What's in a Letter? Using Natural Language Processing to Investigate the Prevalence of Linguistic Biases in Teacher Letters of Recommendation for Postsecondary Applications

Brian Heseung Kim
University of Virginia
brian.kim@virginia.edu

Proposal Draft as of April 29, 2021[1]

Abstract

While scholars have already uncovered many ways that low-income, first-generation-to-college, and racial/ethnic minority students are systematically disadvantaged across the postsecondary application portfolio – from standardized tests to advanced course-taking opportunities – we know almost nothing about whether teacher letters of recommendation advance or impede these students' college aspirations. This blind spot is especially concerning given mounting evidence that recommendation letters in other contexts can contain biased language, that teachers can form biased perceptions of their students' abilities, and that narrative application components more generally may contribute to racial discrimination in selective college admissions. Meanwhile, institutions continue to move away from standardized test scores – positioning recommendations to be even more prominent going forward. In this proposal, I plan to conduct the first system-wide, large-scale text analysis of teacher recommendation letters in postsecondary applications. With application and recommendation data from 1.6 million students, 600,000 teachers, and 800 postsecondary institutions, I will examine the prevalence of "linguistic biases" within these letters: whether students are described by teachers in systematically different ways across racial/ethnic, gender, and socioeconomic groups. By combining rigorous econometric frameworks with sophisticated natural language processing (NLP) and text mining techniques, I can analyze variation in letter characteristics at unprecedented scale and fidelity while accounting for salient confounding factors like student academic and extracurricular qualifications. It is paramount that we better understand the role of these letters in ameliorating or exacerbating inequity, and these analyses will provide urgent insights for college admissions practices, affirmative action litigation, and text-as-data methodologies for education research.

[1] I am deeply grateful for the support of numerous individuals in this work: the tireless support and partnership of Don Yu, Mark Freeman, Preston Magouirk, Dave Tiss, Joe Kek, and other colleagues at the Common Application; the fantastic and generous mentorship of Benjamin L. Castleman throughout my doctoral studies; the kind guidance and feedback from Daphna Bassok, Brian Wright, Vivian C. Wong, Benjamin T. Skinner, Walter Herring, Amanda Lu, Kylie Anglin, and AJ Alvero; and the consistent support of my colleagues in the Center on Education Policy and Workforce Competitiveness (EdPolicyWorks) and the Nudge4 Solutions Lab. This work has been made possible with support from the National Academy of Education and Spencer Foundation Dissertation Fellowship Program, and the Institute of Education Sciences, grant #R305B140026, through the Virginia Education Science Training Pre-Doctoral Fellowship program.

I. Introduction

Researchers have identified numerous ways that low-income, racial/ethnic minority, or first-generation-to-college ("first-gen") students are systematically disadvantaged across nearly every stage of the postsecondary application process (Page & Scott-Clayton, 2016; Hoxby & Avery, 2013), yet one prominent component remains largely unexamined: the teacher recommendation letter. Every year, hundreds of thousands of teachers acrossthe country write a letter of recommendation in support of college-aspiring students. To illustrate, 40% of the 904 colleges using the ubiquitous Common Application ("Common App") require at least one teacher recommendation, alongside 40 of the top 50 national universities and 48 of thetop 50 national liberal arts colleges (as ranked by the U. S. News & World Report in 2019; author's calculations). Available evidence also suggests that these letters carry heft in admissions decisions: 57% of the admissions officers surveyed by the National Association for College Admission Counseling responded that teacher letters were of "considerable" or "moderate importance" in their decision-making – placing them above other salient factors like student extracurriculars, class rank, interviews, and even advanced placement exam scores (Clinedinst & Koranteng, 2017). Given that over 60% of baccalaureate institutions moved to test-optional application policies for 2020, admissions officers are likely to weigh these letters even more heavily going forward (FairTest, 2020; Rosinger et al., 2020).

Despite their rising prevalence and importance, there is very limited evidence on whether teacher recommendation letters ameliorate or exacerbate the obstacles that historically underserved student populations face on their way to college. On the one hand, letters and other "holistic" application components may provide admissions officers a more complete picture of students' assets, backgrounds, and potential contributions to college communities, perhaps offsetting reliance on historically inequitable test scores (Bastedo et al., 2019). On the other hand, the content of these letters may actually reflect and deepen existing inequities by presenting biased depictions of student character and ability. Researchers have previously found that K-12 teachers hold implicit racial biases (Starck et al., 2020) and adjust both their expectations and evaluations of students based on race/ethnicity and gender at other stages of the educational pipeline (Dee, 2005; Grissom & Redding, 2016). If the language, content, and tone of recommendation letters are influenced by such biases, the growing influence of letters in college admissions decisions may serve to further disadvantage racial/ethnic minority, female, and low-socioeconomic status ("SES") students in this increasingly competitive and high-stakes process.

In my dissertation, I will conduct the first system-wide, large-scale text analysis of teacher recommendation letters in postsecondary applications to dissect the role of these letters in supporting or stymieing college accessibility among historically underserved student populations. In partnership with the Common App, I will analyze the universe of teacher recommendation letters submitted via their platform between 2018-2019,[2] as well as the student applications the letters are associated with: a total of approximately 2.3 million unique recommendations and 1.6 million unique students across 800 postsecondary institutions. To make these immense text data tractable for analysis, I will combine rigorous econometric frameworks with sophisticated natural language processing ("NLP") and text mining techniques from the field of data science that facilitate the rapid and consistent coding of linguistic features within the letters themselves. I will then be able to robustly investigate the potential for systematic differences in **(a)** the specific words and phrases used to directly describe students, **(b)** the topical content of letters, and **(c)** the perceived positivity of letter tone, all while accounting for the rich set of student qualifications included in their

[2] Note that my data will also include letters from 2020. Given the exceptional circumstances of the year, I am likely to instead use those data as "training" data for my text analyses, per recommendations from Egami et al. (2018).

applications. I will further explore variation by teacher and school characteristics linked to the prevalence of racial and gender bias. More formally, my guiding research questions are:

1. Are there systematic differences in the **(a)** word choice, **(b)** topical content, and **(c)** tone that teachers use to describe students of varying races/ethnicities, genders, and socioeconomic statuses in their college recommendation letters?
2. Are these systematic differences in letters moderated by observable teacher characteristics linked to bias, such as subject taught, experience in letter writing, and number of courses with the student?
3. Are these systematic differences in letters moderated by school characteristics linked to bias, such as sector, student demographics, standardized testing performance, and regional implicit bias scores?

II. Contribution and Significance

To the best of my knowledge, this will be the first study on teacher recommendation letters spanning more than a single university or state network, offering valuable and timely insight for policymakers, practitioners, and researchers into the system-wide dynamics of implicit bias within these letters (which I will refer to in this manuscript with the shorthand, "linguistic bias") that have previously been infeasible to examine. There currently exists only mixed evidence about the prevalence of linguistic bias in recommendation letters from smaller-scale studies in this context (n=4,792, Akos & Kretchmar, 2016; n=24,000, Schwarz, 2016) and other, related contexts (e.g. job applications in academia and STEM from Madera et al., 2018, or student college application essays from Alvero et al., 2021). And given the accelerating growth of the test-optional and "holistic admissions" movements, as well as on-going, high-profile litigation around affirmative action in admissions where bias in recommendation letters is at issue (e.g. pending U.S. Supreme Court consideration of *Students for Fair Admissions v. Harvard*, 2019; Hong, 2018), the need for more systematic and definitive study in this arena is urgent. It may be the case that letters show no evidence of linguistic biases, making it prudent to encourage greater use of these letters versus more inequitable elements such as GPA and standardized test scores (Kane, 1998). But if they do seem to exhibit bias, we must explore how to mitigate such issues and interrogate the ethics of letter use in admissions practices going forward.

My dissertation will also advance the interdisciplinary literature on linguistic bias, contribute to the development of robust methodological frameworks for quantifying bias using NLP and text mining, and demonstrate the immense promise of these methods for education policy research. My data likely represent the largest known repository of recommendation letters linked to individual qualifications *and* recommender characteristics in *any* context (e.g. graduation school admissions, job applications). Paired with recent advances in the robustness of NLP techniques and the accessibility of high-performance computing, I will be able to explore the phenomena of linguistic bias within evaluation contexts at unprecedented scale, scope, and consistency. Given the novelty of these methods in education research and the delicate nature of this work, my hope is to serve as an exemplar for the effective and responsible application of NLP by transparently modeling the procedures, validity checks, and theoretic frameworks necessary for success.

III. Literature Review

IIIa. Racial and Gender Bias in Evaluations of Students

Central to my study of bias in recommendation letter writing is the prevalence of implicit bias among teachers. Under the framework of implicit bias, the unconscious associations and stereotypes that an individual holds about particular groups go on to shape their perceptions and judgments of people within those groups (Bertrand et al., 2005; Greenwald & Krieger, 2006); the result is an insidious form of discrimination that can be hard to identify as those involved are often unaware it is happening (Devine et al., 2012). Empirical work has shown that these biases are difficult to intervene upon even in the best of circumstances (Lai et al., 2014), and they are particularly impactful on judgment and decision-making in circumstances where individuals are rushed (Payne, 2006), fatigued (Ma et al., 2013), or distracted (Danziger et al., 2011).

Classroom environments often align with these conditions, and the literature indicates that racial and gender biases may indeed be prevalent among teachers. First, we have evidence that teachers can hold negative implicit associations of Black individuals (Starck et al., 2020), and also that the degree of negativity varies by geography, school demographic composition, and regional test score disparities (Chin et al., 2020). We also have evidence that such biases can be consequential: math teachers with more negative stereotypes about female students were more likely to encourage female students to pursue vocational tracks instead of scientific/academic ones (Carlana, 2019) and even grade their exams more harshly (Avitzour et al., 2020). Research on demographic matching between teachers and students also offers additional evidence for the prevalence of such biases: white teachers held systematically lower expectations for Black students than Black teachers of the *same* students (Gershenson, Holt & Papageorge, 2016), with similar patterns for gender-matching (Dee, 2005).

IIIb. Racial and Gender Bias in Recommendation Letters

While we have strong evidence for the prevalence of biased attitudes and expectations among educators, we don't yet have a clear consensus on whether and how these biases manifest in letters of recommendation. There is suggestive evidence that female applicants to STEM-related academic positions are described with less exemplary adjectives and phrases (e.g. "good" versus "phenomenal") after controlling for observable qualifications (n=880; Schmader et al., 2007) and less positivity in general (n=1,224; Dutt et al., 2016). I refer to these phenomena as letter "**tone**." Similarly, researchers have found female applicants can be described with more "communal" terms (e.g. "team player," "helpful") and fewer "agentic" terms (e.g. "leader," "pioneer") than male counterparts (n=624, Madera et al., 2009). Letters for female applicants are also more likely to have "doubt-raising" language: language that is outright negative, preceded by hedging, or irrelevant (Madera et al., 2018). I refer to these phenomena as letter "**word choice**."

Importantly, however, these studies all note that letters were broadly more similar than different, and they often failed to find anticipated differences along other measures. For example, a study of teacher recommendation letters for a single postsecondary institution found that female students had letters that were *more* positive, while racial minority students had letters that were more neutral. But when looking at whether teachers focus on describing different student characteristics across groups (e.g. athletic ability versus intellectual curiosity), the author found very few differences (n=24,000; Schwarz, 2016). I refer to these phenomena as letter "**topical content**." In a similar study at a different institution, researchers found that that female students were more likely, and racial minority students less likely, to have letters using "grindstone" words (e.g. "hardworking," "diligent") – but found no meaningful differences along any other conceptually-relevant categories like achievement (n=4,792; Akos & Kretchmar, 2016).

While research findings on this subject have been mixed, the lack of consensus is likely attributable to variation in analytic methods and settings. Much of the aforementioned evidence comes from smaller-scale studies of letters at individual institutions, so the lack of consistency may reflect lack of statistical power, sample idiosyncrasies,and other contextual differences (e.g. institutional selectivity). Moreover, they deploy a range of text analysis methodologies that trade nuance against scalability, from subjective ratings of letters by trained readers to simple word-count analyses. These studies provide crucial groundwork to identify the ways linguistic bias may manifest in letters (tone positivity, word choice, and topical content), but they also point to the need for studies that can at least partially overcome the steep tradeoff between scalability/generalizability and analytic nuance. A growing literature in education has demonstrated the utility of NLP methods to this end (Anglin, 2019; Fesler et al., 2019), motivating and guiding my application of these approaches for the present study.

For concision, I synthesize the insights of this literature review together into a unified theoretical model for my specific context in Appendix VIIIa.

IV. Data and Sample

My primary data consist of all applications submitted through the Common App platform between 2018-2019. Through several years of collaboration with researchers and leadership at Common App, we have constructed a robust and exhaustive data system that consolidates de-identified versions of all application materials submitted by students, teachers, and counselors via their portal.[3] For the present study, I will limit my sample to first-time applicants (excluding transfers) who submitted at least one teacher recommendation. About 800,000 such students submit an application through the Common App each year, yielding a total of about 1.6 million students and 2.3 million teacher recommendations. These data contain every field self-reported by students in their applications: gender, race/ethnicity, socio-economic status (parental education, low-income status, and benefit receipt), colleges applied to, academic performance (GPA, SAT/ACT), extracurricular involvement (open text response), and high school. These data also contain every field self-reported by teachers in each of their recommendations -- current school, courses with the student, and recommendation letter text -- as well as the total count of recommendations submitted by each teacher, each year, to proxy for their letter writing experience within and beyond the data timeframe. Importantly, I do not observe teacher demographics; as such, I intend to explore teacher-student demographic match dynamics using supplementary state staffing data in future work.[4]

To access school-level and institution-level characteristics, I will link the Common App data to:

[3] In order to filter personally identifiable information (PII) from the text of the recommendation letters, I collaborated with researchers at the Common App to algorithmically search and replace PII text prior to my analysis. Based on the design of the algorithm, this should not have any substantive repercussions for my analysis or its interpretation. Regardless, I will examine the text data closely for any issues related to this process.

[4] While theoretically possible to impute certain teacher characteristics from teacher names, such as gender and race, attempts to algorithmically derive individual demographics in this way carries with it a number of deep ethical and moral concerns (e.g. that advancing such technology could eventually facilitate the systematic discrimination of individuals by their race/ethnicity if utilized by bad actors – or simply uncritically). This concern is especially pressing in my circumstance, as I intend for my work to serve as a model for future researchers, and I plan to open-source my code to the full extent practicable. Even from a straightforward methodological perspective, it is not possible with the provided data to assess the accuracy and potential biases in this prediction process, making it difficult to assess the confidence of any ensuing conclusions. For the totality of these reasons together, I will not attempt imputation analysis of this kind, and will hold this strand of demographic match analyses until supplementary data sources make these analyses possible without imputation.

- *Public/Private Secondary School Universe Survey from the NCES Common Core of Data*for school-level data on student enrollment, student demographics, and school geography.
- *Integrated Postsecondary Education Data System from the U.S. Department of Education*for institution-level data on selectivity, student demographics, and geography.
- *County-Level Educator Implicit Association Test Scores from Chin et al. (2020)* to analyze linguistic biases by regional educator bias score brackets.
- *Stanford Education Data Archive from Reardon et al. (2019)* to analyze linguistic biases by school-level standardized test performance brackets (on the NAEP, specifically).

V. Analytic Approach

My analysis can be divided into two distinct phases: **(1)** deriving quantitative measures of letter characteristics with NLP and text mining, and **(2)** analyzing these letter characteristics in a regression framework.

Va. Text Analysis

In the first phase, I will analyze the text of the teacher recommendation letters using NLP and text mining techniques to construct a series of numeric measures that reflect their characteristics. This will ultimately produce output similar to approaches where researchers manually code text for the incidence of certain phenomena or characteristics, but in a mostly automated and highly scalable way. That said, these text-as-data methods are no substitute for rigorous qualitative document analysis – rather, they offer distinct and complementary insights at a scope particularly appropriate for research questions like my own.

I propose a sequence of three separate NLP techniques to characterize each letter along lines likely to exhibit linguistic biases per my literature review: **(a) Word Frequency Analysis**, which will tabulate the specific words used to *directly* describe students in each letter (word choice); **(b) Topic Modeling**, which will quantify the extent to which each letter discusses various categories of topical content (e.g. coursework, sports, leadership); and **(c) Sentiment Analysis**, which will quantify the estimated tone and positivity of language used in each letter. Because **(a) Word Frequency Analysis** produces high-dimensional output unsuitable for my main regression model, I consider this an exploratory analysis and describe it in Appendix VIIIb for concision.

Through **(b) Topic Modeling**, I will be able to measure differences in the topics that teachers discuss in each letter. In this framework, words that frequently appear together in the same paragraphs of letters across the dataset are thought to be associated with the same abstract "topic" of discussion, which analysts then interpret for their unifying substantive meaning (Blei, 2003).[5] For example, if "leadership," "competition," and "sports" frequently appeared together in paragraphs throughout the dataset, the algorithm would identify them as belonging to the same topic; an analyst might subjectively interpret this word group as representing the substantive topic of "student sports involvement." Importantly, the algorithm identifies several such topics *based on the provided text data* – this allows for uniquely context-sensitive and flexible output compared to methods using predefined

[5] In more technical terms, I am treating a paragraph, not a letter, as a "document" for training the topic model. This is because these letters are all roughly discussing the same broad topic (students and their academic qualifications), and variation at the letter-level will likely be insufficient to identify meaningfully distinct topics. By contrast, paragraphs within the letter are more likely to be varied in topic; e.g. a teacher might write about a student's achievements in a particular class in one paragraph, and then write about a student's involvement in school leadership in the next. A similar adjustment was helpful in improving the interpretability of topical output in Kim et al. (2021). Note also that it is often prudent to include n-gram analysis in topic models, e.g. 2-3 word phrases, in addition to individual words. For clarity of explanation, I use "words" as a stand-in for n-gram for the rest of this document.

word groups (e.g. the LIWC as used in Alvero et al., 2021), but also means that the substantive topics identified by the algorithm are not knowable before analysis. Once the topics are identified, each *letter* can be quantified for the number of words spent discussing each topic, based on the combination of words within each of its constituentparagraphs.[6] To improve the tractability of the topic model output for later regression analyses, I will use a multi-coder process to consolidate them into a smaller number of substantively-relevant "supertopics," thereby reducing their dimensionality (piloted in Kim et al., 2021).[7]

Two salient limitations of this method are that it is highly sensitive to idiosyncrasies in the data, and topic interpretations are ultimately quite subjective. Standard practice in the field is to procedurally adjust the number of topics the algorithm identifies in the data to maximize conceptual cohesion or model fit in the resulting topic groups – only then does the analyst interpret the substantive meaning for each topic. This data-driven approach means that it is not possible to state a priori how many topics I will create, nor which of the identified topics will be most theoretically relevant to linguistic bias. In Appendix VIIIc, I discuss my plan to reduce the subjectivity of this analysis, improve the robustness of my topic interpretations, and evaluate the accuracy of the paragraph topic classifications themselves.

Through **(c) Sentiment Analysis**, I will be able to measure systematic differences in the occurrence (e.g. "He was in my class last year" v. "He thrived in my class last year") and degree (e.g. "good student" v. "fantastic student") of positivity in recommenders' writing. While sentiment analysis is actually a family of related approaches, most modern algorithms work similarly (Vasawani et al., 2017). First, the algorithm "learns" basic linguistic relationships and structure by ingesting enormous quantities of text data and attempting to statistically derive the common mechanics and word relationships for a language. After establishing this language schema and generating a series of context-dependent "definitions" for each word within its vocabulary,[8] the algorithm is then "fine-tuned" to interpret entire sentences for their perceived positivity, negativity, or neutrality using databases of human-generated crosswalks as its point of reference.[9] Ultimately, each sentence can be measured for its positivity on a five-point scale (-2 for very negative, to +2 for very positive), and

[6] In this study, I will leverage the implementation by Roberts et al. (2019) in R known as Structural Topic Modeling (the "stm" package). This implementation offers a number of important methodological advancements over the standard Latent Dirichlet Allocation (Blei, 2003), most notably the ability to allow the prior distribution of topic prevalence in the data to vary based on document metadata and improve the overall fit of the model. I will thus use the array of covariates present in my regression model as "topic prevalence covariates" in the stm model; it is recommended by Roberts et al. (2019) that these models be congenial in terms of covariate inclusion. Intuitively, this risks "baking in" an anticipated relationship (e.g. if student race is allowed to influence the topic measurement process, and we observe a relationship in a certain topic's prevalence and student race later on); I remark on my approach to overcoming this issue towards the end of this section using an adjustment proposed by Egami et al. (2018).

[7] Mathematically speaking, the topic modeling algorithm will output the calculated proportion of words in each letter that come from each topic – derived in a probabilistic manner using Bayesian methods under the hood. Supertopics then represent a simple shorthand for signifying the proportion of words in each letter that come from constituent topic A, *OR* constituent topic B, *OR* constituent topic C, and so on. This form of transformation may be susceptible to issues of multi-modal output in the topic modeling algorithm; thus, I will need to assess the extent to which this is an issue in the data before doing so.

[8] The newest methods on this front, "transformer" neural networks (Vaswani et al., 2017), are unique in their ability to incorporate context before *and* after each word it examines, thus allowing it to change its "understanding" of a given word based on the exact sentence in which it appears. This allows the transformer family of algorithms to understand that the word "bank" in "I went to the bank to cash a check" is different in meaning than in "I sat by the bank and watched the water flow by."

[9] The Stanford Sentiment Treebank ("SST"; Socher et al., 2013) is the most common crosswalk for such a purpose. In brief, they "crowd-sourced" human-generated ratings of word, phrase, and sentence positivity/negativity on a five-point scale using Amazon Mechanical Turk. Unfortunately, their definition for positivity and negativity was intentionally vague and left up to scorers' interpretation, and I was unable to find inter-rater reliability metrics for the source data.

each letter can then be assigned an aggregate value (i.e. total count of very positive sentences, total count of positive sentences, etc.) based on these sentence-level values. Per Kim et al. (2021), I operationalize the definition of sentiment as the *perceived positivity of emotions and ideas presentin a given text*. This definition then is a conglomeration of the writer's stated emotions ("Her lack of effort saddened me..." versus "I was excited to hear..."), communicated intention ("I wish her the best!" v. "I wish she'd try harder"), and, at least to some extent, topical content ("financial hardship" v. "class valedictorian").

These modern approaches to the task allow the algorithm to better interpret more complex linguistic features like negation (e.g. "not bad"), multiple word meanings (e.g. "I *ran* to the mall" versus "I *ran* the code") and the subjunctive tense (e.g. "I wish this were good"), but trade off in transparency and potential for algorithmic bias (see Shah et al., 2020, for a helpful conceptual review). Thus, in Appendix VIIId, I describe: several ways in which I plan to pre-process the text data and *prevent* undue influence of algorithmic bias; the methods by which I intend to *assess* the prevalence of any lingering algorithmic bias in sentiment output (piloted in Kim et al., 2021); my intention for measuring whether the algorithm output exhibits construct validity; and how I weigh the several available methods for sentiment analysis against one another.

Finally, per recommendations from Egami et al. (2018), I will refine my approach to applying each of the aforementioned methods using a random subset of 30% the overall letter data, to then apply and use these methods (without further adjustment) on the remaining 70% of data for use in my ensuing regression analyses.[10] It is conceptually valuable for an researcher to be able to adjust their analytic approach in response to the specific data at hand given how sensitive these methods are to training data and cleaning decisions; Egami et al. (2018) thus recommend this methodological adjustment to preserve this important experimentation process while ensuring that analysts don't inadvertently (or intentionally) "bake in" an anticipated effect by tuning data cleaning and algorithm training steps accordingly. To minimize reductions in the degree of common support in my sample (discussed in more detail in the next section) due to this adjustment, I will randomly sample *teachers* (stratified by region, subject, number of letters written, and average school academic performance on the NAEP) from the overall dataset into the training and validation sets, incorporating all of their letters and students accordingly. I will then also be able to assess training and test data balance on said teacher covariates, as well as corresponding student covariates and letter characteristics.[11]

Vb. Regression Analysis

For the second phase of my analysis, my conceptual goal is to compare the same teacher's writing about two students whose only substantive difference is their race/ethnicity, SES, or gender, to reveal linguistic biases across these groups. That in mind, I use my NLP-derived measures as the outcomes for a series of teacher fixed effects regressions to examine whether the measures – and the writing of the letters by proxy – vary systematically by student race/ethnicity, SES, and gender, even after controlling for additional student characteristics and qualifications that would likely influence letter characteristics. Applying teacher fixed effects will allow me to compare letter characteristics among the group of students for whom a single teacher has written recommendations, thus controlling for any teacher characteristics that are fixed within teachers (e.g. teacher subject area, school characteristics/culture, writing ability, etc.). For student-level controls, I use the intuitive

[10] This is only a tentative split for now. Reducing the proportion allocated for training may reduce the robustness of the ensuing algorithmic output, while increasing it (and thus reducing the proportion allocated for testing/estimation) will

[11] Any differences in teacher or student characteristics would be indicative of a failure in the random train-test splitting process. Any differences in letter characteristics would be indicative of the training set processes being, for some reason, inapplicable or unsuitable for the test set. That latter concern will be of the utmost importance to examine when determining whether this train-test split proportion (30%-70%) is appropriate.

covariates like GPA that are included in each student's application, but I will also use the selectivity and number of colleges applied to ("college application profile") that Dale and Krueger (2002; 2011) showed to be strong proxies for student academic ability.

Equation 1 represents my main regression specification for RQ1. Y_{it} represents any one of the NLP-derived letter characteristics (e.g. number of very positive sentences, or number of words spent discussing a given supertopic) from the writing of teacher *t* for student *i*. λ_t represents the vector of teacher fixed effects, τ_i is the vector of student cohort-year indicators, C_{it} represents the total count of sentences (for sentiment analysis measures) or words (for topic modeling measures) in each letter, and X_i represents the vector of student covariates: GPA, SAT/ACT scores, extracurricular involvement (type and amount), leadership roles, additional student demographic characteristics (e.g. immigrant and military status), and college application profile. ε_{it} represents the idiosyncratic error term.

$$(1) \quad Y_{it} = \lambda_t + \tau_i + X_i + C_{it} + \beta_1 R_i + \beta_2 SES_i + \beta_3 F_i + \varepsilon_{it}$$

My coefficients of interest will be betas 1-3 on R_i, SES_i, and F_i: indicators for each race/ethnicity category, low-SES status, and female-identifying students. I can interpret these as the observed difference in letter characteristic Y (e.g. number of very positive sentences) on average for each student group after controlling for other student characteristics to the greatest extent possible. Importantly, the use of many covariates and granular demographic groups may affect the common support for estimating each coefficient of interest; I discuss the repercussions of this concern in more detail in Appendix VIIIe.

The coefficients from this approach will represent a *weighted average* of linguistic bias across teachers due to the mechanics of fixed effects regressions, but it is likely that these biases also vary meaningfully by teacher and context (Chin et al., 2020). To explore whether observable teacher and school characteristics moderate any observed biases for RQ2 and RQ3, I will re-run Equation 1 on subsamples of the student-teacher pairs to compare the coefficients on R_i, SES_i, and F_i across them. For example, to explore linguistic bias by teacher letter writing experience, I will examine the distribution of total letters written by each teacher and split the sample at the median to identify high- and low-experience teachers. After running Equation 1 on each subsample, I can assess significant differences for my coefficients of interest. For teacher characteristics (RQ2), I will examine subsample analyses by: subject taught (STEM, humanities, other), experience in writing recommendations (median split), and number of courses with the individual student (median split). For school characteristics (RQ3): proportion of racial/ethnic minority students (median split), school urbanicity (urban/rural), NAEP performance (median split), sector (public/private/charter), and regional educator implicit bias scores (median split). That said, the aforementioned specifications will likely need to be adjusted as appropriate pending a better understanding of how these variables are distributed in the data.

Vc. Limitations

While these data, alongside recent advances in NLP methods, offer me the opportunity to examine this research question with uncommon scope and comprehensiveness, there are still a number of meaningful limitations to this approach that should be acknowledged.

First, my results are descriptive in nature. Despite my ability to include many compelling covariates as controls, there will almost always be additional confounders in observational data without a strong quasi-experimental design. Moreover, we know that several critical features of the data generating process in this context make a true apples-to-apples comparison across student-teacher pairs intractable, as we are fundamentally trying to control for multi-dimensional, longitudinal relationships between students and their teachers. In other words, the imperfect proxies I use to describe students, teachers, and their relationship in these data may inadvertently mask or misestimate relevant dynamics that I would otherwise have assumed I captured.

Second, my analysis focuses entirely on identifying systematic bias at the stage of letter writing. There are importantly many inequities and biases at other points in the educational process that contribute to differences in the control variables I utilize (e.g. GPA, extracurriculars, SAT/ACT scores) and selection into this sample of college-aspiring students. My intention is not to ignore these important disparities, but to present the best estimates I can on this individual margin as my contribution to a broader discourse on gender, class, and race in our society.

Third, my exact sample for estimating this letter writing bias is quite specific. That is, I am estimating the systematic disparities in letter content across student demographics *only* among the group of students who meet all of the following criteria:

1. Applied to postsecondary institutions in the U.S. using the Common Application[12]
2. Applied to institutions that required at least one teacher recommendation[13]
3. Successfully obtained at least one teacher recommendation[14]
4. Received a recommendation from a teacher who has written recommendations for students of multiple demographics from 2018-2019 (i.e. fixed effects common support)
5. Successfully submitted a completed application to at least one institution

I intend to explore the repercussions of **2** and **3** using these data in separate work, but until then, we should be cognizant about the potentially idiosyncratic sample of student-teacher pairs here. Arguably, my data still represent the relevant population of students well, and it remains the best available source of data to analyze in these pursuits.

Fourth, my reliance on algorithmic NLP techniques means I am unlikely to reveal all forms of bias potentially present in the letters. While researchers have made great strides incorporating word context and sentence composition in topic modeling and sentiment analysis algorithms, there will always be exceptions and fringe circumstances in language that cannot be accounted for with these approaches.[15]

Fifth, this analysis assumes that the self-reported race, class, and gender of students are salient and perceptible to teachers. It is likely that many students who self-identify in particular ways may be misperceived by their teachers, resulting in a sort of measurement error in my demographic variables (e.g. student reports being Black, but teacher actually perceives white). This has repercussions not just for attenuating my estimates, but also for my interpretation – I am technically estimating the dynamics of teacher *perceptions of* student demographics, not actual student demographics.

[12] Knight & Schiff (2019) find that institutions accepting the Common App tend to be more selective.

[13] My initial descriptive analyses show that, of institutions accepting the Common App, those requiring teacher recommendations are even *more* selective. I argue these institutions remain broadly relevant given concerted efforts across stakeholder groups to improve college access to such selective institutions, especially among low-income and minority students (Hoxby & Turner, 2014; Page & Scott-Clayton, 2016).

[14] This criteria actually stands in for two important dynamics. The teacher recommendation process is logistically complex, requiring substantial planning and coordination on behalf of the students. Further, this recommendation process requires a two-sided match: students must select a teacher, and the teacher must accept. The students who can then successfully obtain even one letter are likely to be meaningfully different in many ways from students who cannot.

[15] For example, none of my approaches would be able to *reliably* capture something subtle like group attribution – a teacher diffusing a student's individual successes across a broader group. E.g. "Brian excelled in my math class" versus "The groups Brian worked with excelled in my math class."

Lastly, I cannot comment on the extent to which any measured disparities in letter language would contribute to disparities in actual admissions outcomes for students. It could very well be the case that linguistic bias exists, but it has a negligible (or, unintuitively, a positive) impact on college admissions. For now, this is a descriptive exploration that may motivate such research in the future.

VI. Conclusion

With this robust dataset and thoughtful application of NLP methods, even given my aforementioned limitations, I hope to offer the most comprehensive evidence regarding linguistic bias in teacher recommendation letters to date – examining letters across student demographics, teacher characteristics, and school contexts. As no dataset of this kind has existed in the United States until just recently, this research endeavor provides an unprecedented chance to better understand the dynamics of linguistic bias in college admissions. As these letters grow in prevalence and importance, they represent both an opportunity and threat to equity in college admissions and to students' college aspirations, nationwide. This in mind, I hope my dissertation research will provide scholars, practitioners, and policymakers urgent insights for grappling with their use going forward.

VII. References

Akos, P., & Kretchmar, J. (2016). Gender and Ethnic bias in Letters of Recommendation: Considerations for School Counselors. *Professional School Counseling*, *20*(1), 1096-2409-20.1.102. https://doi.org/10.5330/1096-2409-20.1.102

Alvero, A., Giebel, S., Gebre-Medhin, B., antonio, anthony lising, Stevens, M. L., & Domingue, B. W. (2021). *Essay Content is Strongly Related to Household Income and SAT Scores: Evidence from 60,000 Undergraduate Applications* (No. 21–03; CEPA Working Papers). Stanford Center for Education Policy Analysis. https://cepa.stanford.edu/content/essay-content-strongly-related-household-income-and-sat-scores-evidence-60000-undergraduate-applications

Anglin, K. L. (2019). Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. *Journal of Research on Educational Effectiveness*, *12*(4), 685–706. https://doi.org/10.1080/19345747.2019.1654576

Avitzour, E., Choen, A., Joel, D., & Lavy, V. (2020). *On the Origins of Gender-Biased Behavior: The Roleof Explicit and Implicit Stereotypes* (SSRN Scholarly Paper ID 3692175). Social Science Research Network. https://papers.ssrn.com/abstract=3692175

Bastedo, M. N., Glasener, K. M., Deane, K. C., & Bowman, N. A. (2019). Contextualizing the SAT: Experimental Evidence on College Admission Recommendations for Low-SES Applicants. *Educational Policy*, 0895904819874752. https://doi.org/10.1177/0895904819874752

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit Discrimination. *The American Economic Review*, *95*(2), 94–98.

Blei, D. M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 30.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics*, *134*(3), 1163–1224. https://doi.org/10.1093/qje/qjz008

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., & Lovison, V. S. (2020). Bias in the Air: A Nationwide Exploration of Teachers' Implicit Racial Attitudes,Aggregate Bias, and Student Outcomes: *Educational Researcher*. https://doi.org/10.3102/0013189X20937240

Clinedinst, M., & Koranteng, A.-M. (2017). *2017 State of College Admission* National Association for College Admission Counseling. https://www.nacacnet.org/globalassets/documents/publications/research/soca17final.pdf

Dale, S. B., & Krueger, A. B. (2002). Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables. *The Quarterly Journal of Economics*, *117*(4), 1491–1527. https://doi.org/10.1162/003355302320935089

Dale, S., & Krueger, A. B. (2011). *Estimating the Return to College Selectivity over the Career Using Administrative Earnings Data* (Working Paper No. 17159). National Bureau of Economic Research. https://doi.org/10.3386/w17159

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(17), 6889–6892. https://doi.org/10.1073/pnas.1018033108

Dee, T. S. (2005). A Teacher like Me: Does Race, Ethnicity, or Gender Matter? *The American Economic Review*, *95*(2), 158–165.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*(6), 1267–1278. https://doi.org/10.1016/j.jesp.2012.06.003

Dutt, K., Pfaff, D. L., Bernstein, A. F., Dillard, J. S., & Block, C. J. (2016). Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature Geoscience*, *9*(11), 805–808. https://doi.org/10.1038/ngeo2819

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). How to Make Causal Inferences Using Texts. *ArXiv:1802.02163 [Cs, Stat]*. http://arxiv.org/abs/1802.02163

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as Data Methods for Education Research. *Journal of Research on Educational Effectiveness*, *12*(4), 707–727. https://doi.org/10.1080/19345747.2019.1634168

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, *52*, 209–224. https://doi.org/10.1016/j.econedurev.2016.03.002

Greenwald, A. G., & Krieger, L. H. (2006). Implicit Bias: Scientific Foundations. *California Law Review*, *94*(4), 945–967. JSTOR. https://doi.org/10.2307/20439056

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Grissom, J. A., & Redding, C. (2016). Discretion and Disproportionality: Explaining the Underrepresentation of High-Achieving Students of Color in Gifted Programs. *AERA Open*, *2*(1), 2332858415622175. https://doi.org/10.1177/2332858415622175

Hong, M. K. and N. (2018, October 16). Harvard Cites Weaker Teacher Recommendations for Asian-American Applicants. *Wall Street Journal*. https://www.wsj.com/articles/harvard-cites-weaker-teacher-recommendations-for-asian-american-applicants-1539721051

Hoxby, C., & Avery, C. (2013). The Missing "One-Offs": The Hidden Supply of High-Achieving, Low-Income Students. *Brookings Papers on Economic Activity*, *2013*(1), 1–65. https://doi.org/10.1353/eca.2013.0000

Kane, T. J. (1998). Racial and Ethnic Preferences in College Admissions. *Ohio State Law Journal*, *59*(3), 971–996.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., … Nosek, B. A. (2014). Reducing implicit racial preferences: A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Ma, D. S., Correll, J., Wittenbrink, B., Bar-Anan, Y., Sriram, N., & Nosek, B. A. (2013). When Fatigue Turns Deadly: The Association Between Fatigue and Racial Bias in the Decision to Shoot. *Basic and Applied Social Psychology*, *35*(6), 515–524. https://doi.org/10.1080/01973533.2013.840630

Madera, J. M., Hebl, M. R., Dial, H., Martin, R., & Valian, V. (2018). Raising Doubt in Letters of Recommendation for Academia: Gender Differences and Their Impact. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-018-9541-1

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591–1599. https://doi.org/10.1037/a0016539

Meet Our Members. (2018). Coalition for College. http://www.coalitionforcollegeaccess.org/members-new.html

National Liberal Arts College Rankings. (2019). US News & World Report. https://www.usnews.com/best-colleges/rankings/national-universities

National University Rankings. (2019). US News & World Report. https://www.usnews.com/best-colleges/rankings/national-universities

Page, L. C., & Scott-Clayton, J. (2016). Improving College Access in the United States: Barriers and Policy Responses. *Economics of Education Review*, *51*, 4–22. https://doi.org/10.1016/j.econedurev.2016.02.009

Payne, B. K. (2006). Weapon Bias: Split-Second Decisions and Unintended Stereotyping. *Current Directions in Psychological Science*, *15*(6), 287–291. https://doi.org/10.1111/j.1467-8721.2006.00454.x

Penner, E. K., Rochmes, J., Liu, J., Solanki, S. M., & Loeb, S. (2019). Differing Views of Equity: How Prospective Educators Perceive Their Role in Closing Achievement Gaps. *The Russell Sage Foundation Journal of the Social Sciences*, *5*(3), 103–127. https://doi.org/10.7758/RSF.2019.5.3.06

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., Jang, H., Chavez, B, Buontempo, J., & DiSalvo, R. (2019). *Stanford Education Data Archive (Version 3.0)*. Stanford University. http://purl.stanford.edu/db586ns4974

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02

Rosinger, K. O., Ford, K. S., & Choi, J. (2020). The Role of Selective College Admissions Criteria in Interrupting or Reproducing Racial and Economic Inequities. *The Journal of Higher Education*, *0*(0), 1–25. https://doi.org/10.1080/00221546.2020.1795504

Schaeffer, B. (2020, August 12). *Three-Fifths of Four-YearColleges and Universities Are Test-Optional for Fall 2021 Admission; Total of Schools Not Requiring ACT/SATExceeds 1,450*. FairTest: The National Center for Fair and Open Testing. https://www.fairtest.org/threefifths-fouryear-colleges-and-universities-are

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants. *Sex Roles*, *57*(7), 509–514. https://doi.org/10.1007/s11199-007-9291-4

Schwarz, J. D. (2016). *Lost in Translation: Elite College Admission and High School Differences in Letters of Recommendation* [Dissertation Manuscript].

Scott-Clayton, J. (2016). *Early Labor Market and Debt Outcomes for Bachelor's Degree Recipients: Heterogeneity by Institution Type and Major, and Trends Over Time* (CAPSEE Working Papers, p. 38). Center for Analysis of Postsecondary Education and Employment. http://ccrc.tc.columbia.edu/media/k2/attachments/early-labor-market-debt-outcomes-bachelors-recipients.pdf

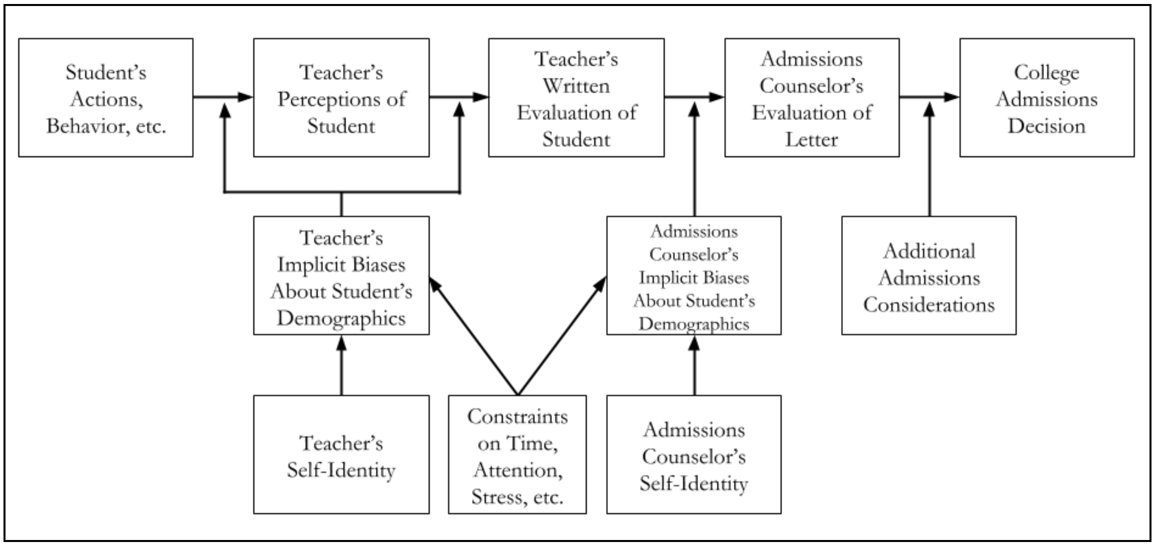
Starck, J. G., Riddle, T., Sinclair, S., & Warikoo, N. (2020). Teachers Are People Too: Examining the Racial Bias of Teachers Compared to Other American Adults: *Educational Researcher*. https://doi.org/10.3102/0013189X20912758

VIII. Appendix

VIIIa. Theoretical Model for Bias in Teacher Recommendations

Drawing from the work I reviewed in Section III, I created a theoretical model in Figure A1 below to concisely articulate my understanding of the role that implicit bias likely plays in the teacher recommendation process. The first row across the top is a simplified flow diagram of how student actions are eventually translated into the letter that admissions counselors review. That is, we begin with a bundle of student actions, interactions, and behaviors that exist in reality. These behaviors are then translated into some interpretation of the student by the teacher; the reality becomes subjective perception and selective memory. The teacher will then transcribe some version of their subjective perception of the student into the text of a recommendation letter. Lastly, the letter is interpreted by the admissions counselor, whose evaluation of the letter is utilized alongside a complex host of additional information and context into an actual admissions decision (Clinedinst & Koranteng, 2017; Schwarz, 2016).

Figure A1. Theoretical Model for Bias in Teacher Recommendations


Student's Actions, Behavior, etc.
Teacher's Perceptions of Student
Teacher's Written Evaluation of Student
Admissions Counselor's Evaluation of Letter
College Admissions Decision
Teacher's Implicit Biases About Student's Demographics
Admissions Counselor's Implicit Biases About Student's Demographics
Additional Admissions Considerations
Teacher's Self-Identity
Constraints on Time, Attention, Stress, etc.
Admissions Counselor's Self-Identity

Importantly, teacher bias can conceptually influence this process at two distinct points: the formation of teacher perceptions about the student's actions, and then the transcription of those perceptions into written language. To illustrate this distinction, imagine a concrete student behavior, such as top performance in a science class taught by the recommending teacher. The work of Akos & Kretchmar (2016) and Madera et al. (2009) would suggest that this behavior, when eventually translated into a letter for a female student, is more likely to be written as the result of her *hard work* instead of some innate ability or *agentic* drive. The extent to which this dynamic occurs hinges on the teacher's own implicit biases about female students in the sciences (Carlana, 2019), which in itself is influenced by the teacher's self-identity as well (Dee, 2005; Gershenson, Holt, & Papageorge, 2016). In this circumstance, the biased language about the student could actually be the result of three distinct scenarios:

I. It could be that a teacher sees a high-performing female student in their science class and, immediately from that stimulus, stores a bias-contaminated perception of the student (*Oh wow, she's working so hard to succeed*). The writing of the letter is then just a "perfect" translation of that already bias-contaminated perception of the student, such that no further bias is introduced in the act of writing ("She works hard to succeed").

II. It could be that a teacher actually stores a bias-free perception of the student from the stimulus (*Oh wow, she is brilliant and performing excellently in my class*), but the bias-free perception of the student is "imperfectly" translated into a bias-contaminated phrasing in the letter ("She performs well because of her hard work").

III. It could be that a teacher both stores a bias-contaminated perception of the student, *and* introduces even more bias through the "imperfect" translation of that perception into writing.

Research on the effects of teacher-student demographic congruence on concrete student outcomes like achievement (Dee, 2007; Egalite, Kisida, & Winters, 2015; Papageorge, Gershenson, & Kang, 2018) make it almost a certainty that **(II)** above is not the case. However, neither I nor prior studies are able to distinguish between **(I)** and **(III).** Despite this limitation, it is conceptually necessary to separate the two points for the sake of future work and intervention. If we can eventually discern with certainty that **(I)** is the case, it may be most prudent to increase efforts to train teachers on implicit bias *in the classroom*. However, if **(III)** is the case, it would also be prudent to increase efforts to train teachers on implicit bias *while letter writing*, as well.

Note that this model relies on the *teacher's perceptions of student demographics*, rather than the student's self-identified demographics. In the framework of implicit bias, we should only expect student race/gender/SES to influence teacher writing if the teacher believes the student to be of a specific demographic and holds implicit attitudes about that demographic. For example, if a teacher does not perceive a student as Black, their implicit attitudes about Black individuals should not meaningfully come into play either in observations of the student, or while writing about the student; this would be indistinguishable in my data from a teacher who does perceive a student as Black, but holds no negative implicit attitudes about Black individuals. In short, teacher misperception or ignorance of student demographics should be kept in mind as a factor while I interpret my analysis.

I also include in this theoretical model an acknowledgement that bias on the part of the letter *readers* may also ultimately influence student outcomes in the admission process, even if I am unable to examine this dynamic in the present study. There is some experimental evidence that untrained readers of letters can actually read the *same letter* and evaluate it differently based on experimentally-manipulated student descriptions the letters are paired with (e.g. race, names; Madera et al., 2018; Morgan, Elder, & King, 2013). Fact-finding from the Students for Fair Admissions v. Harvard University (2019) case made it clear that formal implicit bias training for admissions counselors is rare, even at elite, well-resourced institutions like Harvard; it is then likely that admissions counselors exhibit behavior similar to untrained experiment participants. While some admissions officers might be aware of bias and attempt to account for it in their reading of letters out of personal motivation (Schwarz, 2016), we do not have evidence as to whether counselors are actually successful in this endeavor. Research on the mitigation of implicit bias in other settings suggests that bias is quite pernicious, even in the face of thoughtful and well-devised intervention (Devine, Forscher, Austin, & Cox, 2012; Paluck & Green, 2009).

I would also be remiss not to mention here that broader social and socioeconomic factors play an important role in nearly every node of my theoretical model. That is, they influence student behavior/actions, teacher biases, counselor biases, self-identity, college admissions dynamics, and so on. I exclude this dynamic from the model only for visual clarity (i.e. many overlapping arrows), but recognize that I cannot (and arguably should not) truly extricate anything I observe in my analyses from the broader societal context. My intention is not to ignore these important disparities from other moments in a student's educational journey, but to present the best estimates I can on this individual margin as my contribution to a broader discourse on gender, class, and race/ethnicity in our society.

VIIIb. Word Frequency Analysis Methodology

(a) Word Frequency Analysis is a straightforward endeavor where the occurrence of each word in each letter is counted after adjusting for varied word forms (e.g. verb conjugations, noun plurals, etc., a process known as "lemmatizing") and discarding words without substantive meaning on their own (e.g. "like," "the," "if," commonly referred to as "stopwords"). While one could count all words in a given letter, I intend to restrict the sample of words only to those that *directly* describe a student using additional natural language processing techniques. First, I will identify explicit mentions of the student using a process known as "named entity recognition" – a common NLP task in which occurrences of pronouns or references to them (e.g. "she," "this student") are identified (Sang & De Meulder, 2003). After identifying sentences that mention the student and filtering out mentions of other individuals such as classmates or family members, I can then apply "part of speech tagging" to label the grammatical function of each word in those sentences. Machine learning models designed for this task typically learn how to identify parts of speech using enormous databases of words in context that are manually coded by humans for their correct part of speech, and they do so with exceptionally high accuracy on benchmark datasets (e.g. 98% accuracy on the Penn Treebank test; Bohnet et al., 2018). With these tags, I can identify the specific *verbs* (e.g. "struggles" versus "excels"), *adjectives* (e.g. "phenomenal" versus "competent"), and *nouns* (e.g. "a scholar" versus "an athlete") used in sentences directly describing each student to organize the word frequency output more comprehensibly.

This will result in a simple count of each word used to describe students in each letter, allowing me to explore in aggregate which words are most common for students of each race/ethnicity and gender. I anticipate first constructing a simple table highlighting the 50 most-used verbs, adjectives, and nouns in letters for each race/gender/SES (white, Black, Asian, Hispanic, male, female, etc.), as well as interacted race/ethnicity and gender groups (white male, white female, Black male, Black female, etc.).

I will supplement this basic exploratory analysis by replicating the analysis of Wu (2017) to identify the words that are most *uniquely* used for each demographic. First, I will create word counts at the letter-level for each of the 10,000 most-used words across letters. I will then set binary indicators for each race/gender group (as well as interacted groups) as the outcome of a regularized logistic regression on this vector of word indicators, without additional controls. This will allow me to estimate which words are most predictive of a student being female/Black/white/etc., which can be interpreted as the words *most uniquely used* in letters for that demographic versus others. I can sort the output by which words are most predictive (e.g. largest coefficients); I can alternatively sort the output by actual frequency of word use to explore the relative prevalence of these most predictive words. In either case, this will allow me to directly explore the extent to which students of varying demographics are described using different words by teachers – a potentially major component of linguistic bias as surfaced in my literature review.

VIIIc. Assessing Topic Modeling Accuracy and Validity

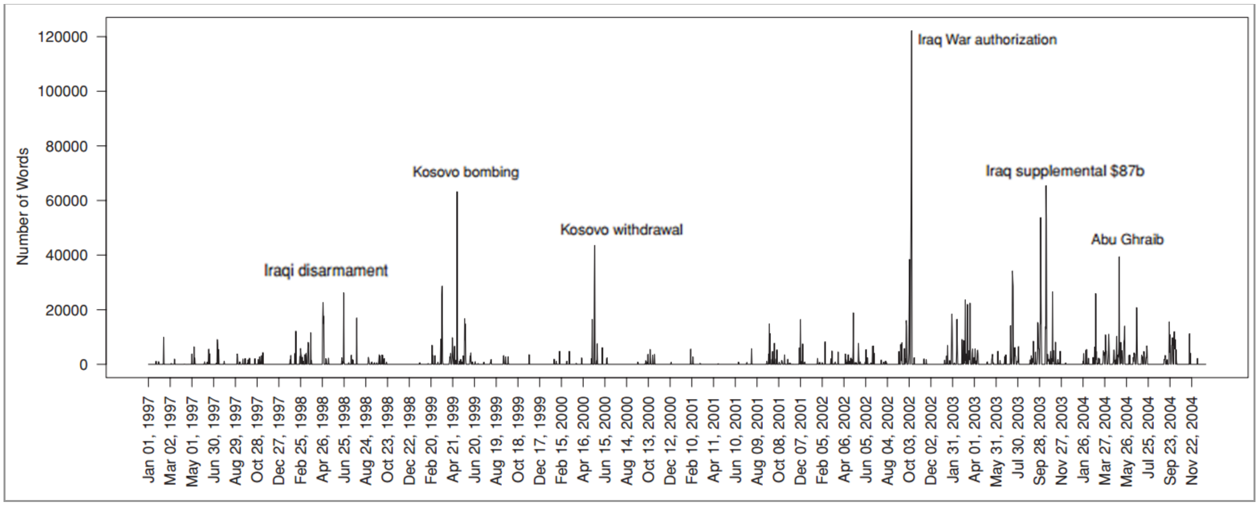
My use of this topic model approach brings up three main questions that I will need to grapple with to improve the robustness of my analysis. I present my intentions below, with the understanding that data, logistics, computing power costs, and time may limit my ability to fully realize these intentions.

- How will I select the number of topics that the topic model algorithm will look for?
    - Researchers who frequently use topic modeling have identified a "topic coherence" measure to evaluate the extent to which the words within each topic are related in substantive meaning (Mimno, Wallach, Talley, Leenders, & McCallum, 2011; Stevens, Kegelmeyer, Andrzejewski, & Buttler, 2012). Without going too deeply into the mechanics, the measure works by examining at the *sentence*-level which words seem to co-occur frequently (whereas the overarching topic model looks at the paragraph-level). By grouping words according to the network of words they most frequently co-occur with (known as "word vector clustering"), we can algorithmically determine an abstract measure of similarity in meaning and use between words (Mikolov, Chen, Corrado, & Dean, 2013). This similarity measure is applied to each topic grouping identified by the topic model to measure within-group similarity, which we can then index across topics to arrive at an overall level of similarity for one set of topics produced by the algorithm. Analysts will systematically re-run the topic model algorithm on random subsets of the text data, adjusting each of its parameters (e.g. number of topics to identify) each time, to compare the topic coherence of each run of the algorithm.They then select the parameters that produce the most reliably high measures of topic coherence as their final model. Given my data and context, this seems to be the most robust and appropriate way to approach this issue without any substantial modifications necessary – though the computational costs of this approach may prevent more exhaustive analyses.
- How will I rigorously interpret the topic model output for substantive meaning?
    - In this endeavor, I will replicate the approach of Penner et al. (2019) to emulate the methods of qualitative researchers conducting thematic analysis (Nowell et al., 2017). Once Penner et al. produced their topic model output, they had four separate analysts interpret each topic grouping for substantive meaning by referencing the most common words in each group, the most representative words in each group (i.e. those with the lowest appearance in other groups), and a random sample of documents identified as highly reflective of each topic. After that process was completed by each analyst in isolation, the four analysts met to discuss any discrepancies and harmonize their interpretations. I intend to replicate this process using a group of 5 raters (including myself), documenting the process, any discrepancies in coding, and any revisions made to topic interpretations as a result of the harmonization process. This process has been piloted in detail in Kim et al. (2021). Because these interpretations are non-numeric, it is not possible to calculate any measure of interrater reliability except a very basic rate of outright agreement per topic. The high-level notes of this process, alongside the pre-harmonization interpretations from each analyst, will be included as an appendix to the main paper for transparency. This same harmonization process will be applied after the topics are interpreted to group them into substantive *supertopics* as well. Again, the pre- and post-harmonization results by coder will be included in an online, open-source appendix for transparency. If it is any indication, both topic and supertopic agreement in Kim et al. (2021) was exceptionally high, with only minor discrepancies in details, and full agreement in overarching concepts.
- How will I measure the accuracy of classifications for each paragraph's primary topic?
    - Once the topics are constructed and interpreted, I can then have this same group of five raters read and hand-classify a set of roughly 100 randomly sampled paragraphs from the recommendation data using the finalized topic interpretations as the set of codes. Because the topic model identifies a number of topics and it is important to assess the

extent to which it accurately classifies paragraphs across each topic, I will stratify this random sample of paragraphs by topic (e.g. if thereare 10 topics identified, I will sample 10 paragraphs per topic). From there, I can more easily calculate the interrater reliability among human raters (I intend to use the adjusted kappa from Byrt et al., 1993, per recommendation of Hallgren, 2012). Then, I can calculate the rate of alignment between the algorithm's classifications against the consensus (if any) of human raters as my main accuracy measure.

Lastly, I can borrow the spirit of the approach that Quinn et al. (2010) utilized in their analysis of U.S. Senate speeches to further assess the validity of my topic modeling output more holistically. They first analyzed 118,000 transcribed speeches to derive their topics and then interpreted these topics for substantive meaning. To check the robustness of these interpretations, they looked at whether the prominence of each topic across speeches trended intuitively with related events in U.S. history. For example, they found a topic that seemed to represent the substantive topic of "defense," and then examined whether the occurrence of words in this topic trended alongside major defense-related events like the Kosovo bombing, the Iraq War authorization, and debates around Abu Ghraib (Figure 2 below). While not conclusive, this approach is one way to provide compelling evidence to support the validity of topic interpretations.

Figure C1. "Defense" Topic Frequency in Senate Speeches Over Time (Quinn et al., 2010)



In my case, I can use some creative correlation analyses to conduct the same conceptual checks for each identified topic. For example, we might interpret one topic as "sports" using the aforementioned multiple-coder approach I described above. Because I can analyze the student's application for explicit mentions of sports activities in their open-response extracurriculars box, I can identify students who actually played sports. Logically, we should expect that students who actually played sports are more likely to have paragraphs in their recommendations classified as being about sports. To check this explicitly, I can run a simple regression of the proportion of paragraphs in a letter written about sports on an indicator for actual student sports involvement. If we return a statistically and substantively significant result, I can provide at least some evidence that our interpretation of the topic trends with substantive student characteristics as expected.

I will benchmark each of these simple correlation analyses with a pair of falsification tests to better guard against spurious correlations. The first falsification test will involve running the same correlation test, but between the proportion of paragraphs in a letter written about a topic, and a binary indicator for whether the student was born on an odd-numbered day of the month. Because odd- and even-numbered birth dates are arguably randomly assigned, we should observe *no* relationship between this indicator and the proportion of paragraphs written about that topic. If we recover a statistically and substantively significant coefficient on the binary indicator for birth date, we would be concerned about whether a significant result for the actual sports involvement indicator in the prior check was just due to a noisy measure.

The second falsification test will be contextual based on the actual topic: I will run the same correlation test, but between the proportion of paragraphs in a letter written about a topic, and a binary indicator for some student characteristic more strongly related to a different substantive topic. To return to the sports example, we would imagine that the proportion of paragraphs in a letter written about sports is much less related to whether a student plays a musical instrument. While arguably still related (e.g. via a common relationship to family financial resources), the coefficient we estimate on a binary indicator for playing a musical instrument should be *much* smaller than the coefficient we estimated on the indicator for actual sports involvement. If the two are close to equal, we should question whether the interpretation of the topic is actually capturing sports per se, or something broader.

VIIId. Sentiment Analysis Model Selection, Validity, Accuracy, and Bias

Because there are a variety of sentiment analysis models, and prior research has shown that these models tend to suffer from low inter-model agreement, the decision for which exact model I will deploy for this analysis is highly consequential (Gonçalves et al., 2013). Moreover, the field of NLP is advancing beyond the “lexical” models (described in the main narrative) and towards black-box “neural network” models that tend to be more accurate and nuanced in their classification processes, but substantially harder to interpret and more prone to algorithmic bias: changing its interpretation/classification of a given text based on irrelevant demographic features like the presence of female pronouns (e.g. classifying “he is assertive” as positive, but “she is assertive” as negative). This is because these algorithms are trained on massive text datasets that implicitly contain the biases of societal writing more broadly, “teaching” the algorithm these same biases (Caliskan et al., 2017; Park et al., 2018; Sun et al., 2019). Bias of this nature would threaten the validity of the entire analysis by *creating* bias in the coding of the letters, rather than *identifying* bias in the letters themselves.

For these reasons, I will conduct the sentiment analysis with a big-tent approach and select a final analytic model using an empirical process that balances accuracy against algorithmic bias. First, I will deploy a range of lexical and transformer models on the data, to include:

- Lexical Models:
  - LIWC (Tausczik & Pennebaker, 2010)
  - VADER (Hutto & Gilbert, 2014)
  - SO-CAL (Taboada et al., 2011)
- Neural Network Models:
  - BERT (Devlin et al., 2019)
  - ALBERT (Lan et al., 2020)
  - XLNet (Yang et al., 2020)
  - Stanford Stanza (Qi et al., 2020)
- Ensemble model that incorporates all of the above together (Kim et al., 2021)

Once the sentiment analysis has been conducted using each method described above (despite seemingly a monumental task, the marginal effort required to run an additional model is nearly zero), I can mirror my approach to measuring the accuracy of my topic model paragraph classifications (Appendix VIIIc above). First, I will randomly sample 1100 sentences (100 common cases, then additional 200 unique cases per coder) from the recommendation data, and I will stratify this sampling across crossed student demographic groups (gender, race/ethnicity, SES). Using the same team of 5 analysts, I will have them read each sentence and code on the five-degree measure of positivity (very negative, negative, neutral, positive, very positive) using the working definition of sentiment as *the perceived positivity of emotionsand ideas present in a given text* (coders will be trained on specific examples, as well).

Again, I can calculate the interrater reliability among human coders (using a two-way model Intraclass Correlation Coefficient for absolute values, again per recommendation of Hallgren, 2012), and calculate the alignment/accuracy of each algorithm's classifications against the consensus of human raters – while also specifically examining accuracy across student demographic groups (per the current literature's proposed measures of algorithmic fairness; see Corbett-Davies et al., 2017 for a helpful review). Statistically significant differences in *accuracy* across student groups would suggest that the algorithm is detecting or reacting to superfluous information in its classifications and resulting in algorithmic bias. While not exhaustive nor conclusive, this process will allow me to benchmark the best algorithm across a variety of relevant measures; I will then utilize the algorithm that *minimizes bias* and *maximizes accuracy*.

Finally, note that several cleaning steps can be taken to the data to minimize at least the extent of gender bias in the letters. Per Kim et al. (2021), I can extract and replace any explicitly gendered pronouns or nouns (e.g. "he," "she," "actress") and examine the extent to which algorithm classifications of sentences change. Any meaningful mismatch would indicate that gender matters to the algorithm, and give us some degree of bounds around the influence of algorithmic bias in this sense.

<u>VIIIe. Issues of Common Support with Teacher Fixed Effects</u>

My primary fixed effects strategy will reduce the effective estimation sample for each coefficient to only those observations with common support. That is, the coefficient on a binary indicator for female students can only be estimated using recommendations from teachers who have written for both male *and* female students. This is unlikely to be a concern for gender or SES, but it could present issues for proportionally small racial/ethnic groups, such as Native American students. While my main specification currently includes binary indicators for each student group, it may end up being more prudent to group certain racial groups together. One obvious option is to instead utilize a binary indicator for racial/ethnic minority status. However, this may wash out meaningful biases that occur in opposite directions depending on the precise group. Instead, I will operate with the intention of keeping white non-Hispanic, Black, Hispanic, and Asian separately specified. Remaining groups may need to be consolidated as needed (e.g. Native American, Pacific Islander, multi-racial).

Another concern this brings up is whether the sample of teachers with common support are meaningfully different from other teachers. That is, there may be a form of selection occurring here: for a teacher to have common support, they must be teaching students of multiple demographics and be *asked* and *willing* to write postsecondary recommendations for these students. Teachers who either teach in schools/classes without substantial college-going populations of students across demographics, or are outright discriminatory (i.e. refuse to write recommendations for certain groups of students), would be excluded here.

I will be able to empirically assess differences between these teacher groups along a set of characteristics like school context/geography (using CCD data). However, I will not be able to assess it along characteristics that require more detailed teacher staffing data, such as race/ethnicity, gender, years of teaching experience, value-added, exact courses taught, and so on. I would argue that my estimation sample includes *most* of the relevant margin of teachers, but I acknowledge this will be a limitation of the data. To the extent we think teachers with less exposure to minority students are more likely to exhibit bias in recommendations (drawing from other literature on bias such as Lai et al., 2014 or Chin et al., 2020), my resulting coefficients will likely end up being underestimates of what they would be with these teachers included.

Lastly, this issue of common support is my primary motivation for considering crossed race/gender categories (e.g. Black females versus Black males) a *supplementary* analysis. While absolutely of interest both theoretically and substantively speaking, my fixed effects strategy makes the region of common support for these interacted categories substantially smaller despite the size of my overall sample. It may be more appropriate to remove the fixed effects specification to approach this analysis, but that would greatly expand the number of confounding factors that I cannot feasibly account for within that model, bringing into question the robustness of any results obtained from such an approach. I intend to still conduct these analyses within the fixed effects specification, but I maintain enough caution and concern that I cannot consider them main analyses.

VIIIf. Appendix References

Bohnet, B., McDonald, R., Simoes, G., Andor, D., Pitler, E., & Maynez, J. (2018). Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings. *ArXiv:1805.08237 [Cs]*. http://arxiv.org/abs/1805.08237

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *ArXiv:1701.08230 [Cs, Stat]*. https://doi.org/10.1145/3097983.309809

Dee, T. S. (2007). Teachers and the Gender Gaps in Student Achievement. *Journal of Human Resources*, *XLII*(3), 528–554. https://doi.org/10.3368/jhr.XLII.3.528

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, *45*, 44–52. https://doi.org/10.1016/j.econedurev.2015.01.007

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. *Proceedings of the First ACM Conferenceon Online Social Networks*, 27–38. https://doi.org/10.1145/2512938.2512951

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *ICWSM*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ArXiv:1909.11942 [Cs]*. http://arxiv.org/abs/1909.11942

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the Conferenceon Empirical Methods in Natural Language Processing*, 262–272. http://dl.acm.org/citation.cfm?id=2145432.2145462

Morgan, W. B., Elder, K. B., & King, E. B. (2013). The emergence and reduction of bias in letters of recommendation. *Journal of Applied Social Psychology*, *43*(11), 2297–2306. https://doi.org/10.1111/jasp.12179

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*(1), 160940691773384. https://doi.org/10.1177/1609406917733847

Paluck, E. L., & Green, D. P. (2009). Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. *Annual Review of Psychology*, *60*(1), 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607

Papageorge, N. W., Gershenson, S., & Kang, K. M. (2018). *Teacher Expectations Matter* (Working Paper No. 25255). National Bureau of Economic Research. https://doi.org/10.3386/w25255

Park, J. H., Shin, J., & Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. *Proceedings of the 2018 Conference on Empirical Methodsin Natural Language Processing*, 2799–2804. https://doi.org/10.18653/v1/D18-1302

Penner, E. K., Rochmes, J., Liu, J., Solanki, S. M., & Loeb, S. (2019). Differing Views of Equity: How Prospective Educators Perceive Their Role in Closing Achievement Gaps. *The Russell Sage Foundation Journal of the Social Sciences*, *5*(3), 103–127. https://doi.org/10.7758/RSF.2019.5.3.06

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Sang, E. F. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *ArXiv:Cs/0306050*. http://arxiv.org/abs/cs/0306050

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://www.aclweb.org/anthology/D13-1170

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961. http://dl.acm.org/citation.cfm?id=2390948.2391052

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *ArXiv:1906.08976 [Cs]*. http://arxiv.org/abs/1906.08976

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Languageand Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Wu, A. (2017). *Gender Stereotype in Academia: Evidence from Economics Job Market Rumors Forum* (No. 2017–09; Working Papers). Princeton University, Woodrow Wilson School of Public and International Affairs, Center for Health and Wellbeing. https://ideas.repec.org/p/pri/cheawb/2017-09.html

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *ArXiv:1906.08237 [Cs]*. http://arxiv.org/abs/1906.08237