Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?

Savannah Larimore
Washington University in St. Louis

Ian Kennedy
University of Washington

Breon Haskett
University of Washington

Alina Arseniev-Koehler
University of California - Los Angeles

Abstract

An abundance of methodological work aims to detect hateful and racist language in text. However, these tools are hampered by problems like low annotator agreement and remain largely disconnected from theoretical work on race and racism in the social sciences. Using annotations of 5188 tweets from 291 annotators, we investigate how annotator perceptions of racism in tweets vary by annotator racial identity and two text features of the tweets: relevant keywords and latent topics identified through structural topic modeling. We provide a descriptive summary of our data and estimate a series of linear models to determine if annotator racial identity and our 12 topics, alone or in combination, explain the way racial sentiment was annotated, net of relevant annotator characteristics and tweet features. Our results show that White and non-White annotators exhibit significant differences in ratings when reading tweets with high prevalence of certain racially-charged topics. We conclude by suggesting how future methodological work can draw on our results and further incorporate social science theory into analyses.

1 Introduction

Hateful and racist language is abundant on social media platforms like Twitter and a growing body of work aims to develop tools to detect such language in these spaces. Such a tool would offer opportunities to intervene, like providing automatic trigger warnings, and would provide a powerful barometer to measure racism. However, these efforts are hampered by low interrater agreement, low modeling performance, and a lack of consensus on what counts as racist language (e.g., Kwok and Wang, 2013; Burnap and Williams, 2016; Waseem, 2016; Schmidt and Wiegand, 2017). These efforts are also largely disconnected from rich understandings of race and racism in the social sciences (but, see Waseem, 2016). Indeed, social scientists have long acknowledged the difficulties of measuring racism, even when using traditional social science methods (e.g., interviews and surveys), due to social desirability biases and the increasingly covert nature of racism (Bonilla-Silva, 2006).

In this paper, we reconsider efforts to annotate for racism in light of sociological work on race and racism. Instead of generalizing our detection of racism, we narrow our scope to focus on anti-Black racism. In other words, we focus on racialized language directed at or centering on Black Americans. Using human annotations[1] for the racial sentiment (positive or negative) of 5188 Tweets, we describe how ratings vary by annotators' own racial identity. Our findings suggest that White raters respond differently to particular types of racialized language on Twitter, as identified by structural topic modeling (STM), than non-White raters. Failing to account for this systematic difference could lead researchers to consider tweets which non-White annotators identify as including negative sentiment as innocuous because more numerous White annotators rate those same tweets as positive or neutral. We conclude by suggesting several ways in which future work can account for the variability in annotations that comes from annotators' own racial identities.

1.1 Annotating for Racism

Collecting annotations is a key step to developing a tool to detect racial sentiment in text data. At the same time, the challenges of this task have been well-documented as ratings depend upon the ability of human annotators to consistently identify the racial sentiment of words and phrases (Zou and Schiebinger, 2018). Empirical work often finds large amounts of disagreement in these annotations, even with carefully designed annotations

[1] We use the terms rating/annotation and rater/annotator interchangeably throughout.

Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pages 81–90
Online Workshop, June 10, 2021. ©2021 Association for Computational Linguistics

schemes (e.g., Bartlett et al., 2014; Schmidt and Wiegand, 2017).

As these efforts have shown, perceptions of racial sentiment are contextual and subjective, making the prevalence of racism in text sources inherently difficult to detect. Recent social scientific work (Bonilla-Silva, 2006; Carter and Murphy, 2015; Krysan and Moberg, 2016; Tynes and Markoe, 2010) has taken that difficulty as its subject, and sought to capture and understand the variation in perceptions of racism or racial sentiment, by showing how individuals' perceptions of racism vary systematically based on the their *own* racial identity. These findings suggest that annotator disagreement is not merely noise to smooth over. Rather, annotator disagreement for racism includes important variation that should be disaggregated and accounted for.

1.2 Varying Perceptions of Racism

Differential attitudes about and perceptions of racism based on an individuals' own racial identity are well-documented. White Americans tend to hold more overtly racist beliefs, are less likely to believe racial discrimination is prevalent in modern society, and are less likely to recognize racial microaggressions than Black Americans (Bonilla-Silva, 2006; Carter and Murphy, 2015; Krysan and Moberg, 2016; Tynes and Markoe, 2010). In addition, Krysan and Moberg (2016) note that White Americans increasingly disregard racial topics on questionnaires, signaling that they have "no interest" in issues of racial inequality. Likewise, fewer White Americans agree that racial inequality is due to overt discrimination, arguing instead that racial discrimination is a thing of the past and that in contemporary society, everyone has equally fair life chances (Bonilla-Silva, 2006). As Carter and Murphy (2015) note, White and Black Americans may differ in their views of racial inequality because White Americans compare contemporary racial inequalities to the past, referencing slavery and Jim Crow and naturally concluding that conditions have improved, while Black Americans compare the present to an imagined future, in which an ideal state of racial equality has been achieved.

These differences also extend to how we perceive racism in online platforms. Williams et al. (2016) find that while White students were equally as likely as students of color to perceive racially-themed internet memes as offensive, students of color tended to rate these same memes as more offensive than White students. That is, while White students could identify that a meme was racist, they rated the level of offensiveness lower than students of color. In addition, Tynes and Markoe (2010) find that European American college students were less likely than African American college students to react negatively to racially-themed party images on social media. Furthermore, European American students reported higher levels of color-blind racial attitudes and students with lower levels of color-blind attitudes were more likely to react as "bothered" by the images, implying that both race and racial attitudes influence perceptions of racism online. Similarly, Daniels (2009) finds that critical race literacy matters more than internet literacy in identifying racially biased or "cloaked" websites (i.e., websites that appear to promote racial justice but are actually maintained by White supremacist organizations). This finding suggests that students who lack a critical race consciousness may be less likely to identify racist materials online and that White students may be particularly susceptible.

The subtlety of racism that pervades social media sites like Twitter may also influence perceptions of racism. As Carter and Murphy (2015) note, Whites tend to focus only on blatant, overt forms of racism (e.g., racial slurs, mentions of racial violence) but are less attuned to microaggressions and other, subtler forms of racism. As such, scholars have also advocated for a methodological move away from "bag of words" approaches to the evaluation of racism on social media (Watanabe et al., 2018) because these approaches reinforce a focus on blatant, overt forms of racism, and neglect more subtle, or contextually racist tweets (Chaudhry, 2015).

Similarly, Kwok and Wang (2013), noting the subtlety of racism pervading social media posts, argue that to get evaluations of tweets that accurately assesses meaning, features of tweets other than the text must be included. Tweet features set the rules of engagement by offering markers of credibility, sarcasm, and persuasiveness (Sharma and Brooker, 2016; Hamshaw et al., 2018). Tweet features such as links, hashtags, and number of comments have been shown to illuminate the context of the tweet's message (Burnap and Williams, 2016). The inclusion of these features in evaluation offers deeper context and more realistic evaluation of tweets allowing for greater attention to the differential evaluations of people in racially marginalized groups engaging with the social media platform.

Here, we expand on previous research by investigating how annotator racial identity and tweet features interact to influence perceptions of racism on Twitter. Our analysis builds on previous research that uses racist speech as a stimulus (Leets, 2001; Cowan and Hodge, 1996; Tynes and Markoe, 2010), either in print or digital media, and calls for renewed attention to variations based on annotator racial identity and how these variations ultimately influence instruments to measure racial sentiment. We extend this body of work by including potentially racist speech sourced randomly from Twitter, rather than developed by researchers, and by including tweet features as well as annotator racial identity in our analyses.

2 Hypotheses

Based on previous work in annotation and a sociologically informed understanding of race and racism, we propose three hypotheses:

H_1: Annotations will vary, on average, based on the racial identity of the annotator: White raters will rate tweets as having a more positive racial sentiment on average compared to non-White raters.

H_2: Annotations of racial sentiment will vary by the racially charged keywords in the tweet and the latent topics in the tweet. Tweets with racialized keywords (e.g., N****r) will be rated as more negative than those without.

H_3: Annotations will vary based on the interaction of text features (racially charged keywords and latent topics) and the racial identity of the annotator: Compared to non-White raters, White raters will interpret particular topics as having a more positive racial sentiment, and interpret other topics as having a more negative racial sentiment.

3 Methods

3.1 Data

We combine data from two separate research projects, producing a final sample of 291 human raters applied to 5188 unique tweets. The first project collected 1348 tweets from Twitter's Streaming API from June 2018 to September 2019. The second project collected 3840 tweets from the Digital Online Life and You Project (DOLLY; a repository all geotagged tweets since December 2011) (Morcos et al.). For both projects, tweets were restricted to those that were sent from the contiguous United States, were written in English, and contained at least one keyword relevant to the analysis. To limit our sample to tweets that concerned Black Americans, we used common hate speech terms, the term "black girl magic", and the same keywords to identify tweets about Black Americans as Flores (2017).

For the second project, tweets were also restricted to a 10% sample of all tweets sent from 19 metropolitan statistical areas between January 2012 and December 2016. While both data collection processes yielded millions of unique tweets, both projects sampled several thousand tweets for annotation based on available funding to compensate annotators or access to undergraduate classrooms (for a similar methodology, see Burnap and Williams, 2016). We then collected human annotations using Amazon Mechanical Turks and college students from two separate classrooms at the same university, for a total on 291 annotators. All annotators were instructed to use the same annotation tool, were provided with brief training, and we applied a coding structure such that each unique tweet was rated by at least 5 annotators.

Annotators also reported their race and gender identities. Race was reported using the following categories: White/Caucasian/European, Black/African American, Asian/Pacific Islander, Latino/Hispanic/Spanish, and Other. Annotators were allowed to select more than one race. For the current analyses and due to sample size restrictions, we collapsed the annotator race into two categories: 1) Non-Hispanic White/Caucasian/European alone (henceforth, "White") and 2) All other racial classifications (henceforth, "non-White"). Gender was reported as woman, man, or other gender.

A total of 52.23% of our 291 raters identified as White, and 46.05% identified as non-White and 1.72% were missing race, respectively. A total of 38.49% of our raters identified as women, 58.73% identified as men, and 1.37% were missing gender. On average, our raters were 27.45 years. Given these characteristics, our raters are similar to Twitter users in regard to age and gender, but are perhaps more likely to identify as White (Perrin and Anderson, 2019).

3.2 Variables

Our outcome variable is a continuous measure for racial sentiment, which we operationalize as how "positively" or "negatively" a tweet was rated. Raters used a 7-point Likert scale to describe the sentiment of the tweet, ranging from "very negative" (i.e., -3) to "very positive" (i.e., +3), with a "neutral" rating at the center of the scale (i.e., 0).

Our key independent variables are two text features of tweets (relevant keywords and latent topics) and the racial identity of annotators. Relevant keywords were inductively identified by the research team from a close reading of the tweets. Keywords of theoretical significance (e.g., mentions of racialized violence; animal epithets) were also considered. This process yielded 8 groups of keywords: 1) keywords with allusions to sex or sexuality (e.g., "sex") 2) keywords about people (e.g., "ppl") 3) animal epithets 4) spelling variations of N***a 5) spelling variations of N***er 6) derogatory words towards women (e.g,. "B***h") 7) spelling variations of F**k 8) keywords about racialized violence (e.g., lyn**). Each of these 8 groups of keywords were treated as a binary variable (1 = any keyword in the group is present in the tweet).

Latent topics were identified using the STM package in R and we used STM's built-in methods to select a model with 12 topics (Roberts et al., 2019). We labeled each topic by 1) examining the words with the highest probability of being generated by the topic 2) examining the top words for the topic based on STM's "FR-EX" measure that uses word frequency and exclusivity within a topic (Roberts et al., 2014), and 3) reading the 20 tweets which have the highest loading onto the topic. Topics were treated as continuous (i.e., the "amount" of a topic in a given tweet). Racial identity was measured as White or non-White, as described previously.

We include several covariates in our models. First, we control for annotator gender identity (woman/man/other) and age (years). Second, we include binary indicators for the following tweet features: if a URL link is present and if there is a mention included, indicating a conversation between users. Third, we control for the length of the tweet, measured in characters.

3.3 Analysis

Our analysis proceeds in three steps. First, we provide descriptive summaries of our data. We summarize our STM results, and provide Krippendorf's alpha coefficients (Krippendorff, 1980) to assess inter-rater reliability for all raters, for White raters, and for non-White raters.

Second, we estimate three linear models, each respectively testing our three hypotheses. Model 1, the "Annotator Race" model, regresses racial sentiment on a binary indicator for annotator racial identity (1 = White). Model 2, the "Text Features" model, regreses racial sentiment on binary indicators for relevant keywords and continuous measures for topics (described in Variables). Model 3, the "Interaction" model, regresses racial sentiment on three interaction terms: one for each statistically significant, theoretically-informed topics identified in Model 1 (i.e., topics 2, 5, and 9), interacted with annotator racial identity. As such, Model 3 treats annotator racial identity as an effect modification variable in the analysis because we expect that annotator racial identity will modify ratings of tweets based on salient topics.

For Models 2 and 3, we include all covariates (described in our Variables section). For Model 3, we additionally control for all keywords and topics included in Model 1. Using the results from Model 3, we also compute predicted racial sentiment ratings based on the amount of a topic in a given tweet and the racial identity of an annotator. We visualize regression coefficients and 95% confidence intervals in Figure 1 for all three models, and we visualize selected results on predicted sentiment ratings by amount of topic in Figures 2-4. Analyses were performed in R (R Core Team, 2017) using α= 0.05 for statistical significance.

Third, we qualitatively describe annotated tweets that illustrate the results of our interaction models based on proportion of a given topic and a high degree of inter-rater disagreement between White and non-White raters.[2]

[2] Specifically, for a given topic, we first selected the tweet with the most amount of a topic by percentile x > 90%. Second, narrowed this candidate list of tweets to the 20 with most disagreement between White and Non-White raters. Finally, among this short list of 20 tweets, we selected the tweet best reflected the differences shown numerically in Figures 2, 3, and 4.

4 Results

4.1 Descriptive Analysis

Table 1 provides a summary of our 12 topics. This summary includes topic labels, the seven most representative words (by "FR-EX" as described earlier), and the tweet which loads most highly onto each topic. As might be expected from our corpus, many of the 12 topics are relevant to race, such as a topic we label "Racial Arguments" and a topic we label "Police Brutality." To be clear, these topics may be mentioned in positive, neutral, or negative lights. For example, a tweet containing the topic "Police Brutality" might be reporting on successful efforts to minimize police brutality. We refer to the topic by the numeric ID it was assigned by STM and the title we assigned the topic.

Overall agreement on racial sentiment was low among raters (Krippendorf alpha = 0.39), but higher within White raters (Krippendorf alpha = 0.44). Among non-White raters, agreement was also low (Krippendorf alpha = 0.34). This low agreement echoes prior work on the challenges of annotating racially charged language (e.g., Bartlett et al., 2014; Schmidt and Wiegand, 2017), as described earlier. Importantly, the goal of this study was *not* to arrive at annotations with high agreement (i.e., for training a predictive model); instead our goal was to examine patterns of agreement and disagreement in annotations.

4.2 Regression Analysis

The results of our regression analysis are shown in Figure 1. We present the results for each consecutive model, showing the main effects for annotator racial identity and text features in Models 1 and 2, respectively, before turning to the interaction terms in Model 3. We do so to highlight changes across models as predictors are introduced and to confirm effect moderation but note that constitutive terms for the interactions in Model 3 should not be interpreted as unconditional marginal effects (Brambor et al., 2006).

In H_1, we expected a difference in average racial sentiment rating between White and non-White raters. Using Model 1, we find that the association between annotator racial identity (as White) and sentiment is positive and statistically significant, but small (β=.071, p<.001).[3] This suggests that while, on average, White raters tend to rate racial sentiment of tweets as higher than do non-White raters, this difference may not influence our annotations in a substantial way. Thus, we conclude that we find limited support for this hypothesis.

In H_2, we expected that text features of the tweet would significantly influence sentiment ratings. Using Model 2, we find strong support for this hypothesis: all of our topics and keyword features are significantly associated with sentiment rating (see Figure 1). This result confirms the intuition that raters are responding to a variety of racially coded language as they make their annotations. Notably, we also observe that the effect of these text features on raters' sentiment is far greater than the main effect of racial identity.

In H_3, we expected that the difference between White and non-White raters' ratings depends on how much of a topic was present in a tweet. We tested this using Model 3, where we interacted topics and rater racial identity. We find that the interaction terms for seven of our topics are significantly associated with racial sentiment (see figure 1), providing strong support for H_3.

We illustrate several of these interaction effects more directly in Figures 2-4 for three of the topics (Topic 2: Police Brutality, Topic 5: Empowering History, and Topic 9: Antiracist Politics). These figures show the *expected* racial sentiment rating of a tweet in our model against how much the tweet loads onto a given topic, among White and non-White raters.[4] The x-axis of each plot ranges from the 1st percentile of topic values in our data to the 100th percentile. These figures show that for the topics with significant interactions, substantial differences by annotator race arise when certain topics are very prevalent in the tweet. Thus, while Model 1 suggests that White and non-White raters have small differences in rating across *all* tweets, Model 3 shows that for tweets about *certain* topics, White and non-White raters in fact rate tweets quite differently. We examine examples of this di-

[3] We note that similar conclusions may be reached with Model 2, where the coefficient for annotator racial identity is also significant but small (β=.042, p<.01).

[4] Because the vector of topics in an STM sum to one, increasing one topic implies decreasing others, which are also included in our models and therefore will also influence the estimated rating. Moreover, the topics are often correlated, so just decreasing the other topics evenly would produce a potentially misleading result. For these plots, as we increased our focal topic–Topic 2 in Figure 2–we adjusted the other topics based on their average in our data for that value of Topic 2. That means that the plots reflect the estimated tweet rating for White and non-White raters for various values of Topic 2 and the average values of other topics at those values of Topic 2, with other covariates held at their means.

Table 1. Topic Titles and Top Words (by "FR-EX")

Topic Title	*Top Words*	*Example Tweet*
Topic 1: Breaking Stereotypes	chick, til, retail, wut, firework, camp, gramma	People salty cause they never seen a black guy work at PacSun before
Topic 2: Police Brutality	man, teacher, fool, nicki, pride, doctor, histor	Lorenzo Clerkley, a 14 year old black kid who was with friends playing with a BB gun in broad daylight was shot 4 times by an officer after being given 0.6 second warnings
Topic 3: Racial Arguments	shit, use, word, poor, stupid, respect, mexican	I'm sorry to intrude on this but that's kind of a screwed up concept to say that white people are inherently racist? Races and ethnicities of all kinds have conquered and enslaved others throughout mankind but only one group gets it all pinned onto them?
Topic 4: Black Women and Girls	get, let, twitter, amaz, els, seem, bro	In 1968, Shirley Chisolm (1924-2005) was the 1st Af-Am woman elected to Congress (D-NY). In 1972, she was the 1st woman to seek the Democratic nomination for POTUS. A staunch women's rights activist, she delivered this speech in support of the ERA in 1970.
Topic 5: Empowering History	girl, king, magic, martin, luther, celebr, sell	Pioneers : African American surgeon, Daniel Hale Williams, opened the first interracial Hospital in Chicago in 1891 and performed the first documented open-heart surgery in 1893.
Topic 6: Monkey	atwitterhandl, your, aint, shes, season, theyr	atwitterhandle atwitterhandle some of his stuff is alright I guess but overall I cant stand that cheeto dread monkey
Topic 7: Empowering Information	african, american, definit, murder, dog, student, leader	#FridayFeeling new #exhibit open at #DunnMuseum features local #AfricanAmerican history of Booker T. Washington Progressive Club. Open January 11 - February 24.
Topic 8: Irreverent Interactions	fuck, back, big, turn, damn, wtf, light	Driving on the highway past the big black dude who was grinding up weed while driving the kids bouncy house truckgtgtgtgt you do you dawg
Topic 9: Antiracist Politics	like, look, color, start, lot, run, everyth	Associates with white nationalists and open bigots. He's the architect of the Muslim ban and cruel policies that separate children from families at the border. If it looks like a duck, walks like a duck, quacks like a duck - it's a duck.
Topic 10: Black the Color	awebsit, shop, doesnt, widow, coffe, disgrac, hot	Drinking a Catskill Mountain Black IPA by Gilded Otter Gilded Otter Brewing Company awebsite photo
Topic 11: Debates about Race and Racism	mentionplacehold, minor, bait, control, societi, kkk, negro	[50 mentions] And that's a bullshit statement you know I really don't know why you white leftist hate your own race so much ... what you just said is no different than saying black people can't be racist
Topic 12: Honest Opinions	world, next, month, learn, post, honest, number	I'm a 41 year old African American born in Minneapolis and have lived close to Seattle my entire life. Everytime I hear this manufactured crisis by the media I change the channel. If the media fails us again/2016 this country will never recover,?media will never be trusted again.

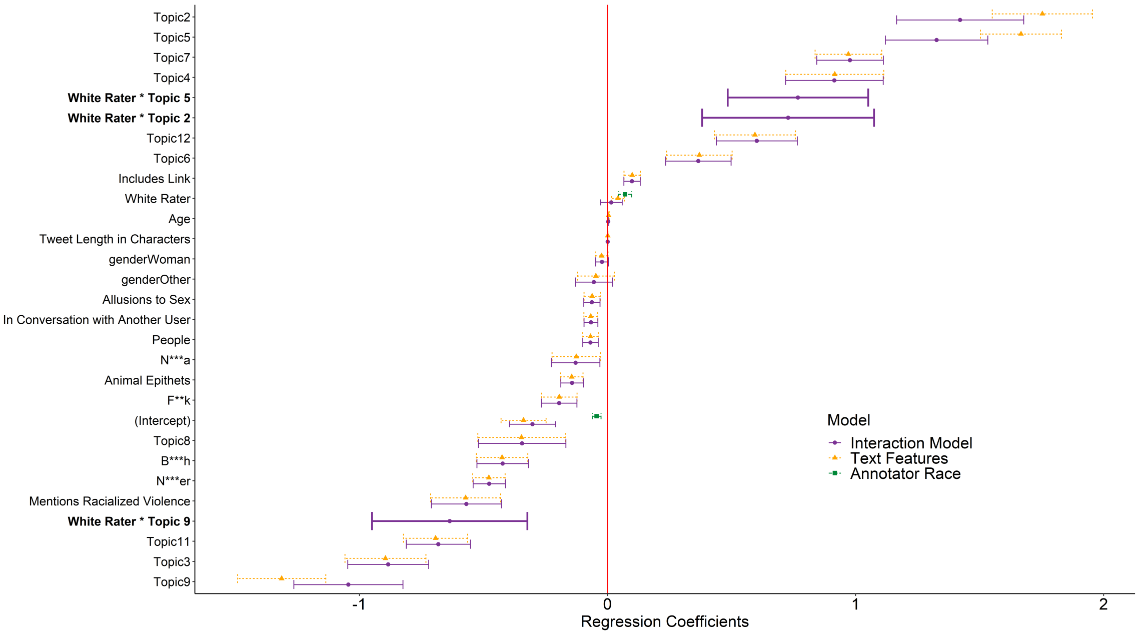
vergence in our qualitative analysis.

4.3 Qualitative Analysis

To provide qualitative examples of our findings, we identify exemplary tweets in Table 2 from each of the three topics displayed in Figures 2-4. For topic 2: Police Brutality, we find that White raters considered this tweet "moderately positive", with an average racial sentiment rating of 2.0. In contrast, non-White raters considered this tweet closer to neutral, with an average racial sentiment rating of 0.4. This difference is particularly striking, as this tweet makes reference to wrongful detention, something that sociological and computational research would suggest White raters would also consider negative. While this tweet suggests an attempt to make amends within the news story presented in the tweet, the perception of the racial sentiment is quite different across raters.

For Topic 5: Empowering History, we find White raters consider this exemplary tweet "moderately positive", while non-White raters consider this tweet "neutral", on average. This tweet quotes Dr. Martin Luther King Jr. and suggests that references to historical figures may signal different things for White and non-White raters. Further, the connection to a specific Christian observance of Lent signals little racialized content for some. For Topic 9: Anti-racist politics, we find that White raters consider this exemplary tweet "neutral" while non-White raters consider it "moderately positive". This implies that the White raters who viewed this tweet may not have considered anti-racist work to have a positive racial sentiment.

Figure 1: Regression Coefficients and 95% Confidence Intervals For Three Regression Models



Note: The "Annotator Race" model regresses racial sentiment on annotator racial identity. The "Text Features" model regresses racial sentiment on annotator racial identity, racially charged keywords, latent topics and covariates. The "Interaction" model regresses racial sentiment on all terms in Model 2 as well as interaction terms between selected topics and annotator racial identity.

Figure 2: Topic 2 Estimated Tweet Rating Including Annotator Race Interaction

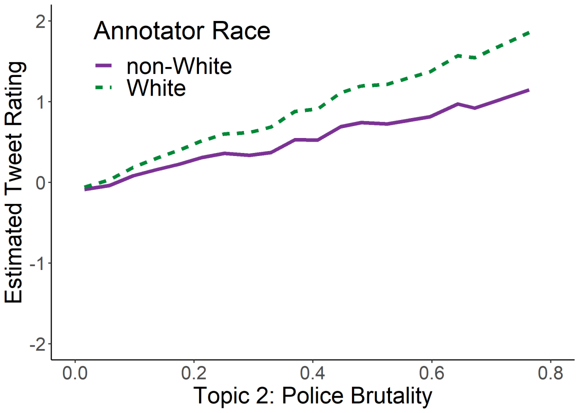
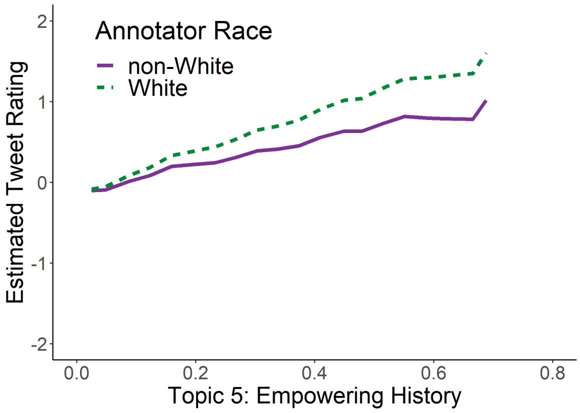


Figure 3: Topic 5 Estimated Tweet Rating Including Annotator Race Interaction



In contrast to our results from H_1, which show that the average difference in sentiment rating by annotator race is small, these qualitative results add further evidence to support H_3: that raters of different racial identities interpret topics differently. These qualitative examples illustrate that differences between raters are not just statistically significant but also practically meaningful.

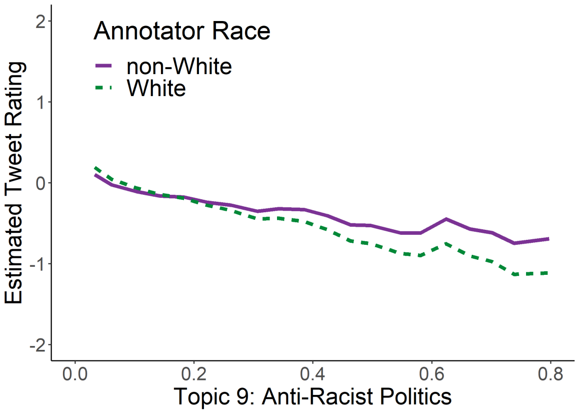
5 Conclusions

The goal of our analysis was to determine if and how annotator racial identity influenced perceptions of the racial sentiment of a tweet regarding Black Americans. When we examine the mean difference between White and non-White annotators (Model 1), we find a small but significant difference in sentiment ratings. When we consider White and non-White raters responses to different amounts of topics in the tweets (Model 3) we find strong evidence that annotator racial identity *does* inform perceptions of sentiment towards Black people for seven of our twelve topics. Our results suggest that White and non-White raters interpret these seven topics differently and as these

Table 2. Exemplary Tweets of Interaction Model, by Latent Topic

Topic Title	***White Raters***	***Non-White Raters***	***Exemplary Tweet***
Topic 2: Police Brutality	2.0	0.4	The meeting is in response to a incident earlier this month in which an African American man was detained shortly by police while cleaning outside his home in Boulder.
Topic 5: Empowering History	2	0	Forgiveness is not an occasional act it is a permanent attitude Dr Martin Luther King JrHow about this for Lent
Topic 9: Antiracist Politics	0	2	Sounds like good is locked in battle with perfect. I am a white person trying to fight white supremacy, and I will never not be flawed. I don't need your cookie, but it would be nice not to take friendly fire.

Figure 4: Topic 9 Estimated Tweet Rating Including Annotator Race Interaction



topics increase in tweets (Figures 2-4), this gap in interpretation widens.

Notably, the topics we identify as most divisive are some of the very topics which social scientists may be most interested in analyzing: references to police brutality, references to historical figures or events, and discussions of anti-racist politics. Our descriptive qualitative analysis suggests that White annotators may not be as attuned to the nuances of these topics in tweets. Future work might expand on these results to investigate raters' rationale for their ratings.

Given that perceptions of racism vary by annotator's own race, it is crucial that future work considers whose interpretations are reflected in annotations for racism. Indeed, annotators' interpretations end up being a gold standard from which models learn to detect what counts as racist or not. While we focus on the role of annotators' racial identities, many other dimensions of annotators' identities likely also influence their responses on annotation tasks more generally. This issue extends beyond annotation tasks: across the disciplines, there is growing recognition that much of the social scientific knowledge produced to-date is specific to the population from which we most often draw participants (Henrich et al., 2010).

We suggest several takeaways for future research. First, researchers should use purposeful sampling (Palinkas et al., 2015) to gather annotations from diverse populations of annotators. This may be challenging given that platforms for collecting annotations may include a particular demographic of workers. In particular, young, white, and well-educated workers are over-represented on MTurk (Hitlin, 2016). Second, research using human annotation might collect (and report) annotator demographics, in order to be explicit about whose interpretations the annotations do (or do not) reflect. Third, given the many possible identities, researchers might consider several possible strategies to focus on a particular demographic of annotators. Researchers might focus on the populations for whom the gold standard is most important, or might be most divisive. We suggest that the gold standards for racist language should reflect the interpretations of who is impacted most the standard. For example, annotations for anti-Black racism should ideally reflect how Black individuals interpret the data. Perhaps annotations could be weighted when training classifiers to detect racist language, so that annotators whose identities are most affected by the gold standard have stronger influences on the gold standard.

Human annotation lies at the crux of many advances and tools in computer science. Our work also fits into a broader, growing body of scholarship which reconsiders how researchers' choices and assumptions around human annotation shapes the tools and information that annotation is used to produce (e.g., Sap et al., 2019; Al Kuwatly et al., 2020; Wich et al., 2020; Blodgett et al., 2020). Annotations for racist and hate speech must be reflexively collected and used to avoid contributing to other forms of biases along the way.

Acknowledgements

This work benefited from support from the Summer Institute for Computational Social Science (SICSS) for initial annotator funding, a Graduate Student Research Funding Award from the Sociology Department at the University of Washington and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650604. We also extend our gratitude to Matthew Zook and Ate Poorthuis, for maintaining and sharing the DOLLY data for this project, and the 291 annotators for providing ratings.

Ethical Considerations

This study was approved by the University of Washington Institutional Review Board. We only collected public tweets, and the annotation tools that we used did not display the profile IDs of the Twitter user who authored the tweet. We also reviewed tweets to make sure that authors were not members of groups at an elevated risk of harassment or doxing (e.g., transgender persons). Prior to beginning the annotation task, annotators were informed that the task may involve reading offensive content and were required to provide consent. In addition, we did not collect any identifying information from annotators and only report demographics in the aggregate. Finally, annotators were given the opportunity to exit the task at any time and allowed to write a debriefing response at the end of the task.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190.

Jamie Bartlett, Jeremy Reffin, Noelle Rumball, and Sarah Williamson. 2014. Anti-social media. *Demos*, (2014):1–51.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Eduardo Bonilla-Silva. 2006. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers, Oxford, UK.

Thomas Brambor, William Roberts Clark, and Matt Golder. 2006. Understanding interaction models: Improving empirical analyses. *Political analysis*, pages 63–82.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5(11):1–15.

Evelyn R Carter and Mary C Murphy. 2015. Group-based differences in perceptions of racism: What counts, to whom, and why? *Social and Personality Psychology Compass*, 9(6):269–280.

Irfan Chaudhry. 2015. # hashtagging hate: Using twitter to track racism online. *First Monday*, 20.

Gloria Cowan and Cyndi Hodge. 1996. Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology*, 26(4):355–374.

Jessie Daniels. 2009. *Cyber racism: White supremacy online and the new attack on civil rights*. Rowman & Littlefield Publishers, Plymouth, UK.

René D Flores. 2017. Do anti-immigrant laws shape public sentiment? a study of arizona's sb 1070 using twitter data. *American Journal of Sociology*, 123(2):333–384.

Richard JT Hamshaw, Julie Barnett, and Jane S Lucas. 2018. Tweeting and eating: The effect of links and likes on food-hypersensitive consumers' perceptions of tweets. *Frontiers in public health*, 6:1–12.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Paul Hitlin. 2016. Research in the crowdsourcing age, a case study' pew research center. Technical report, Pew Research Center.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Los Angeles, CA.

Maria Krysan and Sarah Moberg. 2016. A portrait of african american and white racial attitudes. *University of Illinois, Institute of Government and Public Affairs*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the twenty-seventh AAAI conference on artificial intelligence*, pages 1621–1622.

Laura Leets. 2001. Explaining perceptions of racist speech. *Communication Research*, 28(5):676–706.

Mark Morcos, Ate Poorthuis, and Matthew Zook. The dolly project (digital online life and you). Technical report, Floating.Sheep.

Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research*, 42(5):533–544.

Andrew Perrin and Monica Anderson. 2019. Share of u.s. adults using social media, including facebook, is mostly unchanged since 2018. Technical report, Pew Research Center.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2019. stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2):1–40.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*, pages 1–10.

Sanjay Sharma and Phillip Brooker. 2016. #notracist: Exploring racism denial talk on twitter. In *Digital sociologies*, pages 463–485. Policy Press.

Brendesha M Tynes and Suzanne L Markoe. 2010. The role of color-blind racial attitudes in reactions to racial discrimination on social network sites. *Journal of Diversity in Higher Education*, 3(1):1–13.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199.

Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of internet memes. *Computers in Human Behavior*, 63:424–432.

James Zou and Londa Schiebinger. 2018. Ai can be sexist and racist—it's time to make it fair. *Nature Publishing Group*, 559:324–326.