Abstracts

00 **Richard Jean So and Edwin Roland**, Race and Distant Reading

This essay brings together two methods of cultural-literary analysis that have yet to be fully integrated: distant reading and the critique of race and racial difference. It constructs a reflexive and critical version of distant reading—one attuned to the arguments and methods of critical race studies—while still providing data-driven insights useful to the writing of literary history and criticism, especially to the history and criticism of postwar African American fiction, in particular James Baldwin's *Giovanni's Room*. Because race is socially constructed, it poses unique challenges for a computational analysis of race and writing. Any version of distant reading that addresses race will require a dialectical approach. (RJS and ER)

Race and Distant Reading

RICHARD JEAN SO AND EDWIN ROLAND

RICHARD JEAN SO, assistant professor of English and cultural analytics at McGill University, is a coauthor of the essay "Race, Writing, and Computation: Racial Difference and the US Novel, 1880–2000," which appeared in the *Journal of Cultural Analytics* in 2019, and the author of the *PMLA* article "All Models are Wrong" (2018). His book *Redlining Culture: A Data History of Race and American Fiction* is forthcoming from Columbia University Press in 2021.

EDWIN ROLAND is a doctoral student in English at the University of California, Santa Barbara. His areas of research include American literature and new media. Previously, he was the coordinator for digital literary study with the Digital Humanities at Berkeley initiative. He holds an MA in the humanities from the University of Chicago.

Introduction

THIS ESSAY AIMS TO BRING TOGETHER TWO METHODS OF cultural-literary analysis that have yet to be fully integrated: distant reading and the critique of race. *Distant reading* is a term coined by Franco Moretti to describe the use of quantitative methods to study large, digitized corpora of texts. The goal of this work is to identify textual patterns of content and form at the scale of thousands of texts, invisible to the (closely reading) human eye. The use of numbers to analyze literature, of course, is not new. We can trace its origins to an earlier cohort of literary critics and historians who used numbers and statistics to study literary style and history. This cohort includes book historians affiliated with the *Annales* school of socioeconomic history, the poet and critic Josephine Miles, the cultural sociologist Janice Radway, and the literary scholar John F. Burrows. But the publication of Moretti's *Distant Reading* in 2013 helped to galvanize the most recent surge of interest in the use of computers and statistics to study literature. In this study, Moretti defines distance as "a condition of knowledge: it allows you to focus on units that are much smaller or larger than the text: devices, themes, tropes—or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, Less is more. If we understand the system in its entirety, we must accept losing something" (48). Since then, a number of scholars, such as Natalie Houston and Tanya E. Clement, have extended and problematized this method in fields such as nineteenth-century British poetry and Anglo-American literary modernism.

Much of this work, particularly research that has emerged from the Literary Lab at Stanford University, has focused on belletristic

concerns. For example, a recent pamphlet from the lab has used distant-reading methods to think about "style at the level of the sentence" (Allison et al.). Work by scholars like Andrew Piper and Eva Portelance have pushed distant reading to consider questions of cultural capital, identifying the dynamics of power in the literary field that determine, for example, what types of books win literary prizes (Piper and Portelance). But for the most part, scholars—those following in Moretti's footsteps and Moretti himself—have shied away from using large-scale quantitative methods to explore questions of identity, particularly questions of racial identity and minority discourse. The reasons for this are not difficult to fathom. Distant reading requires quantification. Reading race distantly thus requires the quantification of racial identity or racialized language. One need only invoke terms like *bell curve* or *eugenics* to recall the long and ugly history of the use of ostensibly objective quantitative methods to degrade persons of color and, more generally, to authorize and reinforce racial stratification. By contrast, cultural and historical approaches to race emphasize its social constructedness. Race is a category that escapes measurement or simply renders it untenable.

Scholars such as Tara McPherson and Kim Gallon have voiced particular concern over how "the digital" potentially encodes racial bias. McPherson reminds us that racial difference is bound up with the history of technological innovation that produced the modern computer in the 1960s and 1970s. Contemporary data-mining approaches, such as natural language processing—methods increasingly common in the digital humanities—are heirs to that technological revolution and thus must bear the weight of racial discrimination that is its partial effect (McPherson). Gallon also points to the dangers of uncritically employing these methods. She calls for a "black digital humanities" that prefaces any applied use of computation for

the study of culture or history with a racial critique of computation.

A simple exercise illustrates the potential absurdity of bringing data science to bear on questions of race and literature. If we gave you a short passage from a novel and only told you that the text was written by an American author and published between 1950 and 2000, could you guess that author's racial identity?

In just asking this question, we see how preposterous and likely offensive this exercise is. Any attempt to associate diction and syntax, or style and narrative, with race will likely end up reproducing stereotypes. *Black writers only write about X or Y, while white writers write about A or B*. These are claims rooted in reductive and racist assumptions about identity. Further, the entire exercise is based on faulty assumptions. For example, "racial identity" is partly performative, and how it is expressed in language is contingent on local context; the use of a word, such as *Negro*, will have a different significance in different contexts for different authors, even if the authors all identify as African American or black. Also, not all black novelists write "black" literature. And one could argue that the very category of black literature is too incoherent to be viable. Yet this is precisely how, for example, online advertisers use classification algorithms to understand the relation between our racial identity and how we use words.

The challenges of bringing distant reading and the study of race together are clear. In addition to McPherson and Gallon, digital humanists such as Lauren Klein and Roopika Risam have precisely documented the way that many existing digital corpora and tools fail to register racial otherness. In this essay, we extend this work to develop a more inclusive version of distant reading, one that is reflexive about the forms of power that enable it to look at things from a distance. Following Klein, we explore not only what distant reading reveals but also what it conceals. Specifically, we develop a critical form of distant

reading that integrates critiques of race and computation into its experimental design, while also producing results potentially useful for the writing of literary criticism. We begin with a naive approach to computation to introduce a common form of distant reading. We then problematize and transform our model through the lens of critique to do two things: to illustrate the limits of standard computational methods for the analysis of race and to produce a series of results that nonetheless advance our understanding of the texts and authors under investigation. We argue that a dialectical approach to distant reading—an interplay between criticism and computation—allows for a reading of race that does not just restage racial stratification. In fact, exposing the racial limitations of computation can reveal things otherwise occluded within literary history.

Data and Method

Can computational approaches help us study racial discourse in textual artifacts that have been formally identified as "literature," such as novels? To answer this question, we first need to produce a corpus of texts that corresponds to normative distinctions of race. Here, we decided to focus on the postwar American novel. We constructed a corpus that is evenly split between novels by black authors and those by white authors. We are interested in studying the similarities and differences between these two corpora as a means to understand the distinctions between novels written by authors of different racial identifications.

Our corpus of novels by black authors was produced in the following way: a group of scholars and librarians at the University of Kansas, working at the Project on the History of Black Writing, directed by Maryemma Graham, have spent the past twenty years identifying and acquiring physical copies of every novel written by an author identified or self-identified as black. In total, these scholars and

librarians have found approximately 1,200 such novels published between 1880 and 2000. With the help of their colleagues at the University of Chicago, they have so far digitized 220 of these novels, which represent a random sample of the overall corpus. Working with a team of research assistants, we then identified the gender of each author in the corpus. Our standard for doing this work was rigorous: to tag authors by gender, we had to find reliable scholarly sources that identified the authors' gender or evidence of the authors themselves identifying their gender. If we could not find evidence that met this standard, we left authors unmarked and set them aside from our corpus. These 220 works include a range of canonical and noncanonical texts, such as, respectively, Toni Morrison's *Beloved* and Melvin Van Peebles's *True American*. While we believe scholars at the University of Kansas have produced a sound representation of black literature, we acknowledge and emphasize that this category is a contested one and that our corpus presents just one version of it.

To build the corpus of novels by white authors, we first identified the twenty thousand American novels published from 1950 to 2000 held in the most libraries in the United States based on Worldcat records. Next, we acquired electronic copies of approximately nine thousand of these texts. Then, with a group of research assistants, we identified the gender and race of each author on the list, using the same methods and standards of evidence described above. This reduced our list of novels to approximately 5,900 texts. We then identified the genre of each text, such as literary fiction or the detective novel. Our standard here again was high; if the novel did not identify its genre, we found a scholar who did. This further reduced our list, to about one thousand texts. Finally, we randomly drew fifty-five novels from each of the most-represented genres in our corpus: best seller, prizewinner, science fiction and detective. We limited this corpus to 220 novels because our computa-

tional method requires that the corpora we compare be the same size, and we currently have 220 digitized novels by black authors.

We restricted our corpus to novels published between 1950 and 2000 because this time period is long enough to identify chronological dynamics yet short enough to avoid major shifts in what specific, racialized words, such as *colored* or *Negro*, signify. Following historians like Matthew Jacobson and Nell Painter, we contend that race is talked about and represented through language in a relatively coherent way during this period in United States history. We broke our corpus of novels by white authors down by genre to determine whether any distinction we might find between our white and black corpora was animated by genre instead of race. As cultural historians such as Eric Lott argue, whiteness is expressed differently in different literary genres. Finally, we identified the gender of each author in our two corpora to determine if gender is a significant factor in how whiteness or blackness is expressed in novels.

While the process by which we identify writers as white or black is based on the body of academic scholarship surrounding each author, this process still risks reifying racial identity as a category. The racial ontology of an author is not stable; what it means to be white or black changes over time and place. For example, most scholars today identify the mixed-race writer Nella Larsen as black. But this categorization has evolved since the 1920s, when she wrote—in that era, Larsen was commonly referred to as a "mulatta." At the same time, labeling authors white or black risks erasing dynamics of intersectionality. A novelist like James Baldwin is labeled black in our corpus, but Baldwin also identified himself as a gay man and as a writer. One cannot understand one form of identity without the others. We revisit these concerns in the final part of our essay, where we complicate the ontologies of racialized authorship we are provisionally treating as fixed.

Next, we need a method, or "model," to analyze our two corpora. For our first-pass analysis, we will use a classification algorithm.[2] The purpose of algorithmic classification is to predict the identity of a text given two possible categories to which that text might belong. For example, each day our e-mail account distinguishes "real" e-mail from "spam," blocking what it identifies as spam from our inbox. To learn how to make this distinction, a machine is shown tens of thousands of examples of real e-mails and spam. The machine will study the textual features of both types of texts, noting which ones tend to appear more in one than the other. This includes diction (spam tends to use words like *Viagra*), syntax (spam tends to use incorrect syntax), the use of first versus second person (spam tends to use the latter more than the former), and so forth. It will then quantify such tendencies—for example, spam uses the word *Viagra* ten times more often than real e-mails. These tendencies (a list of features with quantitative weights) become the machine's "mental catalogue," which allows it to distinguish our two e-mail classes. In scientific terms, we refer to this as a "language model." Finally, the machine will test how good its model is in performing classification. A human will give the machine one hundred new and unlabeled e-mails and ask it to predict whether they are real or spam based on its mental catalogue. If it correctly predicts the identity of a given percentage of e-mails—say, ninety-five percent—we can assert that our language model is generally accurate and our model is sound. If it falls below that threshold, the machine refines its catalogue—shifting weights, removing features—and runs this test again, altering itself until it performs at or above the desired threshold.

Reading this description of the algorithm, one is likely to raise a number of immediate objections from a humanistic standpoint. One major objection, we imagine, might be that the difference between spam

and nonspam e-mails is not equivalent to the difference between novels written by white authors and novels written by black authors. *Spam* and *nonspam* are technical descriptions of discretely defined objects (e-mails). *White* and *black* are socially constructed categories of racial identity. Thus, machine classification is inappropriate for literary criticism. The goal of machine classification is to identify and label objects. The point of minority-discourse analysis is, in part, to critique and problematize the very idea of categories. Moreover, we could add two more concerns: first, the machine assigns textual features to each category, which in the case of racial categories simply reifies racial identity, such as blackness; and, second, the machine relies on the assumption that its initial categories are coherent and real and that they exist meaningfully within a binary relation, but scholars have long argued that race defies binary categorization. We do not disagree with these objections and, as our essay unfolds, we will alter and, in some cases, deform our model to account for these challenges. Specifically, to qualify our main approach, we will make two arguments. First, the machine is a relational, not ontological, thinker. It does not impute essential qualities to classes of objects or texts; it simply marks the line that divides them, as well as the strength or weakness of that line. Second, while the machine must start with nominal categories to do its work, native to the machine itself are methods to test the integrity of those initial categories and to explore their potential contingency or tenuousness. But for now we take a naive, first pass at classification to set up an analysis of its inevitable limits, as well as to produce an initial signal regarding patterns in the data, which we revise and extend.

We constructed a model that we believe summarizes the overall textual properties of the novels in our corpora. While such a model cannot capture the particularities of every text, it allows us to compare texts at a

large scale. Our model includes three types of features: diction, syntax, and narrative.[3] First, our model simply counts how many times certain words appear in a text. Word frequency cannot of course capture things like metaphor or irony, but it provides a way to glimpse the content of a novel. Second, our model counts how often various parts of speech, like nouns and adverbs, appear in a text. Here, too, simply measuring their frequency can only tell us so much, yet while it cannot reveal the paragraph-level syntactic ambitions of a text, it can tell us about its sentence-level syntactic habits. To that end, we counted how often certain "bigrams," or pairs of parts of speech, like a noun followed by an adverb, occur in a novel. Finally, the model computes statistics relevant to the represented world of the story. These include the ratio of dialogue to narration, the number of characters in the story, the average amount of narrative attention paid to each character, the ratio of manufactured to natural objects in the novel, and the average number of locations or settings in it. We refer to this third category of features as "narration." While our model's analysis of these narrative features cannot remotely capture the full complexity of the concept of narrative as asserted by literary theorists, it is a useful, albeit coarse, account of how novels use the building blocks of narration: dialogue, characters, objects, and setting.

Now that we have a language model to characterize the texts in our corpora, we can use it to try to distinguish between novels by white authors and novels by black authors:

Comparison	Classification Accuracy
All novels by white authors versus all novels by black authors	92%
Best-selling novels by white authors versus all novels by black authors[4]	90%
Prizewinning novels by white authors versus all novels by black authors	94%
Detective novels by white authors versus all novels by black authors	93%
Science-fiction novels by white authors versus all novels by black authors	92%

A few things stand out. First, the machine is excellent at distinguishing novels written by white authors from those by black authors. Ninety-two percent is an extraordinarily high rate of accuracy. And while our list of features is simple and one might wish for greater complexity, the machine does not need more nuanced features to do its work. This is how different our two classes of texts are. Next, genre is insignificant. Whether a white author has written a best seller or a work of science fiction does not make that work much more or less likely to differ from fiction written by black authors.

Moreover, the machine can report which features are most significant in classifying our texts. In addition to our syntactic and narrative features, each word, in a sense, is a feature. Out of the million potential features that might significantly contribute to the model, only a handful actually do[5]:

Corpus	Examples of Most Distinctive Features
White	*it*, *enormous*, *might*, *absolutely*, *matters*, *estimate*, *nonsense*, *identified*, *impassive*, *forty*, *presumably*, *interesting*, *unlikely*, *assuming*, *slot*
Black	*white*, *before*, verb_noun, *colored*, *woman*, preposition_noun, *people*, *snatched*, verb_verb, *black*, *lie*, *jazz*, adverb_noun, *freedom*, *floor*

The outlines of a story emerge: novels by black authors tend to use more verbs and nouns (i.e., action and object-based language) than novels by white authors. And the latter tend to use more qualifying language (such as "enormous" and "interesting") than the former. The second claim is a bit more tenuous because, overall, novels by white authors did not contain on average more adverbs and adjectives than novels by black authors, but some subsets of novels by white authors, such as prizewinning novels, did contain more modifiers than novels by black authors. In any case, we have a signal as to semantic and syntactic distinctions. And we find that narrative features are not significant.

Analysis and Model Critique

It is tempting to use these results as the basis for a broader analysis. For example, we might develop a reading that argues that literary whiteness is in part defined by an attention to linguistic qualification—the constant deferral of meaning. Or we might claim that literary blackness privileges things and action over description and that if it does describe things, those things tend to be racialized ("white," "colored"). Or, bringing these readings together, we might argue that the narrative worlds of white and black authors are highly distinct, preoccupied by different concerns defined by the words they tend to use. A list of most distinctive features can be the basis for making arguments about the corpora they represent.

But making any of these moves implies that we accept the machine's results and, more broadly, its approach, as valid. And perhaps we do not—perhaps our initial concerns still weigh too heavily. Here, we directly take up these concerns by working through the logic of the machine itself. Our first concern is that the machine understands racial difference in a binary (white/black) frame. Yet scholars such as Gary Okihiro have long argued that race often exceeds or resists binary classification. White and black are not cogent, monolithic classes of identity. Our machine performs as if it were.

We can, however, manipulate our results to decompose these apparently solid categories. The key is thinking about how each category expresses its features, and how that expression varies by category. We recall that our machine relies on a set of features that are particular to texts written by white or black novelists in order to distinguish one kind from the other at the remarkably high rate of ninety-two percent. Let us call these "white" and "black features." If our main categories of authorship are actually commensurable, monolithic entities that are perfect opposites,

they should express these features in commensurable ways. For example, novels by black authors should express white and black features consistently as a category. But if they do not, what we might be calling novels by black writers might simply be a collection of many smaller categorical divisions. This category might be constituted by difference within difference, and thus it might not be commensurable with our other category.

An analogy will help to explain the machine's approach to the way our groups of texts relate to each other. Imagine two classrooms, each containing ten students. If we want to know whether students in both classrooms are comparably tall—because we want to know whether the desk heights in classroom A will be appropriate for the students in classroom B—we could ask what the average height of the students is in each class and compare the two values. Let us say the average height in both classes is five feet six inches. If the students in both classrooms are all close to that height, the average will show that the height of the desks in room A will suit the students in room B. Now imagine a second scenario, in which classroom A is the same as before but classroom B is evenly split between students who are five feet and six feet tall. In both of these classrooms, again, the average height is five feet six inches, yet in this case none of the students from classroom B will be able to sit properly at the desks in classroom A. These two scenarios demonstrate that the statistic of average height is not meaningless but potentially deceptive. Fortunately, we can evaluate the average's usefulness by taking a second measurement: variance. In the first example, the variance of heights in both classrooms was about the same—near zero—since most students were close to the average. In the second example, the variance was low in one classroom but high in the other, indicating that the students' heights are not comparable on the basis of their averages. In the model we built for this study, as in many real-

world applications, the machine assumes that both categories—in this case, texts by white and black authors—will have about the same variance in the distributions of their features and thus are comparable.

To gauge how monolithic each of our initial categories is, we tested that variance in our own corpora. We wanted to know whether texts by white and black writers use their predictive features in the same way, and whether they use the other class's features comparably. Following the analogy above, each novel is a classroom and each feature is a student. In fact, we found that the averages for these values were comparable. On average, novels by black authors use the model's predictive black features at about the same rate that texts by white authors use its white features; similarly, texts by black authors use white features at about the same rate that texts by white writers use black features. However, we still want to know whether these features have comparable variance.

This is what we found. The variance in how novels by black authors use black features is about the same as the variance in how novels by white authors use white features. However, there is a significant difference in how novels by white and black authors use each other's features. Specifically, the variance in how novels by black authors use white features is forty percent greater than the variance in how novels by white authors use black features. That is, novels by black authors show a wider variety of engagements with the features that allow us to distinguish between novels by white and black writers—the very basis of our model. This engagement constitutes an internal differentiation that greatly exceeds what we find in texts by white writers.[6]

What can we conclude about these results? Categories of novels by white and black authors are not commensurable. The former is far more coherent than the latter. Not unlike our classroom split between five- and six-foot-tall students, the category of novels by black

authors is better described as a collection of subgroups. This conclusion supports scholarship by black studies scholars who argue that recent dynamics of diaspora put immense pressure on the ostensible coherence of African American or black literature, as well as arguments by Kenneth Warren, who contends that the category is coherent only as an effect of Jim Crow. But our results go even further, suggesting that the category has been highly diffuse since the 1950s and perhaps since even earlier. Similarly, our results extend canonical arguments in whiteness studies. Refuting claims that whiteness is "unmarked" or "contentless," whiteness, we find, is relatively coherent (Morrison). But indeed, in keeping with the claims of Eric Lott and Morrison, this coherence is derived entirely from its relation to blackness. Novels by white authors cluster within a tight band of expression, but only through their expression of blackness.

Nonetheless, despite this decomposition of our categories, the analysis thus far still relies on the assumption that white and black are meaningful categories, even if one has been shown to be less coherent or stable than the other. If one rejects this view, one rejects the analytic consequences of these categories. For example, if novels by white and black authors do not exist, then so-called white and black features do not exist. We can, however, extend our process of decomposition further.

Machine classification allows us to more fully deconstruct the categories of white and black novels. Instead of following a strictly binary logic, the machine assigns a probability between 0 and 1 to each text, where 0 corresponds to the likelihood of the text's having been written by a black author and 1 corresponds to the likelihood of its having been written by a white author.[7] The values 0 and 1 are arbitrarily assigned to their respective groups, and switching the assignments would not change the outcome of the experiment. If a text has a score above 0.5, the machine labels that text white, and if a text has a score

below 0.5, the machine labels that text black. Thus, identity in this framework exists on a spectrum, even if each text must ultimately be assigned a binary value. Few novels have a score of 0 or 1. Most have scores somewhere in between, and several fall at the 0.5 level—the domain of total indeterminacy.

The graph in figure 1 dramatizes the machine's own unstable conception of these categories. Each marker represents a text; novels by white authors are circles and novels by black authors are triangles. The y-axis corresponds to the machine's prediction of each text's probability of being by a white or black author. The higher its position—the closer its probability is to the arbitrarily selected value of 1—the more likely it is to be white; the lower its position, the more likely it is to be black. Last, we add two straight lines ("lines of best fit" derived from a linear regression analysis) to visualize the relation between all the points and their general direction over time. Now, what is striking is not only the magnitude and stability of the distinction between novels by white and black writers but also the fact that a number of texts appear where they should not be according to their binary labels. That is, a handful of circles fall near the black-author trend line and a large number of triangles fall near the white-author line. These unexpectedly placed markers represent novels that the machine has misidentified (or misclassified) as white or black.

Typically, computer scientists think of misclassifications as simple errors—like an algorithm's misidentification of spam as a real e-mail. But when we begin with the belief that our categories may very well be tenuous, and we are interested in testing that possibility, misclassified texts provide information about what makes those categories unstable. Misclassified texts indicate that the machine has become confused; but rather than take that confusion as a sign that the machine has made a mistake, we can read the misclassified texts as marking the limits of what the

machine can reliably understand. Misclassified novels mark the threshold at which our categories become unreliable.

So far, following the more conventional approach to classification, we have thought of the machine as a device to mark difference: the distinction between novels written by white and black authors. An attention to misclassified texts, though, allows us to invert our analytic orientation. Rather than track the distinctness of our categories, we can track its opposite: their indeterminacy. How indeterminate is the boundary between white and black authorship in this period, as well as over a larger stretch of time? Is it increasing or decreasing? Are novels by white writers more likely to be indeterminate than novels by black authors? We tested these hypotheses.[8] The number of misclassified texts by year is stable; thus, the indeterminacy of white and black is unchanging during the period we studied. However, we found that novels by black authors are far more likely (nearly five times more likely) than novels by white authors to be misclassified. This means that if the distinction between white and black is tenuous, this indeterminacy is animated more by our corpus of novels written by black authors.

Close Reading: *Giovanni's Room*

Our discussion of misclassification allows us to think through the limits of our model and how that model can be manipulated in order to respond to those limits, while also producing new insights about our material. Still, our analysis remains rather coarse. We talk about a "threshold" between white and black authorship in our model, but what does this threshold look like at the level of a single text? Concretely, what do we mean when we say that our categories are contingent? The category of the misclassified allows us to pivot back to the text. What exists in general form at the scale of the entire corpus can attain granularity at the level of a specific novel. If the general

class of the misclassified points to the erosion of the machine's initial binary understanding of white and black, a close analysis of a single misclassified text can reveal what precisely motivates that ontological undoing.

In this final section, we analyze a specific misclassified text: *Giovanni's Room* (1956), a novel written by an identified black author, James Baldwin. Why did we choose this text? Because black authors are far more likely to be misclassified than white authors, they drive the machine's color-line confusion. Of the black authors who have more than one novel in our corpus, six have at least one novel that is classified correctly and at least one that is not. These are the authors that particularly perplex the machine, existing at once on both sides of the color line. These writers include James Baldwin, Robert E. Boles, Samuel Delany, Hugh Holton, Charles Johnson, and Nora DeLoach. Baldwin stands out because the machine believes with unusual certainty that he is both a white and a black author. It predicts that there is a 99.9 percent likelihood that *Go Tell It on the Mountain* was written by a black author but also an 87.4 percent likelihood that *Giovanni's Room* was written by a white author. Analyzing the latter novel can help us understand how the line that otherwise effectively divides white from black can dissolve.

Readers even passingly familiar with Baldwin's novel will hardly be surprised that this postwar literary text has confounded the machine's classification. *Giovanni's Room* features a nearly all-white cast and takes place primarily in Paris. It tells the story of David, a white American living in Paris in the early 1950s, who has an affair with Giovanni, an Italian man. The story focuses on David's sexual passing as a straight man; despite having several affairs with men, such as Giovanni, he retains a heterosexual relationship with an American woman. The central tension of the story is David's inability to reconcile the pressures of his life in the United States, which centers on postwar norms of middle-

class domestic life, with his desire to have romantic relationships with men in Europe. The novel's frank depiction of homosexual intimacy provoked controversy when it was first published in 1956, and since then it has become celebrated as a pioneering work of queer literature.

Academic scholarship on the novel in the late 1990s generally struggled to understand the novel as an explicit work of black literature; Mae Henderson, for example, valorizes the novel's engagement with "paradoxical subjectivity" but generally finds that the novel "erases" blackness through its attention to sexual identity (313). More recent scholarship, however, argues that *Giovanni's Room* "analogizes" racial difference through its tropes of sexual difference (Abur-Rahman 477). Racial passing appears in the novel in displaced form as sexual passing. What concerns Baldwin, according to this scholarship, is the normative expression of power in society, and that power, as our close readings will show, flows from white, European ideals of desire. Both racial blackness and homosexuality are at odds with such ideals. Thus, the text articulates a critique of whiteness and a valorization of racial difference, even as it is nominally displaced by sexuality. Aliyyah I. Abur-Rahman writes, "[M]y focus is on Baldwin's critique of whiteness, specifically through his subtle allusions to the *racializing* effects of queerness" (480).

This analogical thinking in recent scholarship on *Giovanni's Room* takes the novel's cast as its starting point: the absence of black characters compels an interrogation of its white characters' sexual, national, gender, and class identity. Much of this scholarship thus focuses on the displacement of blackness in the text and, by extension, the ciphering of identity that whiteness enables. However, the text also displaces whiteness through its thematic deployments of the features our model uses to construct white authorship.

How do we read Baldwin's novel closely as a misclassified text? The model identifies fifty-seven features that are statistically significant in telling apart our two categories of texts. By observing how the model uses those features to make a prediction about *Giovanni's Room*, we can better understand the threshold that divides novels written by white and black authors. But which specific features contribute to the novel's misclassification? We can take the numeric weight the model had assigned to each feature and multiply it by how often it appears in *Giovanni's Room*. This operation returns a ranked list of features that the machine had used to make its decision about the novel, which, in turn, suggests a counterfactual exercise: what if we changed each of these ranked features to the average of their frequencies in all the other novels by black authors (i.e., the values that we expect to see in these texts)? How many features would we have to change in order to get the machine to predict *Giovanni's Room* was written by a black author? We found that six out of the model's fifty-seven features had to be altered for the machine to reclassify Baldwin's novel as a novel written by a black author. These features, sorted by their contribution to the text's misclassification from highest to lowest, are the words *absolutely*, *very*, *course*, *appalled*, *might*, and *white*. The first five features are characteristic of texts written by white authors and appear frequently in *Giovanni's Room*, while *white* is predictive of black authorship but appears infrequently in the text.

Before proceeding, we should dismiss an incorrect reading of this output. It would be false to interpret this result as suggesting that the text's heavy use of certain diction makes it "sound" like a novel by a white person and that we have "unwhitened" the novel to make it sound more "black." To reiterate, the machine is not an ontological thinker. It does not say that all novels by white authors are defined by the use of such and such features, and vice versa. Rather, the machine

simply measures the robustness of the social constructedness of these given categories and points out what gives them such vigor. In the analysis that follows, we try to understand why a very short list of words means so much to Baldwin's novel—and how they contribute to our understanding of its status as a work of black literature.

That the word *white* should appear in this list is striking for two reasons. First, in distinction to previous scholarship that has focused on the novel's absence of blackness, the machine is concerned with the absence of *white*. This output does not directly contradict the existing scholarship, which often dwells on the coproduction of whiteness and blackness. Instead, we offer an alternative point of entry. Second, it is the absence, not the presence, of the word *white* that the machine interprets. The word appears just twenty-six times, meaning that it appears not only less often than it does in most novels by black authors but also less frequently than it does in the average novel by a white author in our corpus. This is not simply a case of whiteness going unremarked. Rather, in this novel, the absence of *white* is strongly felt.

How do we read the absence of *white*? We first need to study how the word is used. It can help Baldwin masterfully render Paris in vivid terms, the text often exploding in color and material descriptions, of tables, walls, radios, shirts:

> Indeed there were young people, half a dozen at the zinc counter before glasses of red and *white* wine, along with others not young at all. A pockmarked boy and a very rough-looking girl were playing the pinball machine near the window. There were a few people sitting at the tables in the back, served by an astonishingly clean-looking waiter. In the gloom, the dirty walls, the sawdust-covered floor, his *white* jacket gleamed like snow. Behind these tables one caught a glimpse of the kitchen and the surly, obese cook. He lumbered about like one of those overloaded trucks outside,

> wearing one of those high, *white* hats, and with a dead cigar stuck between his lips.
> (Baldwin 50; our emphasis).

This scene takes place early in the story. It features Giovanni, David, and Guillaume—an older homosexual man—eating at a bohemian restaurant in Paris. The scene uses adjectives to create a series of contrasts between the clean and the dirty, the pure and impure. At first pass, the use of *white* seems to help generate the novel's light and dark symbolism.

Reading on, we find that this scene dramatizes the sexual economy of Giovanni's community. As critics have noted, the social backdrop to the story is one in which wealthy, older Parisian men exchange food for sex with young, impoverished men. Yet we find that the operations of this exchange are subtle and not reducible to a naive reading of exploitation. The boys at the bar have agency: each one sizes up the protagonists as they enter the restaurant, "having already calculated how much money he and his copain would need for the next few days . . ." (53). Though financially constrained, the young men in this scene care for one another and consciously negotiate their relationships with the older men.

Careful attention to the white objects in the passage reveals how this economy works. The jacket's whiteness "astonish[es]" David with its contrast to the room's "dirty walls" and "sawdust-covered floor," and this momentary flash reveals the animating dichotomy of the scene (pure/impure, young/old). Beyond its symbolism, the jacket makes visible how this economy becomes material. The older men's purchase of white wine, and later a meal for the boys, enables the face-to-face interactions that follow. After the boys have calculated Guillame's "value," they think about what they want: "The only question left was whether they would be *vache* with him, or *chic*, but they knew that they would probably be *vache*" (53). The cycle of exchange goes on. The passage's white objects help to initi-

ate this process, but very quickly that cycle moves past those original things.

Indeed, what is most striking about the word *white* is how rapidly it disappears from the novel. By the end of the story, the word has completely vanished from the text. As we see in the above passage, white objects often appear in the text to prompt a set of social interactions, but the text quickly stops paying attention to those objects in order to pay more attention to the interactions. This tendency helps to explain the relative prominence of intensifiers in the text—they serve to elaborate and define those social interactions: *absolutely*, *very*, and *of course*. In fact, *very* is perhaps the most prominent word in the novel, appearing over two hundred times—more than three times as frequently as it does in the average novel by a white author, the category of text this word predicts. The word appears in the dialogue of all the main characters and in David's narration, addressed to the reader. For example, in two consecutive sentences, Giovanni tells David that he is "a *very* charming and good-looking and civilized boy" and that this will make maintaining their relationship, even after David's fiancée returns to Paris, "*very* simple." David, on the other hand, rebuffs this vision of their life together with curt responses to his questions. Does he anticipate that he will visit other people without Hella? "Of *course*." Does she make him confess all he does apart from her? "Of *course* not" (47; our emphasis).

Or consider a key moment in David's early life when David finally recognizes his father as a fellow human being rather than as a parental antagonist: "And my father's face changed. It became terribly old and at the same time *absolutely*, helplessly young. I remember being *absolutely* astonished, at the still, cold center of the storm which was occurring in me, to realize that my father had been suffering, was suffering still" (19; our emphasis). This process of recognition hinges on the use of a specific intensifier (*absolutely*),

which appears twice in rapid succession. David and his father essentially become joined through this word: *absolutely* is used to describe first the face of his father and then David's own reaction and feelings. David and his father interact meaningfully through this mirroring of intense emotion.

The machine has picked up on the way Baldwin's novel begins with a nominal attention to whiteness as a potential description of things or places yet rapidly displaces that type of description with other kinds of narrative attention—namely, intensifying how something is described ("very," "of course") rather than just describing it. This movement tracks the disappearance of whiteness. But whiteness does not simply vanish. It reappears in ciphered form. Consider the final word on our list of terms that led to the novel's misclassification: *appalled*. The word occurs just once in the text, but it carries an outsize influence in misclassifying *Giovanni's Room* as a novel by a white author. A single word can make such a difference if it relates to all the other words in the novel differently from the way it does when it appears in the other novels in the corpus. Here, the word comes at a crucial moment in the novel's depiction of David's personality. Echoing mid-twentieth-century psychological discourses on homosexuality, the novel focuses closely on David's vexed relationship with his parents, and again it is his father who instigates a powerful realization of self, but now a negative one: "We were not like father and son, my father sometimes proudly said, we were like buddies. I think my father sometimes actually believed this. I never did. I did not want to be his buddy; I wanted to be his son. What passed between us as masculine candor exhausted and *appalled* me" (17; our emphasis). David lacks an appropriate role model for masculinity in his father. He has come to accept this (as we see above), but the problem now is that what once functioned as a viable father-son relationship based on friendship has now become frustrat-

ingly ineffective, even repulsive. The etymology of the word is telling: *appall* derives from a Middle French term, *apalir*, meaning "to grow pale, make pale" ("Appall"). Here, the moment David develops a troubled relation to normative masculinity is also the moment he becomes "white." But whiteness itself is not and cannot be directly named. It is merely alluded to, as an effect of David's failing relationship with his father. Though it cannot be directly pointed to, it is there.

Our close readings underscore the degree to which whiteness—perhaps as much as blackness—is displaced and, eventually, disappears from the text. But they also identify its semantic mutations, the way in which whiteness emerges as there but not there, latent in David's *appalled* horror at his father. Such readings contribute to the existing scholarship by revealing otherwise invisible linguistic effects that expose a mutation or displacement of whiteness. We have not read any scholarship that has noted the importance of the word *appalled* in this novel. Without the machine's aid, we also would never have noticed it. Yet, a word so seemingly incidental turns out to expose a decisive repatterning of whiteness in the text.

More readings of *Giovanni's Room* can be done using this computational method, but the purpose of this exercise is to underscore the following: a machine can affirm that novels by white and black authors are, as socially constructed categories, remarkably distinct. But the line that divides them—what initially appears so strong (with ninety-two percent accuracy)—is at the same time tenuous and deformable. A mere six features out of fifty-seven needed to be changed in order to unravel the method's binary logic. In particular, a single word, like *appalled*, can appear to have an outsize impact on how we understand categories of white and black authorship, both at the scale of an entire corpus, and on the page.

Conclusion

The results of our close readings offer one final opportunity for a reflexive consideration of the limits of our distant-reading method. Here, we return to our earlier concern that our categories of white and black authorship erase the historical and intersectional way that identity is defined. Our reading's destabilization of the machine's logic of white and black arises directly from the novel's expression of queerness. By queering the machine's color line, Baldwin's novel challenges our initial classifications of the novels as white or black, which had necessarily effaced a more sophisticated, intersectional view of social identity. In their current form, our data and model are not robust enough to handle this kind of intersectionality.

But we can imagine an improved version of this experiment, one that looks at both racial and sexual identity, building these two aspects into the data and the model at the same time. What we would then track is not only the contingency of the machine's categories of white and black but also the contingency of its categorization of works as straight and gay—or, more broadly, queer—and how those contingencies interrelate. Indeed, the goal of this essay is to begin the hard work of developing a critical version of distant reading appropriate for the analysis of race and racial discourse in literature. We envision a reflexive method that is able to identify its own elisions while also pointing to new insights and opportunities for research.

Notes

1. Do 220 novels per category provide enough data to generate meaningful results? Compare that, say, to the tens of thousands of novels published by white authors during the period we examine. While a theoretical discussion of sampling is beyond our scope, we point to recent applications of similar machine-learning tech-

niques that find robust results while using corpora of approximately our size. See Piper and Portelance; Long and So; and Underwood and Sellers. In each of these cases, the authors include no more than two hundred texts or around fifteen million words to represent each category. Our corpus surpasses both of these thresholds.

2. Our statistical model relies on logistic regression for its binary classification. We used the implementation made available in the *scikit-learn* package for Python. In order to ensure the model's generalizability, we employed l1-regularization, which selects a small number of features to use for its predictions, and an approximately optimal regularization coefficient (C=1.0) was determined through tenfold cross validation.

3. These features were produced thanks to the BookNLP pipeline, developed and distributed by David Bamman. Term frequencies were tabulated from lowercased entries in the pipeline's "originalWord" output. Although stopwords are often removed at this stage, we included them in the model, since they are understood to mark genre and authorial style. Part-of-speech tags were tabulated from the pipleline's "pos" output, which reports tags in the Penn Treebank format and relies on the Stanford POS tagger. These were then counted as bigrams of consecutive tags within sentence boundaries. Narrative features included tabulated frequencies of NER and Super Sense tags, from the "ner" and "sst" output columns. This pipeline uses the Stanford NER tagger and Wordnet Super Sense Tagger (SST). For example, these tag references to any PERSON or LOCATION and OBJECT or ACTION in the novel. In addition to these, we counted the share of the text that consists of dialogue (from the "inQuotation" output), the share of the text that consists of character mentions (i.e., total character space, from the "characterId" output), and the number of unique characters normalized by text length (also from the "characterId" output).

Note that, before any model was built, each type of feature was l1-normalized, in order to minimize the effect of text length. All features were then transformed into standard units so as to be comparable with one another.

4. In cases where we divided the larger corpus of novels by white authors into genre (bestseller, prizewinner, detective, science fiction), each of these subcorpora comprised 180 novels, compared against a sample of 180 novels by black authors.

5. The significant features to which we refer have nonzero weights in the model and pass a z-test, indicating that they have different mean values among novels by white and black authors. In this case and all others tests of statistical significance in this paper, we employ a 95-percent confidence threshold. Moreover, in all cases we consider, significance is tested in a multiple-comparison setting, so we adjusted our measure of confidence by the conservative Bonferroni correction. In effect, we require $p = 0.05 / [\#\ of\ observations]$.

6. Black authors used white features at an average frequency of about -0.32; white authors used black features at an average frequency of -0.31. Note that feature frequencies are in standard units, so the above values indicate that both sets of authors use the other group's features at a frequency slightly below the average of the corpus. A z-test was unable to reject the null hypothesis that black authors use white features at a different average frequency than white authors use black features ($p = 0.7$).

However, a z-test rejected the null hypothesis at a high level of confidence when testing the variance of white features in black texts against the variance of black features in white texts ($p << 0.01$). In that case, the variance of white features in black texts ($\sigma^2 = 0.42$) was found to be about forty percent greater than the variance of black features in white texts ($\sigma^2 = 0.3$).

7. The probability that a text belongs to a given category is a native feature of logistic regression, which previous literary scholarship has embraced and which we emphasize here. Each text was assigned its probability through leave-one-out cross validation, where the texts by a given author are set aside during training and afterward receive predictions.

8. The test was performed using logistic regression. We created a binary dummy variable corresponding to a novel's pre-/post-1975 publication, and along with authorial race we regressed these over another binary dummy variable corresponding to whether the novel had been correctly classified by the model (i.e., DATE + RACE ~ CORRECT). Only authorial race was found to be significant ($p << 0.01$), and it has a likelihood ratio where misclassification increases by a factor of 4.58 for novels by black authors.

Works Cited

Abur-Rahman, Aliyyah I. "'Simply a Menaced Boy': Analogizing Color, Undoing Dominance in James Baldwin's *Giovanni's Room*." *African American Review*, vol. 41, no. 3, 2007, pp. 477–86.

Allison, Sarah, et al. "Style at the Scale of the Sentence." *Stanford Literary Lab*, pamphlet 5, June 2013, litlab.stanford.edu/LiteraryLabPamphlet5.pdf.

"Appall, *Verb*." *Merriam-Webster Unabridged*. Merriam Webster, 2019, unabridged.merriam-webster.com/unabridged/appall.

Baldwin, James. *Giovanni's Room*. 1956. Random House, 2013.

Clement, Tanya E. "'A Thing Not Beginning and Not Ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans*." *Literary and Linguistic Computing*, vol. 23, no. 3, 2008, pp. 361–81.

Gallon, Kim. "Making a Case for the Black Digital Humanities." *Debates in the Digital Humanities*, edited by Matthew K. Gold, U of Minnesota P, 2016, pp. 42–49.

Henderson, Mae. "James Baldwin: Expatriation, Homosexual Panic, and Man's Estate." *Calalloo*, vol. 21, no. 1, 2000, pp. 313–27.

Houston, Natalie. "Towards a Computational Analysis of Victorian Poetics." *Victorian Studies*, vol. 56, no. 3, 2014, pp. 498–510.

Jacobson, Matthew Frye. *Whiteness of a Different Color: European Immigrants and the Alchemy of Race*. Harvard UP, 1999.

Klein, Lauren. "Distant Reading after Moretti." *Arcade: Literature, Humanities, and the World*, arcade.stanford.edu/blogs/distant-reading-after-moretti.

Long, Hoyt, and So, Richard Jean. "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning." *Critical Inquiry*, vol. 42, no. 2, Winter 2016, pp. 235–67.

Lott, Eric. *Love and Theft: Blackface Minstrelsy and the American Working Class*. Oxford UP, 1993.

McPherson, Tara. "Why Are the Digital Humanities So White? or, Thinking the Histories of Race and Computation." *Debates in the Digital Humanities*, edited by Matthew K. Gold, U of Minnesota P, 2012, pp. 139–60.

Moretti, Franco. *Distant Reading*. Verso, 2013.

Morrison, Toni. *Playing in the Dark: Whiteness in the Literary Imagination*. Harvard UP, 1992.

Okihiro, Gary. *Margins and Mainstreams: Asians in American History and Culture*. U of Washington P, 2014.

Painter, Nell Irvin. *The History of White People*. W. W. Norton, 2010.

Piper, Andrew, and Eva Portelance. "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading." *Post45*, 10 May 2016, post45.research.yale.edu/2016/05/how-cultural-capital-works-prizewinning-novels-bestsellers-and-the-time-of-reading/.

Risam, Roopika. "Navigating the Global Digital Humanities: Insights from Black Feminism." *Debates in Digital Humanities 2016*, edited by Matthew K. Gold and Lauren F. Klein, U of Minnesota P, 2016, pp. 359–67.

Underwood, Ted, and Sellers, Jordan. "The Longue Durée of Literary Prestige." *Modern Language Quarterly*, vol. 77, no. 3, 2016, pp. 321–44.

Warren, Kenneth. *What Was African American Literature?* Harvard UP, 2012.

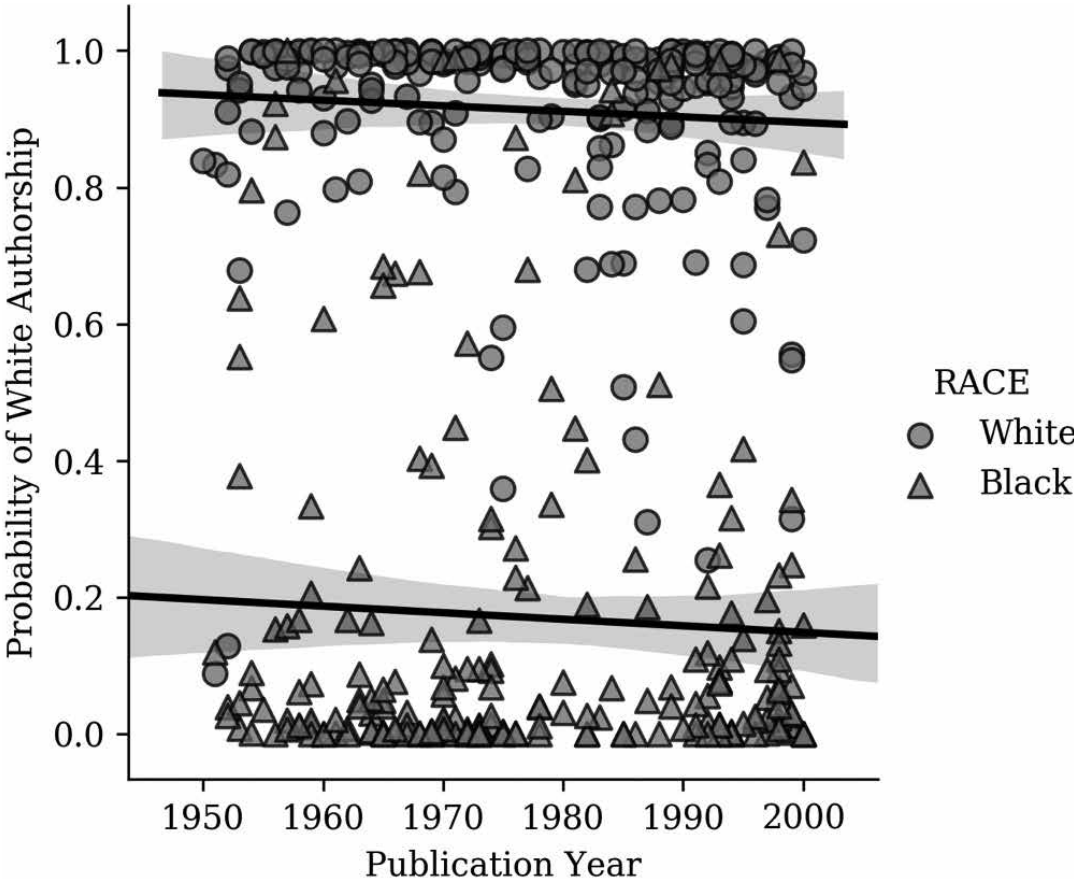

FIG. 1

A predictive model of authorial race, 1950–2000