Racial Bias in Insult Detection

Eric Mc Lachlan
ericmclachlan@gmail.com

Abstract

Insult detection classifiers are being used to detect insulting posts in social media. This paper investigated the degree to which the Perspective API demonstrates racial bias when performing insult detection on English texts. The results of this experiment show that there is little measurable racial prejudice in the identification of insults using the Perspective API when evaluating performance against the Wikipedia Talk Corpus for Personal Attacks; however, these results are not deemed to be significant. Instead, this research revealed that the Wikipedia Talk Corpus for Personal Attacks is likely incredibly under-representative of language used by communities who identify as African American.

1. Introduction

In this paper, I quantified the degree to which the Perspectives API demonstrates racial bias when performing text classification for insult detection on English texts. Previous analysis of the Perspective API has demonstrated racial bias in toxicity detection (Sap, et al., 2019). This work extends the work of Sap et al by analyzing racial prejudice in insult detection in the Perspective API in English texts.

This work finds no evidence to suggest that the Perspective API has racial bias in the identification of Insults in English; however, these results are to be interpreted with caution because, out of the 111,352 samples included in our analysis, only 31 samples were predicted to be from a population consisting of 80% or more African-Americans. Therefore, I caution that in order to substantiate any claim related to racial bias, further research is required. In particular, more generalizable results could be obtained using a benchmark corpus with greater representation of African Americans.

In this paper, I overview some of the literature related to the quantification of racial bias. Having established a framework, I discuss the Wikipedia Talk Corpus, a corpus of texts annotated for personal attacks and harassment. I substantiate why I believed this corpus to be a good benchmark against which to measure Insult prediction by the Perspective API. I also substantiate how the TwitterAAE classifier, developed by Blodget et al (2016) can be used as a proxy for race and, more specifically, to predict whether texts originate from majority white or majority African American communities. Ultimately, this will be used to quantify the degree to which a classifier demonstrates racial bias. Finally, I will discuss the results of the experiment and draw attention to racial disparities in the Wikipedia Talk Corpus' Personal Attack dataset.

2. Background

With Facebook alone reaching 2.91 billion active users in 2021, it's clear that social media adoption has become prolific (Statista Research Department, 2021). Furthermore, the large amount of content posted on social media each day has motivated social media companies to find ways of automating the moderation of posts to prevent abuse and keep their platforms a place their users want to be.

One strategy for automating moderation used by social media companies is to use machine learning to detect social media posts that have problematic characteristics. By using text classification for toxicity, for example, social media companies can focus their moderation

efforts on the most potentially harmful posts; reducing the reaction time for any interventions, and thereby helping to minimize harm (Jigsaw, 2021).

To this end, Google conducts research via Jigsaw, a unit of Google that "explores threats to open societies, and builds technology that inspires scalable solutions" (Jigsaw, 2020). In particular, Jigsaw and Google Counter Abuse Technology team have developed a product called the Perspective API, which "is a free API that uses machine learning to identify "toxic" comments" (Jigsaw, 2021).

In previous work, Sap et al (2019) proved that the Perspective API demonstrates racial bias when performing text classification for toxic posts. This was done by using dialect as a proxy for race and showing that Tweets characterized by markers of African American English (AAE) were significantly more likely to be reported by the model as toxic. To demonstrate this bias, Sap et al divided their test data into two groups, one characterized by markers of AAE and another characterized by markers of "white-aligned" English. Using this methodology, Sap et al were able to demonstrate that the Perspective API was significantly more likely to predict a text as Toxic if it were characterized by markers of AAE than the white-aligned group. (Sap, et al., 2019)

In this paper, I will build on the foundation laid by Sap et al (2019) to measure whether the bias demonstrated in the Perspective API extends to Insults; another of the categories predicted by the Perspectives API.

In order to do this, I will need:

- The Perspective API to generate predictions for the Insult category.
- A gold-standard corpus of texts annotated with labels for Insults against which to evaluate the Perspective API's predictions.
- A proxy for race in order to retrospectively apply a grouping of the test samples along demographic lines in order to quantify racial bias.

Each of these components will be discussed in more detail in the sections below.

2.1 Perspective API

The Perspective API performs multilingual text classification using machine learning to predict whether texts fall into one of the predefined categories: Identity Attack, Insult, Profanity, Severe Toxicity, Sexually Explicit, and Threat. According to the documentation, Insults are defined [1]as "Insulting, inflammatory, or negative comment towards a person or a group of people" (Jigsaw, 2021). The Perspective API provides a programmatic interface for classifying texts. For each text sent to the Perspective API, a prediction is returned for each of the threat categories mentioned above. I will measure the accuracy of these predictions using a human annotated benchmark as the gold standard for evaluation. More specifically, the sample will be grouped by majority-white and majority-African American samples, and an analysis will be performed to measure whether these two groups are treated equally or whether there is racial bias.

2.2 Gold Standard

I used a subset of the Wikipedia Talk Corpus' Personal Attack dataset as the gold standard for evaluating the performance of the Perspective API. This dataset consists of 100,000 labeled comments from English Wikipedia. On average, each comment was labelled by 10 crowdworkers (Wulczyn, et al., 2017). Crowdworkers were prompted with the heading:

"Does the comment contain a personal attack or harassment?"

Below this heading, crowdworkers could select multiple of the following options:

- Targeted at the recipient of the message (i.e. you suck).
- Targeted at a third party (i.e. Bob sucks)
- Being reported or quoted (i.e. Bob said Henri sucks.)
- Another kind of attack or harassment.
- This is not an attack or harassment.

[1]https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

The term "personal attack and harassment" here is vaguely defined; however, the interpretation would undoubtedly have been guided the provided example "Bob sucks", which is clearly an insult. When framed in this way, I interpret "personal attack and harassment" as inflammatory or derogatory statements intended to induce discomfort or inflict psychological harm on the recipient, or to harm the reputation or good standing of a third party.

2.3 Comparability

The Perspective API can be argued to be compatible with a subset of The Wikipedia Talk Corpus' Personal Attacks dataset. As negative test samples, I included all samples of the Wikipedia Talk Corpus' dataset annotated with the label "This is not an attack or harassment". As positive test samples, I included samples from the Wikipedia Talk Corpus reported as personal attacks or harassment that was "targeted *at the recipient*" or "targeted *at a third party*". This subset of with Wikipedia Talk Corpus is well matched to the Perspective API's definition of Insult, which we reiterate is "Insulting, inflammatory, or negative comment *towards a person* or *a group of people*" (Jigsaw, 2021). In this way, I've tried to exclude samples of the Wikipedia Talk Corpus that might fall outside the class boundaries of the Perspective API's classifier.

2.4 Proxy for Race

Following the precedent set by Sap et al (2019), I used the classifier developed by Blodgett et al (2016) to categorize English texts into two groups that act as proxies for race.

While Sap et al (2019) frames their paper in terms of dialect markers, I frame this paper in terms of majority-white communities and majority-African American communities because the underlying classifier developed by Blodget et al (2016) actually predicts the constituencies of communities; rather than dialect markers. While there is a clear relationship between the demographic constituents of communities and language markers, I prefer to frame the conversation using the original paradigm.

The classifier developed by Blodget et al (2016) uses machine learning to predict the demographic constituency of a community associated with a specified text. The community is assumed to consist of "whites, non-Hispanic blacks, Hispanics (of any race), and Asian[s]". The classifier was trained using 60 million geolocated Twitter posts in conjunction with demographic information from the Census' 2013 American Community Survey. This classifier is thus able to provide "a rough proxy for likely demographics of the author and the neighborhood they live in". (Blodgett, et al., 2016)

Using this classifier, I was able to select from the test samples those samples predicted to be representative of majority-African American and majority-white communities. These two groups of texts will be used to quantify the degree to which texts written by white communities and African American communities are treated differently by the Perspective API.

3 Methodology

There are two major components to the methodology applied in this paper. The first is software written to gather all the data necessary to perform our analysis. The second component relates to the analysis of the gathered data.

The software and analysis of the data can all be found on GitHub[2].

Reproducing the results of the experiment involves i) downloading the source code from GitHub, ii) adding your Google credentials to your environment variables, iii) running the init.sh script, iv) running the run_experiment.sh script, v) reperforming the analysis of the results.

The software consisted of a number of bash scripts and Python code written to download, extract, and enrich the Wikipedia Talk Corpus Personal Attack corpus. This involves, downloading the corpus, downloading the TwitterAAE library from GitHub, applying a small patch to update the TwitterAAE project to Python 3, running all samples through the TwitterAAE classifier and augmenting the data with the predictions made by the TwitterAAE classifier. Then, the samples are sent to the Perspective API. The software then adds the predictions from the Perspective API to our dataset. The final output of the software is a tab-separated values (TSV) file called corpus.with_perspective.tsv, which contains the

[2] https://github.com/ericmclachlan/lets-do-a-science

original texts and the annotations from TwitterAAE and the Perspective API.

The second component of the methodology is a more manual analysis of the gathered data. The corpus.with_perspective.tsv was imported into Microsoft Excel. From there, certain data was excluded (as detailed in the section related to Data Exclusions, below). Finally, Excel was used to determine the distribution of remaining data in the dataset and perform some additional analyses.

3.1 Data Inclusions

The Wikipedia Talk Corpus' Personal Attack dataset is used as our benchmark against which to test the performance of the Perspective API. The original Wikipedia Talk Corpus' Personal Attack dataset consists of 115,864 samples. In a series of steps, each of these samples was augmented with additional data.

Firstly, each of the samples were annotated by approximately 10 annotators. Majority annotation was used to assign to each sample a true or false value indicating whether or not the sample represents i) a personal attack or harassment directed at a person, a personal attack or harassment directed at a third party, or neither a personal attack nor harassment.

Each of these samples were passed through Blodget et als classifier. The dataset was augmented with predictions of the demographic constituency of the speakers community; adding columns for "African American", "Hispanic", "Asian", and "White", where each category corresponds with the racial designations on the Census' 2013 American Community Survey.

Furthermore, each of the samples were sent to the Perspective API in order for it to make a prediction for Insult. The dataset was augmented with this prediction.

This raw data and the corresponding analysis have been included in the GitHub repository and can be downloaded from here: https://github.com/ericmclachlan/lets-do-a-science/blob/550f922a291bf22371a7bf1fedd71ea646b8663d/corpus.with_perspective.xlsx

3.2 Data Exclusions

In order to simplify the analysis, samples were excluded for a variety of reasons.

To maximize that compatibility of the test samples with the Perspective API, samples of the Wikipedia Talk Corpus' Personal Attack dataset were included in our analysis if they were labeled as 1) samples targeted at the recipient of the message, 2) samples targeted at a third party, or 3) samples marked as neither a personal attack nor harassment. In this way, 4,218 samples were excluded from the analysis because they represented attacks or harassment that were neither targeted at the recipient nor a third party; making it difficult to reconcile with the Perspective API's class boundary. From the remaining 111,646 samples, we removed a further 4 samples for which the Perspective API was unable to generate a prediction. Removing these samples resulted in a total of 111,642 samples. Finally, we removed any sample for which we were unable to get a prediction from Blodget et al's classifier. Reasons for failure to produce a prediction seem to include too few tokens in the text or spelling and grammatical errors that reduce the ability of the model to make meaningful predictions. These technical limitations resulted in the exclusion of a further 290 samples. This resulted in a total of 111,352 samples available for analysis.

3.3 Distribution of Samples

The distribution of positive and negative samples included in our analysis is illustrated by Table 1, below.

	Count	Percent
positive samples	9,341	91.61%
negative samples	102,011	8.39%
Total	111,352	100.00%

Table 1: Distribution of positive and negative samples from the Wikipedia Talk Corpus' Personal Attack dataset included in the analysis.

In order to measure the degree to which texts written by majority-white communities and majority-African American communities were treated differently, it is necessary to identify samples representative of these communities. Therefore, samples were grouped into two groups: majority-white communities, and majority-African American communities. Samples were only included in these groups if they were predicted by Blodgett et al's classifier to have a constituency of 80% or more of a particular racial group, which is consistent with the methodology

applied by Blodgett et al (2016) and Sap et al (2019).

Table 2 show the composition of our test samples after grouping the samples into majority white, majority African American (AA), and Other communities.

Samples	Majority White	Majority AA	Other	Total
Positive	192	16	9,133	9,341
Negative	14,761	15	87,235	102,011
Total Count	14,953	31	96,368	111,352
Total Percent	13.43%	0.03%	86.54%	100%

Table 2: The number of positives and negative samples for each of the groups.

4 Results

For each of the 111,352 samples included in the analysis, I compared the prediction made by the Perspective API to the majority human-annotated label provided by the Wikipedia Talk Corpus. Table 3 illustrates the performance of the Perspective API as measured against the gold standard annotations in terms of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions for each of the groups.

Predictions	Majority White	Majority AA	Other	Total
TP	167	15	8,612	8,794
TN	13,704	11	79,258	92,973
FP	1,057	4	7,977	9,038
FN	25	1	521	547
Total Count	**14,953**	**31**	**96,368**	**111,352**

Table 3: The predictions of the Perspective API categorized by true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

From these figures, we are able to calculate the following analysis of the performance of the Perspective API against our gold standard labels for each of the groups. These figures are illustrated in Table 4:

	Majority White	Majority AA	Other	Total
TPR/Recall	86.98%	93.75%	94.30%	94.14%
TNR	92.84%	73.33%	90.86%	91.14%
Bal. Acc.	89.91%	83.54%	92.58%	92.64%
Precision	13.64%	78.95%	51.91%	49.32%
F1 Score	23.59%	85.71%	66.96%	64.73%

Table 4: The true positive rate (TPR), true negative rate (TNR), balanced accuracy (bal. acc.), precision, and f1 score for each of the groups.

Finally, we calculated the Pearson Product Moment Correlation r between the predictions generated by the Perspective API and the predicted African American constituency made by Blodgett et al's classifier for all 111,352 samples and determined that $r = 0.10986$, which is a very weak positive correlation. This implies that the Perspective API is slightly more accurate at predicting insults originating from members of majority African American communities.

5 Discussion

Unfortunately, it seems that the Wikipedia Talk Corpus was not a good selection for evaluating racial bias in the Perspective API − not because of the incompatibility if the class boundaries for what constitutes an insult to the Perspective API and what constitutes a personal attack or harassment in the Wikipedia Talk Corpus − but because of underrepresentation of African American communities in the Wikipedia Talk Corpus. Blodget et al (2016) and Sap et al (2019) used a threshold of 80% to group communities into majority white communities (where the white constituency is predicted to exceed 80%) or majority African American communities (where the African American constituency is predicted to exceed 80%). Applying this conventional methodology, only 31 of the 111,352 samples are predicted to be ascribed to communities with 80% or more African Americans. That is a measly 0.03% of the corpus. In contrast, 14,953 samples (13.43% of the corpus) are ascribed to majority white communities.

Even when reducing this threshold to 50%, the corpus only reports that 528 samples (0.47%) are likely to be from speakers from communities where 50% or more identify as African Americans. In contrast, 81,059 samples (72.80%) are ascribed to communities where more than 50% identify as white.

Unfortunately, this extreme polarity in the Wikipedia Talk Corpus only became visible after running the samples through Blodget et al's classifier.

This raises a few questions: Are Wikipedia's forums really so incredibly under representative of African American communities? Or were there criteria in the data selection that inadvertently excluded texts from majority African American communities from being included in the Wikipedia Talk Corpus? Or are their sociocultural factors (like code switching) that affect the way African American communities interact on Wikipedia's forums?

Another interesting observation from our results is the massive disparity between balanced accuracy and the f1 scores; particularly for texts attributed to majority white communities. In this demographic, balanced accuracy, which is the average of the true positive rate (TPR) and the true negative rate (TNR), is 89.91%, whereas the F1 score for this same group is only 23.59%. This is because balanced accuracy does not take into consideration false predictions and the model predicted 6 times more false positives (1,057) than true positives (167) for this demographic.

Another interesting observation is that, for texts attributed to majority white communities, there are 14,761 negative samples and 192 positive samples; a ratio roughly 77:1, whereas for African American communities, the data is almost 1:1 with 15 negative samples outweighed by 16 positive samples.

These observations lead me to believe that the Wikipedia Talk Corpus Personal Attack dataset seems to be extremely skew along racial lines; however, this is very difficult to see without performing some kind of race-related analysis.

6 Ethical Considerations

This paper involves a number of ethical considerations, each of which will be presented briefly below.

In terms of the environmental cost of conducting these experiments, I estimate that it cost about 2500 Watts of electricity to produce this paper. This is largely due to a limitation imposed by the Perspective API, which only accepts one request per second. Therefore, in order to process all 115,864 samples, it was necessary to run the application for approximately 33 hours; consuming about 1452 Watts of electricity. The remainder of the cost is attributed to hours spent using a computer to write this paper.

As it relates to crowdworkers, Wulczyn et al (2017) did not mention anything related to how well crowdworkers were paid to annotate the Wikipedia Talk Corpus. Furthermore, the release of worker IDs has been included in the demographic information of crowdworkers who annotated their data and it has been demonstrated that these kinds of identifiers can sometimes be used as a way to re-identify anonymized crowdworkers (Shmueli, et al., 2021).

Racial profiling of any kind should always give the community pause as we carefully consider the motivation for creating such a tool. In this case, I had hoped to identify the degree to which the predictions of a production-level classifier is affected by race in order to motivate further research on the topic and incentivize companies to minimize these racial incongruencies. Because the underlying data used to train Blodgett et al's classifier is based on user-disclosed census data, the racial information is self-disclosed in the sense that the determination is made by the citizen themselves and no attempt has been made to ascribe a race to a person. In this sense, I hope that we've respected that race is a social construct and a matter of personal identity and specifically not a determination that can be ascribed at a individual level. Having said that, ascribing race at an aggregate level can be helpful in finding racial disparities, such as those discovered in the Wikipedia Talk Corpus through this work.

On a more personal level, as a person who grew up in apartheid South Africa, the idea of using machine learning to perform racial profiling is incredibly uncomfortable for me. In particular, I worry that being involved in this kind of work reinforces prejudices about white South Africans who have a history of racism and who have a history of using racial profiling to develop "a science" to justify apartheid. Furthermore, while my intention is to apply objective analysis to identify racial bias in the Perspective API, I acknowledge that my lack of lived experience as a person of color – particularly in an American context - may limit my ability to treat this subject with the sensitivity it deserves. If I fall short of doing so, I sincerely apologize.

Finally, the finding of this paper is that there is no significant racial bias in the Perspective API when evaluating it using the Wikipedia Talk Corpus's Personal Attack dataset. I reiterate, these findings are not significant, and this does not

imply that that the Perspective API is unbiased; it merely means that we were unable to prove bias using this corpus.

7 Limitations

While the Perspective API allows for multilingual classification, this study only investigated English texts.

The substance of this work rests heavily on the virtues of the classifier built by Blodgett et al (2016). This classifier was trained on Twitter data and here we have applied this classifier to non-Twitter text; the Wikipedia Talk Corpus. I have assumed that there is sufficient similarity between these registers to make reasonable comparisons; but proving was deemed to fall outside of the scope of this work.

8 Conclusion

This work has investigated the degree to which the Perspective API differentially predicts Insults for majority-African American and majority-white communities. While I found no significant racial bias in the Perspective API as it relates to the Insult detection, I did discover significant asymmetries in the representation of texts ascribed to majority white and majority African American communities. This illustrates that, while performing studies on race puts the research subject and even the researcher in a vulnerable position, performing such research is incredibly important for revealing asymmetries in data before it becomes the seed for machine learning models that internalize these asymmetries.

Acknowledgments

Special thanks to Emily Bender for challenging me to set a higher scientific standard for this project. Furthermore, I'd like to express my appreciation to Michael K Scott for providing feedback on a very rough and incomplete draft of this paper. Your contribution undoubtably improved the quality of this paper. Finally, I would like to thank the class of Ling 575 (Autumn 2021) at the University of Washington for shaping my understanding of societal impacts of NLP.

References

Blodgett, S. L., Green, L. & O'Connor, B., 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. *Association for Computational Linguistics,* p. 1119–1130.

Jigsaw, 2020. *Jigsaw.* [Online] Available at: https://jigsaw.google.com/ [Accessed 21 11 2021].

Jigsaw, 2021. *Attributes & Languages.* [Online] Available at: https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages [Accessed 21 11 2021].

Jigsaw, 2021. *How it Works.* [Online] Available at: https://www.perspectiveapi.com/how-it-works/ [Accessed 21 11 2021].

Jigsaw, 2021. *Perspective API.* [Online] Available at: https://www.perspectiveapi.com/ [Accessed 21 11 2021].

Sap, M. et al., 2019. The Risk of Racial Bias in Hate Speech Detection. *Association for Computational Linguistics,* p. 1668–1678.

Shmueli, B., Fell, J., Ray, S. & Ku, L.-W., 2021. Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. *NAACL.*

Statista Research Department, 2021. *Number of monthly active Facebook users worldwide as of 3rd quarter 2021.* [Online] Available at: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ [Accessed 21 11 2021].

Wulczyn, E., Thain, N. & Dixon, L., 2017. *Wikipedia Talk Corpus.* [Online] Available at: https://doi.org/10.6084/m9.figshare.4054689.v6 [Accessed 21 11 2021].