Darling or *Babygirl*? Investigating Stylistic Bias in Sentiment Analysis

Judy Hanwen Shen [*1] **Lauren Fratamico** [*1] **Iyad Rahwan** [1] **Alexander M. Rush** [2]

Abstract

Sentiment analysis is increasingly used for a range of applications from customer service to opinion mining. Stylistic bias arises when text generated by different groups of people expressing the same underlying content receive disparate treatment. Using three lexical alignment techniques, we find that standard sentiment models produce significantly different sentiment scores for word pairs that mainly differ stylistically. We suggest a simple align and substitute method to automatically generate examples of potentially undesirable biases in black-box models in order to better facilitate identification and mitigation of differential treatment based on stylistic variation.

1. Introduction

Natural language processing techniques have rapidly advanced in accuracy over the past several years. Standard text analysis libraries are also becoming increasingly simple to integrate into any application. These advancements have enabled governments, corporations, and researchers to employ various off-the-shelf content analysis tools to understand their constituents. With the growing body of readily available data from social media, businesses that traditionally understand customer feedback on an individual basis can now solicit customer reactions to new products by the millions. Similarly, politicians can now understand entire groups of people through Twitter reactions to news events and trending topics.

Sentiment analysis is a frequently used tool to automate opinion collection. However, examples of bias in sentiment analysis have recently gained attention as words such as *gay* and *jew* have been reported to elicit negative sentiments[1]. While identifying bias in language *describing* different demographics is important, it is also important to detect disparities in language *generated* by different populations. Texts containing the same underlying content but exhibiting stylistic variation should, in theory, be classified similarly. If sentiment analysis is being used to understand groups of people, disparate treatment of language generated by different populations would lead to both a misunderstanding of groups of people and an under-representation of groups of voices.

This work investigates stylistic bias in sentiment analysis models. We first propose three different probabilistic methods for computing lexical alignment between different groups, in particular looking for stylistically different words with roughly the same content. We apply this alignment to datasets that differ along three different socio-economic attributes: gender, race, and political affiliation. We then assess how pre-trained sentiment analysis models differ in the presence of these identified stylistic variations. Experiments show that synonymous words pairs often elicit salient differences in model output. Since these words are used with different frequency across socio-economic groups, disparate treatment can emerge as a result of using these sentiment analysis models in production.

Related Work We broadly decompose normative bias in natural language classification and analysis into two categories: descriptive and stylistic.

Descriptive bias occurs when certain identities are more closely associated with a set of attributes or concepts than others. While differences in semantic association naturally occur through corpus co-occurrence, these differential associations can become problematic when they occur across socio-economic divides which reflect stereotypes. Bolukbasi et al. (2016) demonstrate that vector representations of words can produce biased associations (e.g. $Man - programmer + woman = homemaker$). While the corpus bias likely arises from the unconscious biases and language use of humans themselves, steps can be taken to increase the fairness of a model trained on biased data. Zhao et al.

[*] Equal contribution [1] MIT Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA [2] School of Engineering and Applied Sciences, Harvard University. Correspondence to: Judy Hanwen Shen <judyshen@mit.edu>, Lauren Fratamico <fratamico@mit.edu>.

5th Workshop on Fairness, Accountability, and Transparency in Machine Learning Stockholm, Sweden 2018. Copyright 2018 by the author(s).

[1] https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias

(2017a) show that models can magnify a dataset's biases and propose interventions at the corpus level to debias these models.

Another aspect of bias occurs due to variation in style within the same underlying content. Stylistic bias relates to the choices of writers when describing the same content (e.g. African Americans use the word *missin'* more frequently while White Americans use *missing* more frequently (Jørgensen et al., 2016)). In sociolinguistics, the lexical, phonological, syntactic patterns have been shown to vary across social class, ethnicity, gender, and age (Lawton, 1963; Jørgensen et al., 2016; Schler et al., 2006; Johannsen et al., 2015). Blodgett et al. (2016) studied the disparate mistreatment of African-American English (AAE) by dependency parsing and language identification tools. They find that both classes of tools can exhibit lower accuracy when classifying AAE than text associated with Standard American English (SAE). Hovy & Søgaard (2015) find that part-of-speech taggers trained on traditional news corpora exhibit better performance on text written by older authors. Previous work in the area of stylistic bias has focused on disparate mistreatment in tasks with a more objective ground truth (e.g. language identification, dependency and part-of-speech tagging). Sentiment labels are more impacted by pragmatic concerns. Since sentiment is highly tied to lexical choice, corpus-level lexical distributions may be crucial for precise identification. This can lead to unintended biases against groups that use a significantly different lexical distribution.

2. Approach: Identifying Lexical Variation

We proceed with our examination of sentiment analysis algorithms motivated by the ethical salience of one population *always* being perceived more positively in sentiment than another. Evaluating sentiment analysis bias of corpora originating from different groups is challenging due to the difficulty of disentangling the effect of linguistic style from the effect in the underlying content. There is no readily available corpus of content-aligned pairs of sentences which only vary in style. However there has been recent research that aims to isolate stylistic variance through neural generated semantic representations (Mikolov et al., 2013), unaligned style transfer (Shen et al., 2017; Zhao et al., 2017b) or unaligned translation models (Conneau et al., 2017).

Specifically we propose to create stylistic variation by inducing a controlled lexicon of common substitutions. If we can identify a lexical alignment across groups of interest, we can then test the stylistic biases of various sentiment analysis models by simple word replacement. While this method cannot test all factors of stylistic variance, it does focus in on a particularly important factor in sentiment analysis, word choice.

Table 1. Final size of each dataset

DATASET	GROUP	TOTAL USERS	TOTAL TWEETS
GENDER	MALE	3,839	383,900
	FEMALE	3,617	361,700
RACE	SAE	6,205	625,000
	AAE	6,652	665,200
POLITICS	TRUMP	10,768	1,076,800
	CLINTON	11,965	1,196,500

2.1. Stylistically Varying Groups

Our datasets are comprised of tweets that vary by authorship across different socio-economic attributes: gender, race, and political affiliation[2]. For gender, we use a set of female and male human annotated Twitter users from Crowdflower[3]. For race, we use Blodgett et al. (2016)'s dataset containing users labeled by probability of being of a given race based on U.S. census data. We only include users given a probability score greater than .8 of being African American (labeled AAE) or White (SAE). For political affiliation, we use a dataset of Twitter users collected during the 2016 U.S. presidential election (Vijayaraghavan et al., 2017). It contains the screen names of all Twitter users that exclusively followed either `@realDonaldTrump` or `@HillaryClinton` in the six months preceding the election from which we take $15,000$ users from each of the Trump and Clinton followers. For users in all datasets, we scrape up to their latest $3,200$ tweets.

To reduce the probability that a user was a bot, we screened each through the Botometer API (Davis et al., 2016) and remove those with a bot likelihood score of .8 or higher, removing fewer than 1%. Since the number of tweets per user varies between 1 and $3,200$, we mitigate overrepresenting individual user style by randomly sampling 100 English tweets from each user that had at least that many. The final size of the datasets can be seen in Table 1.

2.2. Inducing Lexical Alignments

We heuristically model stylistic variation by identifying semantically similar words. For each corpus pair (e.g. male-female), we construct a word frequency distribution and use Jenson-Shannon divergence between the two distributions (e.g. between male and female distributions) to extract the top 1000 most differing words in each group. We then use the following methods to find similar words in the opposite corpora (e.g. use the male list to find female synonyms).

[2] None of these attributes are binary. Our scope in this work is limited to two categories per attribute.

[3] `https://www.crowdflower.com/using-machine-learning-to-predict-gender/`

Table 2. Summary of aligned substitution process: number of pruned word pairs, number of tweets modified, and average number of substituted words per tweet

Dataset	Method	Word Pairs	Tweets	Avg. sub. words
Gender	WordNet	301	326K	2.06
	Word2Vec	379	397K	2.02
	MUSE	306	100K	1.55
Race	WordNet	209	456K	1.55
	Word2Vec	441	848K	2.55
	MUSE	233	391K	1.47
Politics	WordNet	217	530K	1.35
	Word2Vec	410	407K	1.28
	MUSE	239	208K	1.19

Method 1: WordNet We use WordNet to extract synonyms (Miller, 1995). Using each list of top words, we consider all word lemmas and all synnets of each lemma in order. We choose the first synonym which occurs more often in the opposite corpora than the current word-list corpora.

Method 2: Word2Vec We train a 100-dimensional Skip-Gram model using both corpora in each category (e.g. male and female) to build a common vector embedding space (Mikolov et al., 2013). For each word in the list of top words, we find the closest word in cosine distance which appears more often in the opposite corpora. We set a maximum cosine distance threshold to .5.

Method 3: MUSE This method builds off of research in unaligned machine translation. Conneau et al. (2017) propose a method to align word embedding spaces of two monolingual corpora to bootstrap a translations model. For each of our three demographic corpora, we treat each group (e.g. male and female) as separate "languages". First we learned the word vector representation of each group with FastText (Bojanowski et al., 2016) resulting in 300-dimension vectors obtained using the Skip-Gram model. Then we used the Multilingual Unsupervised or Supervised word Embedding library (MUSE) created by Conneau et al. (2017) to learn the alignment of our two vector spaces. Then, for each of the 1000 top words in each group, we found the closest word in the vector space used by the other group, discarding the pairs where both words were the same.

Once we obtained pairs of words from each of the three methods described above, we apply preliminary filtering to remove non-stylistic pairs. We prune the lists to remove words that often appear in similar contexts, but are not necessarily synonyms, by removing: names, sports teams, digits, special characters, and plurals (pairs where one word was the plural of another).

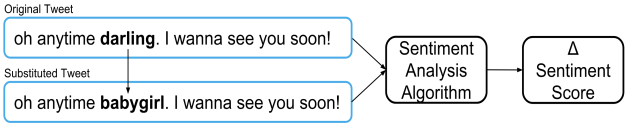
Original Tweet
oh anytime **darling**. I wanna see you soon!
Substituted Tweet
oh anytime **babygirl**. I wanna see you soon!
Sentiment Analysis Algorithm
Δ Sentiment Score

Figure 1. Aligned substitution for the word-pair: (darling, babygirl). This is a three-step process: substitute the paired word into the original tweet, obtain sentiment scores for the original and substituted tweets, and use the difference in scores for regression analysis

Table 3. Percent of Word2Vec word pairs that had a regression coefficient (i.e. sentiment score change) statistically significantly different from 0

Dataset	VADER	TextBlob	DCNN	LSTM
Gender	26.91%	23.22%	50.13%	42.48%
Race	32.43%	22.68%	49.89%	64.85%
Politics	24.69%	11.25%	41.81%	52.57%

3. Experiments

3.1. Word Pair Substitution

With the filtered list of words, we substitute the aligned word pairs into tweets in which the word pair appears over the entire dataset, as illustrated in Figure 1. For example, we use the male-female list to replace all the male words in male tweets with the corresponding female words and similarly with the female tweets. Table 2 summarizes the substitution statistics.

3.2. Sentiment Analysis Algorithms

To test the effect of stylistic lexical changes on sentiment score, we select four publicly available pre-trained models that we treat as black-box sentiment analysis algorithms:

- VADER: Rule-Based model purposed for social media data (Gilbert, 2014)
- TextBlob: Popular Python library with sentiment analysis based on Naïve Bayes[4]
- Dynamic-CNN Model: Convolutional model trained on Twitter data (Kalchbrenner et al., 2014)[5]
- LSTM Model: Recurrent model trained on 82 million Amazon product reviews (Radford et al., 2017)

We first investigate differences in mean sentiment score. For each pair of groups, there was a 1-2% difference in mean

[4] http://textblob.readthedocs.io/en/dev/

[5] We use an implementation that was trained on Twitter data from https://github.com/xiaohan2012/twitter-sent-dnn

Table 4. Examples of word pair substitutions and the regression coefficient of sentiment score change on various models. Values are bolded when statistically significant

		Regression Coefficients			
Dataset	Word Pair	VADER	TextBlob	DCNN	LSTM
Gender (Male → Female)	spectacular → glorious^{w2v}	**0.37**	**-0.24**	-0.04	**-0.19**
	embarrassing → awkward^{wn}	**0.14**	**-0.34**	**-0.21**	**0.19**
	alright → okay^{wn}	**-0.01**	**0.24**	**0.16**	**0.12**
Race (AAE → SAE)	big → huge^{w2v}	**0.14**	**0.17**	**0.20**	0.00
	yea → yeahh^{w2v}	**0.19**	0.00	**0.17**	0.00
	babygirl → darling^{w2v}	**0.32**	-0.01	**0.27**	**0.02**
Politics (Clinton → Trump)	disgusted → triggered^{w2v}	**0.39**	**0.58**	0.04	**0.37**
	marching → protesting^{w2v}	**-0.28**	0.00	**-0.13**	0.00
	clinton's → hillary's^{MUSE}	-0.03	-0.01	**-0.10**	**0.11**

sentiment, and the distributional difference between each set of corpora is statistically significant ($p < 1e^{-5}$).

To determine the effect of each word pair substitution on sentiment polarity, we build count vectors for substitutions for each tweet. If a substitution for $(w, w')_i$ is made, we count that as a 1 in the ith column. This creates count vectors the same size as the total number of word pairs per category. Using the difference in sentiment as the output label, we fit a linear regression model on these count vectors to decompose the effect of each word pair substitution.

3.3. Results

Table 3 shows the percentage of word pairs that lead to a significant change in sentiment for the different models and datasets. Specifically, we compare the regression coefficients across different word pairs as a proxy for the average change in normalized sentiment score (range -1 to 1) produced by this substitution pair. After applying Bonferroni correction, we report word pairs with regression coefficients statistically significantly different from 0. The values in this table are for the Word2Vec word pairs, but we found similar trends in terms of model performance for the other methods. For all cases, the DCNN and LSTM models resulted in the greatest number of word pairs with large sentiment differences after word substitution.

Table 4 shows some examples of similar words which elicit a large difference in sentiment score (regression coefficient) across various sentiment analysis algorithms. For example, substituting *babygirl* for *darling* increases the sentiment score across three of the models. Although both could be perceived as similar terms of endearment, *darling* is five times as likely to appear in the SAE corpus than the AAE corpus. Additionally, the effect of substituting the same word pair is not uniform across all models. For example, changing *spectacular* to *glorious* produces a positive change in VADER, but a negative change in the other three models. Our methods also generated pairs of words that appear in similar contexts across both populations, but that are not always synonyms. For example, *triggered* and *disgusted* can often be used interchangeably to express negative sentiment. However, *triggered* has other definitions outside its colloquial use where *disgusted* does not make an accurate synonym. Depending on the application, algorithmic auditors could then inspect the list generated to identify whether these biases are ethically salient.

4. Discussion

We suggest a simple aligned substitution method for testing the behavior of sentiment analysis algorithms in the presence of stylistic variation. We find similar word pairs produce significant differences in sentiment score and inconsistent behavior across different models. These inconsistencies could introduce unintended disparate treatment when these or other black-box models are used in practice.

A main limitation to our work is validating the robustness of the synonym pairs produced. Due to the ever-evolving possible word senses, it is difficult to evaluate whether words are intended to convey the same sentiment without knowledge of context. Furthermore, since style varies across different populations, human annotators could introduce significant bias in the processes of evaluating the quality of synonym production. A representative group of evaluators labeling each word pair appearing in different contexts is required to empirically conclude whether one method of generating synonyms is better than the others. Future work could utilize phrase-to-phrase style transfer and stylistic disentanglement techniques. Since these are currently open areas of research, attributing disparate treatment to purely stylistic variation remains highly challenging. Despite limitations, we identify an area of natural language bias that could be investigated as techniques in style transfer further evolve.

References

Blodgett, Su Lin, Green, Lisa, and O'Connor, Brendan. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics.

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, and Mikolov, Tomas. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y, Saligrama, Venkatesh, and Kalai, Adam T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

Conneau, Alexis, Lample, Guillaume, Ranzato, Marc'Aurelio, Denoyer, Ludovic, and Jégou, Hervé. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Davis, Clayton Allen, Varol, Onur, Ferrara, Emilio, Flammini, Alessandro, and Menczer, Filippo. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 273–274. International World Wide Web Conferences Steering Committee, 2016.

Gilbert, CJ Hutto Eric. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

Hovy, Dirk and Søgaard, Anders. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pp. 483–488, 2015.

Johannsen, Anders, Hovy, Dirk, and Søgaard, Anders. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pp. 103–112, 2015.

Jørgensen, Anna, Hovy, Dirk, and Søgaard, Anders. Learning a pos tagger for aave-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1115–1120, 2016.

Kalchbrenner, Nal, Grefenstette, Edward, and Blunsom, Phil. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

Lawton, Denis. Social class differences in language development: A study of some samples of written work. *Language and Speech*, 6(3):120–143, 1963. doi: 10.1177/002383096300600302.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Miller, George A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Radford, Alec, Jozefowicz, Rafal, and Sutskever, Ilya. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

Schler, Jonathan, Koppel, Moshe, Argamon, Shlomo, and Pennebaker, James W. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pp. 199–205, 2006.

Shen, Tianxiao, Lei, Tao, Barzilay, Regina, and Jaakkola, Tommi. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pp. 6833–6844, 2017.

Vijayaraghavan, Prashanth, Vosoughi, Soroush, and Roy, Deb. Twitter demographic classification using deep multi-modal multi-task learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pp. 478–483, 2017.

Zhao, Jieyu, Wang, Tianlu, Yatskar, Mark, Ordonez, Vicente, and Chang, Kai-Wei. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989, Copenhagen, Denmark, September 2017a. Association for Computational Linguistics.

Zhao, Junbo Jake, Kim, Yoon, Zhang, Kelly, Rush, Alexander M., and LeCun, Yann. Adversarially regularized autoencoders for generating discrete structures. *CoRR*, abs/1706.04223, 2017b. URL http://arxiv.org/abs/1706.04223.