Addressing Age-Related Bias in Sentiment Analysis

Mark Díaz[*], Isaac Johnson[*], Amanda Lazar[†], Anne Marie Piper[*] Darren Gergle[*]
[*]Northwestern University, [†]University of Maryland
{mark.diaz, isaac.j}@u.northwestern.edu, lazar@umd.edu, {ampiper, dgergle}@northwestern.edu

ABSTRACT
Computational approaches to text analysis are useful in understanding aspects of online interaction, such as opinions and subjectivity in text. Yet, recent studies have identified various forms of bias in language-based models, raising concerns about the risk of propagating social biases against certain groups based on sociodemographic factors (e.g., gender, race, geography). In this study, we contribute a systematic examination of the application of language models to study discourse on aging. We analyze the treatment of age-related terms across 15 sentiment analysis models and 10 widely-used GloVe word embeddings and attempt to alleviate bias through a method of processing model training data. Our results demonstrate that significant age bias is encoded in the outputs of many sentiment analysis algorithms and word embeddings. We discuss the models' characteristics in relation to output bias and how these models might be best incorporated into research.

Author Keywords
Sentiment analysis; older adults, algorithmic bias; aging.

ACM Classification Keywords
H.5.m. Information interfaces and presentation: Miscellaneous

INTRODUCTION
Although the concept of ageism was identified several decades ago [12], negative attitudes and stereotypes about growing older are only now receiving worldwide attention. The World Health Organization has recently called for a "global campaign to combat ageism," given the association between negative views about aging and decreased health and longevity [57]. Age discrimination and age bias are topics that have also begun to receive attention within HCI where work highlights the ways that researchers and designers tend to treat aging as a "problem" with technology as a solution, rather than viewing aging as a complex and natural part of the lifespan [74]. To help counter age-related stereotypes around technology use, prior work has emphasized cases of older adults going online to actively create and share content [9,10,31], learn to program [30], and even form social movements around ageism [47].

CHI 2018, April 21–26, 2018, Montreal, QC, Canada
© 2018 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-5620-6/18/04.
https://doi.org/10.1145/3173574.3173986

An adjacent but growing area of interest concerns how the tools and techniques used to understand online behavior may propagate social biases against certain groups, particularly those that may be underrepresented or stigmatized [40,41,64]. Sentiment analysis in particular is a popular computational approach to understanding attitude, affect, and opinion in text [58]. It is often used to measure opinions in product reviews or financial markets [32], which can inform and drive branding decisions, political campaign strategies, and automated financial trading systems [26]. Some computational algorithms have been shown to exhibit social biases, however, and tools for measuring sentiment vary widely in their implementation, from computing values of component words and phrases within a document (lexicon-based models) to using labeled example text to train a machine learning classifier (supervised, corpus-based models) [69] to hybrid models integrating both approaches [66,76]. In the case of age-related bias, automated methods of opinion polling surrounding issues related to old age may falsely report more negative attitudes toward political issues or financial investments regarding age-related concerns, such as Medicare and Social Security. Though bias may stem from many sources, we bring particular attention to addressing bias rooted in training data. For many sentiment analysis tools, the output of algorithms or machine learning models is still largely dependent on these annotated datasets [70]. Computational algorithms are sensitive to not only the size and quality of the underlying datasets but also to human social bias that exists within them.

In this paper, we focus on age-related social bias in sentiment analysis as a case of using computational, algorithmic tools to study underrepresented attitudes and opinions. There is a growing awareness of age discrimination worldwide (e.g., [45,57]), and age-related bias in particular has not been studied with regard to popular sentiment analysis tools that are used to make strategic decisions about products, politics, finances, social services and employment [22,44,54,59]. Nor has there been much work specifically aimed at addressing or reducing age bias in algorithms. Potential underlying bias around age has implications for the appropriateness of these tools in contexts where attitudes towards age matter, as well as the ways that subtle forms of age discrimination manifest in technologies that pervade everyday life.

The primary questions motivating the present study are whether age bias manifests in the output of machine learning models and, if so, what this bias looks like across commonly-

used sentiment analysis models in a realistic research context. Our analysis focuses both on the treatment by sentiment analysis methods of words that are explicit encodings of age (e.g., "old" or "young") as well as words that are implicit encodings of age (as determined through word embeddings). Given that context is deeply tied to algorithmic bias, researchers have called for technologies to be studied in the contexts in which they operate [16]. We evaluate the impact of these techniques on a text-based corpus of discussions of aging to observe how age bias may manifest in this naturalistic context.

In this paper we contribute: (1) a systematic analysis of age-related bias in a large number of popular sentiment analysis tools and word embeddings. In doing so we find significant age bias in algorithmic output, for example sentences with "young" adjectives are 66% more likely to be scored positively than the same sentences with "old" adjectives; (2) a nuanced understanding of how the technical characteristics of various sentiment analysis methods impact bias in outcomes – particularly that tools validated against social media data exhibit increased bias; and (3) a case study in attempting to reduce bias in training data where, with a relatively straightforward approach, we successfully reduce age bias by an order of magnitude. We conclude with critical reflection on the use of these tools for studying social movements and underrepresented populations.

RELATED WORK

There is a growing interest in issues of social justice in HCI, as evidenced by new frameworks and agendas [2,3,23,39,47,48,63,65] that attempt to shift power balances between researchers, society, and marginalized groups. These frameworks tackle diverse domains but converge on several points. One of these points is that science, technology, and design are not neutral or valueless; rather, they perpetuate certain points of views or ways of thinking. Work in critical algorithm studies embraces this view, and some have described algorithms as "the new power brokers in society" [22]. In addition to these critiques, a number of studies have focused on understanding the underlying mechanisms that drive bias in algorithms.

Critical Algorithm Studies

Critical algorithm studies is an emerging area of research that spans computer science, sociology, science and technology studies, communication, legal studies, and other fields. Much work in critical algorithm studies examines algorithmic bias, which can be defined as "systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" [27]. Prior work analyzes algorithmic bias in search engines [36,38,53], surveillance systems (e.g., Facial Recognition Systems) [37], and social media [20,55,71]. For example, Introna and Nissenbaum describe the ways that biased search engines diminish access to information as well as individuals' abilities "to be seen, and heard" [36]. While a growing body of work calls attention to algorithmic bias as an instance of technology embodying social, ethical, and political values [56], others have focused on understanding the sources of bias and identifying ways to diminish it.

Understanding Bias in Algorithms

Researchers have described the algorithms that drive many of the systems we use as "black boxes" (e.g., [22,59]). Given the opaqueness of the ways algorithms operate, one area of study in HCI has been "folk theories," or how individuals interpret algorithms. Researchers have examined how people may attempt to reverse engineer systems based on their understanding of algorithms [5], how altering the design of system features affects an individual's understanding of algorithms [21,24], and how understanding algorithms affects user behavior [25]. User behavior can also lead to bias in systems. Liao et al. examined the ways Twitter users with minority opinions used design features such as the "retweet" button to amplify their views and introduce bias [50]. Other researchers have attempted to systematically sort out user bias (e.g., keywords that are put in) from system bias [43]. As part of the larger discussion of algorithmic bias, recent work has begun to analyze the design and underlying mechanisms of algorithms that contribute to bias, with a call for more empirical studies [42]. Nissenbaum stated that "Fastidious attention to the before-and-after picture, however richly painted, is not enough" [56]. Instead, she states that what engineers and computer scientists can contribute to the field is "a fine-grained understanding of systems - even down to gritty details of architecture, algorithm, [and] code," as these are essential to "explaining the social, ethical, and political dimensions of information technologies" [56]. Some researchers have directly manipulated open-source algorithms to reveal the extent of structural biases [40]. However, given that many algorithms are proprietary, researchers have also attempted to decipher algorithms by interpreting output while varying inputs [15,22] – We make use of both approaches in this paper.

Algorithmic Bias in Text Processing

Natural language processing techniques, including sentiment analysis, are a primary point for inquiry among both social scientists and machine learning researchers [11,18,33,51]. Sentiment analysis, or opinion mining, can be understood as the "computational treatment of opinion, sentiment, and subjectivity in text" [58]. Many sentiment analysis tools are *lexicon-based*, which involves using sentiment values of component words and phrases within a document to calculate a sentiment value for the whole. Another common approach (*corpus-based*) is to employ classifying techniques using labeled example text to train a machine learning algorithm [69]. Other tools are hybrids, using some combination of lexicon-based and machine-learning techniques [66,76].

Bias in natural language processing tools can arise from a variety of sources. Some work has focused on word embeddings [6,14], which map words and phrases to vectors of real numbers and are used to capture the semantic relatedness of terms within a dataset. Bias also emerges in

algorithmic decision-making, which is often opaque to users of technologies [8]. Instances of bias in algorithmic decision-making include the auto-complete function of search engines [1], advertisements based on search terms [68], and image search results [41], which propagate harmful racial and gender stereotypes. While bias manifests in the presentation of search results, for example, it may be rooted in the human-authored data used to train algorithms. Caliskan et al. trained a popular machine learning model on a standard text corpus and found that human biases toward race and gender in a text corpus emerge as semantic biases in word embeddings [13]. Similarly, Sen et al. describe how gold standard datasets produced by Mechanical Turkers are significantly different than gold standard datasets produced by people in other communities. The authors conclude that algorithms should be evaluated based on how well they work for a given community [64], which is a view we take in this paper.

To investigate whether age-related bias might be present in sentiment analysis methods, and to understand how various characteristics of sentiment methods influence this form of bias, we study several lexicon-based and corpus-based tools, the type of data they were validated against, as well as word embedding models upon which many algorithmic tools are built. We view each of these features as an area where bias can be introduced, amplified, or potentially diminished.

RESEARCH APPROACH

We use a three-phased approach to understand whether and how different sentiment analysis methods produce bias in their results with respect to age. First, we examine the extent to which popular sentiment analysis tools exhibit age bias around *explicit encodings* of age contained within actual sentences sampled from a realistic research context – a community of older adult bloggers. In other words, we sampled cases when age is clearly and unambiguously mentioned in written language (e.g., "It's starting to be a trend to lay off *older* workers"). Next, we explore the extent to which the same sentiment analysis methods may demonstrate an age bias derived from more *implicit encodings* of age. To achieve this, we make use of commonly-used word embeddings to produce a set of "older" and "younger" analogs for common adjectives (e.g., for the word "unique", the "older" analog is "distinctive" while the "younger" analog is "innovative"), and then compare outputs from sentiment analysis tools in a similar fashion to that used in the first analysis. Finally, we train a custom sentiment analysis model by selectively sampling an existing Twitter dataset in an attempt to address bias in training data. Taken together, our approach provides a robust assessment of age-related bias in sentiment analysis models across a number of tools, with a variety of sentence types and forms, and begins to assess methods for addressing bias.

PHASE 1: EXPLICIT ENCODING OF AGE

The goal of our first phase of analysis is to determine whether sentiment analysis tools treat explicit indications of age (e.g., "old" and "young") differently.

Model	Type	Validation Data
AFINN	Lexicon	Social Media
EmoLex	Lexicon	Other
HappinessIndex	Lexicon	Other
NRC Hashtag	Lexicon	Social Media
Opinion Lexicon	Lexicon	Other
OpinionFinder	Hybrid	Other
PANAS	Lexicon	Social Media
Sasa	Classifier	Social Media
Sentiment140	Classifier	Social Media
SentiStrength	Hybrid	Social Media
Sentiwordnet	Hybrid	Other
SOCAL	Lexicon	Other
Stanford	Hybrid	Other
Umigon	Lexicon	Social Media
VADER	Lexicon	Other

Table 1. The fifteen different sentiment analysis methods examined, and their corresponding type and validation data used when building the model. Validation data that is not social-media-based is predominantly based on movie or product reviews or news corpora.

Method

We perform our analysis using fifteen popular sentiment analysis tools used in practice (Table 1 describes the tools and associated annotations that we use) [62]. Exploring multiple sentiment analysis tools minimizes the likelihood of reporting idiosyncratic findings from a single tool, and allows us to compare common implementation techniques that may influence bias. We use 15 of the 20 sentiment analysis models implemented in SentiBench [62] that span a variety of computational techniques, domains, and levels of complexity. We exclude the remaining five models due to a lack of variance in output scores and because one model only accepts emoticons as input. In line with how sentiment analysis models are often used, the models are standardized to produce one of three sentiment outputs: negative (-1), neutral (0), or positive (+1). In our analyses, we also code each sentiment analysis tool according to its computational method (unsupervised, lexicon-based approach vs. supervised, corpus-based approach) and the training and validation data used in building the model (social media vs. other sources). Because our method is an initial probe for sources of bias, we do not code for all model characteristics, however our work sets the stage for more detailed analyses.

Statistical Methods

We test the sentiment tools for age-related bias by examining the sentiment output scores using multinomial log-linear regressions (via the R package *nnet* [73]). We build two types of multinomial log-linear regressions: 1) a single full model for each phase of analysis that includes the data from all of the sentiment analysis tools in order to test for the presence of age-related bias across the models (Table 2), and, 2) individual models for each sentiment analysis tool (15 in total) in order to assess which specific tools demonstrate age-related bias (Table 4).

Sentiment Output	AdjectiveYoung		CorpusBased		ValDataSocialMedia		Young x CorpusBased		Young x SocialMedia		Intercept	
	e^(coef)	95% CI	e^(coef)	95% CI	e^(coef)	95% CI	e^(coef)	95% CI	e^(coef)	95% CI	e^(coef)	95% CI
Positive	**1.66****	[1.30-2.11]	**2.56****	[1.99-3.29]	**0.51****	[0.39-0.65]	**0.59****	[0.42-0.85]	0.84	[0.60-1.18]	**0.76****	[0.63-0.90]
Negative	0.88	[0.68-1.15]	**2.73****	[2.17-3.45]	1.14	[0.91-1.42]	1.21	[0.87-1.70]	0.98	[0.71-1.35]	**0.70****	[0.59-0.84]

Table 2. Regression results for explicit age analysis. The models include data from all sentiment analysis tools and are multinomial log-linear regressions, resulting in a model for positive sentiment and a model for negative sentiment. The reference categories are: neutral sentiment, "old" adjectives (i.e., "old" or "older"), lexicon-based approaches, and non-social-media validation data. Exponentiated coefficients (i.e., e^coef) provide relative risk (e.g., the sentiment analysis models were 1.66 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from the "older" adjective to a "younger" adjective"). Note: *p<0.05; **p<0.01

The dependent variable is the sentiment output (nominal: negative, neutral, positive). Our primary independent variable of interest is the relative age of the adjective in the sentence ("old" vs. "young"). We also examine how the regression coefficients vary across the different sentiment analysis tools according to the type of sentiment tool used (lexicon-based vs. corpus-based), and the model's validation data (social media vs. other data).

Regarding regression result interpretation, the exponentiated coefficients represent relative risk (because the models are log-linear and produce multinomial logit coefficients), or the change in the model prediction (i.e. to positive or negative sentiment holding all other variables constant). Neutral sentiment is the reference category used by the regressions for all but one of the tools[1]. Exponentiated coefficient values greater than one indicate that the regression model's sentiment is more likely and a neutral sentiment prediction is less likely, and exponentiated coefficient values less than one indicate that the regression model's sentiment is less likely and a neutral sentiment prediction is more likely.

Context of Study and Testing Data

It is important to understand the impact of these techniques and potential bias within a particular topic of study [64]. The opportune context in which we study age bias stems from research that examined a community of older adult bloggers to understand blogging as a form of online participation among older adults [10] and analyzes online blog-based discussions of age discrimination in the U.S. and U.K. [47]. Applying computational techniques to understand sentiment within these discussions, in which contributors are attempting to challenge negative views of aging, is a relevant case to understand the implications of underlying age bias.

We source sentences for the analysis by scraping 4,151 blog posts from a prominent "elderblogger" community [47] as well as 64,283 comments on each post created between 2004 and 2016[2]. Each researcher then independently, randomly samples posts and comments containing sentences with the word "old". Of these posts, we extracted 162 unique sentences. Because we are particularly interested in the use of the term "old" to describe people and aging, we exclude sentences using "old" to modify other nouns (e.g., "old things", "old movie") and as a general descriptor of age (e.g., "the 32-year-old"). We also exclude sentences that contain the word "young" or other youth-related terms as well as complex sentences with embedded clauses or unusual grammar or structure. Example sentences include, "We live in a culture that deliberately hides and ignores older folks." and "Old age is worth waiting for." Although the term "old" appears 86,145 times across our corpus, our exclusion process results in 121 sentences from our initial sample.

In each of the 121 sentences, we replace the term "old" (as well as "older" and "oldest") with the term "young" (as well as "younger" and "youngest") to provide a comparative dataset (242 sentences total). Our goal in doing this is to understand if sentiment analysis tools provide equivalent sentiment measures if the content from this blog were to describe younger people and youth instead of old age.

By using a standardized set of sentences and varying only the age-related terms, we are able to attribute any observed changes in sentiment score to the particular words we vary. Example sentences include, "It also upsets me when I realize that society expects this from <old/young> people."; "But it is not <old/young> folks who should be ashamed and embarrassed; it is the culture at large." We test 242 sentences in total, running each through all 15 sentiment analysis tools, which gave 3,630 sentiment analysis outputs[3].

Results

In this first stage of our analysis we aim to understand whether sentences featuring keywords related to older age ("old", "older", "oldest") are on average scored more negatively than the same sentences with words related to youth ("young", "younger", "youngest"), and whether this difference varies depending on the particular type of model (lexicon-based or corpus-based) and form of validation data (social media or other) used by the various sentiment analysis methods. Age-related terms (i.e. "old" versus "young") are our independent variable of interest.

Our findings for this phase of analysis are threefold. First, the results of the regression (see details in Table 2) revealed that across all of the sentiment analysis tools, sentences containing *young* adjectives (*AdjectiveYoung*) were 66%

[1] The Sentiwordnet model did not classify any of the sentences in phase 2 as "neutral." For this reason, we used "negative" as the reference category for the Sentiwordnet multinomial model (see Table 4).

[2] For more dataset information, please visit: https://github.com/markjdiaz

[3] We conducted an identical analysis using 558 researcher-generated sentences, varying age-related adjectives as well as gender (e.g., "man", "woman", "person"). We found similar results to those reported here.

Embedding	Source	Vocabulary
WG-6B-50D *WG-6B-100D* *WG-6B-200D* *WG-6B-300D*	English Wikipedia 2014 text and Gigaword 5 (7 sources of English-language newswire)	400K words, uncased
CC-42B-300D	Common Crawl of the Internet	1.9M works, uncased
CC-840B-300D		2.2M words, cased
TW-27B-25D *TW-27B-50D* *TW-27B-100D* *TW-27B-200D*	2 billion Twitter tweets	1.2M words, uncased

Table 3. Details on the 10 GloVe models. The first part of the name references the source, the second part of the name gives the number of tokens contained in the source (e.g., 6B = 6 billion), and the third part of the name gives the number of dimensions of the word vectors (e.g., 300D = 300-dimensional vectors for each word in the vocab). Further details at https://nlp.stanford.edu/projects/glove/

more likely to be scored positively than the same sentences containing *old* adjectives, when controlling for other sentential content.

Second, examining the type of sentiment analysis tools, supervised learning-based tools (*CorpusBased*, as opposed to lexicon-based) were more likely to indicate either positive or negative sentiment (rather than neutral) compared with unsupervised, lexicon-based tools, indicating a polarizing effect. Because supervised learning-based tools had a polarizing effect on the likelihood of both positive and negative indications *and* because the sentiment analysis methods were more likely to indicate positive for "young" sentences, these two trends had a disproportionate effect on pushing "young" sentences toward positive sentiment.

Third, sentiment analysis tools validated against social media data (*ValDataSocialMedia*) were less likely to rate sentences as positive (rather than neutral) compared with tools validated against other forms of data.

These findings may be explained by the fact that machine learning classifiers are necessarily trained with chunks of data larger than a single word or phrase. Training on larger chunks of naturalistic text allows classifiers the ability to learn subtle biases in human language use. Much of this training data is social media-based (e.g., Twitter) as well.

Analyzing the data from all 15 sentiment models revealed a significant interaction between age and the type of sentiment method (lexicon-based vs. corpus-based). The corpus-based method moderated the likelihood of a positive outcome for "young" sentences compared with the lexicon-based method.

Considering the individual regression model results provides a more complete picture by partitioning out the results by each sentiment analysis tool (see Table 4 for details). These results reveal that 4 (of 15) sentiment methods were significantly more likely to indicate positive sentiment when a sentence contained a young-age-related adjective. Three of these tools were lexicon-based and one was corpus-based. Two tools were less likely to indicate negative sentiment for young-age-related sentence and one was *more* likely to indicate negative sentiment for a "young" sentence[4].

Taken together, the results of both the full regression and the individual regressions indicate significant age-related bias with respect to explicit encodings of age, with corpus-based tools presenting a more polarizing effect and those trained on social media data skewing less positively across all sentences. Yet, questions remain around whether similar bias exists for implicit encoding of age, such as through the word embeddings that underlie many of these tools.

PHASE 2: IMPLICIT ENCODING OF AGE

Age-related bias may seep into computational approaches in various ways. In order to better understand sources of potential bias, we now turn to analyze whether age-related bias may be rooted in how word embeddings encode implicit associations with age and aging.

Method

We again manipulate specific words in sentence templates, but now we generate the adjectives inserted into the templates by taking a list of common English adjectives and skewing them "old" or "young" through the use of vector math on word embeddings.

Word embeddings are multi-dimensional vectors (often 100-300 dimensions) where each vector represents a specific word and the values for each dimension are learned based on the context (i.e., surrounding words) within which that word commonly appears. One of the most salient emergent properties of word embeddings is that they have been shown to encode analogies (e.g., "king" – "man" + "woman" = "queen") [52]. Thus, word embeddings can be used to transfer the relationship between two words (e.g., between "man" and "woman") onto a different word (e.g., "king") and provide a reasonable semantic analog (i.e., "queen").

While word embeddings are effective at capturing semantic and syntactic properties of words, they also have been shown to latently encode stereotypes and human biases (e.g., "computer programmer" – "man" + "woman" = "homemaker") [7]. We explore this in the context of age and generate "older" and "younger" analogs of common adjectives. We start with the 500 most common English adjectives [19] and then generate "older" and "younger" analogs for each adjective. For example, we find in one embedding that "stubborn" – "young" + "old" gives "obstinate" while "stubborn" – "old" + "young" gives

[4] For instance, the individual regression model for SOCAL output a significant positive coefficient of 0.950 for age. Coefficients, when exponentiated represent relative risk, meaning the SOCAL model is 2.56 times more likely to indicate positive for young-age-related sentences. Meanwhile, the Sasa tool is 0.597 times as likely (i.e. 1.66 times *less* likely) to indicate *negative* for young-age-related sentences.

	Sentiment Analysis Tool		Positive	Intercept	Negative	Intercept
Lexicon	AFINN	e^(coef)	1.000	0.956	1.000	0.733
		95% CI	[0.553 - 1.807]	[0.63 - 1.451]	[0.53 - 1.887]	[0.468 - 1.149]
	EmoLex	e^(coef)	**3.180***	1.119	**0.368****	0.762
		95% CI	**[1.732 - 5.839]**	[0.738 - 1.695]	**[0.141 - 0.958]**	[0.481 - 1.208]
	Happiness Index	e^(coef)	**3.743****	**1.972****	**2.373***	**0.389****
		95% CI	**[1.852 - 7.565]**	**[1.319 - 2.947]**	**[1.458 - 6.435]**	**[0.21 - 0.721]**
	NRC Hashtag	e^(coef)	1294.656	7122.400	0.001	8.871
		95% CI	[0 - 6.9E+25]	[0-3.7E+26]	[0 - 5.3E+19]	[0 - 4.0E+27]
	Opinion Lex	e^(coef)	1.000	**0.600***	1.000	**0.600****
		95% CI	[0.544 - 1.84]	**[0.39 - 0.923]**	[0.544 - 1.84]	**[0.39 - 0.923]**
	Panas	e^(coef)	N/A	N/A	1.000	**0.034****
		95% CI	N/A	N/A	[0.244 - 4.093]	**[0.013 - 0.093]**
	SOCAL	e^(coef)	**2.586***	**1.933****	1.043	**1.394***
		95% CI	**[1.109 - 6.031]**	**[1.036 - 3.605]**	[0.654 - 1.663]	**[1.001 - 1.941]**
	Umigon	e^(coef)	0.999	**0.243****	1.000	**0.392***
		95% CI	[0.482 - 2.071]	**[0.145 - 0.407]**	[0.545 - 1.836]	**[0.255 - 0.602]**
	VADER	e^(coef)	1.000	**0.238****	1.000	**0.202****
		95% CI	[0.502 - 1.994]	**[0.146 - 0.388]**	[0.479 - 2.09]	**[0.12 - 0.341]**
Corpus	Opinion Finder	e^(coef)	0.985	**0.323****	0.957	**0.538****
		95% CI	[0.491 - 1.975]	**[0.198 - 0.528]**	[0.534 - 1.716]	**[0.357 - 0.813]**
	Sasa	e^(coef)	1.034	**2.578****	**0.597***	**2.113****
		95% CI	[0.467 - 2.291]	**[1.519 - 4.376]**	**[0.38 - 0.937]**	**[1.453 - 3.072]**
	Sent140	e^(coef)	1.307	**6.501****	0.963	**52.98****
		95% CI	[0.162 - 10.56]	**[1.466 - 28.835]**	[0.133 - 6.972]	**[13.07 - 214.72]**
	Senti Strength	e^(coef)	1.000	**0.634****	1.000	0.692
		95% CI	[0.539 - 1.854]	**[0.41 - 0.982]**	[0.548 - 1.825]	[0.452 - 1.059]
	Stanford	e^(coef)	1.111	0.834	0.797	**3.209****
		95% CI	[0.496 - 2.486]	[0.46 - 1.51]	[0.421 - 1.51]	**[2.029 - 5.077]**

Table 4. Individual regression results for explicit age analysis. The results from each sentiment analysis method were fit to a multinomial log-linear regression model, resulting in a model for positive sentiment and a model for negative sentiment for each sentiment analysis method. The reference categories for each model are: neutral sentiment and "old" adjectives. Coefficients that are not significant at p<0.05 are greyed out. Exponentiated coefficients (i.e. e^coef) provide effect sizes for relative risk (e.g. the EmoLex model was 3.18 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from "old" (or "older" or "oldest") to "young" (or "younger" or "youngest") holding all else constant. Note: *p<0.05; ***p<0.01

NOTE: *The Sentiwordnet model (corpus-based) is not included in this table because it did not classify any of the sentences in phase 2 as "neutral." Instead, we used "negative" as the reference category for the Sentiwordnet multinomial. This model was **4.121** times more likely to indicate positive for a "young" sentence compared to an "old" sentence (p<0.01, 95%CI: [2.390, 7.106])*

			Positive			Negative		
	Sentiment Analysis Tool		Adj-Young	Adj-Old	Intercept	Adj-Young	Adj-Old	Intercept
Lexicon	AFINN	e^(coef)	1.01	**0.908****	**1.347****	**0.891****	0.969	**1.17****
		95% CI	[0.979 - 1.042]	**[0.878 - 0.939]**	**[1.308 - 1.387]**	**[0.862 - 0.921]**	[0.937 - 1.002]	**[1.134 - 1.207]**
	EmoLex	e^(coef)	**1.134****	**0.848****	**0.739****	**0.783****	**0.953***	**0.112****
		95% CI	**[1.103 - 1.166]**	**[0.825 - 0.872]**	**[0.72 - 0.758]**	**[0.737 - 0.832]**	**[0.902 - 1.007]**	**[0.106 - 0.118]**
	Happiness Index	e^(coef)	1.017	**0.08****	**7E+07****	0.973	**0.082****	**3E+07****
		95% CI	[0.205 - 5.054]	**[0.025 - 0.26]**	**[3E+07 - 2E+08]**	[0.196 - 4.837]	**[0.025 - 0.266]**	**[1E+07 - 7E+07]**
	NRC Hashtag	e^(coef)	**1.058****	1.008	**3.347****	0.991	1.020	**2.088****
		95% CI	**[1.005 - 1.113]**	[0.958 - 1.008]	**[3.218 - 3.481]**	[0.957 - 1.009]	[0.967 - 1.02]	**[2.004 - 2.176]**
	Opinion Lex	e^(coef)	**1.110****	**0.929****	**1.033***	**0.882****	**1.077****	**0.968***
		95% CI	**[1.076 - 1.145]**	**[0.9 - 0.959]**	**[1.003 - 1.064]**	**[0.853 - 0.912]**	**[1.044 - 1.111]**	**[0.94 - 0.997]**
	Panas	e^(coef)	**5.876****	**0.157****	**0.004****	**0.677****	**0.741****	**0.01****
		95% CI	**[4.994 - 6.914]**	**[0.105 - 0.236]**	**[0.003 - 0.005]**	**[0.586 - 0.783]**	**[0.643 - 0.853]**	**[0.009 - 0.011]**
	SOCAL	e^(coef)	**1.124****	**0.805****	**1.752****	**1.041***	**0.941****	**1.536****
		95% CI	**[1.087 - 1.162]**	**[0.779 - 0.832]**	**[1.698 - 1.808]**	**[1.005 - 1.078]**	**[0.91 - 0.973]**	**[1.489 - 1.585]**
	Umigon	e^(coef)	0.997	**0.935****	**1.126****	**0.950****	1.030	**1.207****
		95% CI	[0.964 - 1.031]	**[0.904 - 0.967]**	**[1.093 - 1.16]**	**[0.921 - 0.98]**	[0.998 - 1.063]	**[1.172 - 1.243]**
	VADER	e^(coef)	**1.071****	0.973	**0.525****	**0.626****	**1.089***	**0.059****
		95% CI	**[1.042 - 1.101]**	[0.947 - 1]	**[0.512 - 0.539]**	**[0.579 - 0.677]**	**[1.017 - 1.166]**	**[0.055 - 0.063]**
Corpus	Opinion Finder	e^(coef)	**1.581****	**0.613****	**0.026****	**0.787**	0.996	**0.057****
		95% CI	**[1.47 - 1.7]**	**[0.56 - 0.671]**	**[0.024 - 0.028]**	**[0.733 - 0.844]**	[0.932 - 1.065]	**[0.054 - 0.06]**
	Sasa	e^(coef)	**1.175****	**0.824****	**0.22****	**0.858****	**0.950****	**0.61****
		95% CI	**[1.132 - 1.22]**	**[0.792 - 0.857]**	**[0.212 - 0.228]**	**[0.832 - 0.885]**	**[0.922 - 0.978]**	**[0.593 - 0.627]**
	Sent140	e^(coef)	1.293	1.287	**483****	1.297	1.309	**237****
		95% CI	[0.825 - 2.025]	[0.823 - 2.012]	**[338.7 - 688.7]**	[0.828 - 2.032]	[0.836 - 2.051]	**[166.2 - 337.9]**
	Senti Strength	e^(coef)	1.030	**0.925****	**1.207****	**0.919****	1.006	**1.195****
		95% CI	[0.998 - 1.063]	**[0.895 - 0.956]**	**[1.172 - 1.243]**	**[0.889 - 0.95]**	[0.973 - 1.04]	**[1.16 - 1.231]**
	Sentiwordnet	e^(coef)	0.979	**0.766****	**3.216****	**0.861****	**0.900****	**2.795****
		95% CI	[0.934 - 1.026]	**[0.731 - 0.803]**	**[3.092 - 3.345]**	**[0.821 - 0.902]**	**[0.859 - 0.943]**	**[2.688 - 2.907]**
	Stanford	e^(coef)	**1.142****	**0.884****	**0.598****	**0.783****	**1.788****	**0.042****
		95% CI	**[1.111 - 1.174]**	**[0.86 - 0.909]**	**[0.583 - 0.613]**	**[0.709 - 0.865]**	**[1.666 - 1.919]**	**[0.039 - 0.045]**

Table 5. Individual regression results for the implicit age analysis. The results from each sentiment analysis method were fit to a multinomial log-linear regression. The reference categories for each model are: neutral sentiment, and "control" adjectives. Exponentiated coefficients (i.e. e^coef) provide effect sizes for relative risk (e.g. the top right coefficient -0. the EmoLex model was 1.134 times more likely to indicate positive sentiment when the adjective in a given sentence was changed from the "control" adjective to an "older" adjective as determined by the word embeddings. Note: *p<0.05; **p<0.01

"courageous". As a control, we also generate the most similar word to each adjective (e.g., in this case, also "obstinate" for "stubborn"). We then substitute these three versions of each adjective into our template sentences (i.e., the control adjective, the "older" adjective, and the "younger" adjective). We test 10 different word embedding models, the common GloVe (Global Vectors for Word Representation) embeddings provided by Pennington et al. [60]. These embeddings differ in the number of dimensions (fewer dimensions encode less information about a word) and text from which they were trained (Wikipedia, a common crawl of the Internet, and Twitter). See Table 3 for a description of each embedding[5]. Similar to phase one, we test word embedding features to probe possible sources of bias.

As in phase one, we classify each sentence according to each of the 15 sentiment analysis tools. In order to keep the number of sentences and sentiment analysis outputs to a computationally tractable level, we used three researcher-generated sentence templates in this analysis ("The <adj><noun> went to the movies", "The <adj><noun> had a lot of trouble understanding. "The <adjective><noun> wrote an amazing novel"). In addition to varying "young" and "old" variants of each adjective, we varied the gendered noun being described (e.g. "man", "woman", "person"). This results in 135,000 sentences in total (3 templates x 500 adjectives x 3 adjective types x 10 word embeddings x 3 nouns); running each through all 15 sentiment analysis tools results in 2,025,000 sentiment analysis outputs.

Results

In line with the results from phase one, which found significant differences in the sentiment of explicit age-related keywords, we also found significant differences in the sentiment of implicitly coded age-related keywords generated through word embeddings.

The full regression results indicated that sentences constructed with implicitly "old" adjectives were 0.91 times as likely to be scored positive, compared with the control adjective ($p<0.01$, 95% CI [.899, .921]). Similarly, sentences with implicitly "old" adjectives were 1.03 times more likely to be scored more negatively compared with the control adjective ($p<0.01$, 95%CI [1.017, 1.045]). Sentences with implicitly "young" adjectives were 1.09 times *more* likely to be scored positive ($p<0.01$, 95% CI [1.075, 1.101]). And sentences with implicitly "young" adjectives were 0.94 times as likely to be scored negatively ($p<0.01$, 95% CI [.926, .952]).

We included all 10 GloVe word embeddings in the full regression[6], and examined whether there was variation in effects across the different word embeddings. Although we could not isolate which embedding source yielded the most bias (due to conflation with dimensionality), the Wikipedia embeddings demonstrated the least amount of bias, whereas Twitter embeddings led to the greatest bias.

When examining the individual regressions (Table 5), we see that 9 of 15 models indicate a significantly greater likelihood of positive sentiment in "young" adjectives as compared to the control adjective (*Adjective-Young*). In contrast, 12 of 15 models exhibit a significantly lower likelihood of indicating positive sentiment for the "old" adjectives compared to the control (*Adjective-Old*). In terms of negative sentiment, 11 of 15 models see a significantly lower likelihood of indicating negative sentiment for "young" adjectives (*Adjective-Young*), but we see mixed results for the effect of old-age-related adjectives on the likelihood of indicating negative sentiment. In sum, the sentiment of young-oriented adjectives generated through the word embeddings are on the whole more likely to be rated positively and less likely to be rated negatively compared to old-oriented adjectives.

PHASE 3: ADDRESSING AGE BIAS VIA TRAINING DATA

Given that the first two phases of our work reveal the existence of age bias in sentiment analysis models the final phase aims to demonstrate a method to diminish that bias so that researchers might still take advantage of these computational approaches to study topics where attitudes toward age matter. In this phase, we modify the training dataset originally used to create the Sentiment140 classifier and train our own custom models with this filtered data. This allows us to conduct a more fine-grained analysis of bias within a single model and from where this bias originates.

Method

First, we build two custom sentiment analysis classifiers. There are two components to each classifier: the model architecture and the data upon which it was trained. Each of our custom models share the same architecture and only vary in the data that we use to train them. This allows us to directly connect output bias to changes in the train data.

The architecture of each of our custom models is a Maximum Entropy bag-of-words classifier, which is a widely-used approach in various text classification problems, including sentiment analysis, that predicts the most likely label (e.g. "positive" or "negative") for a given input using logistic regression. Bag-of-words models convert text inputs to a set of words, disregarding word order and grammar but retaining word frequency. This set of words is used as an input to the model, which then learns how different patterns of words map to the different labels across thousands of inputs. We use the Python SciKit Learn package—a common machine learning package—to create and train our models.

For training data, we needed a dataset of labeled text that we could manipulate for our custom classifiers. We adopt the train data used by Sentiment140 because it is one of only two publicly-available, annotated training datasets used to train a

[5] The GloVe embeddings are available to download for free and are commonly employed in various research and applied contexts [60].

[6] Because no one dimensionality was consistent across all word embeddings, we could not run a single statistical comparison of all GloVe sources.

Train Data	*Original*	*Age-Related*	*Age-Removed*
Increase in likelihood for a "young" sentence to be classified as "positive"	+13.61%	+24.26%	+1.18%

Table 6. The increase in likelihood that a "young" sentence will be classified as "positive" compared to its "old" counterpart. Training the model on the full, original dataset, a "young" sentence was 13.26% more likely to be "positive" compared to its "old" counterpart. There were 169 "old" and "young" sentence pairs.

corpus-based model that we tested. The training data was labeled through an automated process wherein tweets were annotated based on the presence of emoticons [28]. This original training dataset contains over 1 million tweets and corresponding labels.

We split the original Sentiment140 training dataset into two, exclusive subsets to observe whether we can isolate bias in the training phase of creating the classifier. First, we filter the training data to find tweets that include the terms "young" and "old". This leaves us with a training dataset of 13,781 tweets, which we refer to as the "Age-Related" corpus[7]. We use this dataset to determine where bias exists. We then reverse this filtering process to create a second dataset that *excludes* these age-based tweets (referred to as the "Age-Removed" corpus). This dataset allows us to diagnose the extent to which bias in the Age-Related corpus impacts output bias. We also retain the original, unfiltered dataset to implement the "Original" classifier.

Similar to the first phase of our analysis, we run each of our custom-trained models on a test set of sentences sourced from the posts and comments of a blog within the elder blogger community. After randomly selecting 169 sentences containing the term "old", we duplicate the sentences and replaced the term "old" with "young" to double the set to 338 sentences, which are then used to test the custom classifiers for the presence of bias (i.e. difference between output probabilities for "old" and "young" sentences). We increased the sample size to provide greater sensitivity and to help illuminate whether our filtering approaches could be effective. For this phase of analysis, we analyze the outputs from each of the custom-trained models using a paired t-test to determine the extent of bias that results from training on each of the different corpora. Specifically, given an input sentence, our classifiers provide a probability that each possible output category ("positive" or "negative") is correct. Unlike previous phases, we use this continuous probability, rather than a categorical output of 'positive' or 'negative' in our statistical model. For example, our *Original* model (i.e. trained on the unfiltered training data) classifies the sentence, "As much as I work on acceptance of getting old, I don't like it!" as "negative" with a 0.9509 probability (i.e. 95% confidence) and just 0.0491 probability (i.e. 5% confidence) in the remaining outcome ("positive"). We run the paired t-tests on these probabilities for each output category. If there were no bias (i.e. if the classifier treated "old" and "young" as equivalent in sentiment), we would expect an equal number of positive outcomes for "old" and "young" sentences. Although the Sentiment 140 model did not exhibit statistically significant bias in the phase one analysis, the model estimates trended in the expected direction. Additionally, in phase three we use a larger dataset (169 sentences vs. 121 sentences in phase one) and use a continuous probability rather than categorical output, which provide a more sensitive measure for bias. Notably, while the models we test in phases one and two also include "neutral" as a class, our custom implementation only differentiates explicitly between "positive" and "negative."

Train Data	*Original*	*Age-Related*	*Age-Removed*
Mean confidence "young" – "pos"	0.5867	0.5161	0.5671
Mean confidence "old" – "pos"	0.5196	0.4492	0.5608
Mean Difference *[95% CI]*	0.0671 *[0.023, 0.111]*	0.0669 *[0.023, 0.111]*	0.0063 *[-0.038, .050]*
***p*-value**	*p<.0027**	*p<.0028**	*p<.7796*

Table 7. T-test results for custom-trained classifiers. A likelihood above .50 produces a classification of "positive".

By isolating the age-related tweets in our different training corpora, we can determine the source of the output bias and assess whether manipulating examples of "old" and "young" can effectively prevent our custom classifier from exhibiting age-related patterns of bias possibly rooted in these training examples. If we observe the greatest bias in the *Age-Related* and *Original* corpora, this would indicate that the output bias is embedded in the labels of these age-related tweets. If we observe the greatest bias in the *Age-Removed* corpus, however, this would indicate that the output bias results from a dearth of training examples related to age and aging. Finally, if there is no significant difference in bias between the *Original* and *Age-Removed* corpora, this would indicate that the output bias largely derives from other, less contextually relevant tweets in the original dataset. Worth noting is that our approach addresses a reduction in *explicit* age-related bias, rather than *implicit* bias, which may manifest as coded language or stereotyping.

Results

Table 6 shows the increase in likelihood for a sentence to be classified as "positive" when "old" is replaced with "young". Table 7 shows the results of each classifier output side-by-side. Overall, we find the greatest output bias in classifiers trained on the *Age-Related* and *Original* corpora (both of which contain tweets with "old" and "young") and no significant bias in the *Age-Removed* corpora. This indicates that the output bias does indeed originate from bias in the labels of age-related tweets and can be remedied by removing these training examples.

[7] Although "old" and "young" have several definitions and are not always used to describe humans, our custom classifier does not feature word sense disambiguation. However, we created an additional dataset filtered by age-related phrases to isolate uses of "old" and "young" strictly with respect to humans (e.g., "young man", "older people"). This dataset was much smaller than the others we produced (1,550 training examples) and produced outputs similar to those of our dataset filtered on the terms "old" and "young."

The custom classifier trained on the *Original* dataset produced significant bias with respect to the terms "old" and "young" ($p<.0027$) where sentences containing the terms "old", "older", or "oldest" were more likely to be classified as negative. This result is in line with those of our phase one aggregated analysis. The custom classifier trained on the *Age-Related* corpus also produced significant bias ($p<.0028$). The outputs of this classifier were more negative compared to the custom classifier trained on the full Sentiment140 dataset, indicating the age-related tweets in the training data were more negative than the overall corpus.

The custom classifier trained on the Age-Removed corpus did not show significant bias ($p<.7796$). The reduction in bias compared to the classifier trained on the original dataset and the classifier trained on age-related tweets, was statistically significant ($p<.0008$). Notably, the mean gap in likelihood for an "old" vs. "young" sentence to be classified as positive was an order of magnitude lower compared with the other two classifiers (0.0063 vs. 0.0671 and 0.0669). Although the *Age-Removed* model had a slightly higher probability of classifying "young" sentences as positive as compared to "old" sentences as positive, this stemmed from only classifying two (out of 169) sentences containing the term "old" differently than their "young" counterparts. In both of these instances, the classification was negative for the "old" sentence and positive for the "young" sentence.

DISCUSSION

This study demonstrates significant age-related bias across common sentiment analysis tools and word embedding models as well as one approach to diminishing bias in training data. The findings have implications for how researchers interpret sentiment analysis results, the strategies we use to understand and mitigate bias, and the challenges of using these techniques to study online social movements.

Implications of Age-Related Bias

These findings have implications for text-based analyses of content describing older adulthood and aging. We extracted sentences for our analyses from a community of older adult bloggers, which primarily discusses the experience of aging. Discussions here cover a wide range of topics, such as politics, health, government, pop culture, and news, in relation to the experience of an older person. Thus, when the aforementioned sentiment analysis tools are applied to understanding the views, opinions, and experiences reported in this corpus, the sentiment output is less positive simply because the sentences describe an older person taking part in an interaction. For example, the statement "This old guy was 3 or 4 feet from the tide line and the tide was going out" was rated less positively than the sentence "This young guy was 3 or 4 feet from the tide line and the tide was going out." This is problematic when we examine sentences from this corpus that may be mined by algorithms to understand attitudes towards products ("I love seeing older non-professional women modelling clothes."), health information ("The older adults' brain scans showed activity in the same area"), and learning ("Life and learning does not end in old age"). Decisions by researchers and companies can be influenced by the relative sentiment of older adults' experiences compared to younger people, potentially affecting the products and services available to them. Additionally, researchers using sentiment analysis to understand attitudes across the lifespan would find that statements describing older adulthood ("Every one is telling the world what it's really like to get older") are inherently less positive than those describing youth, even when this language is designed to promote positive associations with growing older.

Strategies for Addressing Bias

The analysis of our custom trained, maximum entropy models highlighted that we could reduce bias in our research context by resampling training data from a larger dataset. This relatively simple change to the training data reduced the age-related bias to statistically insignificant levels and highlights one way in which researchers can begin to account for some kinds of bias in datasets as well as how they might adapt available datasets to their particular research context. However, our approach may not work similarly for other types of machine learning models such as those built on recurrent neural networks, which are sensitive to word order and syntax, and it does not address subtler instances of bias, such as the association of broader topics with gender (e.g., women and relationship- and family-related topics) [75].

Our approach is particularly relevant with regard to studying underrepresented populations. When data pertaining to a particular participant population is sparse or difficult to obtain, adapting a large, existing, annotated dataset may be more feasible than collecting sufficient data and annotating it to train a model. Many other approaches also point to underlying bias in the ways certain algorithms operate and generate output. While some researchers consider quantitative approaches to artificially remove bias from a dataset [6], such an approach would be difficult to employ across all instances of social bias and neglects the fact that social bias rarely exists along a single dimension (i.e., the notion of intersectionality [17]). While datasets can be tailored by sampling from certain communities, the complexity of language makes it virtually impossible to create a dataset free of social bias along all dimensions.

Given this complexity, contextualizing how we apply, interpret, and report algorithmic outputs is an important step toward avoiding conclusions that a given algorithmic output is ground truth, free of social bias or universally agreed upon. Instead, researchers should view the output score of a machine learning sentiment model as an approximation of the subjective opinion of individuals represented in the training data. In the context of our study, this means that for the classifiers trained on Twitter data, sentiment outputs are a determination of how that particular sample of Twitter users would interpret the input text rather than approximating how the socially underrepresented group would interpret the text. One explanation for why the corpus-based models

exhibited age-related bias is because the underlying datasets were also drawn from a predominantly younger demographic present on social media sites [61]. Relatedly, we identified a significant source of age bias in our training dataset after retraining our custom classifier specifically on the *Age-Related* corpus. However, we know that older adults also internalize ageist messaging [49], which means that training on data from older people may still result in bias.

Regardless, researchers and organizations creating machine learning models can consider adding context to algorithmic outputs by describing the data used in training and the population who generated it. Currently, the origins of datasets are not always explicit or available (e.g., IBM's Alchemy, Microsoft's Cognitive Services), and the models and datasets may not be modifiable, as was the case for the majority of sentiment analysis models we studied. In fact, only one of the corpus-based models we tested in phase one features both publicly available training data *and* a model that others can re-train on custom datasets. For this reason, it is particularly critical to rethink "off-the-shelf" use of these tools – that is, the use of sentiment analysis models that have not been tailored to the particular context of use. Of course, even in the rare instance that a sentiment model can be modified for a specific context, getting the necessary training data and training a model require sufficient expertise.

Challenges of Studying Social Movements

Using computational techniques to study social movements and the emergence of non-dominant narratives situated within particular cultures is becoming increasingly common (e.g., [67,72]). However, a central aspect of social movement formation involves using language strategically to destabilize dominant narratives in society and calling attention to underrepresented social perspectives. That is, language use is changing and evolving along with the emergent social movement. The shifting terms and specialized uses of language present a clear challenge for sentiment analysis models trained on static data that do not reflect evolving attitudes and language usage. For example, a recent analysis showed that as part of forming a social movement recognizing ageism, older adult bloggers reframed the aging experience as a positive and natural aspect of life [47]. One way in which they did this was by "reclaiming" words related to age that may have negative connotations, such as "gray," to have positive or alternate meanings (e.g., "Go gray!" is a common phrase used within this blogging community to positively promote changing hair color with age). The misalignment between static training datasets and evolving contemporary language makes understanding bias increasingly complex.

Engaging members of these communities in annotating train data for machine learning models is one way of addressing this misalignment between the views of a particular community and those represented in the underlying dataset. Yet, annotation methods vary widely. Some feature automatic tweet annotation based on the inclusion of specific emoticons [28], while others employ trained annotators using criteria for removing inconsistent or inaccurate annotations [35]. Nevertheless, all annotated datasets have limitations [64]. Even with custom annotated datasets, retraining models requires the means to acquire sufficient data, technical skills to preprocess data, as well as time and computational power. Future work could examine how to lower barriers and provide incentives to support community-oriented participatory data annotation. This strategy draws on Baumer et al.'s case study for adapting the Delphi Method for CSCW [4], which is one strategy for involving respondents in validating researchers' interpretations of data.

While there are challenges to applying computational techniques to text-based corpora, one advantage is the ability to observe phenomena of bias at a societal level that may be difficult to detect on an individual or case study basis. Within research on aging, tools such as the Implicit Association Test [29] and Measurement of Aging Anxiety [46] are used to assess attitudes towards older adulthood (e.g., [34]). While these tests are useful for understanding individual attitudes, they probe overtly about issues of age bias and would require an extremely large sample to understand broader views on aging in any specific cultural context. In contrast, sentiment analysis methods can be a lens for understanding underlying bias that is difficult for humans to detect overtly themselves. Indeed, age-related bias is a global phenomenon that has until recently been largely neglected in discussions of social justice and equality [12,57]. Sentiment analysis could be used as a barometer to understand broader attitudes towards various social dimensions of society, such as aging.

CONCLUSION

This paper systematically compares a number of popular and diverse sentiment tools, with respect to age-related bias. We find significant age-related bias among a variety of tools and commonly-used word embeddings and successfully reduce bias in a custom-built classifier. While we provide a first step in understanding how the technical characteristics of sentiment algorithms affect bias and identify one technique for reducing bias, our analysis is not exhaustive. Future work should consider additional characteristics of algorithmic models, such as the type of classifier implemented and richer model parameters. Further, researchers should consider the unique challenges of using computational techniques such as sentiment analysis to study underrepresented groups and social movements. As the "new power brokers in society," [22] algorithms affect many aspects of life, including hiring, social policy, and finance; all of which are domains where age discrimination is common. In addition to understanding social bias in algorithms, we can use them as a lens to understand how unrecognized social bias operates at scale.

ACKNOWLEDGMENTS

This work was supported in part by NSF grant IIS-1551574. We thank the bloggers who made their discourse and experience with aging publicly available online.

REFERENCES

1. Paul Baker and Amanda Potts. 2013. "Why do white people have thin lips?" Google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies* 10, May 2015: 187–204. https://doi.org/10.1080/17405904.2012.744320
2. Shaowen Bardzell. 2010. Feminist HCI : Taking Stock and Outlining an Agenda for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, 1301–1310.
3. Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, 675–684.
4. Eric P.S. Baumer, Xiaotong Xu, Christine Chu, Shion Guha, and Geri K. Gay. 2017. When Subjects Interpret the Data: Social Media Non-use as a Case for Adapting the Delphi Method to CSCW. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*: 1527–1543. https://doi.org/10.1145/2998181.2998182
5. Michael S Bernstein, Eytan Bakshy, Moira Burke, Brian Karrer, and Menlo Park. 2013. Quantifying the Invisible Audience in Social Networks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*: 21–30.
6. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Quantifying and Reducing Stereotypes in Word Embeddings. *arXiv preprint*.
7. Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*.
8. Danah Boyd, Karen Levy, and Alice Marwick. 2014. The Networked Nature of Algorithmic Discrimination. *Data and Discrimination: Collected Essays. Open Technology Institute.*
9. Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. 2016. "Why would anybody do this?": Understanding Older Adults' Motivations and Challenges in Crowd Work. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*: 2246–2257. https://doi.org/10.1145/2858036.2858198
10. Robin Brewer and Anne Marie Piper. 2016. "Tell It Like It Really Is": A Case of Online Content Creation and Sharing Among Older Adult Bloggers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 5529–5542. https://doi.org/http://dx.doi.org/10.1145/2858036.2858379
11. Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 32, 1.
12. Robert N. Butler. 1969. Age-ism: Another form of bigotry. *Gerontologist* 9, 4: 243–246. https://doi.org/10.1093/geront/9.4_Part_1.243
13. Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora necessarily contain human biases. *Science* 356: 183–186. https://doi.org/10.1126/science.aal4230
14. Aylin Caliskan-islam, Joanna J Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *arXiv:1608.07187v2 [cs.AI] 30 Aug 2016*: 1–14.
15. Le Chen, Alan Mislove, and Christo Wilson. 2015. Peeking Beneath the Hood of Uber. *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*: 495–508.
16. Kate Crawford. 2016. Can an Algorithm be Agonistic ? Ten Scenes from Life in Calculated Publics. *Science, Technology, & Human Values* 41, 1: 77–92. https://doi.org/10.1177/0162243915589635
17. Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Stanford Law Review* 43, 6: 1241–1299.
18. Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10)*, 241–249.
19. Mark Davies. 2008. *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*. BYE, Brigham Young University.
20. Michael A Devito. 2016. From Editors to Algorithms. *Digital Journalism*: 1–21. https://doi.org/10.1080/21670811.2016.1178592
21. Michael Devito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms ruin everything": # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), In press*. https://doi.org/10.1145/3025453.3025659
22. N Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. *Tow Center for Digital Journalism: A Tow/Knight Brief*.

23. Lynn Dombrowski, Ellie Harmon, and Sarah Fox. 2016. Social Justice-Oriented Interaction Design: Outlining Key Design Strategies and Commitments. *Proceedings of the Designing Interactive Systems Conference (DIS '16)*: 656–671.
24. Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it , then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 2371–2382.
25. Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*: 153–162.
26. Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM* 56, 4: 82–89.
27. Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3: 330–347. https://doi.org/10.1145/249170.249184
28. Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1, 12.
29. Anthony G Greenwald, Debbie E Mcghee, and Jordan L K Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Soclal Psychology* 74, 6: 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464
30. Philip J Guo. 2017. Older Adults Learning Computer Programming: Motivations, Frustrations, and Design Opportunities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '17)*.
31. Dave Harley and Geraldine Fitzpatrick. 2009. YouTube and intergenerational communication: the case of Geriatric1927. *Universal Access in the Information Society* 8, 1: 5–20. https://doi.org/10.1007/s10209-008-0127-y
32. Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)*, 168–177. https://doi.org/10.1145/1014052.1014073
33. Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1 (ACL '12)* Vol. 1.
34. Mary Lee Hummert, Teri A. Garstka, Laurie T. O'Brien, Anthony G. Greenwald, and Deborah S. Mellott. 2002. Using the Implicit Association Test to measure age differences in implicit social cognitions. *Psychology and Aging* 17, 3: 482–495. https://doi.org/10.1037//0882-7974.17.3.482
35. CJ J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*, 216–225.
36. Lucas D. Introna and Helen Nissenbaum. 2000. Shaping the Web: why the politics of search engines matters. *The Information Society* 16: 169–185. https://doi.org/10.1080/01972240050133634
37. Lucas D Introna and David Wood. 2004. Picturing Algorithmic Surveillance : The Politics of Facial Recognition Systems. *Surveillance & Society: CCTV Special Issue* 2, 2/3.
38. Lucas Introna and Helen Nissenbaum. 2000. Defining the Web: The Politics of Search Engines. *Computer 33*, 54–62.
39. Lilly Irani, Janet Vertesi, and Paul Dourish. 2010. Postcolonial computing: a lens on design and development. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*: 1311–1320. https://doi.org/10.1145/1753326.1753522
40. Isaac Johnson, Connor McMahon, Johannes Schöning, and Brent Hecht. 2017. The Effect of Population and "Structural" Biases on Social Media-based Algorithms – A Case Study in Geolocation Inference Across the Urban- Rural Spectrum. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*.
41. Matthew Kay, Cynthia Matuszek, and Sean a. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, 3819–3828. https://doi.org/10.1145/2702123.2702520
42. Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1. https://doi.org/10.1080/1369118X.2016.1154087
43. Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Krahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported*

Cooperative Work and Social Computing (CSCW '17), 417–432.

44. Nathan R. Kuncel, Deniz S. Ones, and David M. Klieger. 2014. In Hiring, Algorithms Beat Instinct. *Harvard Business Review* May.
45. Joanna N. Lahey. 2010. International Comparison of Age Discrimination Laws. *Research on Aging* 32, 6: 679–697. https://doi.org/10.1126/scisignal.2001449.Engineering
46. K.P. Lasher and P.J. Faulkender. 1993. Measurement of Aging Anxiety: Development of the Anxiety About Aging Scale. *The International Journal of Aging and Human Development* 37, 4: 247–259. https://doi.org/10.2190/1U69-9AU2-V6LH-9Y1L
47. Amanda Lazar, Mark Diaz, Robin Brewer, Chelsea Kim, and Anne Marie Piper. 2017. Going Gray, Failure to Hire, and the Ick Factor: Analyzing How Older Bloggers Talk about Ageism. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '17)*.
48. Amanda Lazar, Caroline Edasis, and Anne Marie Piper. 2017. A Critical Lens on Dementia and Design in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, In press.
49. Becca Levy. 2009. Stereotype Embodiment: A Psychosocial Approach to Aging. *Current directions in psychological science* 18, 6: 332–336.
50. Q. Vera Liao, Wai-Tat Fu, and Markus Strohmaier. 2016. #Snowden: Understanding Biases Introduced by Behavioral Differences of Opinion Groups on Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 3352–3363.
51. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*.
52. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*, 3111–3119. https://doi.org/10.1162/jmlr.2003.3.4-5.951
53. Boaz Miller and Isaac Record. 2016. Responsible epistemic technologies: A social-epistemological analysis of autocompleted web search. *new media & society*: 1–19. https://doi.org/10.1177/1461444816644805
54. Claire Cain Miller. 2015. Can an Algorithm Hire Better Than a Human. *The New York Times*.
55. Karine Nahon. 2015. Where there is Social Media there is Politics. In *Forthcoming in Routledge Companion to Social Media and Politics*, A. Bruns, E. Skogerbo, C. Christensen, O.A. Larsson and G.S. Enli (eds.). Routledge, NYC, NY.
56. Helen Nissenbaum. How Computer Systems Embody Values. *Computer* 34, 3: 118–119.
57. Alana Officer, Mira Leonie Schneiders, Diane Wu, Paul Nash, Jotheeswaran Amuthavalli Thiyagarajan, and John R. Beard. 2016. Valuing older people: Time for a global campaign to combat ageism. *Bulletin of the World Health Organization 94*, 709–784. https://doi.org/10.2471/BLT.16.184960
58. Bo Pang and Lillian Lee. 2006. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval* 1, 2: 91–231. https://doi.org/10.1561/1500000001
59. Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms That Control Money*. Harvard University Press.
60. Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*: 1532–1543. https://doi.org/10.3115/v1/D14-1162
61. Pew Research Center. 2014. Older Adults and Technology Use. April. https://doi.org/202.419.4500
62. Filipe N. Ribeiro, Matheus Ara??jo, Pollyanna Gon??alves, Marcos Andr?? Gon??alves, and Fabr??cio Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1: 1–29. https://doi.org/10.1140/epjds/s13688-016-0085-1
63. Jennifer A. Rode. 2011. A theoretical agenda for feminist HCI. *Interacting with Computers* 23, 5: 393–400. https://doi.org/10.1016/j.intcom.2011.04.005
64. Shilad Sen, Margaret E Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao "Ken" Wang, and Brent Hecht. 2015. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*, 826–838.
65. Thomas Smyth and Jill Dimond. 2014. Anti-Oppressive Design. *interactions* 21, 6: 68–71.
66. Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods*

in natural language processing (EMNLP), 1631–1642. https://doi.org/10.1371/journal.pone.0073791

67. Kate Starbird and Leysia Palen. 2012. (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*: 7–16. https://doi.org/10.1145/2145204.2145212
68. L Sweeney. 2013. Discrimination in online ad delivery. *acmqueue* 11, 3. https://doi.org/10.1145/2460276.2460278
69. Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* 37, 2: 267–307. https://doi.org/10.1162/COLI_a_00049
70. Antonio Torralba and Alexai A. Efros. 2011. Unbiased look at dataset bias. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
71. Zeynep Tufekci. 2014. Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency. *Journal on Telecommunications and High Technology Law* 13: 203–218.
72. Marlon Twyman, Brian C. Keegan, and Aaron Shaw. 2016. Black Lives Matter in Wikipedia: Collaboration and Collective Memory around Online Social Movements. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, 1400–1412. https://doi.org/10.1145/2998181.2998232
73. W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Springer, New York.
74. John Vines, Gary Pritchard, Peter Wright, Patrick Olivier, and Katie Brittain. 2015. An Age-Old Problem: Examining the Discourses of Ageing in HCI and Strategies for Future Research. *ACM Transactions on Computer-Human Interaction* 22, 1.
75. Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science* 5, 1. https://doi.org/10.1140/epjds/s13688-016-0066-4
76. Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, 34–35.