Auditing Sentiment Analysis Algorithms for Bias

Jonathan Lo
jolo@ucsd.edu

Flynn O'Sullivan
fosulliv@ucsd.edu

Elsie Wang
e2wang@ucsd.edu

Mentor: Stuart Geiger
sgeiger@ucsd.edu

Abstract

Sentiment analysis algorithms are steadily being integrated across the web to help with tasks such as content moderation. On many major websites, whether or not a comment is allowed to be posted at least partially depends on the output of a sentiment analysis algorithm on that comment. Three such algorithms are TextBlob, VADER, and Perspective API. In this paper, we audit these three algorithms to test for bias against certain racial and gender groups. More specifically, we develop a dataset of sentences with three levels of hard-coded sentiments: positive, neutral, and negative. After developing a method to replace the racial and gender identity in the sentences, we query each sentiment analysis algorithm with each combination of race and gender. Statistical tests are computed on the output to determine if there is a significant bias in how each algorithm treats the different race and gender combinations. Our results from one-way ANOVA and Tukey's HSD find that Perspective API tended to score sentences with racial or gender identifiers as more toxic, with "white" yielding the highest toxicity scores and "Asian" the lowest. TextBlob showed variations in sentiment scores across different identities, such as assigning more negative sentiment to sentences with "black." VADER, however, did not exhibit significant differences in sentiment scores among identities. Statistical tests confirmed significant differences in sentiment scores among identities for Perspective API and TextBlob, but not for VADER. The study highlights the importance of understanding underlying factors contributing to disparities in sentimental analysis algorithms and suggests future research should focus on using validated datasets and manual processing of identity terms to promote transparency in algorithmic decision-making.

Website: https://dsc180b.lojot.com/
Code: https://github.com/jonathanlo411/dsc180b

1 Introduction

As the volume of data and internet users continues to rise, companies are faced with the challenge of effectively managing and maintaining safe content on digital platforms. The ubiquity of online interactions has prompted the adoption of sentiment analysis models on a vast amount of posts and comments to discern their toxicity levels. For instance, major newspapers use Google's Perspective API to moderate comments of articles for toxicity while VADER and TextBlob stand out as one of the largest and highly used sentimental analysis models in higher education and research. Yet, the inherent opacity of these algorithms raises critical questions about the criteria that defines toxicity and the potential biases hidden within them. This paper aims to address this question by evaluating the fairness of TextBlob, VADER, and Perspective API with self-identifying features such as race and gender. Our hypothesis states that these models are unbiased, and by using sentences that include race and gender terms and seeing how these models assign sentiment by changing these terms, we hope to gain more insight into how these models work.

1.1 Literature Review

Recent studies have shown the possibility for algorithms to learn and display human bias based on race, gender, religion, among others. For instance, one highly-cited study conducted by Sweeney (2013) audited Google AdSense ads and found that searches using black-sounding names were more likely to generate ads suggesting an arrest record compared to searches using white-sounding names.

With the pervasive rise and influence of sentiment analysis algorithms, the importance of understanding these workings and implications is fundamental in maintaining fair and unbiased results. One study that attempted to do this is Kiritchenko and Mohammad (2018). Researchers compiled 8,640 English sentences chosen to tease out race or gender and audited 219 automatic sentiment analysis systems in a shared task on predicting sentiment and emotion intensity in tweets, *SemEval-2018 Task 1: Affect in Tweets* (Mohammad 2018). Average intensity scores for sentences generated from templates were compared with female and male noun phrases and white-sounding and black-sounding names. Results showed several systems had statistically significant bias: higher scores for sadness were assigned to more female noun phrases while higher scores for fear were assigned to more male noun phrases. For race, higher scores for joy and valence were assigned to more white-sounding than black-sounding names. This example, among many others, suggests there may be underlying bias in sentiment analysis algorithms.

Another study also audited for bias in Perspective API on the tweets made by users across many demographic categories (Jiang 2020). This process involved collecting 3,000 active users on Twitter, categorized into age, race, education, and gender groups and focused on how Perspective API assigned toxicity scores. Results showed Perspective API efficiently detected profanity, but the ability to delete harmful tweets was less effective. False positives dominated, emphasizing profanity over identity-related attacks, likely due to the prevalence of profanity in social media data used for training. To audit, they calculated the

cumulative distribution funtions and conducted one-way ANOVA tests for pairs of demographics within the same group. This paper similarly employed one-way ANOVA tests to test for pairs of races, for instance black and white, along with other statistical tests. Additionally, it uses Kiritchenko and Mohammad's method of template sentences on three different sentiment analysis algorithms and a different dataset to determine whether the problem of bias in sentimental analysis algorithms still persists today.

1.2 Sentiment Analysis Algorithms

For our paper, we will be using three well-known sentimental analysis algorithms: Perspective API, TextBlob, and VADER.

Perspective API, developed by Jigsaw, a subsidiary of Google, is a tool that uses machine learning to identify "toxic" comments, which are comments that are considered rude, disrespectful, or unreasonable (Jigsaw 2021). The model scores a phrase from a range of 0 to 1, indicating the probability that a reader would perceive the comment provided as toxic. For instance, a probability score of 0.8 would indicate that 8 out 10 people would perceive that comment as toxic. Thus, Perspective API is largely used for content moderation on online platforms such as the New York Times and other local newspapers, Reddit, Coral, FACEIT, among many others. By integrating Perspective API into our audit framework, we can systematically analyze how Perpsective API categorizes racial and gender comments as toxic or not.

TextBlob is a Python library for processing textual data, including assigning sentiment scores (Loria et al. 2018). The sentiment property returns two scores, a polarity score and an objectivity score. The polarity score ranges from -1 to 1 where -1 is the most negative sentiment while 1 is the most positive sentiment. The objectivity score ranges from 0 to 1 where 0 is very objective and 1 is very subjective. With its simplicity, TextBlob has become a popular choice for researchers and developers delving into the sentiment and linguistic characteristics, including evaluating COVID-19 tweets, restaurant customer reviews, patients' opinions about healthcare, and many other papers (Kaur and Sharma (2020), Laksono et al. (2019), RamyaSri et al. (2019)). From social media monitoring to customer feedback analysis, TextBlob will allow us to uncover any bias that may be present in the tools used for academic research.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based general sentiment analysis tool for social media text (Hutto and Gilbert 2014). To calculate a compound score, the model sums up the valence scores of each word in the lexicon and then normalizes them to be between -1, which is the most extreme negative, or 1, which is the most extreme positive. It is particularly meant for handling informal text such as emoticons and slang, and thus, has been used many times to analyze text from social media. For instance, VADER has been used for predicting customer response sentiment, analyzing sentiment of bitcoin tweets during COVID-19, government responses to forest fires, among others (Borg and Boldt (2020), Pano and Kashef (2020), Mustaqim, Umam and Muslim (2020)). By evaluating VADER's capacity to process racial or gendered text, we aim to assess whether the sentiment analysis tool exhibits any biases, particularly when employed

as a means to gain insights into public opinions and sentiment trends.

1.3 Data Description

For our dataset, we curated 152 template sentences explicitly mentioning a gender and race. The racial categories include Asian, white, and black, while gender includes female and male. Each sentence is accompanied by labels of 1, 0, and -1 to indicate whether they have positive, neutral, or negative sentiment, respectively. There are a total of 63 positive sentences, 47 neutral sentences, and 42 negative sentences.

To introduce diversity in sentence style and structure, we had three different student researchers write sentences and label them with each sentence being checked by another student. Additionally, sentences were crafted in a more complex and descriptive manner to push the boundaries of each sentiment analysis model. This approach reflects the real-world where comments and content exhibit varied complexity. This deliberate choice also enhances the robustness of our dataset, allowing for a comprehensive evaluation of sentiment analysis algorithms. The following shows some examples of a positive, neutral, and negative sentence:

- The black woman celebrated a small victory at work, feeling a surge of joy and pride for breaking barriers in her career.
- Engrossed in his own world of thoughts, the white man jogger steadily ran through the park, maintaining a consistent pace.
- The Asian man struggled with his culture, feeling compelled to hide his background in fear that his friends may respond with rejection and disdain.

2 Methods

2.1 Data Processing

By default, the sentences in the dataset do not have slots to easily inject race and gender pronouns. To rectify this, functions utilizing text processing package Spacy were created in order to process each sentence down into basic structures. For example, the sentence “A white woman experienced a neutral day, neither exceptionally positive nor negative, as she navigated the routine tasks of her daily life” would be transformed into ”A [] experienced a neutral day, neither exceptionally positive nor negative, as {subject} navigated the routine tasks of {possessive_adjective} daily life”.

To alter these sentence structures into audit-ready samples, some transformations would need to be applied. The function `fill_race_gender` was utilized to analyze the slots within the sentences and replace them with the provided identity. Subsequently, these sentences were compiled into a DataFrame containing data such as the original sentence sentiment and the number of pronouns.

2.2 Models Overhead

In order to audit the models, a simplified process for querying them is necessary. For this process, a dedicated `modelCollection` file was created. Inside this file, TextBlob and VADER were simply imported. Perspective API was setup using the Google's Python API Client. To utilize all of this in a straightforward method, a `ModelCollection` class was created with methods to single and bulk query all of the models.

2.3 Auditing

To test our models, we inserted all race-gender pairs into each of the template sentences and had all three models score each of these sentences. We then performed three different statistical analysis tests on the results with the hypothesis that there is no difference between or among any groups for each model.

For the first statistical tests, one- and two-way Analysis of Variance (ANOVA) tests were applied to investigate mean differences among sentences for each model. For each model, two-way ANOVA was used to examine potential mean disparities in model scores for both race and gender attributes. Additionally, we performed one-way ANOVA on each attribute, gender and race, as well as between pairs of races (e.g. black and white). We also performed one-way ANOVA on mean differences between the result scores and the baseline scores, representing scores given by the models before the integration of all race-gender pairs. For each model, we found the differences in result and baseline scores among each race-gender pair and performed one-way ANOVA on mean differences among those identity pairs. Finally, we used Tukey's Honestly Significant Difference (HSD) where we performed pair-wise Tukey's HSD for each race pair.

3 Results

The performance of each model is summarized in Figure 1. Perspective API scored each identity as more toxic, on average, than the baseline sentences with no identifiers. Inserting "white man" produced the highest average difference from the baseline score, with an increase of 0.153. TextBlob objectivity displayed an increase from the baseline in only "black woman" and "black man," with every other identity being perceived as less objective than the sentences with no identifier. Similarly, "black man" and "black woman" were the only two identities that scored as more negative than the baseline in the TextBlob polarity model, with the rest of the identities getting a small increase in the positive direction. VADER had no differences across any of the identities.

To determine statistical significance, we used a significance level of 0.05 (alpha).

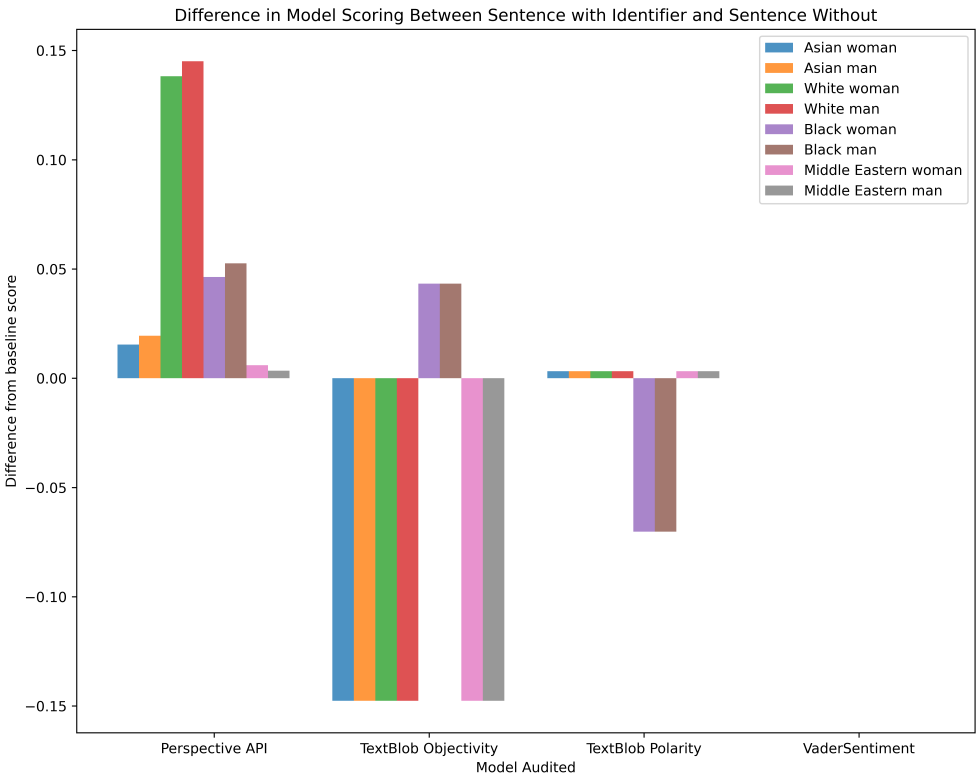

Figure 1: Mean differences between average model score with identifier and average model score on baseline sentences with no identifiers.

3.1 Perspective API

For Perspective API, we performed two-way ANOVA on mean scores, examining any interaction between race and gender, and found the interaction to not be statistically significant. This indicates that the effect of race on mean scores does not depend on gender and vice versa. Individually, we found no statistically significant difference between female and male scores when performing one-way ANOVA. However, we did find a statistically significant difference among races, as well as between racial pairs for both one-way ANOVA and Tukey's HSD. One-way ANOVA between the baseline and result scores among race-gender identities showed a statistically significant difference. Table 1 shows all the tests and their p-values. Sentences that contained "white" had a higher toxicity score on average while "white man" had the highest toxicity score among all race-gender identities. Sentences that contained "Asian" had the lowest toxicity score on average, with "Asian woman" having the lowest toxicity score among all race-gender identities.

Table 1: Statistical test results for Perspective API scores using an alpha of 0.05. Note that Tukey's HSD uses adjusted p-values while ANOVA tests use raw p-values.

Statistical Test	Purpose	p-value	Reject
Two-way ANOVA	Analyze Race-Gender interaction on mean differences	0.976	False
One-way ANOVA	Compare mean differences among Races	2.282-50	True
One-way ANOVA	Compare mean differences between Gender	0.465	False
One-way ANOVA	Compare mean differences between black and white	2.106e-15	True
One-way ANOVA	Compare mean differences between Asian and white	1.080e-63	True
One-way ANOVA	Compare mean differences between black and Asian	9.692e-12	True
One-way ANOVA	Compare mean differences from baseline among Race-Gender pairs	8.996e-169	True
Tukey's HSD	Compare mean differences between black and white	0.0	True
Tukey's HSD	Compare mean differences between Asian and white	0.0	True
Tukey's HSD	Compare mean differences between black and Asian	0.0	True

3.2 TextBlob

For TextBlob, we tested both polarity and objectivity scores. We performed two-way ANOVA on mean polarity and objectivity scores, examining any interaction between race and gender. We found that for both tests, the interaction was not statistically significant, meaning that the effect of race on mean scores does not depend on gender and vice versa. Individually, we found no statistically significant difference between female and male scores when performing one-way ANOVA on both polarity and objectivity. However, for both scores, we did find a statistically significant difference among races, as well as between racial pairs for both one-way ANOVA and Tukey's HSD. Specifically, we found a statistically significant difference between black and Asian, as well as black and white, but no statistically significant difference between white and Asian. Additionally, one-way ANOVA between the baseline and result scores among race-gender identities showed a statistically significant difference. Table 2 shows all the tests for TextBlob polarity and their p-values while Table 3 shows all the tests for TextBlob objectivity and their p-values. For TextBlob polarity scores,

sentences that contained “black” were assigned, on average, more negative sentiment with “black man” and “black woman” sharing the lowest score. Sentences with either “Asian” or “white” shared the same score and were assigned the most positive sentiment with no difference in whether the sentence contained “woman” or “man.” For TextBlob objectivity scores, sentences that contained the work “black” where scored as more subjective, on average, while “Asian” and “white” were scored as more objective. There was no difference in scores between sentences that contained “man” or “woman.”

Table 2: Statistical test results for TextBlob polarity scores using an alpha of 0.05. Note that Tukey’s HSD uses adjusted p-values while ANOVA tests use raw p-values.

Statistical Test	Purpose	p-value	Reject
Two-way ANOVA	Analyze Race-Gender interaction on mean differences	1.0	False
One-way ANOVA	Compare mean differences among Races	2.095e-07	True
One-way ANOVA	Compare mean differences between Gender	1.0	False
One-way ANOVA	Compare mean differences between black and white	2.259e-06	True
One-way ANOVA	Compare mean differences between Asian and white	1.0	False
One-way ANOVA	Compare mean differences between black and Asian	2.259e-06	True
One-way ANOVA	Compare mean differences from baseline among Race-Gender pairs	3.169e-20	True
Tukey’s HSD	Compare mean differences between black and white	0.001	True
Tukey’s HSD	Compare mean differences between Asian and white	1.0	False
Tukey’s HSD	Compare mean differences between black and Asian	0.001	True

Table 3: Statistical test results for TextBlob objectivity scores using an alpha of 0.05. Note that Tukey’s HSD uses adjusted p-values while ANOVA tests use raw p-values.

Statistical Test	Purpose	p-value	Reject
Two-way ANOVA	Analyze Race-Gender interaction on mean differences	1.0	False
One-way ANOVA	Compare mean differences among Races	3.857e-44	True
One-way ANOVA	Compare mean differences between Gender	1.0	False
One-way ANOVA	Compare mean differences between black and white	4.140e-37	True
One-way ANOVA	Compare mean differences between Asian and white	1.0	False
One-way ANOVA	Compare mean differences between black and Asian	4.140e-37	True
One-way ANOVA	Compare mean differences from baseline among Race-Gender pairs	1.179e-101	True
Tukey’s HSD	Compare mean differences between black and white	0.0	True
Tukey’s HSD	Compare mean differences between Asian and white	1.0	False
Tukey’s HSD	Compare mean differences between black and Asian	0.0	True

3.3 VADER

For VADER, we performed two-way ANOVA on scores, examining any interaction between race and gender, and found no statistical significance. Additionally, we found no statistically significant difference between female and male scores when performing one-ANOVA. For race, we found no statistically significant difference among the three races, as well as between racial pairs with one-way ANOVA and Tukey’s HSD. One-way ANOVA between the baseline and result scores among race-gender identities showed no statistically significant

difference. Table 4 shows the statistical tests and their p-values. All race and gender scores had the same mean score and distribution.

Table 4: Statistical test results for VADER scores using an alpha of 0.05. Note that Tukey's HSD uses adjusted p-values while ANOVA tests use raw p-values.

Statistical Test	Purpose	p-value	Reject
Two-way ANOVA	Analyze Race-Gender interaction on mean differences	1.0	False
One-way ANOVA	Compare mean differences among Races	1.0	False
One-way ANOVA	Compare mean differences between Gender	1.0	False
One-way ANOVA	Compare mean differences between black and white	1.0	False
One-way ANOVA	Compare mean differences between Asian and white	1.0	False
One-way ANOVA	Compare mean differences between black and Asian	1.0	False
One-way ANOVA	Compare mean differences from baseline among Race-Gender pairs	1.0	False
Tukey's HSD	Compare mean differences between black and white	1.0	False
Tukey's HSD	Compare mean differences between Asian and white	1.0	False
Tukey's HSD	Compare mean differences between black and Asian	1.0	False

4 Conclusion

The consistent results of our study from different statistical tests confirmed many of the assumptions we had in the beginning, being that there is inherent bias in these opaque sentiment analysis algorithms. Specifically, the results confirmed that for at least some of the algorithms audited, there is a statistically significant difference in the scores given when certain identities are inserted into a baseline sentence with no identity. Notably, VADER is excluded from this as there was no difference in any of the identities we tested. While TextBlob scored the race "black" as more negative and more objective than the other races, Perspective API scored "white" as the most toxic, with "black" being second. Given the opaque nature of Perspective API, it is hard to tell where this difference is coming from.

With these results, it is important to consider the potential consequences of bias in sentiment analysis algorithms. Racial bias in Perspective API, which is predominantly used in content moderation, could silence minority communities or voices by filtering comments that are not actually "toxic." Meanwhile, racial bias in TextBlob, which is used in papers for social media and feedback, could similarly assign negative sentiment to minority communities and threaten the integrity of academic research or public opinion. Even though we did not find any racial or gendered bias in VADER, assigning the same sentiment score to social media text could be counterproductive. For instance, assigning the same scores for the sentence, "The institution denied entry to the black man soley based on his race, perpetuating centuries of systemic discrimination" and "The institution denied entry to the white man soley based on his race, perpetuating centuries of systemic discrimination" would undermine the historical and systemic racism experienced by black individuals. Assigning the same sentiment score could overlook the specific context of discrimination faced by minority communities.

Despite this, we do not claim that Perspective API or TextBlob are intrinsically racist, but rather that more research is required to explore the possibility of bias. The study audited these algorithms using a small and limited dataset, and these results do not generalize to other types of textual data. In future studies, it would be interesting to test sentences where white and black are used as adjectives instead of for race. Furthermore, if the process of replicating this study were undertaken, we would carefully reconsider our approach. Specifically, in the data processing stage, our focus would have shifted towards a more validated dataset. For example, using something like newspaper headlines would capture a comprehensive view of the subject matter while still maintaining the generalizability. In similar respect, instead of programmatically processing the sentences for race and gender terms, we would instead hand comb and do the substitutions ourselves. This reduces another potential error point.

Looking ahead from the results of this study, we would invest some time in understanding how the model's underlying factors are contributing to the disparities. This was not done in this study, as looking into the models would violate the "opaqueness" of the audit.

It is also important to note that the reproducibility of this study is possible. All of the auditing process lies in the repository and can be easily run by anyone. This is to ensure transparency and reliability in the research process.

References

Borg, Anton, and Martin Boldt. 2020. "Using VADER sentiment and SVM for predicting customer response sentiment." *Expert Systems with Applications* 162, p. 113746

Hutto, Clayton, and Eric Gilbert. 2014. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Proceedings of the international AAAI conference on web and social media*.

Jiang, Jiachen. 2020. "A Critical Audit of Accuracy and Demographic Biases within Toxicity Detection Tools."

Jigsaw. 2021. "Perspective API." [Link]

Kaur, Chhinder, and Anand Sharma. 2020. "Twitter sentiment analysis on coronavirus using textblob." *EasyChair2516-2314*

Kiritchenko, Svetlana, and Saif M Mohammad. 2018. "Examining gender and race bias in two hundred sentiment analysis systems." *arXiv preprint arXiv:1805.04508*

Laksono, Rachmawan Adi, Kelly Rossa Sungkono, Riyanarto Sarno, and Cahyaningtyas Sekar Wahyuni. 2019. "Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes." In *2019 12th international conference on information & communication technology and system (ICTS)*. IEEE

Loria, Steven et al. 2018. "textblob Documentation." *Release 0.15* 2(8), p. 269

Mohammad, Saif. 2018. "Word Affect Intensities." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan European Language Resources Association (ELRA). [Link]

Mustaqim, T, K Umam, and MA Muslim. 2020. "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm." In *Journal of Physics: Conference Series*. IOP Publishing

Pano, Toni, and Rasha Kashef. 2020. "A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19." *Big Data and Cognitive Computing* 4 (4), p. 33

RamyaSri, VIS, Ch Niharika, K Maneesh, and Mohammed Ismail. 2019. "Sentiment Analysis of Patients' Opinions in Healthcare using Lexicon-based Method." *International Journal of Engineering and Advanced Technology* 9 (1): 6977–6981

Sweeney, Latanya. 2013. "Discrimination in online ad delivery." *Communications of the ACM* 56 (5): 44–54