5-2020

A Critical Audit of Accuracy and Demographic Biases within Toxicity Detection Tools

Jiachen Jiang

A Critical Audit of Accuracy and Demographic Biases within Toxicity Detection Tools

by

Jiachen Jiang

Advisor: Professor Soroush Vosoughi

A thesis submitted in partial satisfaction of the

requirements for the degree of

Bachelor in Science

in

Computer Science

Dartmouth Computer Science Technical Report TR2020-890

Abstract

The rise of toxicity and hate speech on social media has become a cause for concern due to their effects on politics and the growth of extremist internet communities. The tools currently used to identify and eliminate harmful content have received widespread criticism from both the public and the academic community for their inaccuracies and biases. In our research, we set out to audit the performance of Perspective API, a toxicity detector created by research teams at Google and Jigsaw, on the language of users across a variety of demographic categories. We draw from Crenshaw's framework of intersectionality to discuss the unique harms that result from existing at the intersections of marginalization and examine existing computational models of disparate impact and proxy discrimination. In addition, we conduct A/B testing on Amazon's Mechanical Turk, a popular crowd-sourcing platform for data annotation within research communities, to identify and discuss biases that arise from human demographic prediction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Information Age has generated unprecedented levels of human innovation and discovery, but it has also led to the rise of a startling variety of digital injustices. Some of them are simply new forms of long-existing human inequities, like algorithms that claim to identify the best candidates for university admissions or a particular job listing [25]. The vast majority, however, are distinct phenomena that require particular approaches and state-of-the-art solutions. After all, flawed technologies are harmful not simply because they magnify existing human biases, but because they do so in ways that are extremely difficult - if not impossible - for humans to predict and regulate. People have motives and intents, but ultimately, computational systems are merely the unbiased vehicle of a biased world.

1.1 Related Work

One well-known case study is the polarizing example of COMPAS, a recidivism prediction instrument (RPI) used in courts around the United States to determine the length of sentence for convicted criminals. A widely popularized investigation conducted by a team at ProPublica [1] concluded that the instrument, which predicts a person's likelihood of offending again using demographic information, disproportionately predicting higher rates of re-offending for Black defendants. Specifically, the authors found that a nonrecidivating black defendant is nearly twice as likely to be assessed as high risk as a white defendant and similarly, a recidivating black defendant is nearly half as likely to be assessed as low risk as a white defendant. In technical terms, their findings indicated that the COMPAS instrument violated the fairness criterion of *error rate balance*. Specifically, COMPAS generates significantly higher false positive rates (FPR), the improper indication of a condition, and lower false negative rates (FNR), the improper indication of a lack of presence, for black defendants than for white defendants.

The investigation faced much criticism from the Northpointe corporation and within the academic community, primarily for the criteria the researchers had used to determine fairness. In its statement, Northpoint [24] argues that the fairness of the COMPAS instrument had been incorrectly evaluated and should have instead been validated by *predictive parity*, a fairness criterion that COMPAS does satisfy. Flores et al. [15] argues for yet another metric of *calibration*, which COMPAS also satisfies.

The issue is that there is an incredibly large number of computational definitions for fairness, the vast majority of them mutually exclusive [8]. In this case, predictive parity is incompatible with error rate balance when prevalence differs across groups. The COMPAS RPI is approximately well calibrated and satisfies predictive parity, provided that the high-risk cutoff s_H is 4 or greater. The instrument fails on both false positive and false negative error rate balance across the range of high-risk cutoffs, however. The error rate imbalance of COMPAS is not just a logical outcome of its current design; it cannot be mitigated while in its present context.

When the recidivism prevalence — or the base rate $P(Y = 1 \mid R = r)$ — is different for each group, any instrument that satisfies predictive parity at a given threshold s_H will have unequal false positive or false negative error rates at that same threshold. Given a particular choice of s_H, we can describe an instrument's performance in terms of a confusion matrix, as shown in Table 1.1.

	Low risk	High risk
Y = 0	TN	FP
Y = 1	FN	TP

Table 1.1: T/F denote true/false and N/P denote negative/positive. For instance, FP is the number of false positives: individuals who are classified as high risk but who do not reoffend.

The fairness metrics described in the previous section impose constraints on the values in this table. Another constraint is imposed by the recidivism prevalence within groups. Chouldechova shows that the prevalence (p), PPV, and false positive and negative error rates (FPR, FNR) are related through this equation:

$$FPR = (1 - FNR)\frac{p}{1-p} \times \frac{1 - PPV}{PPV} \ .$$

From this expression, we see that an instrument can satisfy predictive parity, but if the prevalence differs between groups, the instrument cannot achieve equal FPRs and FNRs across those groups. As the recidivism rate among black defendants in the data is 51%, compared with 39% for white defendants, we know some level of imbalance in the error rates must exist in order for the instrument to achieve predictive parity. As not all of the fairness criteria can be satisfied at the same time, it becomes all the more important to understand the individual consequences of failing to satisfy particular criteria.

Broad incompatibility in the literature

The case of the COMPAS RPI is a microcosm of the broader debate occuring in the field of algorithmic fairness. A growing body of work seeks to comprehensively evaluate the performance of different formal definitions of fairness and similar measures.

Friedler et al. [17] identify the axiomatic differences at the heart of the debate and develop a mathematical theory of fairness driven by transformations between different categories of space. The work highlights more spaces implicitly involved in the decision-making process than typically specified, and it argues that much of the confusion and disagreement within the literature occurs when scholars conflate them. To create a framework with which to study fairness criteria, they reinterpret notions of fairness, structural bias, and non-discrimination as quantifying the way that spaces are transformed to each other. Finally, they show that the majority of methods which propose to address algorithmic fairness hinge on implicit assumptions about the nature of these spaces and how they interact with each other.

A more recent paper [18] argues that although different algorithms tend to prefer specific formulations of fairness preservations, many of these measures strongly correlate with one another. Particularly, they find that fairness-preserving algorithms are sensitive to fluctuations in dataset composition, which suggests that fairness interventions may be less adaptable than hoped. In response, the paper presents a test-bed for determining which algorithm has the best performance under a fairness or accuracy measure, as well as what types of algorithmic interventions tend to be the most effective in the long run.

Hutchinson et al. [21] explore 50 years of research centred on machine learning fairness to determine historical lessons and areas of potential further work. The paper goes even further than similar work by calling for more actionable work on the causes of unfairness as opposed to directionless conjecture on the nature of fairness. They also emphasize the clear articulation of assumptions and choices through incorporating quantitative factors for the balance between fairness goals and other considerations, such as a value or ethics system.

In the absence of concrete right and wrong, it appears that the only way forward is to determine and choose the best approach for the particular circumstance. That is, one must consider and address the costs and gains of the choices made in the design of the underlying system when using, scaling, or drawing conclusions. In the fight against disparate impact, it does not matter that an algorithm can behave perfectly as much that an algorithm can behave clearly, able to explain why a particular approach was determined to be the most fitting for the situation.

1.2 Hate Speech and Toxicity Detection

These questions of intent and impact are especially important in the field of hate speech and toxicity detection, which existed only until recently as an afterthought for moderators of digital communities who felt a need to limit profanity and obscenity in their corner of the Internet. With allegations of election interference and the rapid growth of dangerous extremist groups in Internet communities, however, industry and academia have begun to take a much harder look at how toxic language and otherwise harmful content can be detected and eliminated in large swathes. Research has revealed that traditional toxicity detectors, which check new content against existing databases of profane or harmful text, ignore nuances of context and produced biased results [27].

The issue is that a majority of terms associated with hate speech are identifiers for marginalized identities or words that have since been reclaimed, such as ”queer” [33]. When algorithms remove content that contains these words, they disproportionately remove content by and about members of those same communities, who use the terminology to self-identify or share their personal experiences. One such incident was the subject of a 2019 lawsuit against YouTube for restricting all LGBT+ content to mature audiences [3]. Similar complaints have been levied against Facebook, where users have been given suspensions and had their posts removed for discussing discrimination that they have faced, even resulting in the deletion of entire communities [20]. Toxicity detectors also run into general obstacles of natural language processing tools, such as human biases within word embedding [4] and imperfect training data that do not paint an accurate picture of the world. For example, research has shown that tweets in African-American English are twice as likely to be labeled as offensive than others by models trained on popular hate speech data sets, which perpetuates African-American stereotypes about aggression and vulgarity and excludes voices of a marginalized community from research [27].

1.3 Perspective API

One notable toxicity detection tool is the Perspective API, a free tool developed by Jigsaw and Google's Counter Abuse Technology team in a collaborative research project. Its primary and oldest offering is a machine learning model trained on hundreds of thousands of pieces of human-annotated text [10] to predict the perceived harm of a comment on its readers. At its conception, Perspective operated very similarly to other toxicity detection tools, "[providing] a score from zero to 100 on how similar the new comments are to the ones identified as toxic." Currently, they offer a variety of additional tools ranging from real-time feedback for commenters to grouping comments by topics for readers [26]. While Perspective only offers production models for 'Toxicity' and 'Severe Toxicity' at the moment, experimental models like 'Flirtation' and 'Identity Attacks' are tentatively available and suggest that future versions may contend with problems of sexual assault and hate speech - high-profile issues not only for their specificity but for their strong precedence in the legal field.

Since its debut in February of 2017, Perspective API has been adopted by news media worldwide like The New York Times [31] and Spain's El Pais [14]. It has received its share of criticism as well, with users pointing out that shorter text inputs were more likely to be rated as toxic by the model and identified a worrying correlation between identity words of marginalized communities and higher scores, regardless of the actual position and content of the message[19].

Some patterns can be seen easily in Table 1.2, particularly the increase of toxicity with additional identity words. Other relationships between text and toxicity scores are more complicated. Note that "I am a woman" receives twice the toxicity score of "I am a man," but "I am a gay woman" is only 9 points higher than "I am a gay man." There is no linear relationship between the identity words and their toxicity scores, or even how toxicity score increases as different identity words are included. It is interesting to note, however, that the inclusion of racial identifiers is more potent than those describing sexual identities ("I am a gay woman" with a toxicity score of 0.66 versus "I am a white woman" at 0.77 and "I am a black woman" at 0.85), even in comparison to a word that has historically been used as a slur ("I am a dyke," which - interestingly enough - is lower than all of the aforementioned with a toxicity score of 0.60.)

Critics pointed out that if and when similar tools are used broadly, entire populations can be silenced and censored, gone without a trace to anyone on the outside. Some warned that these alarming results could be only the peak of the iceberg, suggesting that that even more harm were being caused by black-box automated comment-policing systems already at work in the innards of Facebook and Twitter [19].

Text Input	**Toxicity Score**
I am a man	0.20
I am a woman	0.42
I am a lesbian	0.51
I am a gay man	0.57
I am a dyke	0.60
I am a white man	0.66
I am a gay woman	0.66
I am a white woman	0.77
I am a gay white woman	0.78
I am a black man	0.80
I am a gay white woman	0.80
I am a gay black man	0.82
I am a black woman	0.85
I am a gay black woman	0.87

Table 1.2: Toxicity scores calculated by Perspective API, August 2017 [19]

Three years later, the Perspective API has undoubtedly improved in leaps and bounds in response to the critical response it garnered. Much of the code for the tool is now entirely open-source, with many of the experiments, research data, and models associated with Perspective made publicly available. The Jigsaw team even offers a practicum in debugging issues of fairness and bias within their models, which walks users through a real case of pinpointing and eliminating disparate impact within toxicity detection. The false-positive problem for comments containing identity terms is explained as a consequence of the training data - the majority of comments containing identity terms for race, religion, and gender were labeled toxic, but while these labels were mostly correct in context, the skew nonetheless taught the model a correlation between presence of these identity terms and toxicity. The main issue was not human biases in the training set, rather that the data did not contain sufficient examples of nontoxic identity comments for the model to learn that the terms themselves were neutral and that the context in which they were used was what mattered. According to the practicum, the Jigsaw team balanced the data and eliminated bias in the algorithm by the simple but insightful act of up-weighting negative subgroup examples.

When we queried the current API model, however, we found results that bore great similarity to those in Table 1.2. Some improvements were clear, particularly for the specific examples above, but the API continued to produce a series of false positives and false negatives for common use cases. These results can be found in Table 4.2 and will be discussed in detail in Chapter 4, which is dedicated to our audit of Perspective API.

The implications of these results were concerning. If a research project that prioritizes transparency and fairness can perform so poorly against the human biases deeply ingrained within language, how fares any of the many toxicity detection systems being used by the world's most popular forms of social media? Moreover, these gains in accuracy are complicated by the aforementioned non-linear increases in toxicity with identifier words, which suggests that decreases in toxicity may follow a similarly non-linear path. When 40 percent of bias is eliminated for a particular group, then, it becomes important to understand which 40 percent. Not everyone in a marginalized group experience marginalization equally, and eliminating a particular amount of harm may mean significant improvements for an individual and anything from no effect at all to the exact opposite for another.

This study investigates disparate impact and bias in the context of toxicity detection under the theoretical framework of intersectionality. We first outline our investigation into modeling webs of relations between discriminatory proxies by generating synthetic data with Bayesian networks. Additionally, we will describe our process of validating our experiment design through A/B testing trials conducted on Mechanical Turk. We then discuss our audit of Perspective API, for which we observed its performance on tweets made by users from a variety of demographic categories. Finally, we explore the significance of our results and describe next steps in the project. Our research seeks not to offer concrete computational definitions or a single correct form of classification. Ultimately, we conduct this study to question long-held assumptions surrounding bias and harm (and the elimination thereof) that exist not just within the field of natural language processing but the broader computer science community.

Chapter 2

Intersectionality and Disparate Impact

Countless papers and workshops within the computer science field discuss the gender [6] and racial [13] biases within algorithmic systems and software design, the majority of which tackle the two separately. Very little research has been conducted on the overlaps, however, including disparate impact that does not lie within these often arbitrary boundaries.

The framework of intersectionality, as coined by Kimberlé Crenshaw [11], arose from critical race theory and describes the ways in which individuals who exist at the intersection of multiple marginalized identities experience unique harm. For example, a black woman will not only experience the prejudices directed at women and at black people, but the harm that emerges particularly from being a black woman. We believe that the framework of intersectionality must be leveraged to design algorithmic systems that are truly just, which is different from simply achieving an arbitrary level of fairness or eliminating a certain amount of bias. Though the framework of intersectionality emerged from and is used largely outside of computer science, we ground our research in it to emphasize the importance of basing algorithmic and experiment design choices not just in theory, but theory that originates from seemingly unrelated fields of study.

Binaries and data are the foundations of computer science, however. Given the fundamental limitations of computational research, it is no surprise that researchers have struggled to model, let alone solve the complex and unpredictable issues of society. Below, we detail unique concerns that originate from this crucial mismatch of tools and problems.

2.1 Proxy Discrimination

Though many modern algorithms treat race, gender, and class as protected classes or eliminate them from the training sets entirely, disparate impact emerges nonetheless. Thus is the problem of proxy discrimination, in which data-driven systems inherit biases from subtle correlations within the data between the protected attribute and seemingly innocuous features. In other words, algorithmic systems are able to discriminate based on data they lack.

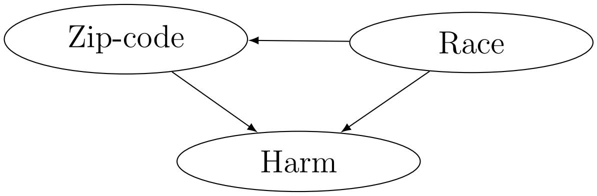


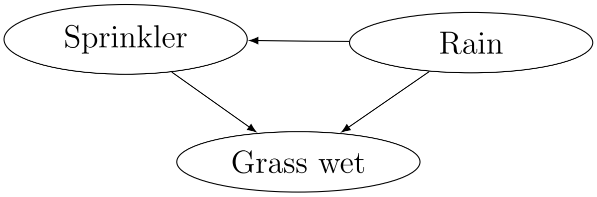
Figure 2.1: Redlining modeled by proxy discrimination

A simple example of proxy discrimination can be seen in Figure 2.1. An algorithm may disregard race and class data of an individual but include information on their zip-code. Zip-codes, however, correlate so strongly with both race and class that bias against zip-codes causes the same disparate impact as bias against people based on their race or class. In other words, zip-codes have become a discriminatory proxy for race and class.

Despite being similar to the real world phenomenon of redlining, in which companies systematically refuse services to certain zip-codes with majority black populations [9], proxy discrimination causes additional harm as a result of its unpredictability and ambiguity in the face of legislation [34]. Technical and legal researchers alike have argued that removing obvious biases within data sets can in fact exacerbate discrimination by creating new types of biases that are much more difficult to detect [2]. While zip-codes are clearly related to race and class, for example, machines can find something entirely unexpected (like amount of carrots consumed per year) that nonetheless correlates with a protected class to make biased decisions. Historically, the issue was tackled by simply removing the discriminatory proxy in question. Every feature within a data set is correlated with other features, however, and continuously eliminating data in this way means that at some point, there will be no data left at all.

We made several forays into computationally categorizing discriminatory proxies, first by investigating the relationship between discriminatory proxies of different degrees, then by magnitude of harm. Our efforts were stymied by the nebulous nature of most protected attributes, which made it difficult to distinguish between a discriminatory proxy and an aspect of the attribute itself. For example, skin color and nationality both correlate but do not cause - in the traditional sense - race, and by that definition, are discriminatory proxies for race. One would be hard pressed, however, to argue that prejudice against people of a certain skin color is significantly different in terms of harm from prejudice against people of a certain race. The problem of defining and drawing distinction between discriminatory proxies is this: either everything is a discriminatory proxy, or nothing is.

	Sprinkler	
Rain	T	F
F	0.4	0.6
T	0.01	0.99



Sprinkler	
T	F
0.2	0.8

		Grass wet	
Sprinkler rain		T	F
F	F	0.4	0.6
F	T	0.01	0.99
T	F	0.01	0.99
T	T	0.01	0.99

Figure 2.2: An example of a simple Bayesian network modeling the relationships between three features.

To accurately portray how discriminatory proxies interact and the complicated relationships between different types of bias, researchers have turned away from concrete calculations of relationships and towards probabilistic causal structures. Recent works measure the effect of a protected attribute by modeling the probability distribution of a class with Bayesian networks, which are able to represent conditional dependencies [23]. A simple example can be found in Figure 2.2.

Notably, a team of researchers extracted the causal structures existing among attributes within a data set of historical decision records in the form of a constrained Bayesian network, on which they performed random walks so to model a variety of anti-discrimination legal concepts (direct and indirect, group and individual) that have been under-considered within computational research [5].

2.2 Methods

We constructed a Bayesian Network with the goal of creating a transparent network of dependencies between different features, with the potential to generate synthetic data that can be used later within our research. Using the Python package *pomegranate* for implementing probabilistic models, we built a basic Bayesian Network modeling a series of relationships between the following features:

- Sentiment (as determined by natural language processing tools)
- Race
- Geographic Location

In this example, race is linked by strong correlation with sentiment and location is linked by strong correlation to race. With race as a protected attribute, location becomes a discriminatory proxy for race. Once the model was trained on known relationships between each pair of features (as expressed by probabilities between 0 and 1, contained within a Conditional Probability Table), it was then prompted to predict the value of a protected attribute - race, in this case - given the known values of sentiment and location.

2.3 Results and Discussion

The model was able to correctly predict the value of the protected attribute in the slight majority of cases, but the results were ultimately inconclusive. Though building the network offered us better insight into how proxy discrimination can be computationally represented, it became clear to us that no synthetic network, even one based off of conditional probabilities, can come close to matching the intricate webs of relationship represented by even the most simple of data sets. Our program only tested one linear instance of proxy discrimination, and the complexity would only increase exponentially as included features increased. From these results, we made the decision to commit both time and monetary resources to obtain real data for analysis rather than to build a more complex network to generate synthetic data. Though the process was helpful for both our understanding of the existing work in the field and allowed us to make an educated decision regarding our next steps, we ultimately had to give up the transparency and clarity that comes with building our own network for the sake of accurately representing the complex relationships between biases that exist within all data sets.

Chapter 3

Human Demographic Prediction

The decision to commit to gathering and building off of real, human-generated data brought with it a variety of additional concerns. Social media data is the primary subject of much of current natural language processing research and has been used for everything from predicting the outcomes of political elections [7] to crime prediction [32]. By its nature, however, it works almost entirely with public data that contains little to no concrete identifiers of identity. Even when they exist, the opt-in nature of sharing information on social media can result in biased data sets as the users who share information about geographic location, for example, is significantly different from the overall population of users [29].

As a result, demographic information must be obtained for hundreds of thousands (and more) of individuals for a single trial and the choice often comes down to relying on either human annotation or prediction by a computational system. The latter option is magnitudes cheaper and faster, which has led to an explosion of research describing how to predict the demographics of Twitter users using everything from regression models built on website traffic data [12] to text analysis of usernames [35] to recursive neural networks [22].

The performance of these models are ultimately validated on a smaller set of Twitter users labeled individually for ethnicity and gender by human annotators. Despite criticism and controversy regarding human subject privacy and inaccuracies in prediction, it is clear that computational demographic prediction is and will continue to be an integral part of natural language processing and broader computer science research. The general consensus in the community seems to be that despite inaccuracies and bias, the ability to predict demographic information is too crucial to research to give up entirely.

But are human beings really the least biased identifiers of race and gender, or even accurate enough to act as a standard for truth to machines? Our literature review revealed a growing reliance of research on digital crowd sourcing and micro-working platforms, particularly Amazon's Mechanical Turk (MTurk), by virtue of being inexpensive and easily accessible. Concerns have been brought up in the community about the dubious legality and ethics of using what some call ”sweatshop labor” where workers earn roughly two dollars an hour [30]. In addition, the vast majority (roughly 70 percent) of all MTurk workers use the service either as a primary or secondary source of income, meaning their demographics are not representative of the general population - specifically skewed towards low-income individuals with less access to higher education [16]. Many of the papers from our literature review do not describe the experiment design or benchmarks used for the data annotations done by humans. Instead, the human demographic annotations were treated as the ground truth against which uncertain computational models were to be compared against. The consequence of this assumption is that if systematic bias exists within those human predictions, the accuracy of the paper results - and possibly, an entire field of research - is called into question.

Due to the time and financial limitations of an undergraduate thesis, our project had to use MTurk for data annotation. It was important to us that despite the ethical and accuracy concerns that came with that choice, however, that we base our experiment design on data as opposed to assumptions. Though there is no such thing as a perfect experiment design, especially when working with such thorny topics of bias and fairness, we wanted to be able to justify why we made the choices we did. In particular, we focused on the question of how much and what kind of information to present to our participants in order to mitigate biases in human annotations.

3.1 Methods

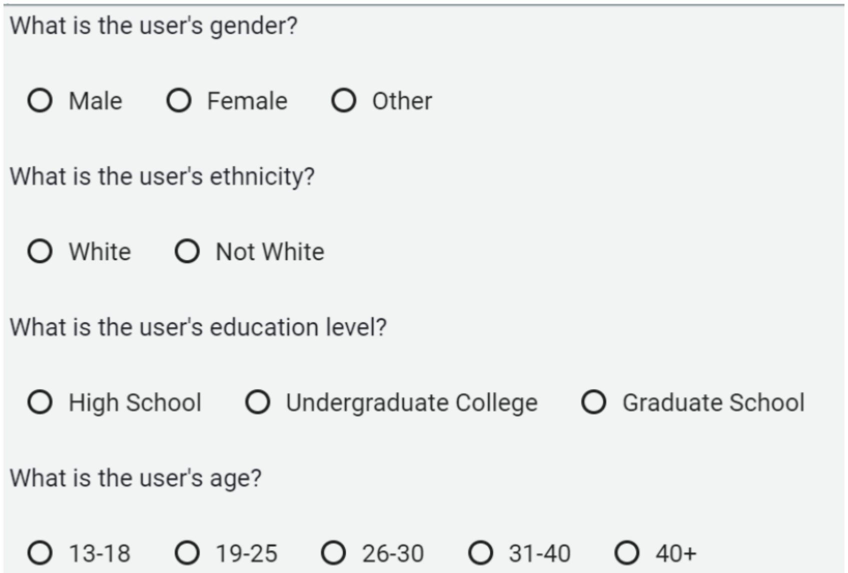
We conducted A/B testing using small collections of personally compiled Twitter accounts with known demographic information and a variety of identities. Our three trials were:

1. Full Twitter profile, minus the exact username so to protect the privacy of the users and to discourage participants from accessing additional information for demographic prediction and biasing our results.

2. Only the text portions of the Twitter profile, which included the display name and description.

3. Only the image portions of the Twitter profile, which included the cover photo and profile picture.

After being presented with the above information, the participant is then given a series of questions to answer regarding how they would predict the identity of the user. The MTurk task presentation is shown below.



What is the user's gender?

○ Male ○ Female ○ Other

What is the user's ethnicity?

○ White ○ Not White

What is the user's education level?

○ High School ○ Undergraduate College ○ Graduate School

What is the user's age?

○ 13-18 ○ 19-25 ○ 26-30 ○ 31-40 ○ 40+

We ran three trials, each with the same ten Twitter profiles, with differing types of information. Ten participants annotated each user, each for a total of 300 annotations.

3.2 Results

Generally, the responses returned by participants were more accurate when only text information was provided. The split in performance across the trials was not straightforward and varied based on ethnicity and gender of the profile (for example, "white" profiles had similar results across trials, while "non-white" profiles had more disparities in annotation in the text and full profile trials), though certain general patterns did arise. Comparing the results for the text-only trials with the full profile, we found that including the image encouraged users to ignore some of the information within the text entirely. For the most part, the results from full profile trials were more similar to those from image-only trials than those from text-only trials. For example, including images resulted in predictions of lower education levels for "female" profiles, but especially so for profiles that were annotated as both "female" and "non-white."

3.3 Discussion

Due to the limitations in the sample size and insufficient results to determine statistical significance, no wide-sweeping conclusions can or should be made based on the data from these experiments. Conducting these trials provided us with context with which to make more educated decisions regarding our final experiment design, however.

Our main takeaway was that images tended to polarize our annotators, possibly because it allowed more room for a human annotator to inject their own biases into their predictions. Though differences in overall accuracy were not immediately clear, the results from the text-only trials appeared more grounded in the information provided. On the other hand, trials that included images had more annotations that were not directly drawn from any of the data provided to the annotator.

With that in mind, we made the decision to conduct our data annotations with only the text data that would be visible on a Twitter profile page. Though we were concerned by the apparent biases and polarization that emerged when images were included, our main consideration was to prioritize transparency in our data annotations. Sociolinguistic identity and actual identity demographics are deeply linked but do not always align, and because the task of toxicity detection falls under natural language processing, biases that emerge will pertain more so to sociolinguistic identity than to the real demographics of the users, which can be more apparent with profile and cover images. Though conducting human demographic predictions using only text may not give us the most accurate results in regards to actual identity of a user, they result in decisions that are based more on information found in the presented data than from an annotator's personal beliefs. Ultimately, we decided that the latter fit more with the scope of our research aims than the former.

Chapter 4

Perspective API Audit

The primary goal of this study is to determine how Perspective API performs on tweets made by users across different demographic categories. We calculated the distribution curves and cumulative distribution functions for each of our chosen demographic categories, then determined statistical significance between pairs of demographics within the same group. Finally, we compiled collections of the most toxic tweets for each demographic category so to gain better insight into what the Perspective API determined to be harmful.

4.1 Methods

We used the Tweepy streaming Python library to pick up the usernames and descriptions, with emojis removed, of 3000 active users which we then split into three batches of 1000 each. Each batch was then submitted to Amazon's Mechanical Turk platform with three annotations requested for each user. The questions provided here were similar to those used in our earlier A/B testing, with the only difference being that ethnicity was broadened from a simple "White/Not White" dichotomy to incorporate the following four categories: White, Black, Asian, and Latinx.

We then parsed through the results to make two lists of users for each demographic category: consensus, to denote the users who were labeled as the particular demographic by all three of their annotators; and chosen, to denote the users who were labeled as the particular demographic by at least

Group	**Demographic**	**Chosen**	**Consensus**
Age	13-18	682	6
Age	19-25	264	71
Age	26-30	276	55
Age	31-40	74	11
Age	40 and up	68	18
Race	Asian	56	8
Race	Black	168	35
Race	Latinx	36	15
Race	White	619	204
Education	High School	214	34
Education	College	584	202
Education	Graduate School	78	14
Gender	Male	593	312
Gender	Female	367	142
Gender	Other	11	4

Table 4.1: Number of accounts by demographic

two out of three of their annotators. Though the number of accounts found per demographic varied somewhat across batches, the general proportions stayed the same. Our data for Batch 1 can be found in Table 4.1 below.

Though we initially hoped to use consensus accounts, it was clear that the small number of consensus accounts in some cases would result in significant bias from even just one account. We then scraped up to 3200 tweets of the most recent tweets for each user, eliminating retweets and quotes, and built a mapping between each user and their data set of tweets to efficiently create data sets of all tweets made by a user that had been a chosen account for a particular demographic. The data sets of tweets were then cleaned by removing all texts that were not in English and replacing shared URLs and mentioned users with the tokens ”URL” and ”USERNAME” so to preserve the original structure of the tweet and increase the accuracy of later processing.

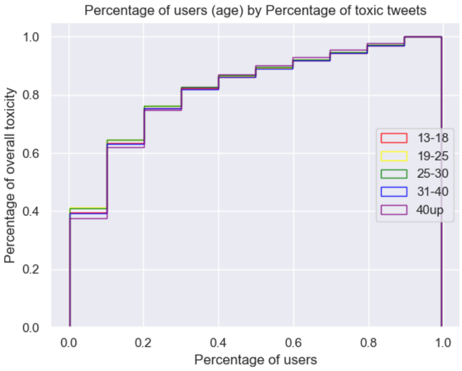
Using the Perspective API, we calculated for each tweet the scores for the following categories:

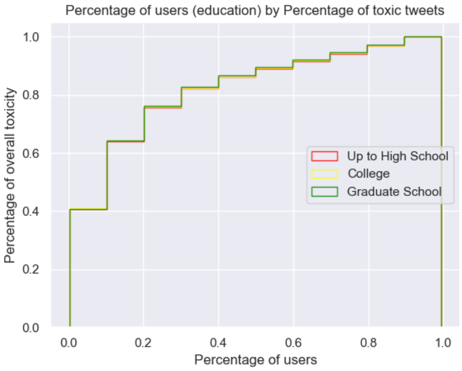
- Toxicity
- Severe Toxicity
- Incoherent
- Inflammatory
- Obscene
- Unsubstantial

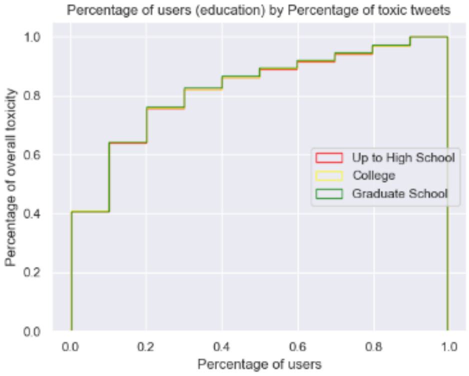
For our data analysis, we focused on the Toxicity category as all others have since been removed from the newest releases of Perspective API. Using the data visualization library Seaborn, we plotted the distribution graphs for each demographic for toxicity scores from 0.5 to 1.0, as well as cumulative distribution functions for each group. A one-way ANOVA test was performed using SciPy on each pair of demographics within the same group, due to our large sample sizes. Our goal was to determine if the distribution of toxicity originated from the same distribution, or in other words, whether the disparities between the toxicity distributions between different demographics were statistically significant. Finally, we outputted data sets of the most toxic (as determined as toxicity scores greater than or equal to 0.8) tweets for each demographic.

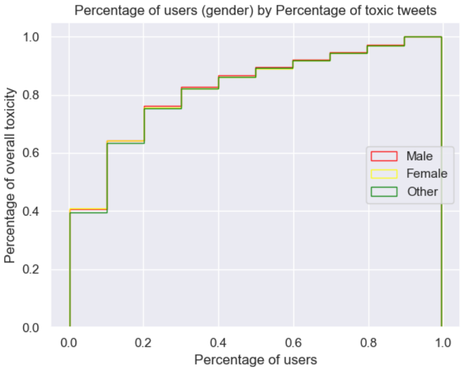
4.2 Results

The distribution and cumulative distribution function (CDF) graphs were broadly similar across all demographics. As seen below, however, the most variation lay in the Age and Race groups, with virtually no difference within the Gender and Education groups.


Percentage of users (age) by Percentage of toxic tweets
Percentage of overall toxicity
1.0
0.8
0.6
0.4
0.2
0.0
13-18
19-25
25-30
31-40
40up
0.0
0.2
0.4
0.6
0.8
1.0
Percentage of users


Percentage of users (education) by Percentage of toxic tweets
Percentage of overall toxicity
1.0
0.8
0.6
0.4
0.2
0.0
Up to High School
College
Graduate School
0.0
0.2
0.4
0.6
0.8
1.0
Percentage of users


Percentage of users (education) by Percentage of toxic tweets
Percentage of overall toxicity
1.0
0.8
0.6
0.4
0.2
0.0
Up to High School
College
Graduate School
0.0
0.2
0.4
0.6
0.8
1.0
Percentage of users



We calculated the cumulative distribution functions and conducted one-way ANOVA tests for pairs of demographics within the same group. For each group, we validated our usage of the ANOVA test by running it on two data sets from the same demographic and receiving a p-value that indicated a lack of significant statistical difference. Our results can be seen below.

	19-25	26-30	31-40	40up
13-18	$3.45e-5$	$3.87e-4$	0.02	0.01
19-25		0.50	$1.35e-11$	$1.85e-6$
26-30			$8.12e-10$	$7.01e-6$
31-40				0.18

Table 4.2: p-values calculated between age demographic pairs. Values lesser than 0.05 denote a statistically significant difference between the distributions.

	Female	Other
Male	$8.48e-8$	0.52
Female		$3.26e-8$

Table 4.3: p-values calculated between gender demographic pairs. Values lesser than 0.05 denote a statistically significant difference between the distributions.

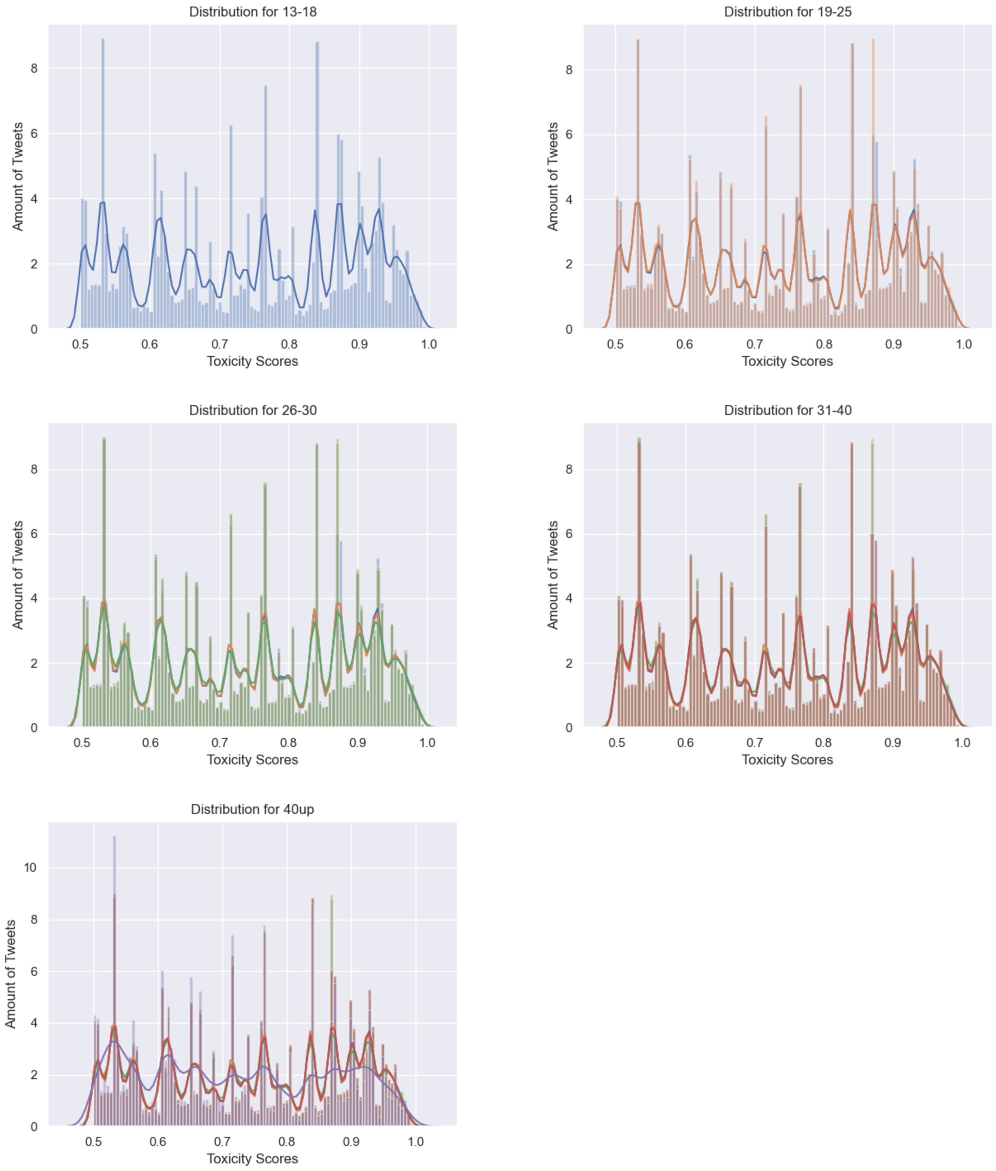
	White	Latinx	Asian
Black	$1.43e-57$	$1.84e-6$	7.12e-65
White		$2.80e-20$	0.01
Latinx			$4.17e-31$

Table 4.4: p-values calculated between race demographic pairs. Values lesser than 0.05 denote a statistically significant difference between the distributions.

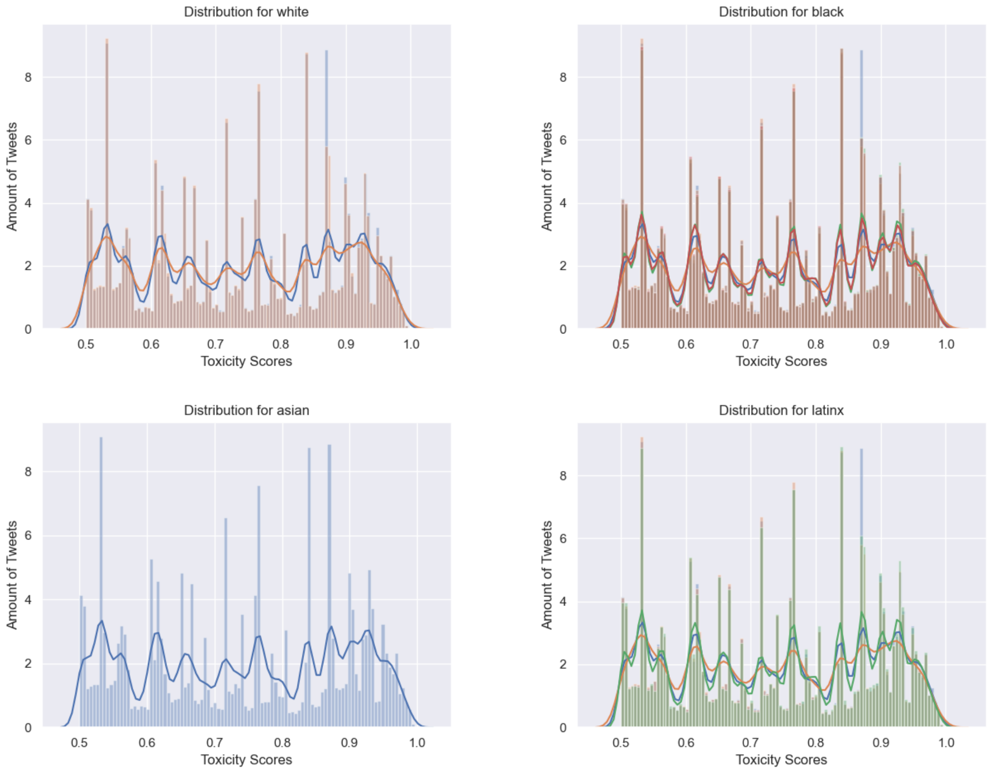
	College	Graduate School
High School	0.015	$1.23e-8$
College		$1.82e-5$

Table 4.5: p-values calculated between education demographic pairs. Values lesser than 0.05 denote a statistically significant difference between the distributions.

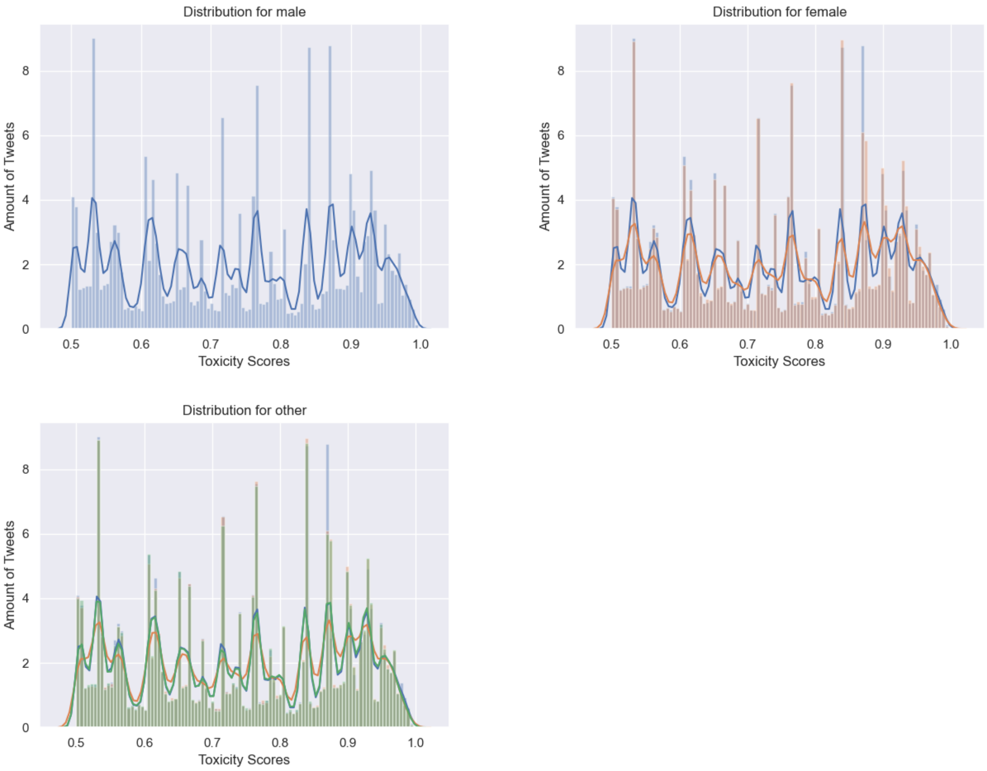
Distribution of Toxicity Scores (0.5-1.0) across Age Groups



Distribution of Toxicity Scores (0.5-1.0) across Race Groups



Distribution of Toxicity Scores (0.5-1.0) across Gender Groups

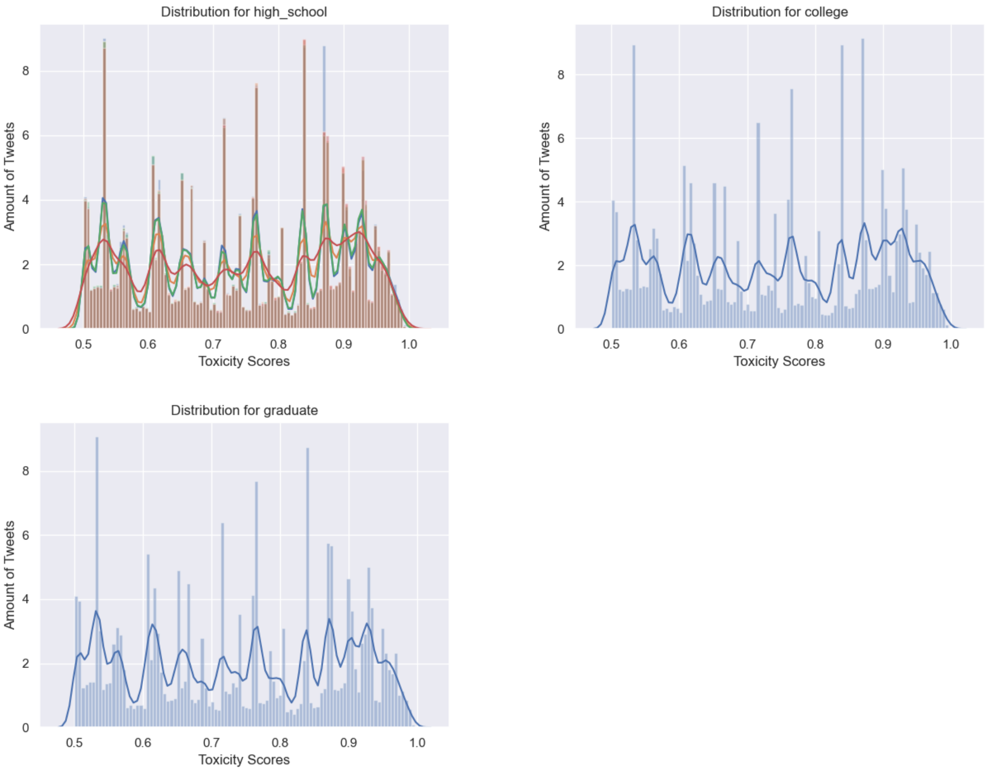


Distribution for male

Distribution for female

Distribution for other

Distribution of Toxicity Scores (0.5-1.0) across Education Groups



Examining the most toxic tweets for each demographic revealed an abundance of false positives and false negatives. Some examples are shown in Table 4.6.

Text	**Score**
Oh shit! Is that a spider web?	0.910651
It's not like Trump has a magic want he can wave and ""poof goes the illegals."" Unfortunately the #DEMONcratParty has a lot of power. Quite frankly, I'm surprised that Trump was able to do what he did for our country, despite Nancy #Bitchlosi and grand coon #MaxineWaters	0.34032694
Me at 20, who gives a fuck	0.91888636
Stay the fuck at home! Hahah	0.9516197
They are not immigrants, they are illegals and they ARE INVADERS, when they come across illegally. You d think someone running for US congressman would know that	0.46589994
Can God please work a miracle and get rid of is pandemic, I'm sick and tired of staying in my room, sick of numbers rising rapidly, innocent people are being killed and cannot sa goodbye to their families, disgusting ass people treating the victims body like shit.	0.92024845
IMAGINE BEING THIS FCKN TALENTED I AM SHOOK	0.83785176
That s drip that can t just be got at the mall!!! Queen shit only	0.8763313
They talking bout we might have to work from home cause of this corona virus shit I m shooked no cap.	0.8376127

Table 4.6: Toxicity scores calculated by Perspective API, 2020

The inclusion of profanity was the strongest indicator for toxicity, with a significant amount of the most toxic tweets being short pieces of text containing one curse word. On the other hand, tweets that attacked particular people or groups without using profanity, or even used racial slurs, received comparatively low toxicity scores. On the other hand, tweets that included slang and linguistic features originating from African American Vernacular English (AAVE) received higher toxicity scores.

4.3 Discussion

Our research sought to find differences between performance across demographic categories, but disparities in performance appeared to be tied much more to the amount of profanity and or contextual toxicity. It is important to note that age and race, the two groups that did show significant differences in distribution and CDF graphs, are both sociolinguistic identities that are influenced strongly by the usage of profanity [28]. The vast majority of tweets that Perspective API scores as highly toxic are false positives due to the prevalence of innocuous profanity usage on social media, which suggests that the model performs inaccurately on much of the social media data it is meant to run on.

Though inaccuracy does not necessarily imply lack of fairness, the two are deeply linked in this case. Disparate impact based on the usage of profanity seems relatively innocuous, as profanity is not a protected attribute like race, gender, or age. Drawing from our previous discussion of proxy discrimination, however, we find that usage of profanity correlates so strongly with the protected attributes of age and race that its inclusion nonetheless results in concerning disparities in performance that are linked to protected attributes.

Overall, our results indicate that while Perspective API is highly effective at identifying tweets that contain profanity, it performs less effectively when finding tweets that cause harm. The overwhelming majority of highly toxic (with scores above 0.8) tweets were short pieces of text input that contained profanity, effectively masking tweets that contained slurs, attacks on identity, and what can legally be defined as hate speech that were given scores ranging from 0.3 to 0.5. It is important to note that the issue of false positives within Perspective API leads directly to that of false negatives in that the model does not score the false negatives as completely harmless, rather that the their toxicity scores appear insignificant and comparatively safe in comparison to the large number of tweets with high toxicity scores. In other words, the model prioritizes profanity over personal and identity-related attacks, likely because there are simply many more instances of the former in any social media data set than the latter, just from the nature of how people communicate on the Internet - including the toxicity data sets used for training by Jigsaw.

4.4 Next Steps

There are a number of next steps for this research as our investigation did not prove its hypothesis as much as it unearthed an alternate angle under which toxicity and hate speech detection should be evaluated.

We found that the performance of Perspective API depended not on conventional categories of demographic information, but on linguistic features and terminology that acted as discriminatory proxies for identity. Instead of focusing on how models perform differently on data generated across demographics, our next steps would focus on disparities across different kinds of language and content. That is not to say that it is more important to discuss censorship of profanity over racial and age biases, but rather to acknowledge that the latter are a by-effect of the former and that more clear results can be generated by studying one degree higher of this proxy-chain relationship.

Chapter 5

Conclusion

The Perspective API has experienced its share of criticism and praise, both of which are justified. Ultimately, the model does not perform badly as much as it simply performs in ways that are undirected. These problems of fairness and equity originate not from technical errors within its code, but from incorrect or oversimplified assumptions made in research and experimental design. The Perspective API models were trained on large data sets to determine 'toxicity', a nebulous term that a data-driven systems cannot define independently. Toxicity can mean sentiments of anger in one context and the usage of racial slurs in another, and the arbitrary conflation of these very different circumstances is what drives systematic disparate impact. Entirely technical solutions like up-weighting parts of a data set do not fully address and resolve the problem. Ultimately, they act as temporary patches over a particular symptom of muddled design that must be added to when another issue arises.

Without outside interference, computers will prioritize efficiency and correlation, not human concepts of harm and fairness. This is not to criticize computational research, more to point out that injustice within technology often arises when data-driven systems are used to solve problems that are out of its scope - whether they are issues that cannot be fully represented by data or just simply too general to allow a computer to define. Technology is most effective when it attempts to solve specific and well-defined problems. When it is forced to fill in the gaps, inaccuracies and biases arise.

This is not to say that technology makes more mistakes than humans do, rather that the consequences of injustice within technology are distinctly harmful in ways that merit our concern. Not only are our current ethical and legal systems designed to manage people and are therefore ill-suited for regulating technologies, but the myth of objectivity that surrounds tech can lead people to put more trust in machines than they would in another person. While the calculations and mathematics within computer science are indeed objective, the design choices that govern them are deeply influenced by human biases. Instead, computational research that attempt to solve human problems should lean into its inherent subjectivity and draw from theory outside of the field of computer science to make informed decisions. The vast majority, if not all of these definitions are imperfect and extremely dependent on context. What is most important for technology right now is not to be perfect, however, but for it to be transparent and accountable.

On that note, the team behind Perspective API appears to be taking a step in the right decision by pivoting away from the monumental task of both computationally defining and finding general toxicity towards more specific and targeted categories of harmful language with a long history in legal literature, like "Flirting" and "Identity Attack." Though it is more likely than not that these models too will have their own problems and imperfections, their specificity and precedence in broader theory allows for regulation and improvements. It is our hope that the decision marks a far-reaching shift within the field of computer science towards prioritizing the transparency and accountability of algorithms, models, and systems over their immediate outcomes.

Bibliography

[1] Julia Angwin et al. *Machine bias: There's software used across the country to predict future criminals and it's biased against blacks.* URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[2] Anya and Daniel Schwarcz. *Proxy Discrimination in the Age of Artificial Intelligence and Big Data.* URL: `https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data/`.

[3] Greg Bensinger and Reed Albergotti. "YouTube discriminates against LGBT content by unfairly culling it, suit alleges". In: *The Washington Post* (2019). URL: `https://www.washingtonpost.com/technology/2019/08/14/youtube-discriminates-against-lgbt-content-by-unfairly-culling-it-suit-alleges` (visited on 12/07/2019).

[4] Tolga Bolukbasi et al. "Quantifying and reducing stereotypes in word embeddings". In: *arXiv preprint arXiv:1606.06121* (2016).

[5] Francesco Bonchi et al. "Exposing the probabilistic causal structure of discrimination". In: *International Journal of Data Science and Analytics* 3.1 (2017), pp. 1–21. DOI: `10.1007/s41060-016-0040-z`.

[6] Shikha Bordia and Samuel R. Bowman. "Identifying and Reducing Gender Bias in Word-Level Language Models". In: *Proceedings of the 2019 Conference of the North* (2019). DOI: `10.18653/v1/n19-3002`.

[7] Pete Burnap et al. "140 Characters to Victory?: Using Twitter to Predict the UK 2015 General Election". In: *SSRN Electronic Journal* (2015). DOI: `10.2139/ssrn.2603433`.

[8] Alexandra Chouldechova. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments”. In: *Big Data* 5.2 (2017), pp. 153–163. DOI: `10.1089/big.2016.0047`.

[9] Fast Company. *High-tech redlining: AI is quietly upgrading institutional racism.* Nov. 2018. URL: `https://www.fastcompany.com/90269688/high-tech-redlining-ai-is-quietly-upgrading-institutional-racism`.

[10] *Conversation AI.* URL: `https://conversationai.github.io/`.

[11] Kimberle Crenshaw. “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics [1989]”. In: *Feminist Legal Theory* (1989), pp. 57–80. DOI: `10.4324/9780429500480-5`.

[12] Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. “Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data”. In: *Journal of Artificial Intelligence Research* 55 (2016), pp. 389–408. DOI: `10.1613/jair.4935`.

[13] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. “Racial Bias in Hate Speech and Abusive Language Detection Datasets”. In: *Proceedings of the Third Workshop on Abusive Language Online* (2019). DOI: `10.18653/v1/w19-3504`.

[14] Pablo Delgado. *How El País used AI to make their comments section less toxic.* Mar. 2019. URL: `https://www.blog.google/outreach-initiatives/google-news-initiative/how-el-pais-used-ai-make-their-comments-section-less-toxic/`.

[15] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. *False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.”*. URL: `http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf`.

[16] Karen Fort, Gilles Adda, and K. Bretonnel Cohen. “Amazon Mechanical Turk: Gold Mine or Coal Mine?” In: *Computational Linguistics* 37 (2011), pp. 413–420. DOI: `10.1162/COLI_a_00057`.

[17] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. *On the (im)possibility of fairness...* Sept. 2016. URL: https://algorithmicfairness.wordpress.com/2016/09/26/on-the-impossibility-of-fairness/.

[18] Sorelle A. Friedler et al. "A comparative study of fairness-enhancing interventions in machine learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* 19* (2019). DOI: 10.1145/3287560.3287589.

[19] *Google's comment-ranking system will be a hit with the alt-right.* URL: https://www.engadget.com/2017-09-01-google-perspective-comment-ranking-system.html.

[20] Jessica Guynn. *Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech.* Apr. 2019. URL: https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/.

[21] Ben Hutchinson and Margaret Mitchell. "50 Years of Test (Un)fairness". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* 19* (2019). DOI: 10.1145/3287560.3287600.

[22] Sunghwan Mac Kim et al. "Demographic Inference on Twitter using Recursive Neural Networks". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2017). DOI: 10.18653/v1/p17-2075.

[23] Koray Mancuhan and Chris Clifton. "Combating discrimination using Bayesian networks". In: *Artificial Intelligence and Law* 22.2 (2014), pp. 211–238. DOI: 10.1007/s10506-014-9156-4.

[24] Northpointe. *COMPAS Risk and Need Assessment System.* URL: http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf.

[25] Cathy ONeil. *Weapons of math destruction: how big data increases inequality and threatens democracy.* Penguin Books, 2016.

[26] *Perspective.* URL: https://www.perspectiveapi.com/#/home.

[27] Maarten Sap et al. "The risk of racial bias in hate speech detection". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 1668–1678.

[28] H. Andrew Schwartz et al. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach". In: *PLoS ONE* 8.9 (2013). DOI: `10.1371/journal.pone.0073791`.

[29] Luke Sloan and Jeffrey Morgan. "Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter". In: *Plos One* 10.11 (June 2015). DOI: `10.1371/journal.pone.0142209`.

[30] Jeremy Snyder. "Exploitation and sweatshop labor: Perspectives and issues". In: *Business Ethics Quarterly* 20.2 (2010), pp. 187–213.

[31] Daisuke Wakabayashi. *Google Cousin Develops Technology to Flag Toxic Online Comments*. Feb. 2017. URL: `https://www.nytimes.com/2017/02/23/technology/google-jigsaw-monitor-toxic-online-comments.html?_r=0`.

[32] Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown. "Automatic Crime Prediction Using Events Extracted from Twitter Posts". In: *Social Computing, Behavioral - Cultural Modeling and Prediction Lecture Notes in Computer Science* (2012), pp. 231–238. DOI: `10.1007/978-3-642-29047-3_28`.

[33] Zeerak Waseem and Dirk Hovy. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter". In: *Proceedings of the NAACL Student Research Workshop* (2016), pp. 88–93. DOI: `10.18653/v1/N16-2013`.

[34] Williams, Brooks, and Shmargad. "How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications". In: *Journal of Information Policy* 8 (2018), p. 78. DOI: `10.5325/jinfopoli.8.2018.0078`.

[35] Zach Wood-Doughty et al. "Predicting Twitter User Demographics from Names Alone". In: *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (2018). DOI: `10.18653/v1/w18-1114`.