a word is closer to (and therefore more biased towards). The closer a word is to a group vector, the higher that word has a bias towards that vector's category.

Rozado (2020) uses two group vectors, considered on either end of a spectrum (i.e. man and woman group vectors), as either end of an x-axis on a graph, while using two other group vectors of opposing polarity (i.e. conservative and liberal) as the y-axis. Projecting a word's position on both axes onto the graph provides a visual representation of more than one bias a corpus might associate with a word. For example, doing so with words associated with negative or positive sentiments on an x-axis of binary gender and a y-axis of the liberal-conservative spectrum showed that positive words clustered towards a female-liberal space, while negative words clustered towards a male-conservative space.

2.2.2 Word Embedding Association Test (WEAT)

Caliskan et al. (2017) define the WEAT based upon the Harvard Implicit Association Test (IAT) (Greenwald et al., 19980701). The IAT measures the association of target and attribute words, by measuring how quickly a participant associates a target with an attribute. Target words could be, for example, professions split by stereotype (engineer, scientist for men, nurse, teacher for women), and attribute words would be the criteria for which the target words are split on (in the above case, the attribute words would be two set of binary gendered words). Theoretically, both target word sets and attribute word sets should have relative similarity towards each other, and one should not bias more (i.e. the distance between *woman* and *engineer* should be the same as the distance between *woman* and *nurse*).

Caliskan et al. (2017) use a modified version of this method, where the closeness of a target word to an attribute word in embedding space is used to decide the attribute that the target word is associated with (instead of how quickly a human participant acknowledges that a target matches an attribute). They see that word embeddings encode not just stereotyped

biases but other associations demonstrated through the IAT, such as flowers being pleasant. They make the point that statistical processing of text allows researchers to derive conclusions that are grounded in the real-life trends and opinions that can be observed from the text.

Rozado (2020) uses the Harvard General Inquirier (HGI) lexicon (Stone et al., 2007), which is much larger than WEAT (3623 words with positive/negative sentiment mapping vs 50 words), and finds that there is some divergence between the projection of WEAT words and HGI words onto different axes — one notable one being that WEAT does not suggest a difference in positivity/negativity valences of words closest to the man or woman group vectors, while the words from HGI project more positive sentiment words towards a female polarity. This difference is explained by the availability of more words to analyze attitudes with.

2.2.3 Sentence Templating

Using contextualized word embeddings from BERT, Kurita et al. (2019) create simple template sentences containing the attribute word for which they want to measure bias, mask the attribute token, try to identify what a target token might be, then determine which target token is predicted as part of the sentence's context (i.e. male and female). Depending on which target token is picked, the assumption is made that the model believes that target token is more likely to be associated with that attribute word. This helps to take context into account when measuring bias. Targets would be something like gendered words, and attributes would be career related words. The steps of their evaluation is as follows:

- 1: Prepare a template: [TARGET] is an [ATTRIBUTE] e.g. [He] is a [doctor].
- 2: Replace [TARGET] with [TARGET-MASK] and compute the probability of TARGET-MASK=TARGET for the sentence e.g. [TARGET-MASK] is a [doctor]
- 3: Replace [TARGET] and [ATTRIBUTE] with [TARGET-MASK] and [ATTRIBUTE-MASK], and compute probability of [TARGET-MASK]=[TARGET] (i.e. *he* or *she*)

and [ATTRIBUTE-MASK]=[ATTRIBUTE] given the sentence [TARGET] is an [ATTRIBUTE] e.g. [He] is a [doctor], [She] is a [doctor].

They define their measure of bias as a score computed by dividing the result of 2 by the result of 3 then taking the log of that value. They find that with this method of measurement, the sentences with higher scores have TARGET and ATTRIBUTE words that reflect underlying biases in the corpora that correspond to social attitudes about gender that also correspond with the findings of Caliskan et al. (2017).

2.2.4 Sentiment Analysis

Sweeney and Najafian (2019) measure bias using positive and negative sentiment words that do not have biased connotations (a set from Hu and Liu 2004). Vectors for these sentiment words are initialized, and then a logistic regression model is trained to predict the probability of any word being a negative sentiment word. Then, a list of neutral identity terms are classified with the model to get positive/negative sentiment for them. They find that certain racial and religious demographic groups that have, in previous literature, been determined to face higher amounts of discrimination, are classified with negative sentiment.

2.2.5 Entity-Centric Contextual Affective Analysis

Field and Tsvetkov (2019) discuss how an analysis of power does not generally happen in NLP bias analysis, and derive a method to obtain power (strength of a demographic characteristic), sentiment (goodness/badness of a demographic characteristic), and agency (activeness/passiveness) scores for an entity as a method of measuring bias.

They use BERT to extract sentence level embeddings (an embedding that represents an entire sentence, instead of just a word) for every instance of every word in a corpus. They then train a regression model using the embeddings as features, in conjunction with the NRC VAD Lexicon (Mohammad, 2018), which contains sentiment, agency, and power annotations for words. Then, a masked version of BERT is used to extract a contextual embedding for

each word, which is then fed to the regression model to obtain power, sentiment, and agency scores. By masking out the target word, this has the model produce embeddings solely derived from the surrounding context of the target word. They found that this method was able to, given an entity (such as the name of a person), predict the power, sentiment, and agency of that person or entity given its context, given the predictions from the regression model.

2.2.6 Takeaways

The approach I intend on taking follows the methods used by Garg et al. 2018 and Rozado 2020 in that I use the concept of group vectors for my representations of binary genders and measure the distance between words and those group vectors to calculate the association a word has to a certain group. In addition, I also use the HGI lexicon and its annotations for categories associated with words, informed by Rozado 2020's usage of the same. Due to the flaws in analogy methods discussed further in 2.3.2, I opted to not use any analogy based methods.

2.3 Concrete Findings

Through various methods like those described above, researchers have been able to find that models do internalize the social biases found in the texts they are trained on, even when certain words should theoretically be neutral (i.e. adjectives that relate to competence should not describe one gender more strongly than another). Not only do models internalize the social biases found in texts, these biases strongly mirror historical gender gaps, occupational gender split percentages, and shifts in census data.

2.3.1 Societal Correlation

Friedman et al. (2017) show that the gender biases in grouped word sets correlate with gender gaps in those sets. For example, they took a group of political terms and computed the

gendered embedding vectors for each country in a list of countries (these were derived from country-specific Twitter data). The graph of a country's gender embeddings and political term embeddings were plotted on the same graph as that of the GGG political empowerment gender gap subindex (World Economic Forum, 2017), and Friedman et al. (2017) found that the countries where the Tweets captured less male bias in political terms correlated with the countries in the gender gap subindex that had greater female empowerment in their political systems. This leads to the theory that cultural biases in language correlate with cultural gender gaps, and bias can be characterized by strength and direction of gaps. However, not all topical word set biases correlate with all gaps, and random word sets do not correlate. Different themed word sets capture different dimensions of gender bias and gaps (i.e. the sets for politics and occupations will capture different gaps).

Garg et al. (2018) find that real world occupation gender-split percentages are correlated with embedding gender bias from news data (segmented by decade). They also note that the effects of real-world events shape the gender-biases found in the word embeddings — for example, there is a sharp shift in sentiment towards women's capabilities in certain occupations before and after the height of the women's movement in the US. Ethnic biases over time can also be quantified through embeddings — for example, they find phase shifts in heightened times of Asian immigration. However, it is worth noting that the historical textual data may not completely reflect popular social attitudes at the time, and the census data embeddings were compared against may not fully capture gender or ethnic associations.

2.3.2 Flaws in Analogy Methods

Schluter (2018) points out that there is a "misalignment in assumptions" when using analogy methods to measure bias in word embeddings. Word embeddings that are generated from raw text are based on the distributional hypothesis, that words can be described sufficiently in terms of their distribution in language. She points out that words that share contextual and distributional information will likely share similar representations and will be grouped together in vector space. She also discusses that the normalization of vectors before analogies

are computed can distort the results.

Nissim et al. (2019) also discuss the downsides to using analogy methods to determine bias, mainly that bias outside of the vectors can be introduced by the human choice element in analogies. Using analogy tasks to evaluate word embeddings is unfair. It is a "party trick", and even setting up an analogy is often done in a biased way. This method of understanding word embeddings and using it as a tool to debias embeddings leaves something to be desired in terms of effectiveness, and does not truly help to mitigate bias, which can mistakenly propagate to downstream applications without people realizing it.

2.4 Importance of Understanding Bias in Literature

In this section, I discuss the impacts that bias in literature has on readers, and why it is important to understand the perspectives that are in text. As children are likely to internalize the social attitudes present in text, it becomes especially important to understand what attitudes are presented to them in assigned books and if the perspectives presented are varied. I emphasize the importance of diverse curriculum selection, and in turn, show why it might be valuable to analyze literary attitudes at a large scale.

2.4.1 Internalization of Narrative Bias by Readers

Male writers tend to be valued more than female writers in the NYT bestsellers list and the US publishing industry. If men and women writers encode different biases in the text as an effect of society and how they perceive their and other genders, these could propagate and reinforce these social biases in the people who read the books (Cima, 2019). Per Rails-back (1993), who comes from the perspective of an education researcher, not all literature accurately and fairly reflects the social values and cultural diversity present in society.

DelFattore (2003) studies instances where people have been opposed to certain literature being involved in curricula (from perspectives of religion, percieved morality, and cultural norms), and asks about the kinds of materials that should be included in school curricula and how they should be selected. Through many given examples, she discusses how books

selected for children that espouse cultural values do not not align with those of the parents of the children (for example, "non-christian" books like *Harry Potter*, or the inclusion of books with LGBTQ+ characters), and how this creates friction between educators and parents. Since narratives convey attitudes, ideals, and philosophies as well as facts, she ends with a few questions: how should a decision be made about what should be included? How important is it for children to be exposed to different perspectives in their school materials? To connect to this and the implications of an author's gender on the attitudes reflected in a work, Luke et al. (1986) discuss and uncover the gender biases present in literature assigned to children, and the flaws in the selection process of literature for curricula. They do this by surveying students in a language arts class within a teaching diploma program, and analyze the works that these students select for a children's curriculum. Luke et al. describe how most of the childrens books they looked at were authored by men, and emphasize the "exclusion or minimization of the female literary voice" in the literature selected for children to read. They also state that "...this is not to argue that gender of author necessarily reflects [a difference] in the ideological form and content of the work selected; it may remain quite possible for authors of one gender to fairly and accurately portray the world view of the opposite gender." (pg. 212). Even so, they still argue that the gender split of authorship and of character portrayals in literature should be carefully considered to understand what attitudes may be passed on to children who read the literature.

Railsback (1993) analyzes literature that children read in school, and finds that the literature does not have adequate representation of women, and an analysis of books often found in curricula usually had male leads over female leads. This is important, because she discusses how gender bias in literature can affect how children see themselves, and cites work on how they reinforce cultural values and how children learn through the models and images presented to them. The promotion of gender stereotypes in literature read by children indirectly sends messages on how they should be acting through social constructs. When analyzed, by Railsback, the writing of young children in classrooms was found to also have high occupational gender bias. However, young female writers tended to write more female

characters.

"The selection of children's literature...is a model of the exclusion of works by and about women and ethnic minorities." says Luke et al. (1986) "...the well documented minimization, distortion, and outright exclusion of women, blacks, and other racial and ethnic minorities in curricular materials is both a reflection and a cause of the relative powerlessness of these groups in the larger society." This shows how fictional texts can present a biased, exclusionist, and error-prone version of the world. The consequence of excluding groups from recommended or required books has been to establish the prioritization of a specific viewpoint, and readers from excluded groups will be especially negatively affected.

This not only has an effect on children, but adults as well. Slater (2003) discusses the impact of narratives used for adult education, particularly cases in which story-telling has been used to influence behaviour. He posits that intentional persuasion using narratives provide a prototype for investigating persuasive effects from narratives without an intentional message. "By arranging characters and events into stories, people are able to develop an understanding of the past, an expectation about the future, and a general understanding of how they should act." (Jacobs, 2003)

Slater (2003) points to examples where narratives have been used in underdeveloped countries to support social and economic efforts such as adult literacy and family planning. He found that narratives influenced attitudes towards certain behaviors. For example, when negative characters were portrayed as passive and chauvinistic, those behaviors were seen by audiences as less desirable, while advocating and kind traits were seen positively. However, Slater notes that though the attitudes of those who read and watched the narratives changed, behaviors necessarily did not, due to set cultural norms and fears.

When aspects, such as cultural factors, that a media consumer relates to are incorporated into narrative, they relate to the content more and internalize messages from it more. Strange (2003) reflects that "...fiction invites readers to adopt perspectives and commitments they would not entertain in their actual worlds... perspectives adopted in narrative experience are internalized and recruited as guides to future action... and the degree to which story-

world commitments carry over to real-world convictions." This is corroborated by Strange (2003) in a historical analysis of the book *Uncle Tom's Cabin* by Harriet Beecher Stowe, and the impact it had on the portrayal and attitudes towards Black Americans at the time. "Indeed the novel is widely accepted to have played a pivotal role in galvanizing public opinion against slavery." Many are convinced that something about Stowe's book had led it to succeed where previous modes of discourse had failed. In this case, having a literary perspective about something that readers were not familiar with, allowed them to engage with the content and learn from the perspectives presented, and even change attitudes about the demographics presented in the books.

In calling attention to the impacts of literary bias on readers, I hope to also draw attention to why it may be useful to use computational methods to understand the biases in large amounts of text, since it has been shown that computational methods can reveal the underlying attitudes present in text. If there is an effective way to calculate the social attitudes present in text, more diverse curricula could be created in classrooms, and could ideally be able to provide a variety of perspectives in the literature assigned to students.

Chapter 3

METHODOLOGY

3.1 Data

I have collected a list of ISBNs for K-12 assigned literature from several sources: the California Department of Education [1], the New York State Education Department [2], the Indiana Education Department [3], the Minnesota Department of Education [4], Newberry Award winners [5], Caldecott Medal winners [6], and books selected by YALSA as notable young adult fiction [7]. All these lists are publically available. Of this masterlist of approximately 10,000 texts, 843 are available in full via HathiTrust using their research center data capsules, and will be used as the dataset for this experiment.

The two books chosen as representative of characterizing demographics in ways outside of the average are *Foundation* by Isaac Asimov, and *Song of Achilles* by Madeleine Miller (Gabler, 2020; Haines, 2012; John, 2013; Britt, 2016; Aulisio; Donoghue; Habit, 2018; Flint, 2020).

---

[1] http://www3.cde.ca.gov/reclitlist/search.aspx, accessed 8/31/2020

[2] http://www.p12.nysed.gov/guides/ela/part1b.pdf, accessed 9/4/2020

[3] https://www.doe.in.gov/sites/default/files/standards/sample-texts-revised-08-05-15.pdf, accessed 8/26/2020

[4] https://education.mn.gov/mde/dse/stds/ela/, accessed 8/26/2020

[5] http://www.ala.org/alsc/sites/ala.org.alsc/files/content/awardsgrants/bookmedia/newberymedal/newberywinners/newpresent.pdf, accessed 8/31/2020

[6] http://www.ala.org/alsc/awardsgrants/bookmedia/caldecottmedal/caldecotthonors/caldecottmedal, accessed 8/31/2020

[7] http://booklists.yalsa.net/, accessed 8/31/2020

3.2 Experiments

To answer the first research question presented in 1.1 (what biases are present in literature assigned to children in grades K-12 in the United States?) I use word embedding models to analyse gender bias in a corpus of K-12 assigned children's literature.

Similarly to the work done in Garg et al. 2018, I will evaluate bias in the trained word embeddings by creating group vectors, which are the average vectors of words considered to be "group words" related to a specific demographic. Gender-specific group words are taken from the appendix of Bolukbasi et al. 2016 to create the group vectors. Each group vector has 20 words that correspond to one-another (e.g. *boy, brother, father* and *girl, sister, mother*, as well as the plurals of these words). Plural forms of words are used in addition to singular forms to gather more information from the models that create seperate embeddings for words, including seperate embeddings for different forms of the same word.

A set of adjective and occupation words from the appendix of Garg et al. 2018 and the Harvard General Inquirer Lexicon (Stone and Hunt, 1963) are used for evaluation. I compute the vectors for these words, and then evaluate the bias of a word by computing how close it is to a group vector. The closer to a group vector, the more biased a word is towards that group. The association between a word and a group is calculated by finding the cosine similarity between that word's vector and the average vector for that group. The cosine similarity does not take the magnitude of vectors into account, just the direction, which is why I choose it over other similarity calculation methods.

In addition, for each of the target adjectives in the evaluation set, the Harvard General Inquirer lexicon's (Stone and Hunt, 1963) sentiment categories are used to categorize them as positive or negative. I then analyze the valence of appearance and competence related adjectives and if there are correlations between categories of adjective groups (such as *positive*, *negative*, *weak*, *strong*, etc.) and the gendered group vectors. As a way to compare the gender biases present in the smaller K-12 corpus with a larger standard corpus, the models trained on the corpus of literary data will be compared with versions of each of the given mod-

els (aside from Word2Vec CBOW, which was not available) trained on the Wikipedia 2017 English Article corpus, downloaded from the University of Oslo's NLPL Word Embeddings Repository (Language Technology Group).

To answer the second question (are different kinds of non-contextual word embedding models sensitive to bias in different ways?), this evaluation will be performed on three different non-contextual word vector models — GloVe (Pennington et al., 2014), Word2Vec (Mikolov et al., 2013), and FastText (Bojanowski et al., 2017). My hypothesis is that if the distance between an adjective vector and group vector conforms to social stereotypes (for example — *man:strong, woman:beautiful, man:angry, woman:caring*), then bias is shown in vectors. When I find that a specfic adjective is particularly close to a group vector for a certain category, I understand that adjective is used in the context of the group it is closest to more than in the context of the other group I am comparing against. Something to note, however, is that a certain adjective, say *feminine*, may be implied as a descriptor for one group, and is found to be closer to the other group because it isn't used in contexts for the group it is implicitly understood to categorize (though I do not analyze this point in this study, I hope to do so in future work).

If the vectors for a model don't show significant distance between adjectives and group vectors, while those from other models do, I infer that specific method does not pick out bias as strongly as the other models. If noticeable distance is present between adjectives and group vectors, but not in a way that conforms to the other models, then I infer that the model introduces its own, potentially random, bias.

To answer the third question (does one book contain enough text to meaningfully determine the gender biases present?), I will look at the word embedding vectors trained on the text of a given book, and compare the biases it encodes using the previously mentioned method, to the collective bias of all the books in the corpus. The books chosen are books that given professional literary critique, I have been given reason to believe represent characters in a certain demographic that deviate from the average (the chosen books, described in section 1.3, are *Foundation* and *Song of Achilles*). If I see that embeddings trained on

solely each book from this selection match the biases expected from the literary critique, I conclude that the text of one book is enough to capture biases in a non-contextual word embedding model.

3.3 Algorithms

3.3.1 Word2Vec

The implementation of Word2Vec used in this project comes from the Gensim Python library (Řehůřek and Sojka, 2010). The model is initialized and given parameters of the sentences to train on, the dimensionality of the vectors, the size of the window between the current word and the predicted word, the minimum count for a word to have a vector created, the number of iterations to take over the training data, and whether or not to use the CBOW or Skip-Gram version of Word2Vec. The default parameters are used for both the CBOW implementation of Word2Vec: a vector size of 100, a window of 5, 5 a minimum word count of 1, and 5 iterations over the training data. For the Skip-Gram model, the window is changed to 10, as suggested by the original Google documentation for the tool (Mikolov et al., 2013).

3.3.2 FastText

The implementation of FastText used is also from Gensim (Řehůřek and Sojka, 2010). The parameters used for FastText are the default vector size of 100 and a window of 5 and minimum word count of 1 to match the parameters used for Word2Vec. The default value of 10 is used for the number of iterations over the training data.

3.3.3 GloVe

The C implementation of GloVe from Stanford is used, with a vector dimensionality of 100, a window size of 5, a min count of 1, and 50 iterations over the text corpus (Pennington et al., 2014).

3.4 Bias Analysis

For each of the pre-trained Wikipedia 2017 models as well as the twelve sets of vectors (Word2Vec CBOW, Word2Vec SG, FastText, and GloVe — each with vectors for all text in the corpus, *Foundation*, and *Song of Achilles*), I create a group vector for "man words" and "woman words" as described in 3.2.

For each word in the set of comparison words (see appendix), I retrieve the word vector for that word from the model and compute the cosine similarity between that vector and the group vector. The distance of a word from the man group vector is subtracted from the distance between the same word and the woman group vector, and the resulting number represents how much closer a word vector is to one group vector than the other. A value closer to 1 means that a word is closer to the man group vector, and a value closer to -1 means that a word is closer to the woman group vector (i.e. 0 - 1). I look at the top N words that are closest to each word vector (in this case, $N = 7$, as I believed $N = 5$ did not provide enough information, and $N = 10$ provided presentation issues). In looking at the words that are closest to each group vector, I hope to gain an understanding of what words are most strongly associated with each of the group vectors, and if information about the social attitudes present in the corpora is conveyed in the meaning of the closest words.

I apply these methods and analyze the resulting data to answer the presented research questions. In the next section, I will describe the results of this analysis.

Chapter 4

RESULTS AND DISCUSSION

In looking at the results of the experiments I performed, I break them down by model type, to more clearly examine if there are significant differences in the valence or sentiment of adjectives/kinds of adjectives that are closest to the group vectors between different models. I see that between all four model types, the words closest to the group vector for women are more positive appearance related than the words closest to the group vector for men, which are generally more negative. In interpreting the graphs, the color red indicates adjectives that are closer to the woman group vector, and the color blue indicates adjectives that are closer to the man group vector.

To answer the first question (what gender biases are reflected in the corpus of K-12 literature?), I look at the evaluative adjectives that are closest to the group vectors in each model (see table 4.1). From personal classification,[1] I interpret that the top adjectives for the K-12 corpus in the FastText model table C.1 show more adjectives with a positive or neutral connotation for the man vector than for the woman vector. The words *gallant, crisp, airy* and *sane* are four out of the seven words closest to the man vector, and are classified by the HGI lexicon as either positive or neutral (positive, neutral, neutral, and positive, respectively), while only two of the words closest to the woman vector have a positive classification: *pretty* and *reasonable*. The rest of the words closest to the woman

[1]The HGI lexicon has categorizations for positivity and negativity, in addition to words relating to emotion, knowledge, weakness, strength, etc. I mainly use the HGI lexicon's classifications for positive and negative valence, and use a combination of personal classification and the adjective categories from Garg et al. (2018) to determine if a word is related to competency or appearance. In addition, the HGI lexicon is robust but not complete. For example, the word *barbaric* is not present in the HGI lexicon, but I believe it is reasonable to assume that it has a negative valence. Any personal classifications I make will be bolded and any judgements from Garg et al. (2018) or the HGI lexicon will not be. See table 4.1 for annotation symbols used.

vector have a negative classification: *stupid, crazy, weak,* and *difficult.* The word *hateful*, is also present as a close word to this group vector, but does not have a direct classification in the HGI lexicon, however, the word *hate* has a negative classification.

The valence of the adjectives between the Wikipedia 2017 FastText model and the K-12 Corpus FastText model shown in table C.1 are relatively similar for the closest words to the man vector, but have more negatively classified words in the K-12 corpus for the words closest to the woman vector. The words closest to the woman vector in the FastText model trained on the Wikipedia corpus are mostly appearance related (based on the classifications from Garg et al. (2018)): *feminine, alluring, voluptuous, attractive, beautiful, sensual*, and *homely*, while as described above, the closest adjectives to the group vectors in the FastText model trained on the K-12 corpus are classified by the HGI lexicon as negative and personality traits. One way to interpret this is that the combined text of the K-12 corpus focuses more on these negative personality traits in relation to women, but this result is not duplicated when I compare the closest adjectives between Wikipedia 2017 and the K-12 corpus in other models. This leads me to believe that the FastText model picks up gender biases differently than the other models in a smaller corpus.

In the GloVe model trained on the K-12 corpus, I see words both positive and negative for the man and woman vector, but the adjectives closer to the man vector are more competency related than those for the woman vector (as classified by the list of competency related words from Garg et al. (2018)). Similarly, in the Word2Vec CBOW model trained on the K-12 corpus, the adjectives closest to the man vector have more of a negative connotation, while the adjectives closer to the woman word vector are appearance focused. This pattern repeats in the Word2Vec skip-gram model. Compared to the Wikipedia 2017 models, I see that all the models trained on the K-12 corpora exhibit different top adjectives, however the valence of the adjectives compared using the HGI lexicon are comparable. Words related to negativity tend to be closer to the man group vector, and words related to positivity tend to be closer to the woman group vector, in addition to the slightly more frequent presence of words related to appearances.

HGI Lexicon Man Group Vector All Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
barbaric	**-**	0.0646	judicious	?	0.2453	regimental	-	0.1826	barbaric		0.0646
gallant	+	0.0576	bald	*	0.2192	treacherous	-	0.1539	gallant	+	0.0576
crisp		0.0562	criminal	-	0.1909	petty	-	0.1494	crisp		0.0562
airy		0.0558	genius	+ ?	0.1884	devious	-	0.1434	airy		0.0558
bald	*	0.0551	benevolent	+	0.1735	criminal	-	0.1385	bald	*	0.0551
sane	+	0.0531	thoughtless	-	0.1722	driving		0.1384	sane	+	0.0531
stern	-	0.0526	civilized		0.1684	daring	+	0.1378	stern	-	0.0526

HGI Lexicon Woman Group Vector All Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
stupid	**- ?**	-0.0573	beautiful	*	-0.2629	circumspect		-0.1702	stupid	**- ?**	-0.0573
crazy	-	-0.0484	tasteless		-0.1957	sweet	+	-0.1570	crazy	-	-0.0484
weak		-0.0484	frivolous	-	-0.1826	gorgeous	+ *	-0.1559	weak		-0.0484
pretty	+ *	-0.0482	smooth		-0.1566	beautiful	+ *	-0.1475	pretty	+ *	-0.0482
hateful	-	-0.0441	soft		-0.1435	precocious		-0.1309	hateful		-0.0441
reasonable	+	-0.0403	realistic	+	-0.1412	soft		-0.1235	reasonable	+	-0.0403
difficult	-	-0.0400	sweet	+	-0.1368	feminine	*****	-0.1214	difficult	-	-0.0400

Table 4.1: Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

Words annotated with + have positive connotations, words annotated with - have negative connotations, words annotated with * have appearance related connotations, words annotated with ? have competency related connotation. Annotations in bold are from personal classification and not from the HGI lexicon or Garg et al. (2018). See footnote 1 for more details on how I classified adjectives.

When looking at more detailed adjective annotations from the HGI lexicon (the additional categories I chose for the analysis set are *Weak, Strong, Vice,* and *Virtue*), in the models

trained on the K-12 corpus (tables C.5 through C.10, as well as tables 4.2 and 4.3), the adjectives closest to the group vectors correspond with the attitudes seen in the models trained on the Wikipedia 2017 data. I am led to believe that the K-12 corpus overall reflects that men are generally written about in terms of their competence and personality traits, while women generally are written about in terms of their appearance and emotions. However, I am not comfortable making this claim strongly without a more consistent way to analyze if an adjective could be classified as describing emotions, personality, or any other trait besides positive-negative valence.

HGI Lexicon Man Group Vector Virtue Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
gallant	+	**0.0576**	benevolent	+	**0.1735**	daring	+	**0.1378**	gallant	+	**0.0576**
sane	+	**0.0531**	courteous	+	**0.1617**	heroic	+	**0.1214**	sane	+	**0.0531**
grand	+	**0.0415**	reliable	+	**0.1530**	loyal	+	**0.1208**	grand	+	**0.0415**
athletic	+ *	**0.0260**	venerable	+ ?	**0.1423**	resourceful	+ ?	**0.0931**	athletic	+ *	**0.0260**
popular	+	**0.0237**	rational	+	**0.1352**	decisive		**0.0916**	popular	+	**0.0237**
dynamic	+	**0.0214**	gallant	+	**0.1347**	progressive	+	**0.0910**	dynamic	+	**0.0214**
versatile	+	**0.0213**	versatile	+	**0.1299**	moderate	+	**0.0858**	versatile	+	**0.0213**

HGI Lexicon Woman Group Vector Virtue Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
reasonable	+	**-0.0403**	pure	+	**-0.0916**	attractive	+ *	**-0.0931**	reasonable	+	**-0.0403**
sincere	+	**-0.0396**	busy		**-0.0748**	elegant	+	**-0.0907**	sincere	+	**-0.0396**
unselfish	+	**-0.0341**	attractive	+ *	**-0.0685**	wholesome	+	**-0.0850**	unselfish	+	**-0.0341**
thoughtful	+ ?	**-0.0316**	friendly	+	**-0.0467**	gentle	+	**-0.0754**	thoughtful	+ ?	**-0.0316**
helpful	+	**-0.0310**	gentle	+	**-0.0426**	gracious	+	**-0.0600**	helpful	+	**-0.0310**
manly	+	**-0.0287**	conventional		**-0.0400**	thoughtful	+ ?	**-0.0587**	manly	+	**-0.0287**
capable	+ **?**	**-0.0256**	gracious	+	**-0.0376**	neat	+	**-0.0559**	capable	+	**-0.0256**

Table 4.2: HGI Lexicon Virtue Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

HGI Lexicon Man Group Vector Weak Adjective Similarities												
FastText			GloVe			Word2Vec CBOW			Word2Vec SG			
bland	-	**0.0510**	foolish	-	**0.1643**	unscrupulous	-	**0.0967**	bland	-	**0.0510**	
fickle	-	**0.0434**	vague	- ?	**0.1532**	desperate	-	**0.0878**	fickle	-	**0.0434**	
brittle	-	**0.0369**	irresponsible	-	**0.1433**	defensive	-	**0.0863**	brittle	-	**0.0369**	
meek	-	**0.0291**	fraudulent	-	**0.1027**	dishonest	-	**0.0632**	meek	-	**0.0291**	
narrow		**0.0253**	withdrawn		**0.0995**	fraudulent	-	**0.0592**	narrow		**0.0253**	
slender	*	**0.0216**	feeble	- *	**0.0935**	foolish	-	**0.0555**	slender	*	**0.0216**	
slim	*	**0.0202**	dull	-	**0.0854**	feeble	- *	**0.0450**	slim	*	**0.0202**	

HGI Lexicon Woman Group Vector Weak Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
desperate	-	**-0.0335**	feminine	*	**-0.1060**	feminine	*	**-0.1214**	desperate	-	**-0.0335**
shy		**-0.0323**	fickle	-	**-0.1035**	delicate	+	**-0.1197**	shy		**-0.0323**
miserable	-	**-0.0319**	hysterical	-	**-0.0988**	sentimental		**-0.1027**	miserable	-	**-0.0319**
ridiculous	-	**-0.0299**	illogical	-	**-0.0909**	bland	-	**-0.0910**	ridiculous	-	**-0.0299**
naive	-	**-0.0295**	brittle	-	**-0.0506**	silly	-	**-0.0845**	naive	-	**-0.0295**
fearful	-	**-0.0293**	insecure	-	**-0.0446**	slender	*	**-0.0787**	fearful	-	**-0.0293**
silly	- ?	**-0.0279**	gentle	+	**-0.0426**	timid		**-0.0761**	silly	-?	**-0.0279**

Table 4.3: HGI Lexicon Weak Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

Looking at Figures B.1 through B.4 as well as 4.1 below, I see that the general trend of distances between the closest adjectives to the man group vector and the woman group vector from the largest to the smallest corpora (in order: Wikipedia 2017, K-12, single books) become increasingly more "clumped together" in the FastText and Word2Vec models than in the GloVe models, where the graphed pattern becomes more "spread out". The "clumped together" pattern implies that the cosine similarity between the group word and the men/women group vectors are quite close, meaning that the "bias" of that word within

the model leans equally towards both groups. In both the FastText and Word2Vec models, the distance between a group vector and its closest adjectives in smaller corpora are less close than the distances between the group vector and it's closest adjectives in larger corpora. However, the opposite applies to the Glove model — the distance between a group vector and it's closest adjectives is much closer in smaller corpora.

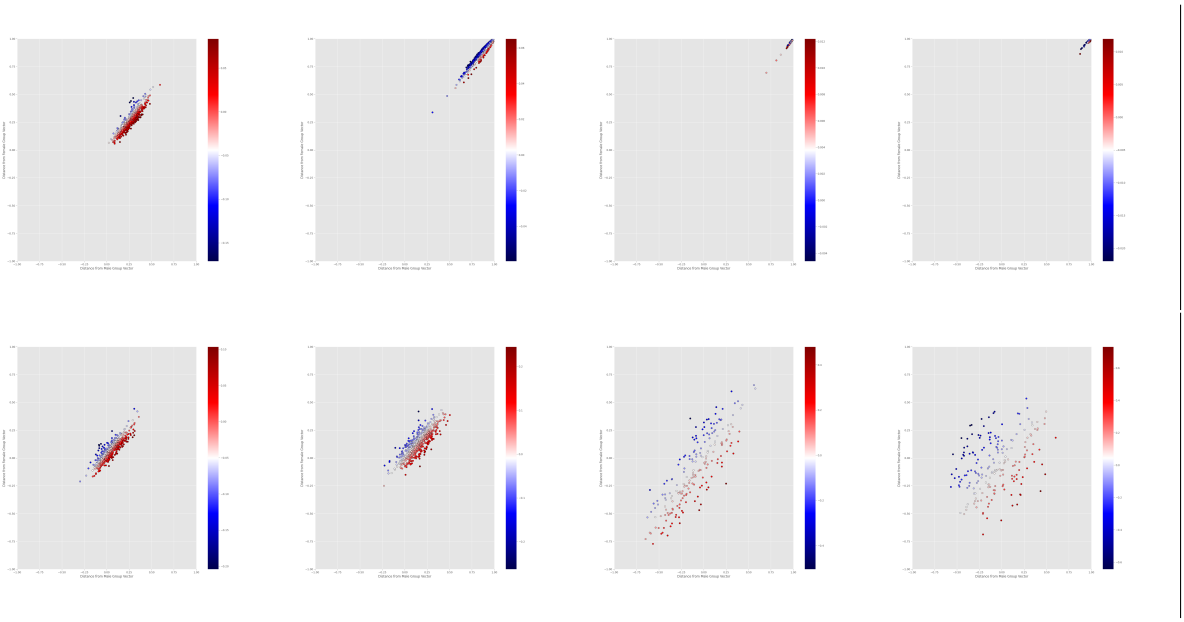

Figure 4.1: Cosine Similarity Comparison of All Adjectives in all FastText and Glove Models From Left to Right, Top to Bottom:
FastText: Wikipedia 2017, K-12, *Song of Achilles*, *Foundation*
Glove: Wikipedia 2017, K-12, *Song of Achilles*, *Foundation*

In answering the second question (how are different models sensitive to bias?) I look at all four models I trained and the differences in the valence or semantics of the adjectives closest to the group vectors in each one. For the FastText model, many of the words closest to the man vector in *Foundation* are some of the closest words to the woman vector in *Song of Achilles* and vice-versa. However, this can also likely be attributed to the distances between group vectors and words (which are comparatively smaller than the distances between words

and group vectors seen in the larger corpora), meaning that these words are not strongly or significantly biased in the direction of either group vector, which corresponds with the lack of spread I see in the graphs and numbers for both FastText and Word2Vec. I then conclude that FastText and Word2Vec do not pick up on social attitudes as strongly in smaller corpora.

I interpret this as the FastText and Word2Vec models pick up the bias of the evaluative adjectives less strongly in smaller corpora, while the GloVe model picks up biases more strongly in smaller corpora. As the vocabulary size and token counts will be lower in smaller corpora, words (such as the adjectives I am analysing) may be less common due to their lower rate of appearance. For GloVe, the co-occurance matrix created will more likely associate these words with the words they co-occur with, while FastText and Word2Vec will associate these words less strongly than GloVe does, as these words are less likely to be predicted in a given-context, or a context is not well-enough known to predict these words.

The GloVe models for both *Foundation* and *Song of Achilles* have a wider spread than that of the K-12 corpus, which itself has a wider spread than the Wikipedia 2017 model, meaning that the word vectors for specific words tend to generally lean closer to one group vector than the other compared to the larger corpora. It is worth noting that the model for *Song of Achilles* is less spread out than the one for *Foundation*.

To answer the third question, I look at the semantics of the adjectives that are closest to the group vectors. I see that the attitudes for the words closest to the woman group vectors do not necessarily reflect negative social stereotypes about women more in the model trained on *Foundation* than in the model trained on *Song of Achilles*, or even the entire K-12 Corpus. However, both individual books have less appearance focused adjectives closer to the woman group vector than the K-12 corpus. The adjectives closest to the man group vector for the K-12 corpus, *Song of Achilles*, and *Foundation* all seem to be around the same valence and from the same categories, spanning mainly competency and personality related adjectives, with less emphasis on physical appearance adjectives. My hypothesis was that the words closest to the group vectors from *Foundation* would reflect more negative stereotypes about

women than *Song of Achilles*, however based on my interpretation of the data, the result of this hypothesis is negative (or at least null) using the approach I took.

Compared to *Foundation*, the distance values seen for *Song of Achilles* (GloVe models for both) are less strong, likely indicating that the strength of the adjectives applying to a certain group vector aren't as strongly or closely associated as those in the *Foundation* model. Though the adjectives closest to the group vectors from both of these books have a similar valence, the closest adjectives are closer to the group vectors in *Foundation* than in *Song of Achilles*. My interpretation of this is that the adjectives we see closest to the group vectors in *Foundation* are used with more frequency than in *Song of Achilles*, which uses adjectives with similar valence to describe men and women, but likely less often (hence the smaller similarity values).

Based on the above, I believe more work needs to be done in analyzing books that may have more stereotypically gendered attitudes than others to come to a definite conclusion on if the text of one book is enough to determine the attitudes in that book using non-contextual word vectors as the measurement tool.

Chapter 5

CONCLUSION

In this thesis, I aimed to answer a few questions. What social biases are reflected in a corpus of K-12 literature? How are different non-contextual word embedding models sensitive to bias? And, how does the size of a corpus affect how models pick up social attitudes reflected in that corpus — is the text of one book enough data for a model to pick up reflected bias, and are the strength of the picked up biases different in corpora of different sizes?

With regards to the first question, when looking at the four different models trained on the corpus of K-12 literature (GloVe, FastText, Word2Vec CBOW, and Word2Vec Skip-Gram), I see that the adjectives that are closest in vector space to the group vector for women tend to be more positive and appearance related, while those closest to the group vector for men tend to be more negative and competency related. I have reason to believe that the words closest to the woman group vector are more emotion related and the words closest to the man group vector are more personality related, but I am not willing to make that claim strongly until I am able to further analyze how adjectives fit into different categories.

In considering how different non-contextual word embedding models are sensitive to bias, there are a couple things I noticed. First, as the corpora got smaller, the adjectives and kinds of adjectives closest to group vectors from the FastText model diverged significantly from the other three models. Secondly, as corpora got smaller, the distance between group vectors and their closest adjectives got smaller in GloVe models, while the distance got larger in other models. All models had completely different words closest to the group vectors compared to the other models in the smallest two corpora (the single books).

When looking at the results for the models trained on a single book and comparing them to the attitudes they have been determined to have from outside critique, I see that

Foundation, which is held to have more stereotypical gender attitudes, has more strong similarities between adjectives and group vectors than *Song of Achilles*, which I interpret as *Foundation* presenting gendered words more strongly than *Song of Achilles*, which presents genders more equally. However, the results of directly looking at the closest words to the group vectors for men and women in both the books are inconclusive, as the closest adjectives to the vectors in both groups have the same valence. I believe more work needs to be done on analyzing group vector and adjective similarity in single books to determine if they are sufficient data for models to interpret their attitudes accurately when trained on.

5.1 Ethical Considerations

One of the first things that needs to be considered in doing an analysis based on gender is *why* gender is being used as a variable in the research, and if it is actually an useful axis for the research being performed (Larson, 2017a). In this work, I wanted to analyze word embedding biases on the axis of gender to understand better what gender stereotypes are found in literature assigned to children, and the strength of those stereotypes. This seemed valuable to me because the research on the impact of literature on children, especially with regards to the gendered stereotypes present in what they read, point to the benefit of varied and diverse perspectives in how people are presented in the assigned curriculum.

It is incredibly important to note that gender is not a binary. However, I only look at binary representations of gender in this work for a couple reasons. First of all, in adopting methods of analyzing attitudes towards group vectors based on adjective sentiment and semantics, it was hard to create a group vector for non-binary genders that was adequatly representative of the concept of genders outside the binary, especially because gender-neutral pronouns such as *they* and *them* are used in contexts outside of a singular person with a non-binary gender. In addition, very few gender-neutral variations on the group words chosen for men and women existed (e.g. using *parent* as an analogous word to *father* or *mother* would be possible, but an equivilent term to *aunt* or *uncle* is not in common usage). Because of these reasons in part, I would not have been able to accurately run the same analysis using

a non-binary group vector.

If methods like the ones I present here are going to be used at all in creating curricula for children in grades K-12, the *way* the information is used needs to be thought about. For example, under what framework should works be chosen to diversify the perspectives that are shown in an already existing curriculum? How is sufficient diversification determined while books are being chosen — i.e. is a threshold for cosine distance enough? What words are being chosen to evaluate the biases present? Though I took the adjectives I used to evaluate group vector similarity on from both previous work and lexicons with sentiment information, how can I be sure that those were still the right words to have run my analysis with? Would there have been other words that might have provided different information to my conclusions, had they been selected? How *should* these evaluative words and measures be chosen? What sort of information *cannot* be derived from an analysis in the vein of what I presented in this thesis?

5.2 Challenges and Reflections

When starting this project, I had a far more ambitious goal. I initially wanted to take a much larger corpus of available books seperated into categories(i.e. awarded for feminist literature, childrens books, written by black authors, various genre, year of publication, etc.), and try to understand general patterns in different categories of literature. I also wanted to look at if there were differences in literature written by different author demographics over time, and how those may correspond with historical events occuring within a specific year, being particularly influenced by Garg et al. (2018). I was hoping to look for patterns in how historical events shape the writing of certain groups, and if that would lend any clues on how the books that a group of people write reflect their attitudes and in-group feelings towards a certain situation.

I struggled majorly with gathering data on book groups and author demographics. The easiest groups to find were book awards and books recommended by certain states to be assigned to K-12 students. Given the availability of the latter, I decided to focus this project

on the 10,000 or so books I confidently found in this category.

The next roadblock I encountered was text availability. I had planned to use the resources provided by Hathitrust to access all the text I needed, as through them, I had access to millions of texts. What I did not expect was that less than 1000 of the texts I had collected were available to me through Hathitrust. I spent some time trying to see if I could access book previews through Google Books or The Open Library, and if I could run my analysis on all 10,000 of my books, but on previews instead of full-text, but that work turned out to be incredibly time consuming and not within the time-scope of this thesis.

Given what I had access to and the questions I originally wanted to answer, I decided to pivot slightly and focus on understanding what my limited data-set could tell me about social attitudes, and in addition, if the limitations of a small dataset had implications on understanding attitudes in text. Given that much of the research I had done in my literature review on the impacts of bias in literature focused on the impact on children, the pivot seemed natural.

5.3 Future Work

In the future, I hope to continue this work in a few ways. I would like to see what the impact is when the parameters regarding the dimensionality of vectors, size of windows, and minimum word counts are changed in these non-contextual models, and I would also like to continue the analysis on single books, finding more books with critiques that determine that they have certain biases, and get more conclusive evidence on if these non-contextual models can be accurately used to determine the attitudes present in the books.

I'd like to do this same analysis using contextual word embedding models instead of non-contextual. The Hathitrust Research Center data capsules that were required to be used for analysis due to text access requirements did not have GPUs, and so were not feasible to run deep learning models such as ELMo and BERT on. Given the research on these models and their ability to capture attitudes deeper than non-contextual word embedding models, I would like to see if my conclusions using this corpus change at all.

I would like to gain access to a higher number of texts, and perform some of the analysis I had originally intended on all 10,000 K-12 books to get a more comprehensive picture of attitudes reflected in these texts as a whole. The ISBNs for this text were categorized into categories such as the state they were recommended by or if they were given an award (such as one for best LGBTQ+ or feminist literature). I would like to see if there are any patterns in different categories of childrens books, and given the research on how children internalize attitudes from literature, if assigning books from different groups or that have been given different awards would diversify the literary attitudes in a curriculum.

I would also like to broaden this further and answer the first research question I posed in my previous section — if historical events shape the writings of certain demographic groups, and what I can learn from the writing of a group over time, especially marginalized and underrepresented groups.

As noted earlier, certain adjectives (such as *feminine*), may be implied as a descriptor for one group, and so might be found to be closer to the other group because it isn't used in contexts for the group it is implicitly understood to categorize. I am interested in studying what these words might be and how that may impact understanding when analyzing gender biases in text. In addition, adjectives aren't the only kinds of words that can encode attitudes about agents in literature. It could be interesting to look at words related to agency (similar to the work done by Field and Tsvetkov (2019) on power, sentiment, and agency of entities in text), and the implications of this on analyzing literature.

Finally, and on a slightly different note, I want to run experiments on literary style transfer and bias. For example, given an author who, through literary critique, is thought to be particularly biased in a certain way, would applying an author-level style transfer of that author's style to another text with little to no biases create biases in that text? And if so, in looking at the neurons on a contextual/deep-learning style transfer model, can I understand where certain biases are encoded in the model?

BIBLIOGRAPHY

Ahmed K Al-Rawi. Foreign Policy and its Impact on Arab Stereotypes in English Popular Fiction of the 1970s-80s. page 23.

American Library Association. The Song of Achilles — Awards & Grants, 2013. URL http://www.ala.org/awardsgrants/song-achilles.

American Library Association. The Song of Achilles — Awards & Grants, 2017. URL http://www.ala.org/awardsgrants/song-achilles-0.

George Aulisio. Madeline Miller to Receive Distinguished Author Award — Royal News: November 26 2019. URL https://news.scranton.edu/articles/2019/08/gen-miller-award.shtml.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1023. URL https://www.aclweb.org/anthology/P14-1023.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. *arXiv:1904.08783 [cs]*, April 2019. URL http://arxiv.org/abs/1904.08783.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv:2005.14050 [cs]*, May 2020. URL http://arxiv.org/abs/2005.14050.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*, June 2017. URL http://arxiv.org/abs/1607.04606.

Danushka Bollegala, Kohei Hayashi, and Ken-ichi Kawarabayashi. Learning linear transformations between counting-based and prediction-based word embeddings. *PLOS ONE*, 12 (9):e0184544, September 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0184544. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184544.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL http://dl.acm.org/citation.cfm?id=3157382.3157584.

Ryan Britt. A Science Fiction Halo Rests Slantedly Over Isaac Asimov's Amiable Head, January 2016. URL https://www.tor.com/2016/01/02/a-science-fiction-halo-rests-slantedly-over-isaac-asimovs-amiable-head/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July 2020. URL http://arxiv.org/abs/2005.14165.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. *Understanding the Origins of Bias in Word Embeddings*. 2018.

CA Dept of Education. Recommended Literature List (CA Dept of Education). URL http://www3.cde.ca.gov/reclitlist/search.aspx.

Caldecott Medal & Honor Books, 1938-Present. Caldecott Medal & Honor Books, 1938-Present, November 1999. URL http://www.ala.org/alsc/awardsgrants/bookmedia/caldecottmedal/caldecotthonors/caldecottmedal.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230. URL http://arxiv.org/abs/1608.07187.

Rosie Cima. The Gender Balance of The New York Times Best Seller List, November 2019. URL https://pudding.cool/2017/06/best-sellers/index.html.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA, July 2008. Association for Computing Machinery. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390177. URL https://doi.org/10.1145/1390156.1390177.

Paul Deane. A Century of Xenophobia in Fiction Series for Young People. *Journal of Youth Services in Libraries*, 3(2):117–27, 1990.

Joan DelFattore. Controversial Narratives Iin the Schools. In *Narrative Impact : Social and Cognitive Foundations*, pages 131–257. Psychology Press, January 2003. ISBN 978-1-4106-0664-8.

Sunipa Dev and M Phillips. Attenuating Bias in Word Vectors. page 9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL http://arxiv.org/abs/1810.04805.

Steve Donoghue. The Song of Achilles by Madeline Miller. URL https://www.stevedonoghue.com/steves-reviews//the-song-of-achilles-by-madeline-miller.

Anjalie Field and Yulia Tsvetkov. Entity-Centric Contextual Affective Analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2550–2560, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1243. URL https://www.aclweb.org/anthology/P19-1243.

Sally Flint. Madeline Miller's The Song of Achilles - Book review and book discussion questions, October 2020. URL http://www.sallyflint.com/1/post/2020/06/madeline-millers-the-song-of-achilles.html.

Joel Escudé Font and Marta R. Costa-jussà. Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. *arXiv:1901.03116 [cs]*, June 2019. URL http://arxiv.org/abs/1901.03116.

Batya Friedman, David G. Hendry, and Alan Borning. A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction*, 11(2):63–125, 2017. ISSN 1551-3955. doi: 10.1561/1100000015. URL http://dx.doi.org/10.1561/1100000015.

Jay Gabler. What to Make of Isaac Asimov, Sci-Fi Giant and Dirty Old Man? — Literary Hub, May 2020. URL https://lithub.com/what-to-make-of-isaac-asimov-sci-fi-giant-and-dirty-old-man/.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/content/115/16/E3635.

Martin Gerlach and Francesc Font-Clos. *A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics*. 2018.

Hila Gonen and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv:1903.03862 [cs]*, September 2019. URL http://arxiv.org/abs/1903.03862.

Siobhan Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. Novel2Vec: Characterising 19th Century Fiction via Word Embeddings. page 13.

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. Exploring the Role of Gender in 19th Century Fiction Through the Lens of Word Embeddings. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*, volume 10318, pages 358–364. Springer International Publishing, Cham, 2017. ISBN 978-3-319-59887-1 978-3-319-59888-8. doi: 10.1007/978-3-319-59888-8_30. URL http://link.springer.com/10.1007/978-3-319-59888-8_30.

Melanie C. Green, Jeffrey J. Strange, and Timothy C. Brock, editors. *Narrative Impact : Social and Cognitive Foundations*. Psychology Press, January 2003. ISBN 978-1-4106-0664-8. doi: 10.4324/9781410606648. URL https://www.taylorfrancis.com/books/9781410606648.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464, 19980701. ISSN 1939-1315. doi: 10.1037/0022-3514.74.6.1464. URL https://psycnet.apa.org/fulltext/1998-02892-004.pdf.

The Book Habit. Review: The Song of Achilles by Madeline Miller, December 2018. URL https://thebookhabit.co.uk/2018/12/08/review-the-song-of-achilles-by-madeline-miller/.

Osian Haines. Reason most absurd: Foundation and Patriarchy, July 2012. URL http://osianh.blogspot.com/2012/07/foundation-and-patriarchy.html.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Newberry Honors. Newbery medals and honors 1922-present.pdf. URL http://www.ala.org/alsc/sites/ala.org.alsc/files/content/awardsgrants/bookmedia/newberymedal/newberywinners/newbery%20medals%20and%20honors%201922-present.pdf.

Indiana Department of Education. Indiana Department of Education. URL https://www.doe.in.gov/sites/default/files/standards/sample-texts-revised-08-05-15.pdf.

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.

Ronald N. Jacobs. The Narrative Integration of Personal and Collective Identity in Social Movements. In *Narrative Impact : Social and Cognitive Foundations*, pages 205–229. Psychology Press, January 2003. ISBN 978-1-4106-0664-8.

Arit John. Science Fiction's Sexism Problem, August 2013. URL https://www.theatlantic.com/culture/archive/2013/08/speculative-fiction-has-sexism-problem/312355/.

Claudia Durst Johnson and Vernon Elso Johnson. *The Social Impact of the Novel: A Reference Guide*. Greenwood Publishing Group, 2002. ISBN 978-0-313-31818-4.

Author keitakurita. Paper Dissected: "Glove: Global Vectors for Word Representation" Explained, April 2018. URL http://mlexplained.com/2018/04/29/paper-dissected-glove-global-vectors-for-word-representation-explained/.

Svetlana Kiritchenko and Saif M. Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *arXiv:1805.04508 [cs]*, May 2018. URL http://arxiv.org/abs/1805.04508.

Sosuke Kobayashi. Soskek/bookcorpus, November 2019. URL https://github.com/soskek/bookcorpus.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. *Measuring Bias in Contextualized Word Representations*. 2019.

University of Oslo Language Technology Group. NLPL word embeddings repository. URL http://vectors.nlpl.eu/repository/.

Brian Larson. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April 2017a. Association for Computational Linguistics. doi: 10.18653/v1/W17-1601. URL https://www.aclweb.org/anthology/W17-1601.

Brian Larson. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April 2017b. Association for Computational Linguistics. doi: 10.18653/v1/W17-1601. URL https://www.aclweb.org/anthology/W17-1601.

Allan Luke, Janine Cooke, and Carmen Luke. The Selective Tradition in Action: Gender Bias in Student Teachers' Selections of Children's Literature. *English Education*, 18(4): 209–218, 1986. ISSN 0007-8204. URL https://www.jstor.org/stable/40172624.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. page 9.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, September 2013. URL http://arxiv.org/abs/1301.3781.

Minnesota Department of Education. Minnesota Department of Education English Language Arts. URL https://education.mn.gov/mde/dse/stds/ela/.

Saif M. Mohammad. R: The NRC Valence, Arousal, and Dominance Lexicon, 2018. URL http://search.r-project.org/library/textdata/html/lexicon_nrc_vad.html.

Malvina Nissim, Rik van Noord, and Rob van der Goot. *Fair Is Better than Sensational:Man Is to Doctor as Woman Is to Doctor*. 2019.

NYSED. NYSED ELA Resource Guide. URL http://www.p12.nysed.gov/guides/ela/part1b.pdf.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv:1802.05365 [cs]*, March 2018. URL http://arxiv.org/abs/1802.05365.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24.

Diane Estelle Railsback. Reading for equality: An examination of gender-bias in children's literature. 1993.

Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

David Rozado. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLoS ONE*, 15(4), April 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0231189. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7173861/.

Natalie Schluter. The Word Analogy Testing Caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2039. URL https://www.aclweb.org/anthology/N18-2039.

Alex SL. PhyloBotanist: Isaac Asimov's Foundation, September 2015. URL http://phylobotanist.blogspot.com/2015/09/isaac-asimovs-foundation.html.

Michael D. Slater. Entertainment Education and the Persuasive Impact of Narratives. In *Narrative Impact : Social and Cognitive Foundations*, pages 157–283. Psychology Press, January 2003. ISBN 978-1-4106-0664-8.

Philip Stone, Robert Bales, J. Namenwirth, and Daniel Ogilvie. The General Inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7:484–498, October 2007. doi: 10.1002/bs.3830070412.

Philip J. Stone and Earl B. Hunt. A computer approach to content analysis: Studies using the General Inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA, May 1963. Association for Computing Machinery. ISBN 978-1-4503-7880-2. doi: 10.1145/1461551.1461583. URL https://doi.org/10.1145/1461551.1461583.

Jeffery J. Strange. How Fictional Tales Wag Real-World Beliefs: Models and Mechanisms of Narrative Influence. In *Narrative Impact : Social and Cognitive Foundations*, pages 263–287. Psychology Press, January 2003. ISBN 978-1-4106-0664-8.

Chris Sweeney and Maryam Najafian. A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1162. URL https://www.aclweb.org/anthology/P19-1162.

CHARLES TEMPLE. "What if Beauty Had Been Ugly?" Reading against the Grain of Gender Bias in Children's Books. *Language Arts*, 70(2):89–93, 1993. ISSN 0360-9170. URL https://www.jstor.org/stable/41482067.

The Hugo Awards. 1966 Hugo Awards, July 2007. URL http://www.thehugoawards.org/hugo-history/1966-hugo-awards/.

Liz Thomson. 2012 Orange Prize Goes to 'The Song of Achilles'. URL https://www.publishersweekly.com/pw/by-topic/industry-news/awards-and-prizes/article/52157-2012-orange-prize-goes-to-the-song-of-achilles.html.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *arXiv:1909.10430 [cs]*, October 2019. URL http://arxiv.org/abs/1909.10430.

Gerhard Wohlgenannt, Ekaterina Chernyak, Dmitry Ilvovsky, Ariadna Barinova, and Dmitry Mouromtsev. Relation Extraction Datasets in the Digital Humanities Domain and their Evaluation with Word Embeddings. *arXiv:1903.01284 [cs]*, March 2019. URL http://arxiv.org/abs/1903.01284.

World Economic Forum. *The Global Gender Gap Report: 2017*. World Economic Forum, Geneva, 2017. ISBN 978-1-944835-12-5.

Wyoming Department of Education. Wyoming Department of Education. URL http://edu.wyoming.gov/downloads/standards/ela_appendix_b.pdf.

YALSA. YALSA Book Finder. URL http://booklists.yalsa.net/.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://www.aclweb.org/anthology/N18-2003.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Contextualized Word Embeddings. pages 629–634, January 2019. doi: 10.18653/v1/N19-1064.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books*. 2015.

Appendix A

GROUP VECTOR AND TARGET WORDS

A.1 Group Vector Words from Garg et al.

A.1.1 Man Group Vector

boy, boys, brother, brothers, father, fathers, he, him, himself, his, male, males, man, men, nephew, nephews, son, sons, uncle, uncles

A.1.2 Woman Group Vector

aunt, aunts, daughter, daughters, female, femen, girl, girls, her, hers, herself, mother, mothers, niece, nieces, she, sister, sisters, woman, women

A.2 Adjectives

A.2.1 All Adjectives

abrupt, accessible, active, adaptable, admirable, adventurous, affected, affectionate, aggressive, agreeable, aimless, airy, alert, alluring, aloof, ambitious, amiable, amusing, analytical, angry, anxious, appreciative, apt, arbitrary, arrogant, artful, articulate, artificial, artistic, ascetic, aspiring, assertive, astute, athletic, attractive, autocratic, awkward, balanced, bald, barbaric, beautiful, benevolent, bewildered, bitter, bizarre, bland, blunt, blushing, boisterous, boyish, brilliant, brittle, brutal, busy, calculating, callous, calm, capable, careless, caring, casual, cautious, cerebral, challenging, changeable, charming, cheerful, childish, circumspect, civilized, clean, clever, clumsy, coarse, cold, colorless, commonplace, compassionate, competitive, complacent, complaining, complex, complicated, conceited, conciliatory, confident, confidential, confused, conscientious, conservative, considerate, constant, contem-

plative, contemptible, contented, contradictory, conventional, cool, cooperative, courageous, courteous, cowardly, crafty, crazy, creative, criminal, crisp, critical, crude, cruel, cultured, curious, cynical, daring, deceitful, decent, deceptive, decisive, dedicated, deep, defensive, deliberate, delicate, demanding, dependable, dependent, desperate, destructive, determined, devious, difficult, dignified, directed, dirty, discerning, disciplined, disconcerting, discontented, discouraging, discreet, dishonest, disloyal, disobedient, disorderly, dissatisfied, dissolute, distrustful, disturbing, dogmatic, dominant, dominating, dramatic, dreamy, driving, dry, dull, dutiful, dynamic, earnest, educated, effeminate, efficient, elegant, eloquent, emotional, energetic, enterprising, enthusiastic, envious, erratic, exciting, expedient, experimental, extraordinary, extravagant, extreme, fair, faithful, faithless, false, fanatical, fanciful, fashionable, fat, fearful, feeble, feminine, fickle, fiery, fixed, flexible, focused, foolish, forceful, forgetful, forgiving, formal, fraudulent, friendly, frightening, frivolous, frugal, gallant, generous, genius, gentle, genuine, gloomy, gorgeous, gracious, grand, greedy, grim, handsome, hasty, hateful, haughty, headstrong, healthy, hearty, helpful, heroic, homely, honest, honorable, hostile, humble, humorous, hurried, hypnotic, hysterical, idealistic, ignorant, illogical, imaginative, imitative, immature, impassive, impatient, impersonal, impractical, impressionable, impressive, imprudent, impulsive, incorruptible, independent, indifferent, individualistic, indulgent, industrious, inert, infantile, informal, ingenious, inhibited, initiative, inoffensive, inquiring, inquisitive, insecure, insulting, intelligent, intense, intolerant, intuitive, inventive, invisible, irrational, irresponsible, irritable, jolly, judicious, kind, kind ,disorganized, knowledge, lazy, leisurely, liberal, logical, loud, lovable, loyal, luminous, lyrical, malicious, manly, mannered, masculine, maternal, mature, mechanical, meditative, meek, mellow, methodical, mild, mischievous, miserable, misguided, mistaken, moderate, modern, modest, monstrous, moody, morbid, muscular, mystical, naive, narrow, nationalistic, natural, neat, nervous, neurotic, neutral, noisy, obedient, objective, obliging, obnoxious, observant, obvious, odd, open, oppressed, optimistic, orderly, ordinary, organized, original, outgoing, outrageous, outspoken, painstaking, passionate, passive, paternal, patient, patriotic, peaceable, peaceful, peculiar, pedantic, persevering, persistent, persuasive, perverse, pessimistic, petty, physical, placid,

playful, pleasant, plump, poised, polished, political, pompous, popular, possessive, practical, praising, precise, precocious, predatory, prejudiced, preoccupied, presumptuous, pretentious, pretty, private, profligate, profound, progressive, protective, proud, providential, prudent, punctual, pure, quarrelsome, queer, questioning, quick, quiet, quitting, rational, reactionary, reactive, realistic, reasonable, rebellious, reckless, reflective, regimental, relaxed, reliable, religious, repressed, resentful, reserved, resourceful, respectful, responsible, responsive, restless, restrained, retiring, ridiculous, rigid, robust, romantic, rude, ruined, rustic, sagacious, sage, sane, sarcastic, scholarly, scornful, scrupulous, secure, sedentary, selfish, sensitive, sensual, sentimental, serious, severe, shallow, sharing, shrewd, shy, silent, silly, simple, sincere, skeptical, skillful, slender, slim, slow, sly, smart, smooth, sober, sociable, soft, solemn, solid, solitary, sophisticated, sordid, spontaneous, sporting, stable, steadfast, steady, stern, stiff, stoic, stout, strict, strong, stubborn, studious, stupid, subjective, submissive, subtle, superficial, superstitious, surprising, suspicious, sweet, sympathetic, systematic, tactful, talkative, tasteless, tense, thankless, thin, thorough, thoughtful, thoughtless, thrifty, tidy, timid, tolerant, tough, traditional, transparent, treacherous, troublesome, trusting, ugly, unaffected, unassuming, unchanging, understanding, unfathomable, unfriendly, ungrateful, unhealthy, unkind, unprincipled, unreliable, unrestrained, unscrupulous, unselfish, unstable, upright, vague, venerable, venomous, versatile, vindictive, voluptuous, vulnerable, warm, wary, weak, whimsical, wholesome, willful, winning, wise, withdrawn, witty, working, worrying, youthful

A.2.2 Competency Adjectives from Garg et al.

adaptable, analytical, apt, astute, brilliant, clever, discerning, genius, imaginative, ingenious, inquiring, inquisitive, intelligent, intuitive, inventive, judicious, logical, luminous, precocious, reflective, resourceful, sagacious, sage, shrewd, smart, thoughtful, venerable, wise

A.2.3 Physical Appearance Adjectives from Garg et al.

alluring, athletic, attractive, bald, beautiful, blushing, fashionable, fat, feeble, gorgeous, handsome, healthy, homely, muscular, plump, pretty, sensual, slender, slim, stout, strong ,

thin, ugly, voluptuous, weak

A.2.4 Negative Adjectives from HGI

abrupt, aggressive, aimless, aloof, angry, arbitrary, arrogant, artificial, autocratic, awkward, bitter, bizarre, bland, blunt, boisterous, brittle, callous, careless, childish, clumsy, coarse, commonplace, competitive, complex, contemptible, contradictory, cool, crafty, crazy, criminal, crude, cruel, cynical, deceitful, deceptive, defensive, dependent, desperate, destructive, devious, difficult, dirty, dishonest, disobedient, disorganized, dissatisfied, distrustful, dull, envious, expedient, extravagant, false, fanatical, fearful, feeble, fickle, foolish, fraudulent, frivolous, gloomy, grim, haughty, homely, hostile, hysterical, ignorant, illogical, immature, impatient, impersonal, impulsive, indifferent, insecure, invisible, irrational, irresponsible, irritable, lazy, malicious, meek, mischievous, miserable, mistaken, monstrous, moody, naive, nervous, neurotic, obnoxious, odd, outrageous, peculiar, perverse, pessimistic, petty, pompous, presumptuous, pretentious, quarrelsome, queer, reactionary, reactive, rebellious, reckless, resentful, restless, ridiculous, rigid, rude, sarcastic, scornful, sedentary, selfish, severe, shallow, silly, skeptical, sly, stern, stubborn, submissive, superficial, superstitious, suspicious, tense, thoughtless, treacherous, troublesome, ugly, unfriendly, ungrateful, unhealthy, unkind, unreliable, unscrupulous, unstable, vague, venomous, wary

A.2.5 Positive Adjectives from HGI

accessible, adaptable, admirable, adventurous, affectionate, agreeable, amiable, appreciative, apt, astute, athletic, attractive, benevolent, brilliant, capable, casual, cheerful, clever, compassionate, confident, conscientious, considerate, courageous, courteous, creative, daring, decent, delicate, dependable, dignified, discreet, dynamic, earnest, efficient, elegant, eloquent, energetic, enthusiastic, extraordinary, faithful, fashionable, fiery, flexible, friendly, frugal, gallant, generous, genius, gentle, genuine, gorgeous, gracious, grand, handsome, helpful, heroic, honorable, humble, humorous, imaginative, impressive, industrious, ingenious, inquisitive, intelligent, jolly, knowledge, logical, loyal, luminous, lyrical, manly, mellow, mild,

moderate, modest, neat, obedient, optimistic, outgoing, painstaking, passionate, patriotic, peaceable, peaceful, persuasive, playful, popular, precise, profound, progressive, protective, proud, prudent, punctual, pure, rational, realistic, reasonable, reliable, resourceful, respectful, responsive, robust, romantic, sage, sane, scrupulous, sensitive, serious, shrewd, sincere, skillful, sober, sociable, sophisticated, stable, steadfast, steady, studious, subtle, sympathetic, thorough, thoughtful, thrifty, tolerant, unselfish, upright, venerable, versatile, whimsical, wholesome, willful, witty

A.2.6 Active Adjectives from HGI

adventurous, aggressive, alert, ambitious, analytical, angry, busy, competitive, creative, decisive, destructive, dynamic, energetic, experimental, helpful, initiative, intense, mischievous, mistaken, obedient, persistent, quick, reactive

A.2.7 Passive Adjectives from HGI

aimless, aloof, bitter, casual, changeable, cool, curious, defensive, dependent, disorganized, dull, faithful, fearful, feeble, gentle, gloomy, humble, hysterical, insecure, intuitive, irrational, irritable, lazy, miserable, naive, nervous, neutral, optimistic, passionate, passive, peaceful, placid, reactionary, sentimental, shy, silent, timid

A.2.8 Emotional Adjectives from HGI

affectionate, aggressive, angry, bitter, callous, cheerful, compassionate, confident, crazy, creative, curious, desperate, enthusiastic, fearful, fickle, friendly, gloomy, haughty, hostile, hysterical, indifferent, irritable, jolly, miserable, moody, nervous, neurotic, passionate, proud, resentful, restless, romantic, scornful, sensitive, sentimental, shy, solemn, sympathetic, tense, thoughtless, wary

A.2.9 Hostile Adjectives from HGI

aggressive, angry, bitter, callous, competitive, contemptible, criminal, cruel, cynical, deceitful, deceptive, defensive, destructive, devious, disobedient, dissatisfied, hostile, irritable, malicious, monstrous, obnoxious, quarrelsome, reactive, rebellious, resentful, sarcastic, scornful, sly, stern, stubborn, suspicious, tense, tough, treacherous, venomous

A.2.10 Knowledge Adjectives from HGI

analytical, astute, complex, informal, ingenious, intuitive, knowledge, logical, neutral, resourceful, sage, subjective, vague

A.2.11 Negative Affect Adjectives from HGI

arbitrary, arrogant, awkward, cruel, feeble, grim, irritable, malicious, obnoxious, reactionary, severe, stern, troublesome, ugly, vindictive

A.2.12 Positive Affect Adjectives from HGI

adventurous, brilliant, compassionate, decent, dependable, enthusiastic, gentle, grand, sympathetic

A.2.13 Strong Adjectives from HGI

adaptable, administrator, adventurous, aggressive, alert, ambitious, apt, arrogant, athletic, autocratic, blunt, boisterous, busy, capable, clever, confident, courageous, daring, decisive, dependable, destructive, dignified, dominant, dynamic, earnest, efficient, energetic, enthusiastic, fiery, gallant, genius, grand, haughty, heroic, impressive, industrious, initiative, instructor, intelligent, intense, knowledge, manager, manly, masculine, methodical, monstrous, muscular, painstaking, passionate, persistent, police, popular, profound, protective, proud, reliable, resourceful, robust, sage, severe, sheriff, shrewd, sober, solid, spontaneous, stable, steadfast, steady, stern, stiff, stubborn, systematic, thorough, tough, upright

A.2.14 Weak Adjectives from HGI

awkward, bland, brittle, defensive, delicate, dependent, desperate, dishonest, disorganized, dull, fearful, feeble, feminine, fickle, foolish, fraudulent, gentle, humble, hysterical, ignorant, illogical, immature, insecure, irresponsible, lazy, meek, miserable, modest, naive, narrow, nervous, passive, ridiculous, sentimental, shallow, shy, silly, slender, slim, submissive, subtle, superficial, thin, timid, unhealthy, unreliable, unscrupulous, unstable, vague, vulnerable, withdrawn

A.2.15 Vice Adjectives from HGI

arbitrary, arrogant, artificial, awkward, bizarre, careless, childish, clumsy, contemptible, crafty, crazy, cruel, cynical, deceitful, destructive, difficult, dirty, disobedient, disorganized, dominant, envious, extravagant, false, fanatical, feeble, foolish, grim, homely, ignorant, impatient, impersonal, impulsive, insecure, irrational, irresponsible, lazy, monstrous, naive, obnoxious, odd, outrageous, peculiar, perverse, petty, presumptuous, pretentious, queer, reckless, ridiculous, rigid, rude, selfish, severe, silly, skeptical, sly, stubborn, superficial, superstitious, timid, treacherous, troublesome, ugly, unfriendly, ungrateful, unkind, unreliable, unscrupulous, unstable, vulnerable, withdrawn

A.2.16 Virtue Adjectives from HGI

accessible, adaptable, admirable, ambitious, amiable, apt, astute, athletic, attractive, benevolent, busy, capable, casual, cautious, clever, compassionate, competitive, conscientious, considerate, conventional, courageous, courteous, creative, daring, decent, decisive, dependable, discreet, dynamic, efficient, elegant, eloquent, extraordinary, faithful, friendly, gallant, generous, gentle, genuine, gracious, grand, handsome, helpful, heroic, honorable, humble, imaginative, impressive, ingenious, intelligent, loyal, manly, moderate, modest, neat, peaceable, peaceful, persistent, popular, precise, profound, progressive, prudent, punctual, pure, rational, reasonable, reliable, resourceful, respectful, sane, scrupulous, shrewd, sincere, skill-

ful, sociable, solemn, sophisticated, spontaneous, subtle, thorough, thoughtful, tolerant, unselfish, venerable, versatile, wholesome

Appendix B

GRAPHS

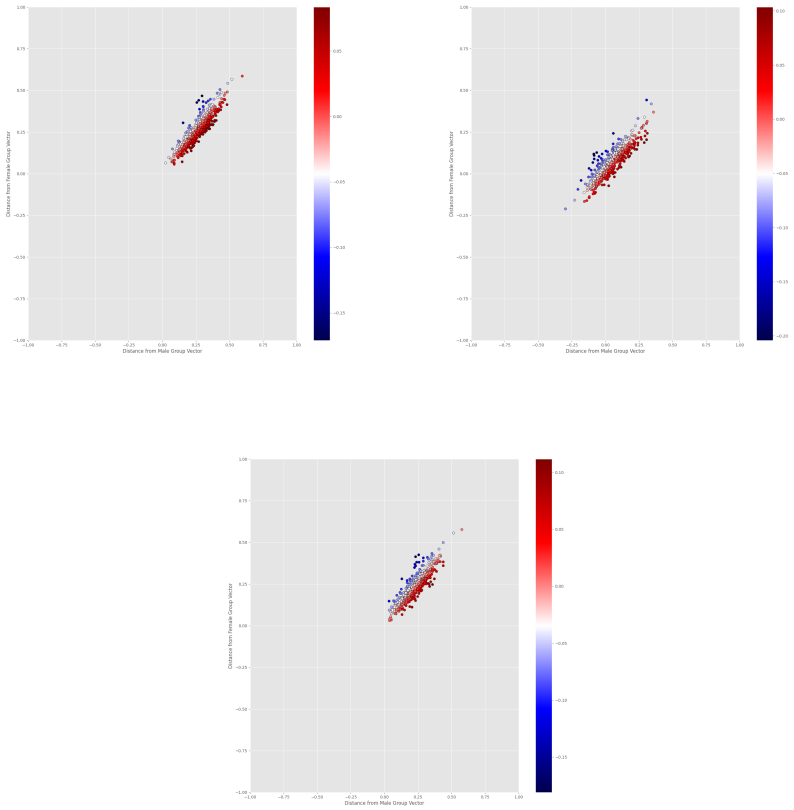

Figure B.1: Cosine Similarities of Adjectives to Gender Group Vectors in Wikipedia 2017 Pretrained Models

From Left to Right, Top to Bottom: FastText, GloVe, Word2Vec SG

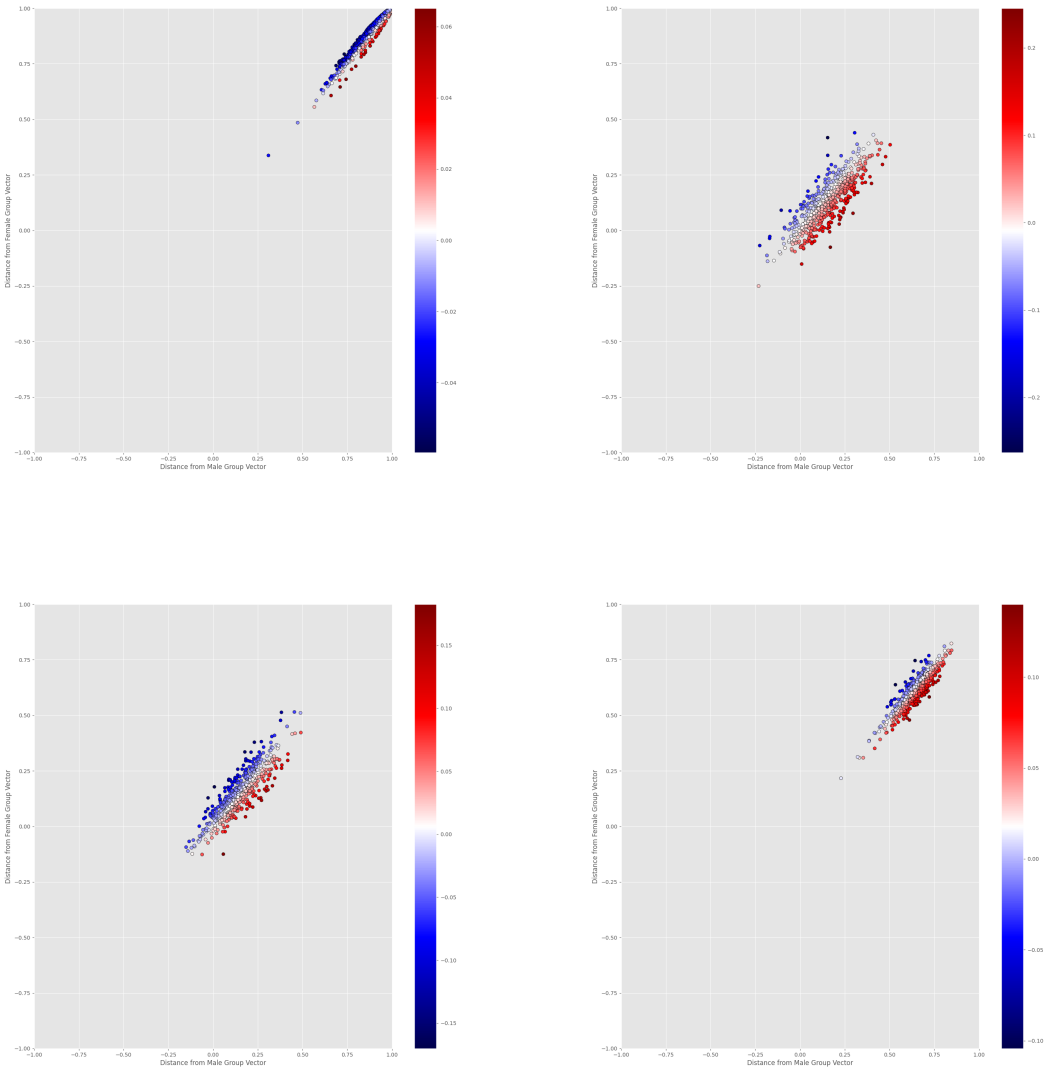

Figure B.2: Cosine Similarities of Adjectives to Gender Group Vectors in the K-12 Corpus From Left to Right, Top to Bottom: FastText, GloVe, Word2Vec CBOW, Word2Vec SG

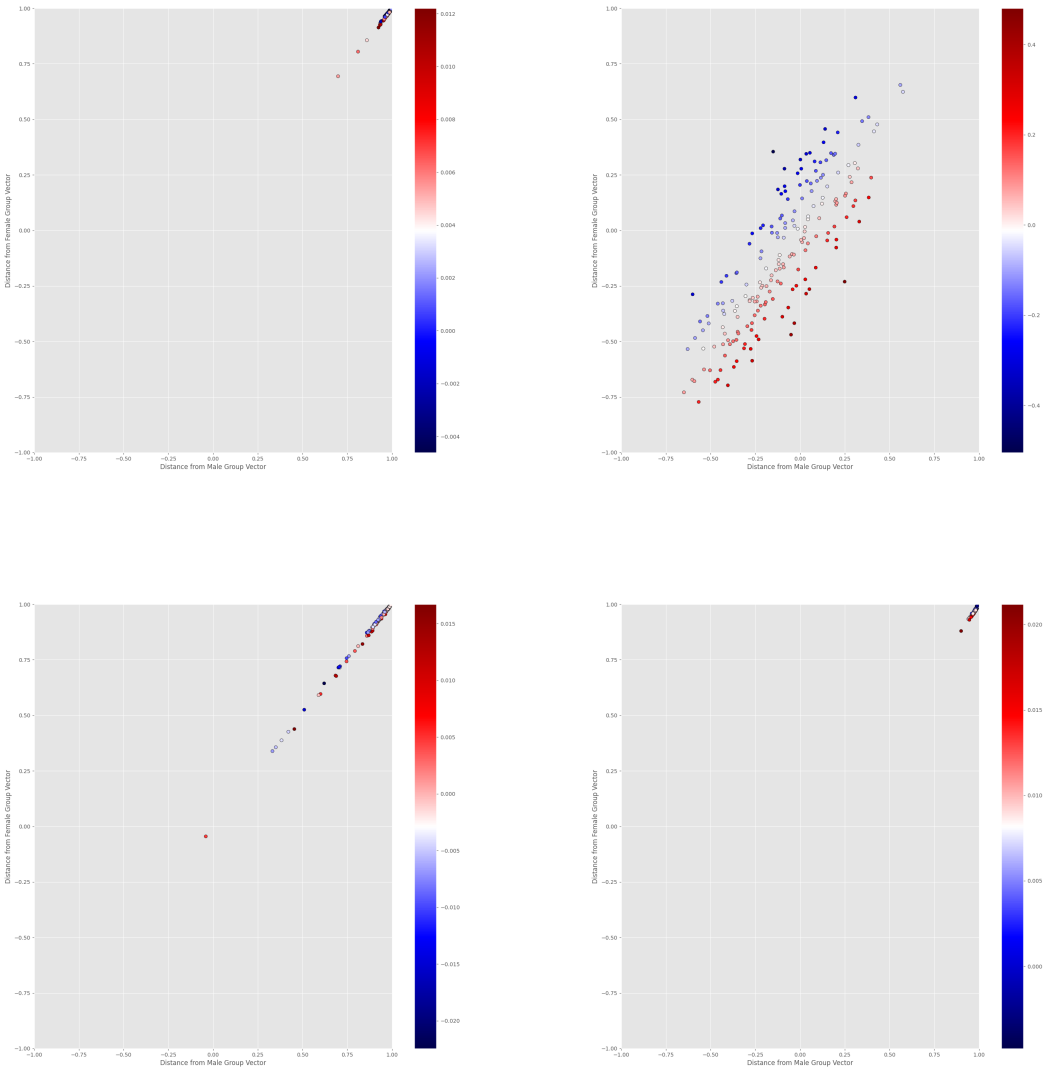

Figure B.3: Cosine Similarities of Adjectives to Gender Group Vectors in *Song of Achilles* From Left to Right, Top to Bottom: FastText, GloVe, Word2Vec CBOW, Word2Vec SG

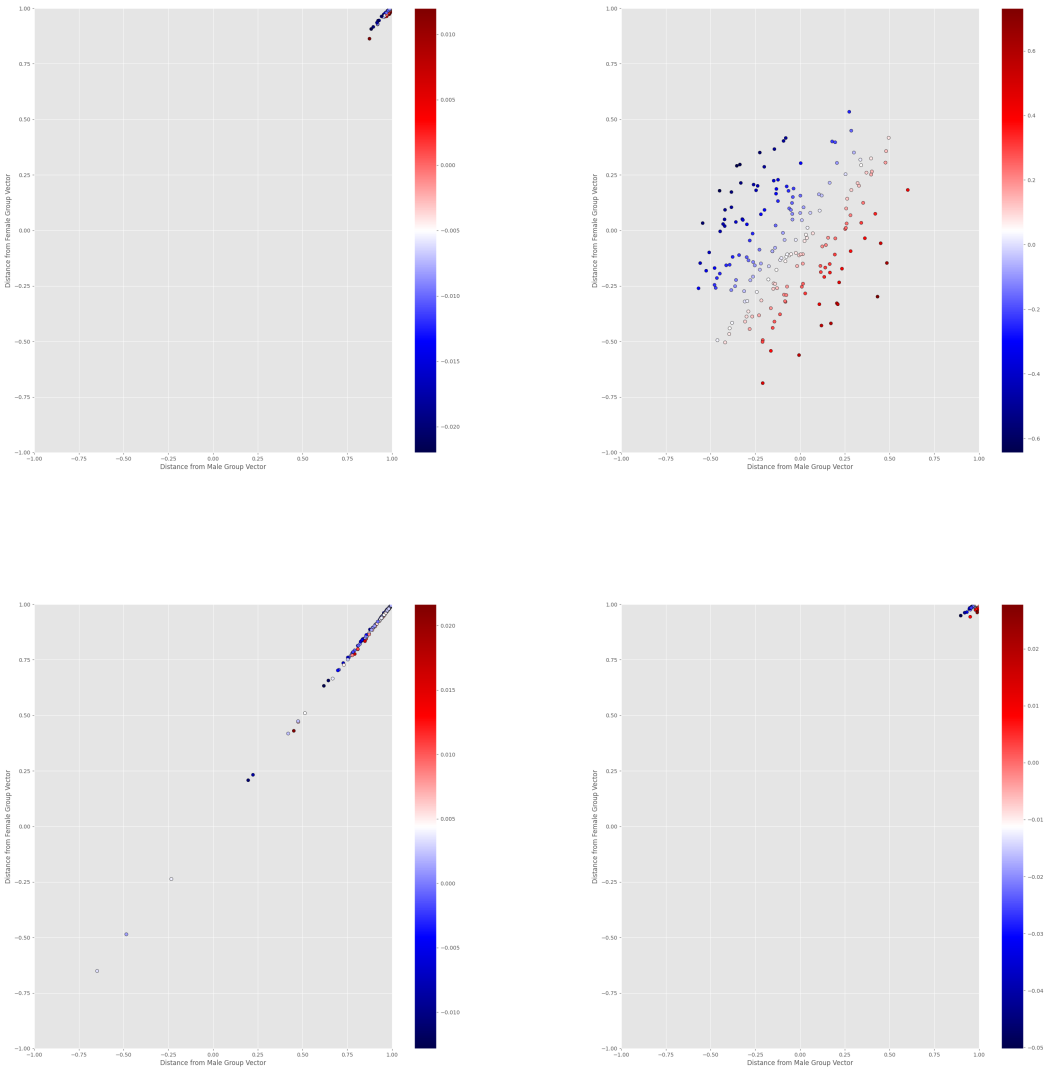

Figure B.4: Cosine Similarities of Adjectives to Gender Group Vectors in *Foundation*
From Left to Right, Top to Bottom: FastText, GloVe, Word2Vec CBOW, Word2Vec SG

Appendix C

TABLES

[1]

FastText Man Vector Adjective Similarities											
Wikipedia 2017			K-12 Corpus			Song of Achilles			Foundation		
heroic	+	0.0835	barbaric		0.0646	willful	+	0.0142	ruined		0.0113
genius	+ ?	0.0830	gallant	+	0.0576	youthful		0.0126	fixed		0.0113
petty	-	0.0829	crisp		0.0562	weak		0.0125	poised	*	0.0104
cowardly		0.0823	airy		0.0558	dull	-	0.0104	odd	-	0.0103
bald		0.0785	bald		0.0551	thin	*	0.0095	cultured		0.0088
judicious	?	0.0781	sane	+	0.0531	skillful	+	0.0094	inhibited	-	0.0086
reckless	-	0.0774	stern	-	0.0526	wise	+ ?	0.0089	oppressed	-	0.0083

FastText Woman Vector Adjective Similarities											
Wikipedia 2017			K-12 Corpus			Song of Achilles			Foundation		
feminine	*	-0.1709	stupid	- ?	-0.0573	inhibited	-	-0.0048	noisy		-0.0199
alluring	*	-0.1699	crazy	-	-0.0484	fixed		-0.0046	willful	+	-0.0197
voluptuous	*	-0.1694	weak	-	-0.0484	educated		-0.0043	thin	*	-0.0196
attractive	+ *	-0.1507	pretty	* +	-0.0482	busy		-0.0043	weak	-	-0.0180
beautiful	*	-0.1288	hateful	-	-0.0441	affected		-0.0043	youthful		-0.0166
sensual	*	-0.1100	reasonable	+	-0.0403	poised	+	-0.0037	bizarre	-	-0.0164
homely	- *	-0.1002	difficult	-	-0.0400	patriotic	+	-0.0034	busy		-0.0150

Table C.1: Adjectives Most Similar to Group Vectors in FastText models

[1] See table 4.1 for information on the adjective classification annotations.

GloVe Man Vector Adjective Similarities											
Wikipedia 2017			K-12 Corpus			Song of Achilles			Foundation		
genius	+ ?	0.1035	judicious	?	0.2453	false	-	0.4805	extreme		0.7317
original	**+ ?**	0.1034	bald	*	0.2192	witty	+	0.4192	logical	+ ?	0.6324
forgiving	+	0.0975	criminal	-	0.1909	warm	+	0.3860	original		0.5914
grand	+	0.0889	genius	+ ?	0.1884	reasonable	+	0.3194	physical		0.5560
stout	*	0.0863	benevolent	+	0.1735	petty	-	0.3191	mild	+	0.5481
dry		0.0782	thoughtless	-	0.1722	silent		0.3169	warm		0.5439
deep	+	0.0751	civilized	+	0.1684	fraudulent	-	0.2941	grim	-	0.5340

GloVe Woman Vector Adjective Similarities											
Wikipedia 2017			K-12 Corpus			Song of Achilles			Foundation		
voluptuous	*	-0.2041	beautiful	+ *	-0.2629	knowledge	+ ?	-0.5045	mystical		-0.6439
hysterical	-	-0.1928	tasteless	**-**	-0.1957	simple		-0.3645	efficient	+	-0.6325
alluring		-0.1909	frivolous	-	-0.1826	cold		-0.3165	transparent		-0.6280
feminine	*	-0.1823	smooth		-0.1566	driving		-0.3161	impressive	**+ ?**	-0.5767
assertive		-0.1713	soft		-0.1435	cultured		-0.3123	upright	+	-0.5750
unkind	-	-0.1599	realistic	+	-0.1412	complicated		-0.3095	narrow	**-**	-0.5560
homely	- *	-0.1528	sweet	+	-0.1368	handsome	+ *	-0.3076	insecure	-	-0.5443

Table C.2: Adjectives Most Similar to Group Vectors in GloVe models

Word2Vec CBOW Man Vector Adjective Similarities								
K-12 Corpus			Song of Achilles			Foundation		
regimental		0.1826	extraordinary	+	0.0167	withdrawn	-	0.0216
treacherous	-	0.1539	impatient	-	0.0149	initiative	+	0.0163
petty	-	0.1494	precise	+	0.0120	sentimental		0.0159
devious	-	0.1434	earnest	+	0.0117	desperate	-	0.0122
criminal	-	0.1385	trusting		0.0108	calm	+	0.0117
driving		0.1384	brilliant	+ ?	0.0097	eloquent	+	0.0109
daring	+	0.1378	relaxed	+	0.0090	sordid		0.0088

Word2Vec CBOW Woman Vector Adjective Similarities								
K-12 Corpus			Song of Achilles			Foundation		
circumspect		-0.1702	subtle	+	-0.0224	casual	+	-0.0128
sweet	+	-0.1570	hurried		-0.0150	contradictory	-	-0.0116
gorgeous	+ *	-0.1559	lyrical	+	-0.0141	stubborn	-	-0.0110
beautiful	+ *	-0.1475	sympathetic	+	-0.0125	mellow	+	-0.0089
precocious		-0.1309	neurotic	-	-0.0121	amusing		-0.0085
soft	+	-0.1235	pure	+	-0.0113	fickle	-	-0.0067
feminine	*	-0.1214	dishonest	-	-0.0104	gallant	+	-0.0065

Table C.3: Adjectives Most Similar to Group Vectors in Word2Vec CBOW models

Word2Vec skip-gram Man Vector Adjective Similarities											
Wikipedia 2017			K-12 Corpus			Song of Achilles			Foundation		
genius	+ ?	0.1116	cowardly		0.1399	relaxed		0.0211	grand	+	0.0278
petty	-	0.1019	regimental		0.1330	trusting		0.0171	open		0.0191
judicious	?	0.0949	crafty	-	0.1153	solid		0.0147	obvious		0.0191
sagacious		0.0921	bald		0.1115	strong		0.0146	capable	+	0.0153
heroic	+	0.0844	dishonest	-	0.1101	ruined		0.0142	knowledge	+ ?	0.0153
jolly	+	0.0821	monstrous	-	0.1074	objective	**?**	0.0139	simple		0.0129
disciplined		0.0817	fiery	+	0.1066	intelligent	+ ?	0.0138	honest		0.0080

Word2Vec skip-gram Woman Vector Adjective Similarities											
Wikipedia 2017			K-12 Corpus			Song of Achilles			Foundation		
voluptuous	*	-0.1809	feminine		-0.1042	loud		-0.0048	honorable	+	-0.0501
feminine	*****	-0.1690	maternal		-0.1033	difficult	-	-0.0038	obedient	+	-0.0411
attractive	+ *	-0.1503	plump		-0.0650	serious	+	-0.0038	thorough	+	-0.0345
alluring	*****	-0.1405	beautiful	+*	-0.0644	knowledge	+ ?	-0.0029	scornful	-	-0.0314
sensual	*****	-0.1403	sweet	+	-0.0609	solitary		-0.0028	eloquent	+	-0.0314
hysterical	-	-0.1255	creative	+	-0.0542	working		-0.0028	mistaken	-	-0.0311
assertive		-0.1241	queer	-	-0.0517	cold		-0.0012	profound	+	-0.0303

Table C.4: Adjectives Most Similar to Group Vectors in Word2Vec Skip-Gram models

HGI Lexicon Man Group Vector Negative Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
stern	-	0.0526	criminal	-	0.1909	treacherous	-	0.1539	stern	-	0.0526
bland	-	0.0510	thoughtless	-	0.1722	petty	-	0.1494	bland	-	0.0510
petty	-	0.0485	foolish	-	0.1643	devious	-	0.1434	petty	-	0.0485
grim	-	0.0470	venomous	-	0.1608	criminal	-	0.1385	grim	-	0.0470
fickle	-	0.0434	vague	- ?	0.1532	reactionary	-	0.1293	fickle	-	0.0434
brittle	-	0.0369	restless	-	0.1498	monstrous	-	0.1113	brittle	-	0.0369
wary	-	0.0348	malicious	-	0.1497	unscrupulous	-	0.0967	wary	-	0.0348

HGI Lexicon Woman Group Vector Negative Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
crazy	-	-0.0484	frivolous	-	-0.1826	unkind	-	-0.1188	crazy	-	-0.0484
difficult	-	-0.0400	fickle	-	-0.1035	bland	-	-0.0910	difficult	-	-0.0400
selfish	-	-0.0357	hysterical	-	-0.0988	silly	- ?	-0.0845	selfish	-	-0.0357
sly	-	-0.0352	ugly	- *	-0.0977	queer	-	-0.0798	sly	-	-0.0352
desperate	-	-0.0335	deceitful	-	-0.0944	homely	- *	-0.0761	desperate	-	-0.0335
ungrateful	-	-0.0333	illogical	-	-0.0909	irritable	-	-0.0729	ungrateful	-	-0.0333
miserable	-	-0.0319	sedentary	-	-0.0760	ugly	- *	-0.0715	miserable	-	-0.0319

Table C.5: HGI Lexicon Negative Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

HGI Lexicon Man Group Vector Positive Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
gallant	+	0.0576	genius	+ ?	0.1884	daring	+	0.1378	gallant	+	0.0576
sane	+	0.0531	benevolent	+	0.1735	genius	+ ?	0.1281	sane	+	0.0531
grand	+	0.0415	courteous	+	0.1617	steadfast	+	0.1216	grand	+	0.0415
sage	+ ?	0.0406	reliable	+	0.1530	heroic	+	0.1214	sage	+ ?	0.0406
mellow	+	0.0372	frugal	+	0.1509	loyal	+	0.1208	mellow	+	0.0372
frugal	+	0.0341	sober	+	0.1433	resourceful	+ ?	0.0931	frugal	+	0.0341
fiery	+	0.0300	venerable	+ ?	0.1423	progressive	+	0.0910	fiery	+	0.0300

HGI Lexicon Woman Group Vector Positive Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
reasonable	+	-0.0403	realistic	+	-0.1412	gorgeous	+ *	-0.1559	reasonable	+ **?**	-0.0403
sincere	+	-0.0396	willful	+	-0.0935	delicate	+	-0.1197	sincere	+	-0.0396
unselfish	+	-0.0341	pure	+	-0.0916	affectionate	+	-0.1111	unselfish	+	-0.0341
thoughtful	+ ?	-0.0316	gorgeous	+ *	-0.0894	luminous	+ ?	-0.1075	thoughtful	+ ?	-0.0316
helpful	+	-0.0310	luminous	+ ?	-0.0853	attractive	+ *	-0.0931	helpful	+	-0.0310
agreeable	+	-0.0299	serious	+	-0.0825	elegant	+ *****	-0.0907	agreeable	+	-0.0299
manly	+ *****	-0.0287	affectionate	+	-0.0751	wholesome	+	-0.0850	manly	+ *****	-0.0287

Table C.6: HGI Lexicon Positive Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

HGI Lexicon Man Group Vector Weak Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
bland	-	0.0510	foolish	- **?**	0.1643	unscrupulous	-	0.0967	bland	-	0.0510
fickle	-	0.0434	vague	- ?	0.1532	desperate	-	0.0878	fickle	-	0.0434
brittle	-	0.0369	irresponsible	-	0.1433	defensive	-	0.0863	brittle	-	0.0369
meek	-	0.0291	fraudulent	-	0.1027	dishonest	-	0.0632	meek	-	0.0291
narrow		0.0253	withdrawn		0.0995	fraudulent	-	0.0592	narrow		0.0253
slender	*	0.0216	feeble	- *	0.0935	foolish	- **?**	0.0555	slender	*	0.0216
slim	*	0.0202	dull	-	0.0854	feeble	- *	0.0450	slim	*	0.0202

HGI Lexicon Woman Group Vector Weak Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
desperate	-	-0.0335	feminine	*****	-0.1060	feminine	*****	-0.1214	desperate	-	-0.0335
shy		-0.0323	fickle	-	-0.1035	delicate	+	-0.1197	shy		-0.0323
miserable	-	-0.0319	hysterical	-	-0.0988	sentimental		-0.1027	miserable	-	-0.0319
ridiculous	-	-0.0299	illogical	-	-0.0909	bland	-	-0.0910	ridiculous	-	-0.0299
naive	-	-0.0295	brittle	-	-0.0506	silly	-	-0.0845	naive	-	-0.0295
fearful	-	-0.0293	insecure	-	-0.0446	slender	*	-0.0787	fearful	-	-0.0293
silly	- **?**	-0.0279	gentle	+	-0.0426	timid		-0.0761	silly	- **?**	-0.0279

Table C.7: HGI Lexicon Weak Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

HGI Lexicon Man Group Vector Strong Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
gallant	+	0.0576	sheriff		0.2644	police		0.2067	gallant	+	0.0576
stern	-	0.0526	police		0.2137	sheriff		0.1892	stern	-	0.0526
grand	+	0.0415	manager		0.2020	daring	+	0.1378	grand	+	0.0415
sage	+ ?	0.0406	genius	+ ?	0.1884	manager		0.1329	sage	+ ?	0.0406
alert		0.0332	reliable	+	0.1530	genius	+ ?	0.1281	alert		0.0332
administrator		0.0317	sober	+	0.1433	steadfast	+	0.1216	administrator		0.0317
fiery	+	0.0300	monstrous	-	0.1358	heroic	+	0.1214	fiery	+	0.0300

HGI Lexicon Woman Group Vector Strong Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
tough		-0.0361	intense		-0.1003	boisterous	-	-0.0638	tough		-0.0361
manly	+	-0.0287	busy		-0.0748	intense		-0.0575	manly	+	-0.0287
proud	+	-0.0263	haughty	-	-0.0520	passionate	+	-0.0482	proud	+	-0.0263
capable	+	-0.0256	capable	+	-0.0353	dependable	+	-0.0436	capable	+ **?**	-0.0256
clever	+ ?	-0.0239	instructor		-0.0342	haughty	-	-0.0356	clever	+ ?	-0.0239
reliable	+	-0.0216	proud	+	-0.0322	instructor		-0.0343	reliable	+	-0.0216
decisive	**?**	-0.0214	boisterous	-	-0.0298	energetic	+	-0.0286	decisive	**?**	-0.0214

Table C.8: HGI Lexicon Strong Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

HGI Lexicon Man Group Vector Vice Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
petty	-	0.0485	foolish	- ?	0.1643	treacherous	-	0.1539	petty	-	0.0485
grim	-	0.0470	selfish	-	0.1495	petty	-	0.1494	grim	-	0.0470
crafty	-	0.0219	irresponsible	-	0.1433	monstrous	-	0.1113	crafty	-	0.0219
arbitrary	-	0.0176	petty	-	0.1404	unscrupulous	-	0.0967	arbitrary	-	0.0176
awkward	-	0.0142	monstrous	-	0.1358	fanatical	-	0.0923	awkward	-	0.0142
odd	-	0.0132	cynical	-	0.1196	reckless	-	0.0829	odd	-	0.0132
fanatical	-	0.0118	treacherous	-	0.1172	crafty	-	0.0712	fanatical	-	0.0118

HGI Lexicon Woman Group Vector Vice Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
crazy	-	-0.0484	ugly	- *	-0.0977	unkind	-	-0.1188	crazy	-	-0.0484
difficult	-	-0.0400	deceitful	-	-0.0944	silly	- ?	-0.0845	difficult	-	-0.0400
selfish	-	-0.0357	dirty	-	-0.0746	queer	-	-0.0798	selfish	-	-0.0357
sly	-	-0.0352	difficult	-	-0.0540	homely	- *	-0.0761	sly	-	-0.0352
ungrateful	-	-0.0333	childish	-	-0.0536	timid		-0.0761	ungrateful	-	-0.0333
ugly	- *	-0.0310	unkind	-	-0.0481	ugly	- *	-0.0715	ugly	- *	-0.0310
unkind	-	-0.0309	insecure	-	-0.0446	childish	-	-0.0654	unkind	-	-0.0309

Table C.9: HGI Lexicon Vice Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

HGI Lexicon Man Group Vector Virtue Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
gallant	+	0.0576	benevolent	+	0.1735	daring	+	0.1378	gallant	+	0.0576
sane	+	0.0531	courteous	+	0.1617	heroic	+	0.1214	sane	+	0.0531
grand	+	0.0415	reliable	+	0.1530	loyal	+	0.1208	grand	+	0.0415
athletic	+ *	0.0260	venerable	+ ?	0.1423	resourceful	+ ?	0.0931	athletic	+ *	0.0260
popular	+	0.0237	rational	+	0.1352	decisive	**?**	0.0916	popular	+	0.0237
dynamic	+	0.0214	gallant	+	0.1347	progressive	+	0.0910	dynamic	+	0.0214
versatile	+ **?**	0.0213	versatile	+	0.1299	moderate	+	0.0858	versatile	+ **?**	0.0213

HGI Lexicon Woman Group Vector Virtue Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
reasonable	+	-0.0403	pure	+	-0.0916	attractive	+ *	-0.0931	reasonable	+	-0.0403
sincere	+	-0.0396	busy		-0.0748	elegant	+	-0.0907	sincere	+	-0.0396
unselfish	+	-0.0341	attractive	+ *	-0.0685	wholesome	+	-0.0850	unselfish	+	-0.0341
thoughtful	+ ?	-0.0316	friendly	+	-0.0467	gentle	+	-0.0754	thoughtful	+ ?	-0.0316
helpful	+	-0.0310	gentle	+	-0.0426	gracious	+	-0.0600	helpful	+	-0.0310
manly	+	-0.0287	conventional		-0.0400	thoughtful	+ ?	-0.0587	manly	+	-0.0287
capable	+	-0.0256	gracious	+	-0.0376	neat	+	-0.0559	capable	+	-0.0256

Table C.10: HGI Lexicon Virtue Adjectives Most Similar to Group Vectors in All Models Trained on the K-12 Corpus

Song of Achilles Man Vector Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
willful	+	0.0142	false	-	0.4805	extraordinary	+	0.0167	relaxed	**+**	0.0211
youthful		0.0126	witty	+	0.4192	impatient	-	0.0149	trusting		0.0171
weak		0.0125	warm		0.3860	precise	+	0.0120	solid		0.0147
dull	-	0.0104	reasonable	+	0.3194	earnest	+	0.0117	strong		0.0146
thin	*	0.0095	petty	-	0.3191	trusting		0.0108	ruined	**-**	0.0142
skillful	**+ ?**	0.0094	silent		0.3169	brilliant	+ ?	0.0097	objective	**?**	0.0139
wise	**+ ?**	0.0089	fraudulent	-	0.2941	relaxed	**+**	0.0090	intelligent	+ ?	0.0138

Foundation Man Vector Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
ruined	**-**	0.0113	extreme		0.7317	withdrawn		0.0216	grand	+	0.0278
fixed		0.0113	logical	+ ?	0.6324	initiative	**?**	0.0163	open	+	0.0191
poised	*****	0.0104	original	**+ ?**	0.5914	sentimental		0.0159	obvious		0.0191
odd	-	0.0103	physical		0.5560	desperate	-	0.0122	capable	**+ ?**	0.0153
cultured		0.0088	mild	+	0.5481	calm		0.0117	knowledge	+ ?	0.0153
inhibited	**-**	0.0086	warm	**+**	0.5439	eloquent	+	0.0109	simple		0.0129
oppressed	**-**	0.0083	grim	-	0.5340	sordid		0.0088	honest	+	0.0080

Table C.11: *Song of Achilles* and *Foundation* Adjectives Most Similar to the Man Group Vector in All Models

Song of Achilles Woman Vector Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
inhibited	**-**	-0.0048	knowledge	+ ?	-0.5045	subtle	+	-0.0224	loud		-0.0048
fixed		-0.0046	simple		-0.3645	hurried		-0.0150	difficult	-	-0.0038
educated	**?**	-0.0043	cold	**- e**	-0.3165	lyrical	+	-0.0141	serious	+	-0.0038
busy		-0.0043	driving		-0.3161	sympathetic	+	-0.0125	knowledge	+ ?	-0.0029
affected	**- e**	-0.0043	cultured		-0.3123	neurotic	-	-0.0121	solitary		-0.0028
poised	*****	-0.0037	complicated		-0.3095	pure	+	-0.0113	working		-0.0028
patriotic	+	-0.0034	handsome	+ *	-0.3076	dishonest	-	-0.0104	cold	**-**	-0.0012

Foundation Woman Vector Adjective Similarities											
FastText			GloVe			Word2Vec CBOW			Word2Vec SG		
noisy		-0.0199	mystical		-0.6439	casual	+	-0.0128	honorable	+	-0.0501
willful	+	-0.0197	efficient	**+ ?**	-0.6325	contradictory	-	-0.0116	obedient	+	-0.0411
thin	*	-0.0196	transparent		-0.6280	stubborn	-	-0.0110	thorough	**+ ?**	-0.0345
weak	**-**	-0.0180	impressive	+	-0.5767	mellow	+	-0.0089	scornful	-	-0.0314
youthful	*****	-0.0166	upright	+	-0.5750	amusing		-0.0085	eloquent	+	-0.0314
bizarre	-	-0.0164	narrow	**-**	-0.5560	fickle	-	-0.0067	mistaken	-	-0.0311
busy		-0.0150	insecure	-	-0.5443	gallant	+	-0.0065	profound	+	-0.0303

Table C.12: *Song of Achilles* and *Foundation* Adjectives Most Similar to the Woman Group Vector in All Models