Intersectional Identities and Machine Learning: Illuminating Language Biases in Twitter Algorithms

Aidan Fitzsimons
Duke University
aidan.fitzsimons@alumni.duke.edu

Thanks to Dr. Jay Pearson and Dr. Kenneth Rogerson for their guidance, support and editing to make this project possible.

Abstract

Intersectional analysis of social media data is rare. Social media data is ripe for identity and intersectionality analysis with wide accessibility and easy to parse text data yet provides a host of its own methodological challenges regarding the identification of identities. We aggregate Twitter data that was annotated by crowdsourcing for tags of "abusive," "hateful," or "spam" language. Using natural language prediction models, we predict the tweeter's race and gender and investigate whether these tags for abuse, hate, and spam have a meaningful relationship with the gendered and racialized language predictions. Are certain gender and race groups more likely to be predicted if a tweet is labeled as abusive, hateful, or spam? The findings suggest that certain racial and intersectional groups are more likely to be associated with non-normal language identification. Language consistent with white identity is most likely to be considered within the norm and non-white racial groups are more often linked to hateful, abusive, or spam language.

1. Introduction and intersectionality

"Intersectionality" was coined by Kimberlé Crenshaw [1] while reviewing various legal precedents that were relevant to both race and gender. The framework of intersectionality is a complicated one, but its core concepts are simple: intersectionality argues that there is a special kind of identity that is born out of a multitude of identities, and that stands on its own. In other words, the specific identity of being a black lesbian has its own set of experiences that cannot be represented merely by the sum of a woman's experiences, a queer person's experiences, and a black person's experiences.

Intersectionality is not a content specialization, nor is it automatically synonymous with feminist-, queer-, or POC-based scholarship [2]. Intersectionality covers the multiplicity of identities that one person, or a collective of people, may possess. Intersectionality aims to represent the intersection of identities in a way that respects each of them individually and how each of them interacts with one another. It is the question of how specific identities co-act in comparison to other groups that makes something intersectionality research.

Within intersectionality, in-group diversity will always be a present factor no matter how many different strata we consider. Considering black lesbians, for example (at the intersection of gender, race, and sexual orientation), the diversity within the sample will always be great because of the implicit differences not covered by the intersectional labels which we are considering. In the example case of black lesbians, we fail to consider how that plays a role in socioeconomic status, ability status, national origin, immigration status, and so many more strata that affect an individual's daily life and are core to their being. This is not to say that this work is futile, but instead to note that it is essential that the intersectional researcher acknowledge what they fail to capture.

1.1. Quantitative intersectionality

Weber and Parra-Medina have asked: "How can a poor Latina be expected to identify the sole—or even primary—source of her oppression? How can scholars with no real connection to her life do so?" [3]. This is precisely why intersectionality is important.

Quantitative intersectionality is a subsection of the intersectionality literature that looks to use big data methods and statistical testing to represent intersectional identities. More specifically, quantitative intersectionality research in the fields of

URI: https://hdl.handle.net/10125/79690
978-0-9981331-5-7
(CC BY-NC-ND 4.0)

psychology [4], political science [5], and epidemiology [6] have emerged by comparing intersectional groups' experiences with healthcare, behavioral science, the legal system, and political structures at large.

Attempting to quantify intersectionality provides a host of methodological challenges. The greatest challenge comes in the availability of information; when researchers collect demographic information about their participants, it is often to ensure that demographic factors are *not* relevant to their research work. In other words, researchers often collect just enough demographic information to rule out that different identities experience a phenomenon in a certain way [7]. Most researchers, especially prior to the last 15 or so years, viewed identities as confounding factors rather than main effects drivers.

Ultimately intersectionality research goes against many of the assumptions implicit in numerical analysis and requires a great deal of care on the part of the researcher to collect meaningful variables related to lived experience prior to analysis.

McCall [8] took on the task of separating the broad field of intersectionality research by methodology. Based on the population studied and the adherence to social labels, we have three methodic categories: the *intracategorical*, the *intercategorical*, and the *anti-categorical*. The intercategorical approach, the focus of this project, is born out of a different central interest than the other two approaches; while the intracategorical and anti-categorical question the validity of the labels used in research, the intercategorical approach highlights real differences in lived experiences quantitatively between multiple identity categories. It considers relationships of inequality between different existing social categories, as imperfect as those categories may be, and attempts to quantify differences in lived experiences between them. The bulk of quantitative intersectionality studies, especially those in the domain of public health, fall into this category.

The study that will follow attempts to quantify differences in multiplicative oppression for different intersectional identity groups. As such, this work is inherently comparative, and thus falls into the intercategorical framework most cleanly. However, we fall short of a truly intersectional analysis (section 2.2) because of the constraints of data availability; this failure to aptly address the multiplicity of identities that one may hold is reflective of the challenges and constraints of intersectional analysis.

Since the inception of intercategorical analysis, a great deal of nuance has been added to the field, both in the questions that researchers are asking, and in the methods employed to answer these more complex questions. A variety of studies fall into two main categories regarding their approach: either they focus on comparing one stratified category (e.g., young black woman) to the broader population, or they use a multitude of categories to compare across a variety of strata. These studies have employed a variety of methods, from a simple bivariate analysis [9, 10] to regression models [11] to more complex multi-level models that use techniques like tree classification to analyze intersections based on heterogeneity [12]. Some methods use even more complex models with mixed methods to analyze many high-dimensional interactions to speak to a variety of intersectional identities at once [13].

2. Social media and intersectional analysis

Since McCall's initial classification of different intersectionality methods, the emergence of platforms like Facebook, Twitter, and Instagram have created a wealth of data than to analyze the human experience. Particularly on platforms like Twitter, users share their thoughts about nearly anything, and thus there is a great deal of information to be gleaned from public posts on social media platforms.

A recent focus of technology ethics literature has been the study of how algorithms may be based on biased training data. In a 2019 study published in Science, Ziad Obermeyer and colleagues worked on studying discrimination in a health prediction algorithm that has real implications for millions of Americans [14]. This interest has extended to the algorithms that big companies use, but since companies like Twitter and Facebook are privately held corporations, many of their activities are kept behind lock and key.

To do this project, we need to know the gender and racial identities of the users. This raises some challenges as users may not share this information or may choose to represent or highlight a variety of identities. While some researchers would argue that pursuit of identification is unethical because of privacy concerns, and we would tend to agree for the sake of people's opportunity to self-identify, a body of research has developed to use different characteristics of text or Twitter profiles to predict race and gender. There are both non-machine and machine learning techniques for gender and race prediction.

2.1 Racial/gender identity on social media

Using a non-machine learning approach, a group of researchers in 2011 attempted to understand who

makes up the Twitter community in the United States and found that Twitter usage is highly dependent on locational and demographic characteristics [15]. The authors note under sampling of certain minority groups, such as black people in the rural south, and oversampling of urban white people. However, much of this analysis was done on an extremely faulty assumption: since Twitter users' demographics are private, these authors used last name to proxy for race. This method, while popular, is extremely imprecise.

Another work, getting closer to the heart of this study, tries to quantify privileges for different intersectional identities on Twitter [16]. To identify users' race and gender, they use a system called Face++ to process Twitter users' profile images. Face++ is an advanced image processing algorithm that can identify race and gender with moderate certainty.

For machine learning, a variety of studies have attempted to quantify intersectional bias in one way or another using Twitter data. The study, titled "White, Man, and Highly Followed: Gender and Race Inequalities in Twitter," uses the Face++ algorithm to determine users' identity factors, and tries to identify which identity groups receive privilege on Twitter in the form of followers and interactions. They find that white men are more likely to interact with one another on Twitter, and as the most populous group on American Twitter, their interaction with one another causes higher incidence and follower rates for white men on Twitter. They also find that racial groups stay within their own "icebergs," to an extent, as the greatest interaction for nearly every intersectional identity group is within their own identity. This is something extremely interesting about the characteristics of Twitter users' behavior, and something that will inform much of the coming analysis.

Another study [17], closer to what we will seek to quantify, uses a "Bag of Words" method to flag and mark hateful terms within Twitter data. The study uses incidences of specific political interest in an identity factor to track identity discourse on Twitter for hatred espoused against certain groups. For example, they used the re-election of Barack Obama to track racial hate and the coming out of pro-Basketball player Jason Collins to track intersectional black/queer hate. They build models for each incident and quantify intersectional discrimination on Twitter through a by-event lens.

Another study that looks in detail at the discrimination of abusive language on Twitter will be foundational to our work going forward in this paper. In fact, the scholars that produced this paper published their annotated dataset, and that's what we'll be using in the coming work to do an analysis of our own. Antigoni-Maria Founta and her colleagues used crowdsourcing techniques to annotate over 80,000 individual tweets for abusive behavior [18]. This dataset provides the basis for a lot of research about abusive behavior on Twitter, including this one as described in the methods section. Of course, human-created labels for hateful or abusive speech leaves the opportunity for human error and bias in the data outcomes, and that's what we'll be investigating.

Training data can have an impact on the systems that they train, and if a human group (with ever-present biases) creates training data, they may train the system to reproduce their own biases without even knowing it.

To address these concerns, researchers [19] have suggested methodological changes like assessing how opinionated an individual worker may be. While much of crowd marking research has addressed objective marking criteria (e.g., identifying whether an image has a car) and issues of fatigue and lack of interest, this issue is particularly poignant when it comes to subjective judgements, like whether a tweet may be identified as spam, hateful, or abusive rather than normal. It is unclear whether Founta and colleagues took much precaution in identifying underlying biases of its crowd marking coders [20, 21].

2.2. Challenges in intersectional data analysis

There are challenges when attempting to find and analyze data to address intersectional questions. The availability of identity-linked data on social media is sparse given concerns of privacy, and when we use technical methods to address a lack of availability of identity-linked data, we bring in new confounding factors.

This study was limited by timeframe and funding. Because of these limitations, we sought a dataset with few specific criteria: first, a measure of interest that describes something meaningful about a person's experience (i.e., health outcomes, technological bias, etc.); second, a dataset with many identity factors for analysis. We failed to find a publicly accessible dataset that met both criteria and wondered how other entities (in academia, business, and the like) predict identity factors about intersectional identity.

Using machine learning and natural language processing to predict identity factors is a common, easily accessible approach to dealing with these uncertainties. However, NLP techniques have not advanced to the point of reliable predictions for identity strata outside of race and gender. A machine learning approach changed the issues considered in three major ways.

First, a lack of predictive methods for different identity strata has limited our ability to address the issues of intersectionality. Intersectionality aims to provide as full a picture as possible and, even with 100% identification reliability (which has not been achieved), addresses only two major identity factors.

Second, adopting a machine learning approach, the possibility of inaccurate identification becomes a major factor. This challenges whether NLP is an effective solution for addressing questions of identity. Though imperfect, these methods are being used by scholars and potentially corporations and real decisions are being made based on these outcomes.

Finally, how NLP algorithms are used after their publication is nearly impossible. While scholars are clearly using these predictive models [22], private companies do not publicize their techniques. Because scholarly work is publicly accessible, we expect that NLP techniques are being used privately as well.

In this study, we will be using the gender and racial predictors described in Sap et al. and Blodgett et al. [23] (described in the methodology) and cross-referencing that with the training data created by Founta. We are interested in whether this training data, created by humans, has an indication of bias within it and if outcomes are different for folks of a certain predicted gender and race. While none of these systems is perfect, they are being used in practice in research and almost certainly in business, and thus it is important to evaluate if the data created by Founta and the prediction algorithms created by Sap and Blodgett could reproduce racist or sexist results.

3. Methodology

In this study, we seek to quantify intersectional oppression between different demographic groups with respect to twitter behavior. Specifically, we are interested in evaluating models that use Natural Language Processing to predict race and gender and assess whether tweets that are flagged as abusive, hateful, or spam are more likely to be attributed by the model to certain racial and gender groups. We will construct our dataset using a variety of open-source research publications who have made their work accessible to the public for research purposes. As described above, three studies will form the basis for our analysis: Founta, Sap and Blodgett.

We will seek to answer questions addressing how the labeling that Founta produced can help us predict race and gender. Specifically, we will evaluate the following hypotheses:

- Predicted race of tweet author will have a meaningful interaction with the likelihood of being labeled as spam, abusive, or hateful.
- Predicted gender of tweet author will have a meaningful interaction with the likelihood of being labeled as spam, abusive, or hateful.
- Predicted intersectional identity of tweet author (race and gender) will have a meaningful interaction with the likelihood of being labeled as spam, abusive, or hateful.

3.1. Why Twitter Data?

In our analysis, we decided to use Twitter data for several reasons. Given its wide accessibility and short-form language, Twitter data is ripe for big data analysis. Twitter's identity within the social media landscape not only allows for significant free expression but also encourages different identity groups to speak to one another in the language most comfortable to their community. The kind of discourse we see on Twitter can be broad or isolated to a specific identity group (see note [16]), and that makes for raw and honest language ideal for our purposes. The academic field has taken a recent interest in Twitter data and how others perceive different tweets with different language styles, and the publicly accessible data allows for a great deal of statistical power.

The dataset comes from a study that takes crowdsourcing annotations to a large scale. Founta sought to create reliable labels based on human interpretation of different tweets. Specifically, the goal of the study was to 1) provide a reliable labeling with meaningful differences between different types of speech that "abuse" the Twitter platform, and 2) create and model a mechanism for crowdsourced annotation on a massive scale. To do so, the researchers recruited a great deal of annotators and trained them in differences between different kinds of abuse speech (abusive, hateful, or spam). For each tweet, between five and 20 reviewers were tasked with classifying it as abusive, hateful, spam, or normal. Once there appeared to be consensus among a tag (typically achieved after five reviewers, but sometimes more), that tag was solidified and attached to a tweet. This procedure was repeated for over 99,000 tweets (N = 99,534) and published for future researchers to gain insight. As noted above, the demographics of a group reviewing a specific tweet may not have been controlled to the degree that the literature recommends and could lead to particularly biased labelling.

3.2. Race and gender-predictive modeling

A model with a variety of racial categories with wide accessibility for short text datasets like Twitter better predicts the race of a tweet's author. For these reasons, we were drawn to Blodgett et al., 2016.

Blodgett builds on the body of Natural Language Processing studies that attempt to predict race based on certain lexical behaviors. While their research was originally structured around a racial binary (Black/White), they began to find terminology consistent with Hispanic and Asian folks' unique linguistic characteristics that they separated out. In their pursuit of predicting racial identity, Blodgett provides a rare opportunity to predict folks' identities more accurately than simply along the Black/White binary. It is worth noting that four racial categories are still not encompassing of the multitude of racial identities and therefore is less likely to correctly attribute authorship to correct identities. It can also completely misattribute the identities of groups not studied.

Blodgett and colleagues do not compare their model to a dataset with known racial identities. Instead, they turn to well-documented phonological and syntactic phenomena to address the question of model validity. Using lexical-level variation, orthographic variation, phonological variation, and syntactic variation, the authors compare results to U.S. Census data and Twitter demographics to conclude that the model performs effectively for large datasets. Note that Blodgett and colleagues only evaluate efficacy along the Black/White binary, despite including other identities, and cite a lack of sufficient information and research regarding other racialized language.

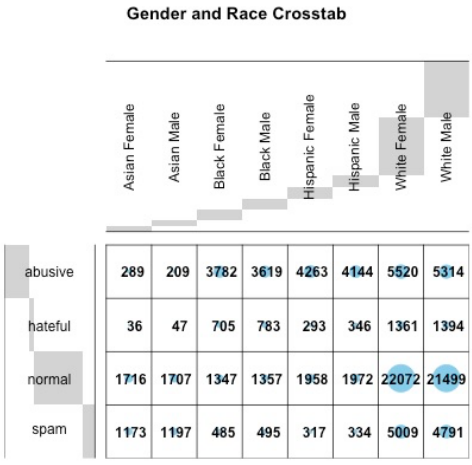
Similarly, we applied a gender-predicting model from Sap et al., 2014, because of its wide applicability to small text samples, high number of samples, and accessibility of methodology to be applied to other datasets. Based on Sap's own evaluation techniques, they report a 91.9% accuracy when it comes to predicting gender. To arrive at this accuracy number, using a dataset with known demographics of Facebook message data, they split their data and trained a model based on roughly 90% of their confirmed data. Next, they used the model on the remaining 10% of their data and calculated the number of correct applications of the model over all applications of the model for a 91.9% accuracy figure. They applied this model to a variety of social media platforms that use short-form language, including Twitter, and did not find significant variation from this 91.9% figure when they varied platform. Finally, to make this data accessible, they published their dictionaries on Github with the code implementing their method and prediction dictionaries.

Following the predictions of race and gender on the nearly 100,000 tweets tagged and published by Founta and colleagues, we are interested in seeing how these predicted identities correlate with the labels that Founta added. More specifically, we are interested in evaluating the likelihood that tweets tagged as "abusive," "spam," and "hateful" are predicted to be a certain gender and/or race. Said another way, we're curious if tweets that are flagged for abuse of some kind are predicted based on machine learning models to be consistent with a certain gender and/or racial lexica. It's important to understand here that we're not looking at the other direction: these data do not tell us whether folks of real identities are more likely to be flagged for abusive behavior. Because of the privacy limitations of Twitter data, we can only ask how these labels help us predict race and gender based on studied racial and gender lexical dictionaries. We seek to answer one key question: how might intersectional identities play into the way others perceive their tweets, particularly tweets that are identified as abnormal?

4. Results and discussion

Figure 1 shows the contingency tables for race, gender, and intersectional identity compared to the abuse labels. These visualizations show the cross-tabulated counts of each interactive term, for example: there were only 36 tweets marked as hateful that algorithms predicted were written by an Asian woman.

Figure 1. Gender and Race Intersection Cross-Tabulated Contingency Bubble Chart



A chi-squared test on each of our contingency tables – for race, for gender, and for intersectional identity – showed that each of our identity interacts significantly with the labels that Founta's crowdsourcing method provided. Table 1 reports the results of the chi-squared tests for each contingency table.

Table 1. Contingency Tables' P-Values

Contingency Table	p-value	Interpretation
Gender Table	0.003982	Gender interacts meaningfully with the label provided
Race Table	< 2.2e-16	Race interacts meaningfully with the label provided
Gender & Race Table	< 2.2e-16	Either gender or race, if not the interaction between them, interacts meaningfully with the label provided

Table 1 shows the p-values which allow us to reject the null hypothesis for each of our three tables, meaning that gender, race, and their interaction, have meaningful correlation with the label provided by Founta and colleagues. There are some interesting things to note here, too: since the size of our dataset is extremely large, we're working with a lot of statistical power. Sample size has a meaningful impact on how statistically significant an effect will show up and seeing that our p-value for the gender table is 0.003982, that should indicate to us that the correlation between predicted gender and the label outcomes that we're interested in is not particularly strong, especially compared to the predicted race and interaction terms which both have p-values less than 2.2e-16.

Testing for basic statistical significance using a chi-squared test on contingency tables allows us to reject the hypothesis that demographic predictions have no interaction with the labels that Founta crowdsourced, but it does not allow us to extend much further. If we wish to comment any further about what is going on, we should perform logistic regressions on each of our labels that deviate from the norm (spam, abusive, hateful) and look at the results of the logistic regression and an ANOVA on the logistic regression.

Next, we used R Studio to create and analyze a logistic regression for each of the non-normal labels attached to different tweets. To run logistic regressions on each of these labelling outcomes, we transform our outcome to a binary response. In practice, this looks like a binary marker for each of our labels of interest (spam, abusive, hateful) attached to each tweet. A logistic regression seeks to use the factors included to predict the binary outcome of interest. To include our two factors of interest (race and gender) and their interaction, we build three elements into our model: race, gender, and the interaction term. A logistic model was created for each label of interest and those outcomes were analyzed separately (full results available from authors). We run regressions using Asian, White, Hispanic, and Black folks as the baselines so that we can compare each identity group with one another.

Additionally, to summarize the performance and results of our model, we used an ANOVA to describe each of the model performances. We recognize that running an ANOVA on this data partially violates the assumption that the outcome of an ANOVA is normally distributed, running an ANOVA is a common way to summarize effects of multiple characteristics within a categorical variable. Thus, in addition to our regression models, we report ANOVA results using a likelihood-ratio test to speak on a factor level rather than by comparing each individual demographic subgroup, as is done in the logistic regression model (full results available from authors).

To showcase the relationships studied, we'll represent our results in several ways. First, we'll use the ANOVA results to ask which identity strata have meaningful relationships with the labels provided. Consider the following significance table:

Table 2. Tags and Demographic Categories

	Gender	Race	Gender : Race
Spam	NS	***	NS
Hateful	***	***	NS
Abusive	NS	***	**

Key: *** > 0.001, ** > 0.01, * > 0.05, NS ≤ 0.05

For all three of our labels, we immediately notice that race has a meaningful relationship with each of the labels provided, and that the statistical threshold for those results is high. In addition to race having a meaningful relationship with all three terms, gender has a meaningful relationship with the hateful tag and the gender race interaction has a meaningful relationship with the abusive tag. To investigate how different races, genders, and intersectional identities experience likelihood of being proscribed these labels differently, we need to look at the logistic regression analysis results. Because there are a lot of relationships to evaluate here, for the sake of simplicity we'll first analyze different racial groups in relation to one another, one label at a time. Consider the following table, which summarizes the different relationships and our interpretations based on statistical values.

Table 3. Racial Interaction Summary

Relationship	Tag	Significance	Statistic Comparison
Black / Hispanic	Spam	***	Z(Black) > Z(Hispanic)

Black / Hispanic	Hateful	***	Z(Black) > Z(Hispanic)
Black / Hispanic	Abusive	**	Z(Hispanic) > Z(Black)
Black / White	Spam	***	Z(White) > Z(Black)
Black / White	Hateful	***	Z(Black) > Z(White)
Black / White	Abusive	***	Z(Black) > Z(White)
Black / Asian	Spam	***	Z(Asian) > Z(Black)
Black / Asian	Hateful	***	Z(Black) > Z(Asian)
Black / Asian	Abusive	***	Z(Black) > Z(Asian)
Hispanic / White	Spam	***	Z(White) > Z(Hispanic)
Hispanic / White	Hateful	NS	NS
Hispanic / White	Abusive	***	Z(Hispanic) > Z(White)
Hispanic / Asian	Spam	***	Z(Asian) > Z(Hispanic)
Hispanic / Asian	Hateful	***	Z(Hispanic) > Z(Asian)
Hispanic / Asian	Abusive	***	Z(Hispanic) > Z(Asian)
White / Asian	Spam	***	Z(Asian) > Z(White)
White / Asian	Hateful	***	Z(White) > Z(Asian)
White / Asian	Abusive	***	Z(White) > Z(Asian)

Key: *** > 0.001, ** > 0.01, * > 0.05, NS ≤ 0.05

Nearly every racial interaction in Table 3 is highly statistically significant. This outcome could be driven by several factors, including our relative statistical power compared to other studies of this kind or the efficacy of the NLP models in comparison to their predecessors. Even so, these results are surprising and very strong. We believe that our process has been accurate, but further studies should approach this question with scrutiny regarding the extreme significance of our results.

Figure 2. Spam Logistic Regression Results

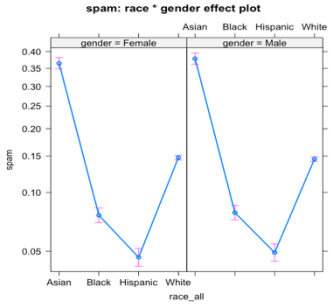


Figure 3. Abusive Logistic Regression Results

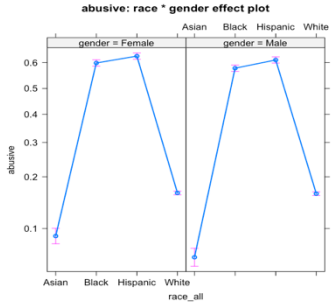
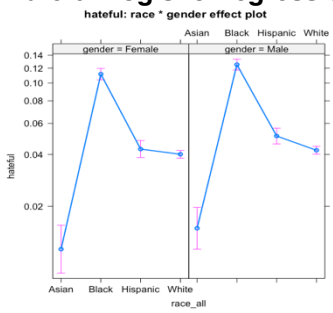


Figure 4. Hateful Logistic Regression Results



For each tag (spam, hateful, abusive), we can rank how strongly associated each racial category is with the label (see Figures 2, 3, 4). Let's look at the hierarchy for, for example, the spam tag; a tweet that has been tagged as spam is least likely to be predicted as written by a Hispanic author and most likely to be predicted as written by an Asian author. The mathematical expressions for each label are, where P(Race) indicates the probability that a tweet will be predicted to be a specific racial identity:

$$Spam: P(Hispanic) < P(Black) < P(White) < P(Asian)$$
$$Hateful: P(Asian) < P(White) \ ? \ P(Hispanic) < P(Black)$$
$$Abusive: P(Asian) < P(White) < P(Black) < P(Hispanic)$$

There are a few interesting things to note here. First, almost every relationship is statistically significant: except for the probability of White versus Hispanic prediction based on the hateful label (represented by a question mark), every hierarchical statement above is (extremely) statistically significant. But even more so, these results are interesting in the way that they reinforce and speak to societal inequities.

The spam tag is defined by Founta as "posts consisted of related or unrelated advertising/ marketing, selling products of adult nature, linking to malicious websites, phishing attempts and other kinds of unwanted information, usually executed

repeatedly" [24]. This tag is most strongly connected with the prediction of Asian authorship by comparison to the other racial groups and is least likely to be attributed to Hispanic authorship. The idea that machine learning algorithms attribute what humans consider to be spam to Asian Twitter users more often is representative of a flaw in our racialized model for linguistics.

There are notable differences between the abusive and hateful labels, however we must again consider the subjectivity of the labelling scheme. The abusive label is applied more generally, described as "any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion" [25]. By comparison, the hateful label is applied specifically to hate speech, defined by Founta and colleagues to be "Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender" [26]. This specific distinction is difficult to apply uniformly, as what one individual perceives as generally abusive another may view as specifically hateful. As such, we expected the outcomes for hateful and abusive language to be largely similar. Looking at the hateful tag and how it relates to predictions for different racialized groups, we see that it has the strongest association with Black-predicted language and the weakest association with Asian-predicted language. Again, without explicating various stereotypes that are harmful to different non-dominant ethnic groups, the disproportionate treatment of some racial groups compared to others is alarming.

Finally, looking at the more general abusive tag, we see a similar pattern to the hateful tag. The most notable difference between the two is where the Hispanic-predicted authors lay in comparison to others; while Hispanic-predicted authors fall in the middle on probability correlated to hateful language, they are even more likely than Black-predicted authors to be correlated with the abusive language tag. We won't waste energy laying academic credence to discriminatory stereotypes that disparage people of color, but the fact that Natural Language Prediction models predict that abusive and hateful language comes more often from underrepresented minority groups is particularly troubling.

Beyond the main effect of race that was clearly notable, there were a few interaction terms that were statistically significant, and we should discuss those few notable differentiations. Because the gender terms were inconsistently (and only barely, if at all) significant across different baseline groups, there is not much interesting to conclude from these data; the interaction terms, however, have yielded a few interesting differences that are worth noting. We've represented the results of the interaction terms with significant statistical values in the table below.

Table 4. Race/Gender Interactions with Tags

Relationship	Tag	Significance	Statistic Comparison
Asian Woman/ Black Man	Abusive	*	Z(BM) > Z(AW)
Asian Woman/ Hispanic Man	Abusive	**	Z(HM) > Z(AW)
Asian Woman/ White Man	Abusive	***	Z(WM) > Z(AW)
White Woman/ Asian Man	Abusive	***	Z(WW) > Z(AM)
Black Woman/ Asian Man	Abusive	*	Z(BW) > Z(AM)
Hispanic Woman/ Asian Man	Abusive	**	Z(HW) > Z(AM)

Key: *** > 0.001, ** > 0.01, * > 0.05

There are several interesting patterns of note shown in Table 4. First, intersectional identity is only significant in the context of the abusive tag, not the hateful or spam tags. Next, we notice that each relationship of statistical significance is between an Asian woman and a non-Asian man or between an Asian man and non-Asian woman. We notice that the Z statistic for each of our non-Asian intersectional identities is greater than our Asian intersectional identities when compared on a one-to-one basis. When we try to understand what to take away from Table 4, we see that all statistically significant race/gender interactions have some common threads and are only predictive when it comes to the label of "abusive".

In summary:

- Race has a meaningful correlation with each of the three tags of interest, and we can form a hierarchy of probability to describe the likelihood of a flagged tweet being attributed by NLP to a certain racial group.
- Except for the relationship between Hispanic-predicted and White-predicted tweets with a hateful tag, all racial differentiations were statistically significant.
- When it comes to abusive tweets specifically, there is not only a racial effect but an intersectional effect when considering Asian-predicted speech.

- Without getting into the weeds of harmful stereotypes, these differences in attribution to certain lexical dictionaries can be troubling.

So, what can we do to combat this attribution of certain negative characteristics and tags on tweets to minority groups? For that discussion, we turn to the question of policy implications and conclusions from this research.

5. Conclusions

Our findings suggest that there are inherent biases within either the natural language processing algorithms or within the training data used to identify abnormal language. While our method of analysis does not allow us to draw conclusions about where this bias is coming from, bias is being reproduced within the construction of the dataset that we have created. The association of white language with normality and non-white language with abuse of the Twitter platform may be driven by either the attitudes of the crowdsourced data creators or the predictive dictionaries used to identify race and gender (or, likely, some combination of the two). While we cannot conclude where this bias is coming from, bias is present in our existing systems of identification. This bias can lead to problematic associations that drive the narrative that whiteness is the norm, and anything outside of whiteness is a deviation from standards.

What do these analyses show on a broader scale? We cannot parse out the biases of the algorithm's creators and trainers compared to real world differences in behavior, but both the biases and real-world differences likely impact the outcomes we've seen in this study. The directionality of real-world differences would be hard to evaluate given that the models we use are likely clouded with bias created by those who constructed them. While we don't want to explicate the harmful stereotypes that our results would reinforce, there are certainly stereotypes at play in the results that we've seen, particularly for non-White folks. In the results we've seen, tweets that are flagged as normal (or rather, are not flagged) are most likely to be associated with White linguistic traits; is that because dominant American society demands whiteness as a precursor to normality? We have many more questions than answers here, but the results presented today are an interesting step in identifying biases in NLP predictions and in the flagging that goes into creating training sets like Founta's.

The main conclusions here are not inherently intersectional, and do not speak directly to the intersectional oppression. As stated in the beginning of this paper, statistical analyses maximizing singular effects minimize their interaction terms: according to Bowleg, "when significant main effects exist, the probability of finding significant first order (a two-way interaction) or higher order interactions (three, four, and *n*-way interactions) decreases because the significant main effects account for the bulk of the variance in the dependent variable" [27]. Perhaps that is why some of our intersectional conclusions were limited. Even so, some interesting intersectional conclusions regarding Asian-identified speech were statistically significant as well, and that's worth thinking about going forward.

For future research, Twitter and NLP are only one piece of the puzzle that make up algorithm development and intersectional discrimination. In a general sense, intersectional data collection is woefully lacking. The challenge of potentially misattributed data based only on prediction would be mitigated if more information about identity were collected in studies for the purpose of addressing identity effects rather than as a confound to be filtered out. More specific to our results, our analysis is correlational and thus does not allow for any causal conclusions, we would recommend future directions of research focus on the development of algorithms and evaluating their bias from the construction period. Studying the people that make algorithms and training data is a future interest of our own research, and subsequently we believe that funding this type of research would yield interesting and meaningful analyses. Only when we commit to truly and accurately depicting the inequities in this country, and investigate the algorithms that drive that inequity, can we start to combat the dangers of the "black box" of computing and build algorithms that result in fair and equitable outcomes.

References

[1] Crenshaw, K. (2015). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, *1989*(1). https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

[2] Hancock, A.-M. (2007). Intersectionality as a Normative and Empirical Paradigm. *Politics & Gender*, *3*(2), 248–254. https://doi.org/10.1017/S1743923X07000062

[3] Weber, L. & Parra-Medina, D. (2003). "Intersectionality and Women's Health: Charting a Path to Eliminating Health Disparities" *in Gender Perspectives on Health and Medicine*. Advances in Gender Research, vol. 7, pp. 204.

[4] Warner, L. R. (2008). A Best Practices Guide to Intersectional Approaches in Psychological Research. *Sex Roles*, *59*(5), 454–463. https://doi.org/10.1007/s11199-008-9504-5

[5] Hancock, A.-M. (2007). Intersectionality as a Normative and Empirical Paradigm. *Politics & Gender*, *3*(2), 248–254. https://doi.org/10.1017/S1743923X07000062
[6] Bauer, G. R. (2014). Incorporating intersectionality theory into population health research methodology: Challenges and the potential to advance health equity. *Social Science & Medicine*, *110*, 10–17. https://doi.org/10.1016/j.socscimed.2014.03.022
[7] Bowleg, L., & Bauer, G. (2016). Invited Reflection: Quantifying Intersectionality. *Psychology of Women Quarterly*, *40*(3), 337. https://doi.org/10.1177/0361684316654282
[8] McCall, L. (2005). The Complexity of Intersectionality. *Signs: Journal of Women in Culture and Society*, *30*(3), 1771–1800. https://doi.org/10.1086/426800
[9] Bambas, L. (2005). Integrating Equity into Health Information Systems: A Human Rights Approach to Health and Information. *PLOS Medicine*, *2*(4), e102. https://doi.org/10.1371/journal.pmed.0020102
[10] Lord, S. M., Camacho, M. M., Layton, R. A., Long, R. A., Ohland, M. W., & Wasburn, M. H. (2009). "Who's Persisting in Engineering? A Comparative Analysis of Female and Male Asian, Black, Hispanic, Native American and White Students". *Journal of Women and Minorities in Science and Engineering*, *15*(2). https://doi.org/10.1615/JWomenMinorScienEng.v15.i2.40
[11] Agénor, M., Krieger, N., Austin, S. B., Haneuse, S., & Gottlieb, B. R. (2014). At the intersection of sexual orientation, race/ethnicity, and cervical cancer screening: Assessing Pap test use disparities by sex of sexual partners among black, Latina, and white U.S. women. *Social Science & Medicine*, *116*, 110–118. https://doi.org/10.1016/j.socscimed.2014.06.039
[12] Cairney, J., Veldhuizen, S., Vigod, S., Streiner, D. L., Wade, T. J., & Kurdyak, P. (2014). Exploring the social determinants of mental health service use using intersectionality theory and CART analysis. *J Epidemiol Community Health*, *68*(2), 145–150. https://doi.org/10.1136/jech-2013-203120
[13] Evans, C. R., Williams, D. R., Onnela, J.-P., & Subramanian, S. V. (2018). A multilevel approach to modeling health inequalities at the intersection of multiple social identities. *Social Science & Medicine*, *203*, 64–73. https://doi.org/10.1016/j.socscimed.2017.11.011
[14] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342
[15] Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2011). Understanding the Demographics of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), Article 1. https://ojs.aaai.org/index.php/ICWSM/article/view/14168
[16] Messias, J., Vikatos, P., & Benevenuto, F. (2017). White, man, and highly followed: Gender and race inequalities in Twitter. *Proceedings of the International Conference on Web Intelligence*, 266–274.
[17] Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 1–15. https://doi.org/10.1140/epjds/s13688-016-0072-6
[18] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1), Article 1. https://ojs.aaai.org/index.php/ICWSM/article/view/14991
[19] Hube, C., Fetahu, B., and Gadiraju, U. (2019). Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* Paper 407, 1–12. https://doi.org/10.1145/3290605.3300637
[20] Ibid.
[21] Barbosa, N. and Chen, M. (2019). Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* Paper 543, 1–12. https://doi.org/10.1145/3290605.3300773
[22] Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., & Schwartz, H. A. (2014). Developing Age and Gender Predictive Lexica over Social Media. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1146–1151. https://doi.org/10.3115/v1/D14-1121
[23] Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic Dialectal Variation in Social Media: A Case Study of African-American English. *ArXiv:1608.08868 [Cs]*. http://arxiv.org/abs/1608.08868
[24] Founta, et. al. (2018), 495.
[25] Ibid.
[26] Ibid.
[27] Bowleg, L. (2008). When Black + Lesbian + Woman ≠ Black Lesbian Woman: The Methodological Challenges of Qualitative and Quantitative Intersectionality Research. *Sex Roles*, *59*(5), 319. https://doi.org/10.1007/s11199-008-9400-z