Addressing Age-Related Bias in Sentiment Analysis *

Mark Díaz[1] , **Isaac Johnson**[1] , **Amanda Lazar**[2] , **Anne Marie Piper**[1] and **Darren Gergle**[1]
[1]Northwestern University
[2]University of Maryland
{mark.diaz, isaacj}@u.northwestern.edu, lazar@umd.edu, {ampiper, dgergle}@northwestern.edu

Abstract

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 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 show significant age bias is encoded in the outputs of many sentiment analysis algorithms and word embeddings, and we can alleviate this bias by manipulating training data.

1 Introduction

Sentiment analysis is often used to measure opinions in product reviews or financial markets [Hu and Liu, 2004], which can inform and drive branding decisions, political campaign strategies, and automated financial trading systems [Feldman, 2013]. 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) [Taboada *et al.*, 2011] to hybrid models integrating both approaches [Socher *et al.*, 2013]. In this paper we focus on the ways in which algorithms are sensitive to, and propagate social biases, particularly age-related bias. In this work, we define algorithmic bias as "systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" [Friedman and Nissenbaum, 1996]. In the case of age-related bias, automated methods of opinion polling may falsely report more negative attitudes toward political issues or financial investments regarding age-related concerns, such as Medicare and Social Security. 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.

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 identical 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 underrepresented populations.

2 Understanding Bias in Algorithms

Researchers have described the algorithms that drive many of the systems we use as "black boxes" [Diakopoulos, 2014]. 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 [Lahey, 2010]. Nissenbaum states that what engineers and computer scientists can contribute to the field is "a fine-grained understanding of systems...down to gritty details of architecture, algorithm, [and] code," as these are essential to "explaining the social, ethical, and political dimensions of information technologies" [Nissenbaum, 2001]. Some researchers have directly manipulated open-source algorithms to reveal the extent of structural biases [Johnson *et al.*, 2017]. Because many algorithms are proprietary, researchers have also attempted to decipher algorithms by interpreting output while varying inputs [Chen *et al.*, 2015] – We make use of both approaches in this paper.

Social bias in NLP tools can arise from a variety of sources. Some work has focused on word embeddings [Bolukbasi *et al.*, 2016] and other work has focused on algorithmic decision-making, including the auto-complete function of search engines [Baker and Potts, 2013], advertisements based on search terms [Sweeney, 2013], and image search results [Kay *et al.*, 2015], which can propagate harmful racial and gender stereotypes. 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 [Caliskan *et al.*, 2017].

[*] Full paper, *Addressing Age-Related Bias in Sentiment Analysis*, was published at ACM CHI 2018.

Similarly, Sen et al. show 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 [Sen *et al.*, 2015], which is a view we take in this paper.

The questions motivating this work are whether age bias manifests in sentiment model outputs and, if so, what this bias looks like across commonly-used sentiment analysis models in a realistic research context. We evaluate the use of sentiment tools on a text-based corpus of blog discussions on aging to observe how age bias may manifest in a naturalistic context. Here, we summarize the high level approach and findings of our detailed study [Díaz *et al.*, 2018].

3 Phase 1: Explicit Encoding of Age

The goal of our first phase of analysis was to determine whether sentiment analysis tools treat explicit indications of age (e.g., "old" and "young") differently. We performed our analysis using 15 popular sentiment analysis tools shown in Table 1 [Ribeiro *et al.*, 2016]. We tested multiple sentiment analysis tools to minimize the likelihood of reporting idiosyncratic findings from a single tool and to compare common implementation techniques that may influence bias. We standardized model outputs to negative (-1), neutral (0), or positive (+1). We also coded each sentiment tool according to its underlying design (lexicon-based vs. corpus-based) and the training and validation data used in building the model (social media vs. other sources).

We made two multinomial log-linear regression models: 1) a single full model for each phase of analysis that included outputs from all of the sentiment tools in order to test for the presence of age-related bias across all models (Table 1), and, 2) individual models for each sentiment tool (15 in total) in order to assess which specific tools demonstrated age-related bias. We report the results of only the full model here. 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 tools according to the type of sentiment tool used (lexicon-based vs. corpus-based), and model validation data (social media vs. other).

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 sentiment tools, type and validation data. "Other" data is predominantly based on product reviews or news.

3.1 Context of Study and Testing Data

It is important to understand the impact of computational techniques and potential bias within a particular topic of study [Sen *et al.*, 2015]. 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 [Brewer *et al.*, 2016] and analyzes online blog-based discussions of age discrimination in the U.S. and U.K. [Lazar *et al.*, 2017].

We sourced sentences for the analysis by scraping 4,151 blog posts from a prominent "elderblogger" community [Lazar *et al.*, 2017] as well as 64,283 comments on posts created between 2004 and 2016. Each researcher independently, randomly sampled posts and comments containing sentences with the word "old". Of these posts, we extracted 162 unique sentences. We excluded sentences using "old" to modify nouns other than people (e.g., "old movie") and as a general descriptor of age (e.g., "the 32-year-old"). We also excluded 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 included, "Old age is worth waiting for." Although the term "old" appears 86,145 times across our corpus, our exclusion process resulted in 121 sentences from our sample.

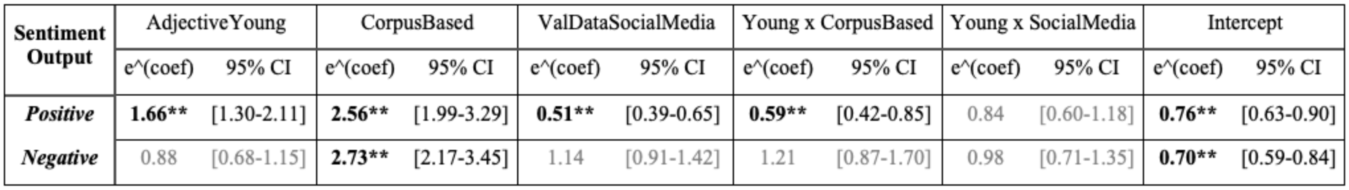
In each of the 121 sentences, we replaced the term "old" (as well as "older" and "oldest") with the term "young" (as well as "younger" and "youngest") to provide a comparative dataset (242 sentences total). By using a standardized set of sentences and varying only the age-related terms, we were able to attribute any observed changes in sentiment score to the particular words we varied.

3.2 Results

First, the results of the regression (Table 2) revealed that across all of the sentiment analysis tools, sentences containing young adjectives (AdjectiveYoung) were 66% more likely to be scored positively than the same sentences containing old adjectives, when controlling for other sentential content.

Second, supervised learning-based tools (corpus-based vs. 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 tools were more likely to indicate positive for "young", there was a disproportionate effect on pushing "young" sentences toward positive sentiment.

Third, sentiment analysis tools validated against social media data were less likely to rate sentences as positive (vs. neutral) compared with tools validated against other data.



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: Results for explicit age analysis. Models include data from all sentiment tools and are multinomial log-linear regressions. Reference categories are: neutral sentiment, "old" adjectives, lexicon-based approaches, and non-social-media validation data. Exponentiated coefficients (i.e., e^coef) provide relative risk. Note: *p<0.05; **p<0.01

4 Phase 2: Implicit Encoding of Age

Next, we analyzed whether age-related bias may be rooted in how word embeddings encode implicit associations with age and aging. We again manipulated specific words in sentence templates, but this time we generated the adjectives using a list of common English adjectives and skewed them "old" or "young" through the use of vector math on word embeddings.

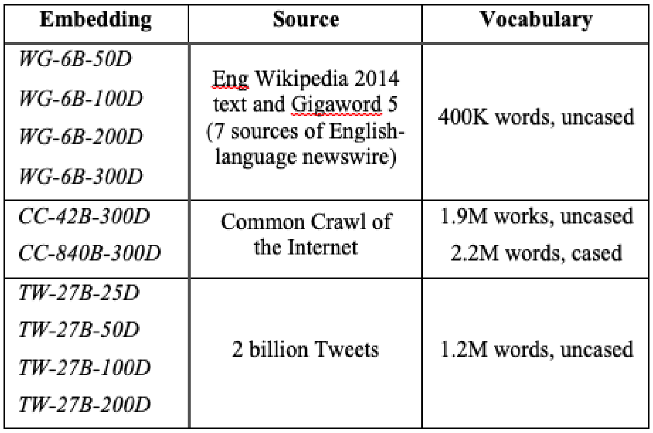
Word embeddings have been shown to encode stereotypes and human biases (e.g., "computer programmer" – "man" + "woman" = "homemaker") [Bolukbasi *et al.*, 2016]. Starting with the 500 most common English adjectives [Mark, 2008], we generate "older" and "younger" analogs for each adjective using the 10 GloVe (Global Vectors for Word Representation) embeddings [Pennington *et al.*, 2014], shown in Table 3. For example, in one embedding "stubborn" – "young" + "old" gives "obstinate" while "stubborn" – "old" + "young" gives "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 "old" adjective, and the "young" adjective).

We classified each sentence according to each of the 15 sentiment analysis tools. To keep the number of sentences and sentiment analysis outputs computationally tractable, we used three researcher-generated sentence templates ("The $\langle adj\rangle$ $\langle noun\rangle$ went to the movies", "The $\langle adj\rangle$ $\langle noun\rangle$ had a lot of trouble understanding. "The $\langle adj\rangle$ $\langle noun\rangle$ wrote an amazing novel"). In addition to varying "young" and "old" adjectives, we varied the gendered noun described (e.g. "man", "woman", "person"). This resulted 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 tools resulted in 2,025,000 outputs.

4.1 Results

As in phase one, sentences with implicitly "young" keywords were more likely to be classified as 'positive'.

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]).



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

Table 3: The GloVe embeddings. Each name references the source, token count (e.g., 6B = 6 billion), and the number of dimensions.

5 Phase 3: Addressing Bias via Training Data

Next, we modified the training data originally used to create the Sentiment140 classifier and trained custom models. This allowed us to conduct a more fine-grained analysis of bias within a single model and locate from where this bias originates. First, we built two Maximum Entropy bag-of-words classifiers. Each of our custom models shared the same architecture and only varied in their training data. This allowed us to connect output bias to changes in the training data.

We use the training data from Sentiment140 because it is one of only two publicly-available, annotated training datasets from a corpus-based model that we tested. We split the dataset of 1 million tweets into two subsets in an attempt to isolate a biased training subset. We filtered the training data to find tweets with the terms "young" and "old". This left a training dataset of 13,781 tweets, which we refer to as the "Age-Related" corpus. We used this dataset to determine where bias exists. We then repeated this process to create a second dataset that *excludes* these age-related tweets (referred to as the "Age-Removed" corpus). This dataset allowed us to diagnose the extent to which bias in the Age-Related corpus impacts output bias. We retained the original, unfiltered dataset to implement the "Original" classifier. By isolating age-related tweets in our different training corpora, we can

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 4: T-test results. Confidence >.50 produces “positive”.

determine the source of the output bias and assess whether manipulating examples of “old” and “young” can prevent our custom classifier from exhibiting age-related patterns of bias possibly rooted in these training examples.

We repeated our phase one test set creation, randomly selecting sentences with the term “old”, duplicating them, and replacing “old” with “young” to double the set. For greater sensitivity, we sampled more sentences (169 to produce 338 total sentences) and analyzed model outputs using a paired t-test. Unlike previous phases, we tested model confidence, rather than the categorical output of ‘positive’ or ‘negative’. We ran the paired t-tests on these confidences for each output category. If there is no bias (i.e. if the classifier treated “old” and “young” as equivalent in sentiment), we would expect equal confidence for “old” and “young” sentences.

5.1 Results

We found 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 indicated that output bias does indeed originate from the labels of age-related tweets and can be remedied by removing these training examples.

The 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 mirrors those of our phase one analysis. The classifier trained on the Age-Related corpus also produced significant bias (p<.0028). The outputs of this classifier were more negative compared to the classifier trained on the full Sentiment140 dataset, indicating the Age-Related tweets in the training data were more negative than the overall corpus.

The 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).

6 Implications of Age-Related Bias

Our findings have implications for text-based analyses of content describing older adulthood. We extracted sentences from a community of older adult bloggers, which primarily discusses the experience of aging. Discussions here cover a wide range of topics in relation to the experience of an older person. Thus, when the aforementioned sentiment analysis tools are applied to understanding the views reported in this corpus, the output is less positive simply because the sentences describe an older person taking part in an interaction.

This is problematic when examining sentences that may be mined for attitudes towards products (“I love seeing older non-professional women modelling clothes.”) or health information (“The older adults’ brain scans showed activity in the same area”). Analyses can be influenced by the measured sentiment of older adults’ experiences compared to younger people, potentially changing product and service decisions.

The analysis of our custom models highlighted that we could reduce bias by resampling training data from a larger dataset. However, our approach may not work for other types of classifiers such as those built on recurrent neural networks, which are sensitive to word order and syntax, and it does not address subtler instances of social bias, such as the association of broader topics with gender (e.g., women and family-related topics) [Wagner *et al.*, 2016].

Our approach is particularly relevant with regard to studying underrepresented populations. When data pertaining to a particular population is sparse or difficult to obtain, adapting a large, existing, annotated dataset may be more feasible than collecting sufficient data and annotating it. While some researchers consider quantitative approaches to artificially remove bias from a dataset, 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 [Crenshaw, 1990]). The complexity of language makes it virtually impossible to create a dataset free of social bias along all dimensions.

Contextualizing how we apply, interpret, and report outputs is an important step toward avoiding conclusions that a given output is ground truth or free of social bias. Researchers should view the output of a sentiment model as an approximation of the subjective opinion of individuals represented in the training data. In our study, this means that, for 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. Researchers can consider contextualizing model outputs, describing the training data and the population who generated it.

7 Conclusion

This paper found age-related bias among popular sentiment tools and commonly-used word embeddings. We successfully reduced bias by creating a classifier built on modified training data. Future work should consider bias with respect to additional technical characteristics of models, and should consider the challenges of using computational techniques to study underrepresented groups.

Acknowledgments

This work was supported in part by NSF grant IIS-1551574. We thank the bloggers who made their experiences publicly available online.

References

[Baker and Potts, 2013] Paul Baker and Amanda Potts. 'why do white people have thin lips?'google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies*, 10(2):187–204, 2013.

[Bolukbasi *et al.*, 2016] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*, 2016.

[Brewer *et al.*, 2016] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. Why would anybody do this?: Understanding older adults' motivations and challenges in crowd work. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2246–2257. ACM, 2016.

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

[Chen *et al.*, 2015] Le Chen, Alan Mislove, and Christo Wilson. Peeking beneath the hood of uber. In *Proceedings of the 2015 Internet Measurement Conference*, pages 495–508. ACM, 2015.

[Crenshaw, 1990] Kimberle Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241, 1990.

[Diakopoulos, 2014] Nicholas Diakopoulos. Algorithmic accountability reporting: On the investigation of black boxes. 2014.

[Díaz *et al.*, 2018] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 412. ACM, 2018.

[Feldman, 2013] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[Friedman and Nissenbaum, 1996] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

[Hu and Liu, 2004] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[Johnson *et al.*, 2017] Isaac Johnson, Connor McMahon, Johannes Schöning, and Brent Hecht. 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 2017 CHI Conference on Human Factors in Computing Systems*, pages 1167–1178. ACM, 2017.

[Kay *et al.*, 2015] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 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*, pages 3819–3828. ACM, 2015.

[Lahey, 2010] Joanna N Lahey. International comparison of age discrimination laws. *Research on aging*, 32(6):679–697, 2010.

[Lazar *et al.*, 2017] Amanda Lazar, Mark Diaz, Robin Brewer, Chelsea Kim, and Anne Marie Piper. Going gray, failure to hire, and the ick factor: Analyzing how older bloggers talk about ageism. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 655–668. ACM, 2017.

[Mark, 2008] Davies Mark. The corpus of contemporary american english (coca): 520 million words, 1990–present, 2008.

[Nissenbaum, 2001] Helen Nissenbaum. How computer systems embody values. *Computer*, 34(3):120–119, 2001.

[Pennington *et al.*, 2014] 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, 2014.

[Ribeiro *et al.*, 2016] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[Sen *et al.*, 2015] Shilad Sen, Margaret E Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao Ken Wang, and Brent Hecht. 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*, pages 826–838. ACM, 2015.

[Socher *et al.*, 2013] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[Sweeney, 2013] Latanya Sweeney. Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*, 2013.

[Taboada *et al.*, 2011] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[Wagner *et al.*, 2016] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1):5, 2016.