REACH
National Center for Research on Education Access and Choice

Online Reviews Are Leading Indicators of Changes in K-12 School Attributes

Linsen Li, Tulane University
Aron Culotta, Tulane University
Douglas Harris, Tulane University
Nicholas Mattei, Tulane University

Technical Report
Published May 29, 2023

Online Reviews Are Leading Indicators of Changes in K-12 School Attributes

Linsen Li
Tulane University
New Orleans, LA, USA
lli23@tulane.edu

Aron Culotta
Tulane University
New Orleans, LA, USA
aculotta@tulane.edu

Douglas N. Harris
Tulane University
New Orleans, LA, USA
dharri5@tulane.edu

Nicholas Mattei
Tulane University
New Orleans, LA, USA
nsmattei@tulane.edu

ABSTRACT

School rating websites are increasingly used by parents to assess the quality and fit of U.S. K-12 schools for their children. These online reviews often contain detailed descriptions of a school's strengths and weaknesses, which both reflect and inform perceptions of a school. Existing work on these text reviews has focused on finding words or themes that underlie these perceptions, but has stopped short of using the textual reviews as leading indicators of school performance. In this paper, we investigate to what extent the language used in online reviews of a school is predictive of changes in the attributes of that school, such as its socio-economic makeup and student test scores. Using over 300K reviews of 70K U.S. schools from a popular ratings website, we apply language processing models to predict whether schools will significantly increase or decrease in an attribute of interest over a future time horizon. We find that using the text improves predictive performance significantly over a baseline model that does not include text but only the historical time-series of the indicators themselves, suggesting that the review text carries predictive power. A qualitative analysis of the most predictive terms and phrases used in the text reviews indicates a number of topics that serve as leading indicators, such as diversity, changes in school leadership, a focus on testing, and school safety.

CCS CONCEPTS

• **Information systems → Web mining**.

KEYWORDS

text classification, education, online reviews

1 INTRODUCTION

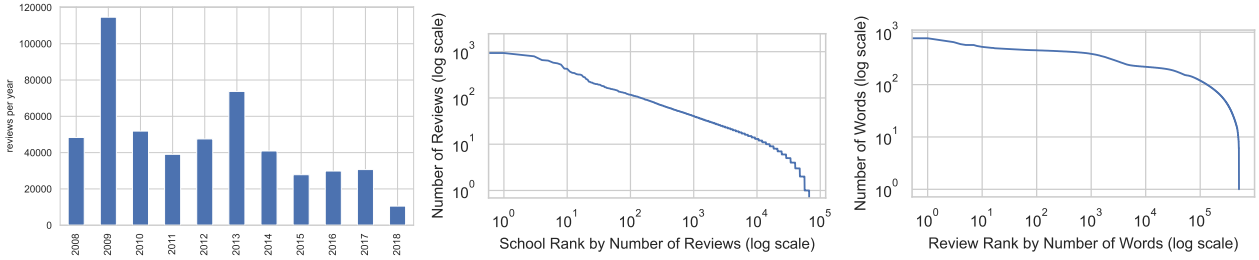
According to the Pew Research Center, 82% of U.S. adults say that they sometimes or always consult online reviews before purchasing products or services [34]. Online product reviews not only reflect but also inform user opinions and perceptions [8]. Understanding both the contents of these reviews and how consumers react to them has been the focus of substantial research in computer science [25, 26, 29], marketing [6], education [11, 15], and many other areas. Within computer science, the typical goals of analyzing these reviews are to extract unifying topics of discussion [19, 26], understand particular aspects of products, and the sentiment towards those aspects [24, 25].

In addition to summarizing the contents of online reviews, researchers have also begun investigating whether they can be used to *predict* what will happen in the future. For example, Alessa et al. [1] mined Twitter to predict where flu outbreaks were likely to occur. This logic has been applied to other outbreaks including COVID-19 [7, 18]. Similarly, Kryvasheyeu et al. [22] mined Twitter to estimate the likelihood that individuals would evacuate in an impending disaster. Within the product review space, prior work has mined reviews for signs that a product may be recalled or exhibit safety hazards [4, 12], mined Yelp reviews to predict health code violations at restaurants [20, 33], and mined social media to detect adverse drug reactions [32, 35].

We are particularly interested in school reviews, specifically from the website GreatSchools,[1] which collects user reviews of public, private, and charter K-12 schools across the United States. How and why parents choose a school is influenced by many factors including social network, distance to a particular school [10], school outcomes, and extracurricular programs [2, 13]. Barnum and LeMee [3] study the effect that GreatSchools reviews have on school selection decisions, and find that these online review platforms may in fact drive White families towards schools with fewer Black and Hispanic students. Hence, understanding how these reviews both shape and inform school choices is important in determining the impact these large scale rating websites have on society. Recently, school review data has been used to understand the factors that enter into parental choice [15], as well as whether or not those reviews are reflective of socio-economic indicators [11]. However, there has not been work on leveraging the reviews posted online to rapidly predict and/or understand changes happening at the school level.

Perhaps more so than most products, reviews on school rating websites are both reflective of community opinions of a particular school and form the perceptions that other parents have *about* that school, i.e., the school's prestige. This trend is especially true at the college and university level [23], but true at the K-12 level as well. Both school perception, as well as other important factors such as housing decisions, affect the eventual choice of whether or not to enroll in a particular school. We focus on the question: *to what extent can the text of school reviews act as leading indicators of both the outcomes and socio-demographic changes at that school?* A deeper understanding of the predictive quality of school reviews

[1] https://www.greatschools.org/



(a) Reviews per year **(b) Reviews per school** **(c) Words per review**

Figure 1: Summary statistics of the review dataset (677K reviews of 84k schools).

provides insights into what perceptible school features precede these changes.

Contribution. We undertake an analysis of textual reviews of K-12 schools in the U.S. Rather than attempt to extract aspects and sentiment of schools in general, we investigate whether or not these text reviews can provide a leading indicator of future changes in both performance indicators (test scores) as well as school socio-demographics. We show that the review text is indeed predictive of these changes on a per-school basis, improving over a baseline of only school level socio-demographic indicators. A qualitative analysis of the words and phrases that are most predictive of future changes reveal that reviews discussing changes of school leadership, a focus on testing, as well as discussions of diversity and school safety provide a strong signal of changes in both test scores and the socio-demographic makeup of a school.

Our results could be used by several distinct groups of stakeholders including social scientists, administrators, and/or parents. From the social science perspective, our work could be used to uncover factors that precede changes in school attributes, e.g., changes in teachers or community perceptions of a school. For administrators, having a predictive model of school outcomes, and having those outcomes tied to topics discussed by parents and/or students could help monitor changing community perceptions of the schools they administer. Finally, for parents, a richer understanding of the the trajectory of a particular school may affect the choice of which school to attend. In this work we focus primarily on the first, as understanding how perceptions drive outcomes is an understudied topic in the school choice literature.

2 RELATED WORK

In addition to the prior work on review mining described in the previous section, there is an emerging line of work using text data in educational research. The most closely related work is that of Gillani et al. [11], who find that the text in online reviews of a school correlates with standardized test scores. Furthermore, they find that the language reflects racial and income disparities in the U.S. education system. To do so, they train a regression model that accepts as input the reviews from a school to (1) predict outcomes describing that school (e.g., test scores, progress scores) and (2) shed light on how important different words and phrases from reviews were in driving that model's predictions. In addition, they also did phrase clustering to discover salient topics.

Similarly, Harris et al. [15] analyze the same text dataset with a set of key words and phrases extracted from the text reviews. These words and phrases are manually coded into a formal taxonomy, i.e., a set of topics about school choice, and the reviews are then labeled for what topics are discussed. They find that reviews of traditional public schools focus on topics of interpersonal relationships while charter and private schools discuss school culture and graduation/post-secondary outcomes more frequently.

Other work considers the role that online review sites have on school choice decisions. For example, Hasan and Kumar [17] find that affluent and highly educated families are better able to take advantage of the new information made available in online school ratings, thereby accelerating racial and ethnic inequities. Similarly, Haber [14] analyzes the text on the websites of thousands of charter schools, finding that they can present different identities to different socio-economic groups. This is conjectured to be a factor that leads to increased consolidation of schools by race and class.

While some of this prior work focuses on correlations between text and school attributes, the primary contribution of the present work is to instead examine the extent to which online reviews serve as leading indicators of school changes along multiple attributes.

3 DATA

3.1 Reviews

GreatSchools is one of the most prominent K-12 school review websites in the United States. Since its national launch in 2003 the site has expanded widely, and GreatSchools reviews and rating data is integrated on major home purchasing websites including Realtor.com and Zillow.com. We collected 677,210 reviews covering 83,795 public, private, and charter schools in the United States between 2002-2019. From this data we find that between 2002-2007 there are less than 400 reviews per year, while starting in 2008 there are more than 20k reviews per year. Hence, we focus on the years with the highest density of data, 2008-2018. To make it have enough reviews to analyze for each school, we restrict to schools with at least 10 reviews. Besides, we drop the private school and restrict the analysis to public and charter schools since we don't have the test score data for private schools. In total, we analyze 276,038 reviews of 69,744 schools. Figure 1a shows the total number

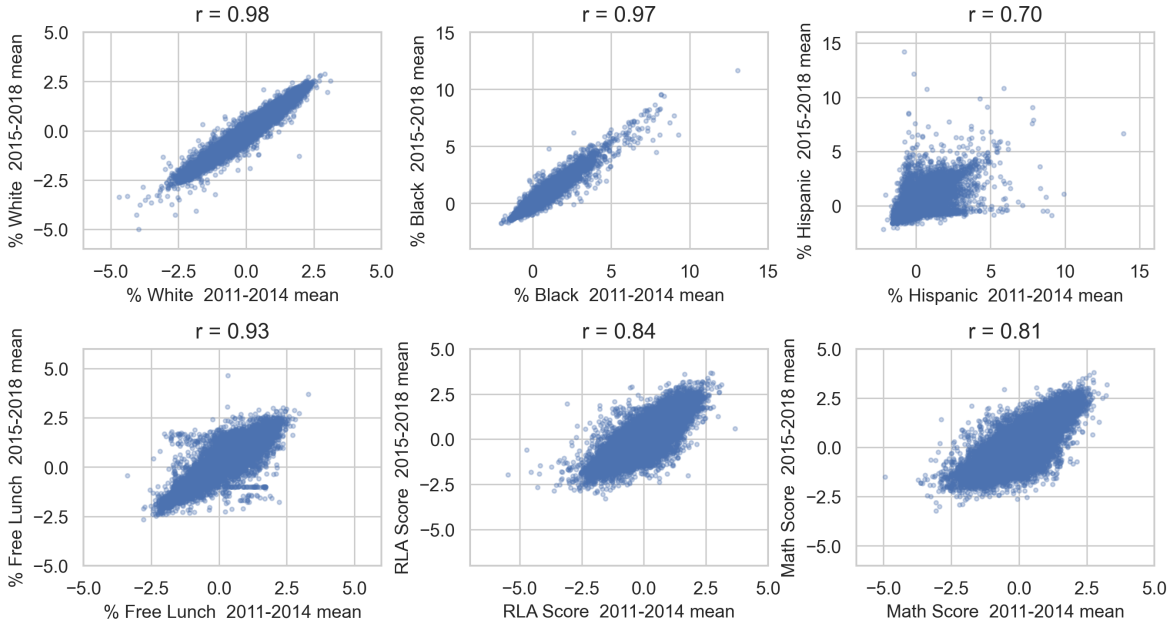


Figure 2: For each school, we plot the state-scaled attributes averaged over the pre-period (2011-2014) versus the post-period (2015-2018), along with the Pearson correlation (r) between the values in the two periods.

of reviews posted per year, which tends to be between 30K-50K, with spikes in 2009 and 2013. These spikes may in part be due to various partnerships and marketing efforts by GreatSchools.org. The low rate in 2018 is due to the fact that we do not have the full year's worth of data; but, this will not affect the analysis below.

Figure 1b shows the total number of reviews per school over the entire time period. As expected, a log-linear pattern shows that most schools have few reviews (less than 100), and a small number of schools have many reviews (~1,000). Figure 1c shows the number of words per review. We can see that reviews are often quite detailed — over 100K reviews are at least 100 words long.

3.2 School Attributes

School level demographic attributes (enrollment by racial/ethnic groups) and the percentage of students on free or reduced lunch programs were gathered from the U.S. Department of Education's Common Core of Data.[2] For test scores, we collect school-level statistics from the U.S. Department of Education EDFacts Data Files.[3] Federal guidelines require states to report achievement data on state assessments. These measurements are assessed against state-specific content standards, with assessments conducted annually in third through eighth grade and at least once in high school. We consider both Math and Language Arts (RLA) scores aggregated for each school, using the value for "the percent of students proficient or above on the state assessment." We collect all values from 2009-2018 academic years, which is the latest available at this time. For school values where ranges are reported (e.g., 85-89%), we use the median value (e.g., 87%).

[2]https://nces.ed.gov/ccd/

[3]https://www2.ed.gov/about/inits/ed/edfacts/data-files/

3.3 Normalizing Values by State

The school attributes described above vary greatly by state. For example, the test scores are derived from state-specific guidelines and assessment criteria. Thus, when a state revises these criteria, these can result in abrupt changes to these values on a year-by-year basis. This can complicate inter-school comparisons when computing changes over time.

To account for these idiosyncrasies, we normalize all school attributes by state by computing state-specific z-scores. Let $a^i_{s,y}$ represent the attribute value for school i in state s in year y. We compute the state-year mean ($\mu_{a,s,y}$) and standard deviation ($\sigma_{a,s,y}$). We then transform the attribute for school i as $a^i_{s,y} \leftarrow \frac{a^i_{s,y} - \mu_{a,s,y}}{\sigma_{a,s,y}}$. Thus, each attribute value represents how many standard deviations from the mean this school's value is for a given state/year pair.

4 METHODS

Our goal is to determine whether the text of a school's reviews in one time period is predictive of changes to a school's attributes in a future time period. We formalize this as a classification task by assigning to the positive class schools that significantly increase in the attribute, and assigning to the negative class schools that significantly decrease. To do so, we must first fix the pre and post time periods of interest. To ensure a sufficient number of reviews per school, we select 2011-2014 as the pre-period and 2015-2018 as the post-period.[4] For each period, we compute the average attribute value for each school. We then compute the difference in the average value for school i between the post- and pre-periods:

[4]In Appendix A.2 we consider variants of this choice; the results are qualitatively similar.

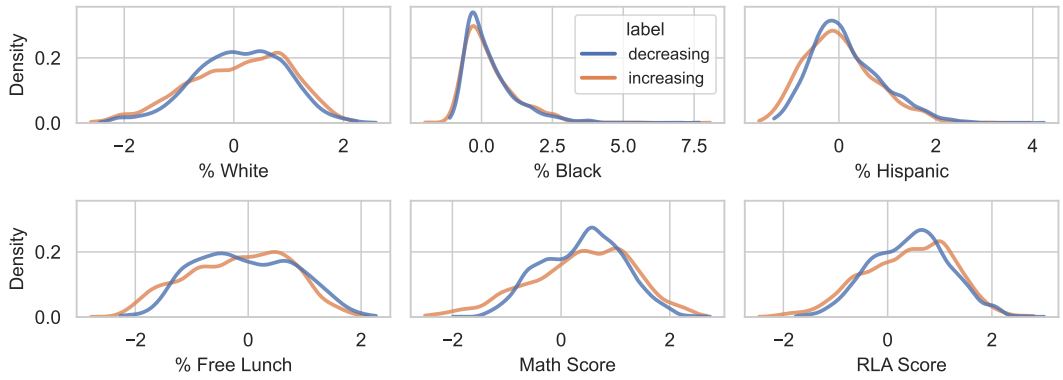


Figure 3: After matching, the distribution of state-scaled school attributes based on the pre-period values are very similar between schools that increase and schools that decrease with respect to that attribute.

$$\Delta_{a^i} = \frac{1}{4}\sum_{y=2015}^{2018} a^i_{s,y} - \frac{1}{4}\sum_{y=2011}^{2014} a^i_{s,y}$$

To visualize this difference, Figure 2 shows scatter plots of average school attributes in the pre-period versus the post-period, along with the Pearson correlation between the two periods. We can see that, as expected, test scores tend to have a greater change between time periods (Math $r = .81$, RLA $r = .84$) as compared to demographics (% White $r = .98$, % Black $r = .97$). An exception is the % Hispanic attribute ($r = .7$), which exhibits large fluctuations over the two time periods. This may in part be due to the overall growth of Hispanic students in schools nationwide. NCES reports that between 2009 and 2020 public school enrollment among White students declined from 26.7 million to 22.6 million, while that of Hispanic students increased from 11.0 million to 13.8 million.[5] This growth appears to be distributed unevenly across schools.

While we could fit a regression model to predict Δ_{a^i} directly, we want to first focus on the most significant changes, to provide a stronger signal to discover salient text features that predict such changes. To do so, for each attribute we sort schools by Δ_{a^i}, and label the schools in the top 20% as **increasing** and those in the bottom 20% as **decreasing** with respect to that attribute. With these labels, we construct a binary classification task: predict whether a school will increase or decrease for a given attribute, based on the content of the reviews written up to the end of the pre-period (2014). In order to have sufficient review content to consider, we restrict our analysis to schools that have at least 10 reviews posted through 2014. (In §5.2.1 we investigate how this threshold influences results.)

4.1 Matching

Given the well-known disparities in U.S. education by socioeconomic variables, a key challenge in this study is to properly "control for" such variables and to isolate the attribute of interest. For example, lower income schools (those with a higher % Free Lunch) are more likely to have fewer White students and lower test scores. When fitting a text classifier to predict changes to the % Free Lunch attribute, we want the classifier to identify terms that are particularly salient for the % Free Lunch attribute, and not those that are predictive only indirectly through the other confounding variables.

This problem is observed in Gillani et al. [11], who present preliminary results using adversarial machine learning methods to control for confounds. In this paper, we take a simpler approach based on nearest-neighbor statistical matching [31]. The idea is as follows: To construct the training set for attribute a, we first annotate the top/bottom 20% of schools as described above to create an initial pool of training samples P_a. Letting $P_a^+ \subseteq P_a$ represent the n "positive" examples and $P_a^- \subseteq P_a$ represent the n "negative" examples, our goal is to identify a subset $\hat{P}_a = \hat{P}_a^+ \cup \hat{P}_a^- \subseteq P_a$, with $\hat{P}_a^+ \subseteq P_a^+$ and $\hat{P}_a^- \subseteq P_a^-$, such that the distribution of pre-period attributes is similar for $\hat{P}_a^+$ and $\hat{P}_a^-$.

To do so, we first represent each school i by a six-dimensional vector of the pre-period school attributes $\mathbf{a}^i = \{a^i_{white}, a^i_{black}, a^i_{hispanic}, a^i_{lunch}, a^i_{math}, a^i_{rla}\}$. Then, for each school $\mathbf{a}^i \in P_a^+$ and $\mathbf{a}^j \in P_a^-$, we compute the similarity between two schools $s(\mathbf{a}^i, \mathbf{a}^j)$ using the cosine similarity measure, resulting in a similarity matrix $S \in \mathbb{R}^{n \times n}$. Finally, to extract the most similar subsets $\hat{P}_a^+, \hat{P}_a^-$, we greedily select the most similar pairs of schools from S. We iteratively select matched pairs, without replacement, until a pre-defined similarity threshold δ is met, such that $s(\mathbf{a}^i, \mathbf{a}^j) \geq \delta$ for each selected pair. The parameter δ determines the matching quality of the resulting training set. For our main results, we set $\delta = 0.8$; §5.2.2 explores alternate settings. The result of this procedure is six matched training sets $\{\hat{P}_a\}$, one per school attribute, used to fit the classifiers described in the following section.

4.2 Classification

For each attribute a, we fit a binary logistic regression classifier with L2 regularization to predict whether a school i is positive ($i \in \hat{P}_a^+$) or negative ($i \in \hat{P}_a^-$) with respect to that attribute. We consider three types of features for this classifier:

(1) **baseline:** We use only the six pre-period attribute values. This classifier indicates how predictive these pre-period values are of post-period changes. Since the matching procedure above by design selects positive and negative instances that

[5] https://nces.ed.gov/programs/coe/indicator/cge/racial-ethnic-enrollment

Table 1: Average classification accuracy (area under the ROC curve) for each attribute and method. Δ = (baseline+text − baseline) represents the additional predictive accuracy provided by text features.

target	baseline	text	baseline+text	Δ
% White	.502 ± .02	**.619 ± .02**	.617 ± .02	+.115
Math Scores	.579 ± .03	.634 ± .01	**.666 ± .03**	+.087
RLA Scores	.609 ± .01	.623 ± .01	**.680 ± .01**	+.071
% Free Lunch	.584 ± .01	.618 ± .01	**.649 ± .01**	+.065
% Black	.536 ± .01	.588 ± .02	**.589 ± .02**	+.053
% Hispanic	**.575 ± .02**	.534 ± .02	.573 ± .01	-.002

are similar in these attributes, these features are not expected to provide much predictive signal.

(2) **text:** We use only the text of the reviews for each school, described below. We restrict these reviews to all those posted up to the end of the pre-period (i.e., through 2014).

(3) **baseline+text:** The final classifier combines the two preceding feature sets.

To represent the review text for a school, we concatenate all reviews posted about the school up to 2014 into a single document. We then compute TF-IDF vectors for each school, using word unigrams and bigrams, with L2 normalization to account for varying document lengths. When processing each document, we remove common stopwords as well as proper names to prevent the classifier from using school names and locations as features.[6]

Our experiments compute the cross-validation accuracy for each of these three classifiers. By examining the difference between **baseline+text** and **baseline**, we can assess the predictive value added by the text features. We use area under the ROC curve (AUC) as our primary evaluation measure.

While we use a simple logistic regression classifier, other more modern machine learning algorithms could be applied as well (e.g., transformer-based language models like BERT [21] and its variants). There are several reasons to use the logistic regression model here: (1) its simplicity makes it easier to identify and interpret the most predictive terms; (2) BERT models are designed for short documents,[7] whereas a "document" in this task may consist of hundreds of reviews; (3) prior work [11] on a similar task found little accuracy difference between linear regression and BERT-based models.

5 RESULTS

In the results analysis below, we investigate the following questions:

(1) How well-matched are the positive and negative instances in the training data?

(2) How much do content-based features improve classification accuracy over the baseline?

(3) What are the qualitative differences among the most predictive terms for each attribute?

(4) How do individual schools change over time with respect to their predicted attribute values?

5.1 Matching Quality

First, we assess the quality of the matching procedure from §4.1. To do so, we plot in Figure 3 the pre-period values of each attribute value for both classes. We can see that the schools in the increasing and decreasing category have very similar distributions of values. While we match along all six attributes, here we show only the match for the primary attribute — the matches for the other attributes are similar.

In Appendix A.1, we present analogous figures when we omit the matching step. We find that the attribute distributions differ more between the class labels, particularly for the test score attributes, where schools with high scores in the pre-period are more likely to increase in the post-period.

5.2 Classification Accuracy

Table 1 displays the main classification results. Each value is the held-out area under the ROC curve value, averaged over five cross-validation folds, along with the standard deviation. As the schools are evenly sampled from the positive and negative classes, a random classifier would achieve an AUC of 0.5. The Δ column is the difference in accuracy between the baseline+text and baseline methods — higher values represent greater predictive value provided by the text features over the pre-period school attributes.

We observe that, for all attributes except % Hispanic, the text features improve classification accuracy substantially. The largest gain is for % White, where accuracy improves 11%, from .502 (essentially random chance) to .617. The smallest gain is for % Black, where accuracy improves from .536 to .589, a difference of over two standard deviations from the mean accuracy of the five folds. These results provide evidence that online reviews in the pre-period provide a more nuanced characterization of the school and its perception than the pre-period school attributes alone.

As for the negative result for predicting % Hispanic, one possible explanation arises from the earlier discussion of Figure 2 in §4. Given the many external factors that have led to the growth of Hispanic enrollment across the country, it is possible that many of these external factors are not reflected in the school reviews. Furthermore, as discussed in Hasan and Kumar [17], there are disparities in how different population groups access and contribute to online school ratings. It may be the case that Hispanic populations are less active on school review sites.

5.2.1 Effect of Number of Reviews. The results above are restricted to schools with at least 10 reviews posted through 2014. As this is a somewhat arbitrary threshold, in this section we conduct robustness checks to determine how this threshold affects results. Furthermore, as Figure 1b suggests, there are many schools with fewer than 10 reviews. Thus, we would like to understand how the method performs for a broader set of schools.

Figure 4 shows the results of varying the minimum number of reviews threshold on the number of schools considered (left panel) and the improvement over the baseline model (center panel). The left panel shows that reducing the review cutoff from 10 to 6 nearly

[6] We use `sklearn`'s TFIDFVectorizer[28] with `min_df=10`, `max_df=0.5`, and `nltk`'s default part-of-speech tagger to identify proper nouns[5].

[7] BERT has a memory requirement quadratic in the number of words in the document. Scaling to long documents is an active area of research [9, 36].

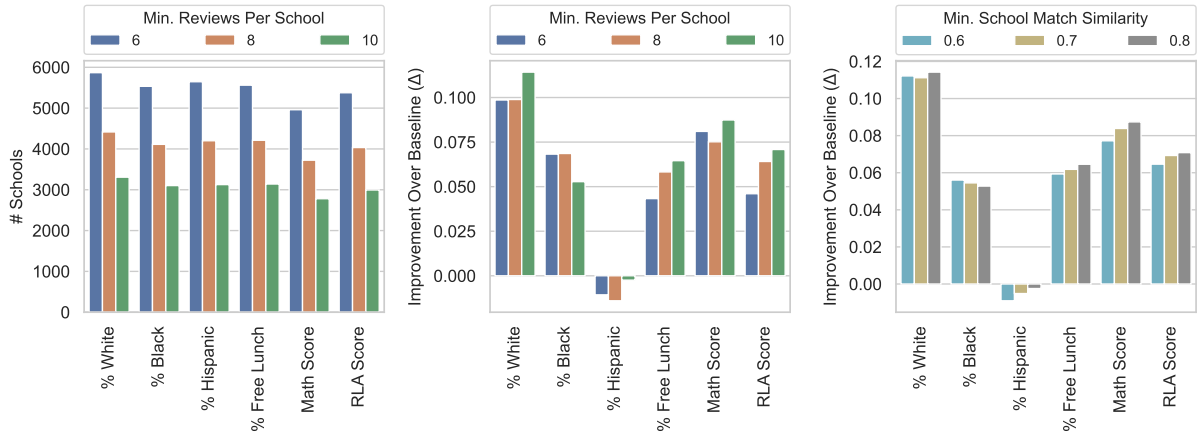


Figure 4: The left and center panels show the effects of restricting analysis to schools with a minimum number of reviews — while the number of schools considered for classification drops, the additional text improves accuracy. The right panel shows that more conservative school matching reduces the accuracy of the baseline model and thus increases the improvement provided by the text features.

Table 2: The words and phrases with the largest classifier coefficients for the positive class for each attribute. Additionally, the Top Categories show the topics that appear most frequently among the top terms from a school choice-related lexicon [15].

attribute	terms with the top coefficients in the classifier predicting an increase in this attribute
White	neighborhood, garden, county, small, magnet, world, beautiful, diverse, incredible, military, engaged, big, terrific, strong, arts, organized, responsive, open, gem, grader, attendance, leadership, loves, heights, chess, major, music, sense community, secret, large, trips, happen, 7th grade, state, community involvement, resources, sense, creative, liberty *Top Categories: overall quality (.09); physical environment (.08); resources (.07)*
Black	academy, security, ms, terrible, average, building, fair, academic, principle, constant, pick, teachers administration, staff members, left, county, pleased teachers, moral focus, fine arts, staff wonderful, talks, club, behavior, problems, events, ratio, graduate, immediately, email, military, anymore, diversity, dismissal, son went, fine, reviews, event, area, improve, participate, motivate *Top Categories: overall quality (.07); school culture (.06); physical environment (.04)*
Hispanic	district, bilingual, sports, preschool, sports programs, positive, students parents, 09, 6th, staff great, willing, quickly, location, noticed, foreign language, rating, understand, teachers outstanding, wrong, music art, everyday, public, honors, problems, providing, wanted, close, handle, help teachers, 45, 5th grader, standards, achieve, communicate, band, spanish, honors classes, language, sons, athletic *Top Categories: overall quality (.08); resources (.07); instruction & learning (.06)*
Free Lunch	graduated, charter, clubs, moral, 4th, great education, counselor, gifted, college, freshman, courses, miss, money, area, faculty, department, great students, senior, awesome, lottery, counselors, rigorous, average, teachers work, advice, safe, moral focus, lived, really care, teams, pay, cafeteria, doing great, help child, french, grandson, current, shown, thank, excellence *Top Categories: resources (.06); school-level features (.06); overall quality (.05)*
Math	neighborhood, district, involvement, campus, science, making, new principal, private, homework, projects, grader, test, parent involvement, fantastic, making sure, middle, diverse, nearly, music, math, field trips, test scores, despite, office, lots, transfer, sad, parent community, bright, kindergarten, principal great, sure, prepared, met, engineering, older, lot parent, competitive, racism, world *Top Categories: overall quality (.1); instruction & learning (.09); school culture (.07)*
RLA	neighborhood, campus, college, sports, test, security, field, white, gifted, senior, small, happier, prepare, succeed, magnet, kid, size, instruction, fairly, field trips, freshman, academic, building, pre, challenging, degree, great deal, zone, seen, scores, girl, gangs, extremely, homework, parent involvement, block, aspects, arts, children currently, scholars, reading *Top Categories: overall quality (.13); instruction & learning (.1); school culture (.06)*

doubles the number of schools considered.[8] The middle panel shows that, except for % Black, having more reviews generally results in a greater accuracy improvement over the baseline. This is mostly expected, as more reviews may provide a more comprehensive characterization of a school. We note that even when we require only at least 6 reviews per school, the text features improve substantially over the baseline for all attributes except % Hispanic, in line with the original results in Table 1.

[8]Note that the number of schools considered varies slightly by attribute, since some are filtered if they do not have attribute observations in both the pre- and post-periods.

5.2.2 Effect of School Matching Threshold. In §4.1, the matching method has a parameter δ, which restricts our dataset to schools who have a matched pair with cosine similarity $\geq \delta$. The results above set $\delta = 0.8$. In this section, we conduct robustness checks to investigate how this parameter influences the results. As δ decreases, we expect matching quality to decrease. This in turn means that the baseline accuracy should improve. On the other hand, this can also increase the size of the dataset, as more schools have a valid match.

The right panel of Figure 4 shows the results of varying δ. We observe that increasing δ generally increases the improvement provided by the text features. This is in part because as we reduce matching quality, the pre-period school attributes become more predictive of post-period attributes. In Appendix A.1, we further consider the case where matching is omitted completely. There, we see even higher overall accuracies for the baseline models, although baseline+text is the most accurate method across all attributes, suggesting that text features add predictive value even in this setting.

5.3 Qualitative Analysis

We now turn to a qualitative analysis to investigate which words and phrases drive the predictive signal. First, Table 2 shows the top 40 unigrams and bigrams for each school attribute according to the classifier coefficients. These represent the terms most strongly associated with schools that increase with respect to each attribute. (Table 4 in the Appendix also shows the words with highest negative association.)

Inspecting these lists reveals several common themes. For example, for % White, terms such as "community involvement," "sense of community," and "engaged" suggest that schools perceived as having strong community engagement are more likely to increase in White enrollment. The additional term "diversity" suggests that some White families are attracted to schools with demographic diversity, although the fact that these schools subsequently increase in White enrollment suggests that racial diversity may decline as a result. These results are also related to ongoing studies of gentrification in K-12 schools [27].

Examining the Math scores, we observe several terms indicating that a math-focus is present in the reviews, such as "science," "math," "engineering." Interestingly, the phrase "test scores" itself appears as a predictive term, suggesting that the reviews reflect a school's focus on test scores. Furthermore, the phrase "new principal" may signal a shift in school leadership that may in turn result in an increase in test scores.

To provide a higher level of abstraction, Table 2 additionally lists the top word categories per attribute. These categories are derived from the word taxonomy of Harris et al. [15], who manually code salient terms from GreatSchool reviews. We map each of the top 100 terms for each attribute to its list of categories in this taxonomy. Table 2 shows the most common word categories and their frequency among the top 100 terms.

The "overall quality" category is prevalent across all attributes, though the specific words vary — e.g., "incredible," "terrific," and "strong," for % White, and "terrible," "average," "fine" for % Black. The category "instruction & learning" is most associated with schools that increase in Math and RLA scores, including terms like "science," "homework,", "rigorous," and "AP." The category "school culture" is most associated with schools that increase in Black enrollment, including terms like "security," "behavior," "dismissal," and "safety." The category "school-level features" is most prominent for % Free Lunch, due to terms like "money," "pay," and "charter."

To provide more context for these terms, Figure 5 shows a sample of reviews with high prediction probabilities (with proper nouns removed). We use the LIME library [30] to extract the most predictive terms per sample, shaded based on how predictive they are of the increasing class (orange) or the decreasing class (blue). It is noteworthy that sentiment alone is insufficient for this task. Figure 5e

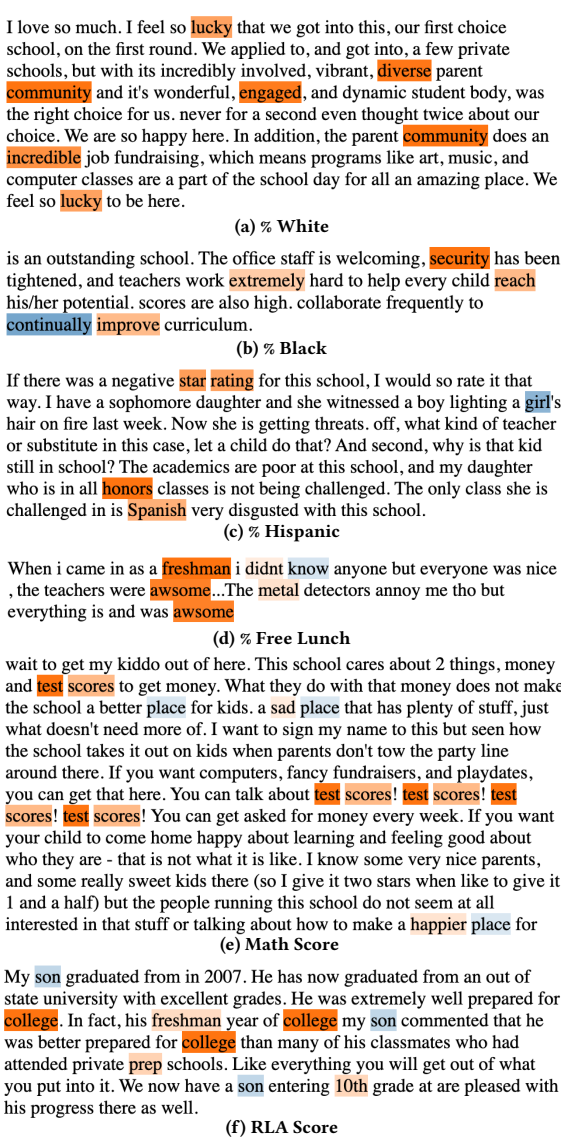


Figure 5: Reviews with high prediction probabilities. Terms are shaded using the LIME library [30] according to how predictive they are of the increasing class (orange) versus the decreasing class (blue).

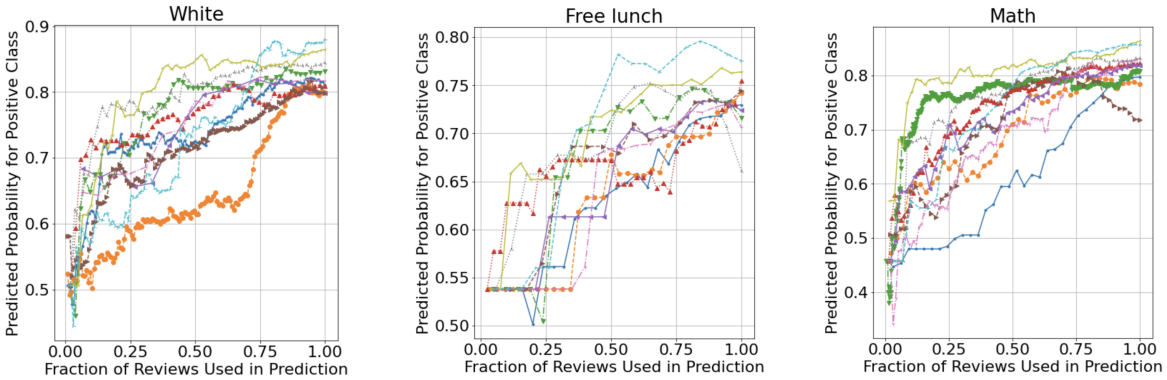


Figure 6: For the ten schools most confidently predicted to increase for each attribute, we plot how the prediction probability changes as we add additional reviews chronologically. Some schools show sudden temporal shifts, suggesting more abrupt changes in school reviews.

shows a review indicative of a school with an increase in math scores. The review emphasizes the school's focus on increasing test scores, though the sentiment is strongly negative.

5.4 Longitudinal Analysis

Finally, we conduct a longitudinal analysis to understand how a school's reviews change over time from the perspective of the classifier. For example, for schools classified as likely to increase in Math scores, can we detect when a shift occurred in the content of the reviews over time that influenced the classification probability?

We address this as follows. For each school, we order their reviews chronologically up to the end of the pre-period (2014). We then apply the classifier with increasing number of reviews as input (e.g., first classify using only the first review, then the first and second review, etc., until all reviews are used). At each iteration, we record the probability of the positive (increasing) class returned by the classifier. Figure 6 shows sample time series from the top ten most confidently predicted schools for three attributes (White, Free Lunch, and Math). Each line represents the trajectory of a single school based on the classification probability.

For each attribute, there is a natural clustering between schools that start with a high probability of the positive class and those that do not. For the latter, we observe an abrupt increase in prediction probability after between 30% and 40% of the reviews are considered. For example, the school corresponding to the bottom, orange line in Figure 6a initially has a low probability of increasing White enrollment based on the first 50% of reviews observed. However, after around 70% of reviews are observed, the probability increases sharply, and continues to rise as with additional reviews. When we examine the reviews for this school before this sharp increase, we find discussion of rising parental involvement ("overwhelmingly impressed with the parent involvement," "This school is on the way up") though critical of some aspects of the school ("Very poor communication from the teachers," "The principal seems indifferent.") During the spike in probability, many reviews begin emphasizing the community aspect of the school ("a true neighborhood community school") as well as specialized programs ("edibles garden," "yoga," "field trip to our state capital.")

While preliminary, this analysis suggest avenues for future work to use the methods proposed here to monitor school reviews in real-time to detect changes in school attributes and/or perceptions.

6 DISCUSSION

Given the growing shift from traditional, neighborhood schools to a market-based model (charter schools, vouchers, etc) [16], understanding how online reviews reflect and influence school perceptions has important implications for the design of Web platforms. Ensuring that such platforms provide informative content while not worsening existing inequities is a challenging socio-technical issue. Existing research into how and when online school reviews are used has shown that these reviews may greatly affect school choice decisions, often driving White families towards schools with fewer Black and Hispanic students. Hence understanding the impact of how these reviews both shape and inform school choices is important to determining the impact of these large scale rating websites on society. At the same time, the vast text data in these reviews provides an opportunity to discover potential factors that influence changes in school attributes. The present work offers quantitative and qualitative evidence that online reviews contain valuable leading indicators of real-world changes in schools.

There are several important limitations to this study. First, the analysis is naturally limited to schools with a sufficient number of online reviews. As prior work has found that engagement with online school reviews varies by demographics [17], a classifier trained on such data may exhibit issues of algorithmic fairness and bias by demographics of the schools and the raters.

Second, while the results suggest that review content provides predictive power beyond pre-period attributes, the overall accuracy is still far from perfect. Unsurprisingly, changes to school attributes are the result of a myriad of social and structural factors. Review content only reflects a small portion of those factors.

Third, while this study suggests that reviews serve as leading indicators of changes in school attributes, it does not attempt to establish causation between review content and school attributes. We speculate that reviews can both serve as indicators of unobserved variables that lead to school changes (e.g., shifting demographics of a neighborhood, new pedagogical strategies) as well as serve as possible factors that contribute to school changes (e.g., by influencing perceived quality that in turn influences school choice). Future work should assess these separately in more detail.

7 CONCLUSION

This study presents evidence that the content of online reviews can serve as leading predictors of changes in K-12 schools. A qualitative analysis of the most predictive words and phrases provides insights into key topics that relate to changes in both the test scores and socio-demographic makeup of a school, including topics such as school leadership, a focus on testing, diversity, and school safety.

8 ETHICS STATEMENT

Our analysis focuses on publicly available school reviews. As Hasan and Kumar [17] point out, disparities in how people access and use school ratings have the potential to further exacerbate inequities in education systems. Similar care should be taken when building predictive models that ingest such data, such as ours, to ensure that the predictions are not systematically biased along dimensions of interest. Although we are primarily interested as social scientists who want to understand factors that precede changes in school attributes our work could also be used by administrators who want to monitor changing perceptions of their school and/or parents who want to understand whether a school is stable or changing. Depending on the application, and whether or not it could potentially lead to adverse selection of some schools over others, is a risk of relying too heavily on our results. However, as mentioned, we feel the analysis here does not create additional risk that was not already present in how and why some parents choose to use information (reviews) already present on the web.

9 ACKNOWLEDGEMENTS

This research was supported in part by the Harold L. and Heather E. Jurist Center of Excellence for Artificial Intelligence at Tulane University. Nicholas Mattei was supported by NSF Awards IIS-RI-2007955, IIS-III-2107505, and IIS-RI-2134857, as well as an IBM Faculty Award and a Google Research Scholar Award.

This research was carried out under the auspices of the National Center for Research on Education Access and Choice (REACH) based at Tulane University, which is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C180025 to The Administrators of the Tulane Educational Fund. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, GreatSchools, or any other organization.

REFERENCES

[1] Ali Alessa, Miad Faezipour, et al. 2019. Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study. *JMIR public health and surveillance* 5, 2 (2019), e12383.

[2] Shannon Altenhofen, Mark Berends, and Thomas G White. 2016. School choice decision making among suburban, high-income parents. *AERA open* 2, 1 (2016), 2332858415624098.

[3] Matt Barnum and Gabrielle LaMarr LeMee. 2019. Looking for a home? You've seen GreatSchools ratings. Here's how they nudge families toward schools with fewer black and Hispanic students. Chalkbeat. *Chalkbeat. https://www. chalkbeat. org/2019/12/5/21121858/looking-for-a-home-you-ve-seengreatschools-ratings-here-s-how-they-nudge-families-toward-schools-wi* (2019).

[4] Shreesh Kumara Bhat and Aron Culotta. 2017. Identifying leading indicators of product recalls from online reviews using positive unlabeled learning and domain adaptation. In *Eleventh International AAAI Conference on Web and Social Media*.

[5] Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. 69–72.

[6] Antoni Serra Cantallops and Fabiana Salvi. 2014. New consumer behavior: A review of research on eWOM and hotels. *International Journal of Hospitality Management* 36 (2014), 41–51.

[7] Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*. 115–122.

[8] Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*. 519–528.

[9] Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems* 33 (2020), 12792–12804.

[10] Danielle Sanderson Edwards and Joshua Cowen. 2021. The Farther You Go, the Closer You Get: Understanding the Roles of Residential Mobility and Distance in Participation in Public School Choice. (2021).

[11] Nabeel Gillani, Eric Chu, Doug Beeferman, Rebecca Eynon, and Deb Roy. 2021. Parents' online school reviews reflect several racial and socioeconomic disparities in K–12 education. *AERA Open* 7 (2021), 2332858421992344.

[12] David Goldberg and Nohel Zaman. 2020. Topic Modeling and Transfer Learning for Automated Surveillance of Injury Reports in Consumer Product Reviews. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

[13] Ellen B Goldring and Kristie JR Phillips. 2008. Parent preferences and parent choices: The public–private decision about school choice. *Journal of Education Policy* 23, 3 (2008), 209–230.

[14] Jaren R Haber. 2021. Sorting schools: A computational analysis of charter school identities and stratification. *Sociology of Education* 94, 1 (2021), 43–64.

[15] Douglas N. Harris, Debbie Kim, Nicholas Mattei, Srihari Korrapati, and Olivia Carr. 2022. A Picture Is Worth 51,930,274 Words: A Text Analysis of Public User Reviews of Schools. In *American Educational Research Association Conference (AERA)*.

[16] Douglas N Harris, John F Witte, and Jon Valant. 2017. The market for schooling. In *Shaping Education Policy*. Routledge, 130–161.

[17] Sharique Hasan and Anuj Kumar. 2019. Digitization and divergence: Online school ratings and segregation in America. Available at SSRN 3265316.

[18] Kia Jahanbin, Vahid Rahmanian, et al. 2020. Using twitter and web news mining to predict COVID-19 outbreak. *Asian Pacific journal of tropical medicine* 13, 8 (2020), 378.

[19] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78, 11 (2019), 15169–15211.

[20] Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1443–1448.

[21] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[22] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. 2016. Rapid assessment of disaster damage using social media activity. *Science advances* 2, 3 (2016), e1500779.

[23] Sook Lim and Nick Steffel. 2015. Influence of user ratings, expert ratings and purposes of information use on the credibility judgments of college students. *Information Research: An International Electronic Journal* 20, 1 (2015), n1.

[24] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[25] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.

[26] Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems* 20 (2007).

[27] Francis A Pearman. 2020. Gentrification, geography, and the declining enrollment of neighborhood schools. *Urban Education* 55, 2 (2020), 183–215.

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.
[29] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108 (2016), 42–49.
[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[31] Donald B Rubin. 1973. Matching to remove bias in observational studies. *Biometrics* (1973), 159–183.
[32] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics* 54 (2015), 202–212.
[33] John P Schomberg, Oliver L Haimson, Gillian R Hayes, and Hoda Anton-Culver. 2016. Supplementing public health inspection via social media. *PloS one* 11, 3 (2016), e0152117.
[34] Aaron Smith, Monica Anderson, and Dana Page. 2016. Online shopping and e-commerce. (2016).
[35] Ming Yang, Melody Kiang, and Wei Shang. 2015. Filtering big data from social media–Building an early warning system for adverse drug reactions. *Journal of biomedical informatics* 54 (2015), 230–240.
[36] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems* 33 (2020), 17283–17297.

A APPENDIX

A.1 Results without matching

Here we present results that omit the matching procedure of §4.1. Figure 7 shows that the distribution of pre-period attribute values vary substantially between schools that increase versus decrease for each attribute. This is most pronounced for the test score figures, which show that schools that increase in test scores are more likely to already have high scores in the pre-period.

This is further supported by the accuracy results in Table 3. Overall, we can see that the baseline method is more accurate without matching, as expected since the school attributes in the pre-period are predictive of attributes in the post-period. Here again the test scores appear most affected by the matching procedure — with matching, the Math and RLA scores are .579/.609, while without matching they are .717/.715.

It is notable that even without the matching procedure the text features still provide predictive value, as the baseline+text method outperforms the baseline approach on all attributes. It thus appears that the text features are not merely characterizing the pre-period school attributes, but are providing additional nuance into school characteristics and perceptions.

A.2 Effect of time period of study

In our main results, we selected 2011-2014 as the pre-period and 2015-2018 as the post-period. While this decision was made based on the volume of reviews available, we wanted to explore how sensitive the results are to the selected time periods. To do so, we considered two other time ranges: (2009-2012) → (2013-2016) and (2010-2013) → (2014-2017). Figure 8 shows the improvement of baseline+text over baseline for each attribute.

While there does appear to be some variation by time range, the text features improve upon the baseline in all cases except % Hispanic for the latest time range. The variation most likely stems from two factors: (1) the number of reviews available in the pre-period; and (2) the influence of unobservable external factors on the attribute changes in those years.

Table 3: Average classification accuracy (area under the ROC curve) for each attribute and method. Δ = (baseline+text − baseline) represents the additional predictive accuracy provided by text features. These results do not use the matching procedure of §4.1, and thus the baseline approach has higher accuracy. However, adding text features still improves in this setting.

target	baseline	text	baseline+text	Δ
% White	.554 ± .01	.626 ± .02	**.633 ± .02**	+.079
Math Scores	.717 ± .01	.659 ± .02	**.756 ± .01**	+.039
RLA Scores	.715 ± .01	.645 ± .02	**.754 ± .01**	+.039
% Free Lunch	.656 ± .03	.632 ± .02	**.700 ± .03**	+.044
% Black	.615 ± .01	.615 ± .02	**.647 ± .02**	+.032
% Hispanic	.656 ± .02	.551 ± .02	**.657 ± .01**	+.001

A.3 Top words that predict a decrease in target attributes

We also have top words that predict a decrease in target attributes, but cut them for space. We add it to the appendix. There is a fair bit of overlap between negative cases and their complement positive class. For example, the top terms in white enrollment going up are similar to those in black enrollment going down.

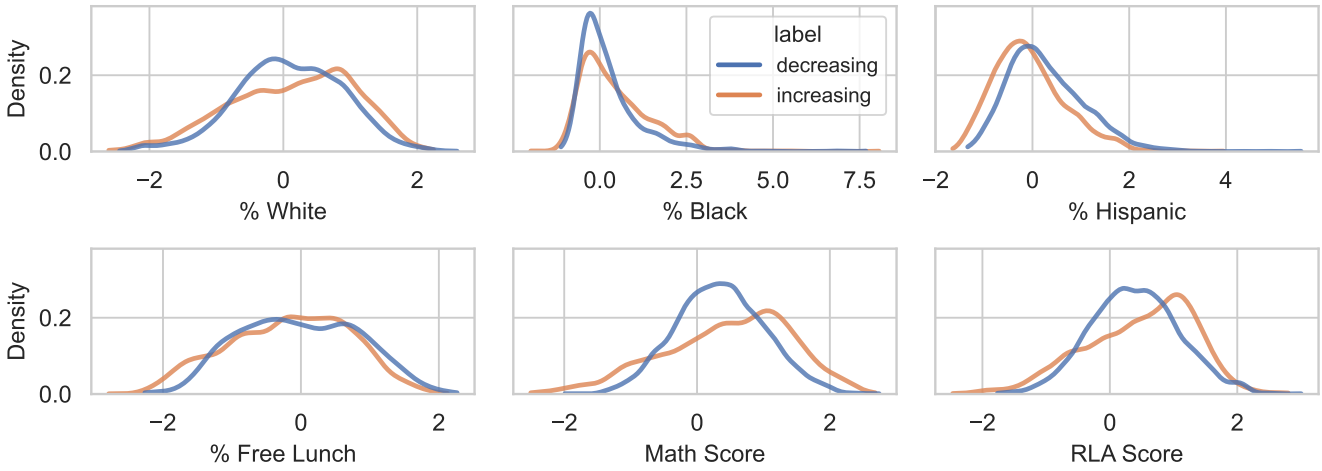


Figure 7: Without the matching procedure of §4.1, the distribution of state-scaled school attributes based on the pre-period values exhibit greater variation between schools that increase and schools that decrease with respect to that attribute. (c.f., Figure 3

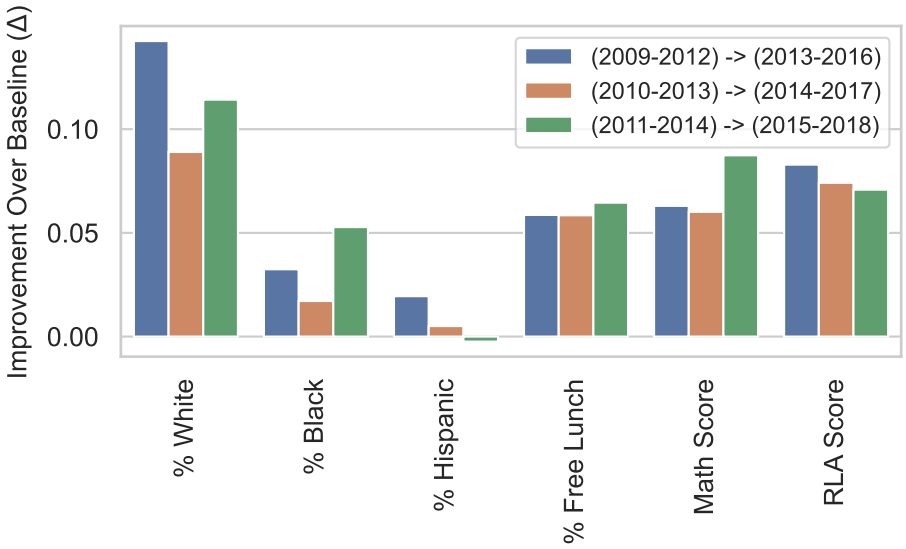


Figure 8: The effect of selecting different pre- and post-periods on the improvement over baseline.

Table 4: The words and phrases with the largest classifier coefficients for the negative class for each attribute. Additionally, the Top Categories show the topics that appear most frequently among the top terms from a school choice-related lexicon [15].

attribute	terms with the top coefficients in the classifier predicting a decrease in this attribute
White	campus, homework, district, kid, left, extra, safe, grade level, traffic, parking, worth, love teachers, parking lot, 1st grade, awesome, complete, curricular, charter, helps, started, giving, college, curricular activities, lunch, students class, aftercare, teachers awesome, happened, student parent, enroll, cut, idea, girls, rules, staff helpful, office, opened, place work, trying, grandchildren, average, 4th, ratings, difference, helping children, excelling, spelling, regular, abilities, end *Top Categories: physical environment (.08); other (.08); instruction & learning (.06)*
Black	sports, 6th, campus, neighborhood, public, try, small, smaller, 6th grade, college, music, beautiful, immersion, spanish, able, culture, 21st, live, music program, lab, district, incredible, forward, junior, large, outside, strict, world, languages, dual, love community, teacher teacher, came, hard work, genuinely, friends, negative, art, grader, trips, jeans, assignments, spend, academically, child education, gang, academics, shy, ball, impressive *Top Categories: resources (.12); physical environment (.12); instruction & learning (.12)*
Hispanic	space, philosophy, math, 2nd grade, moved, strongly, dynamic, instead, projects, warm, mandarin, couldn ask, people say, county, teachers administration, taught, 2nd, meetings, testing, fourth, students come, 13, magnet, option, currently attending, ask better, word, weeks, afterschool, hasn, counselor, past, kids happy, fantastic, great staff, active, bright, families, olds, drugs, offered, complain, ask, leadership, girl, learners, teachers parents, neighborhood, development, spend *Top Categories: instruction & learning (.1); school culture (.08); other (.06)*
Free Lunch	small, diverse, fantastic, neighborhood, district, families, test, strong, responsive, immersion, room, score, backgrounds, spend, parent community, enrichment, 6th, welcoming, committed, public, communication, incredibly, wait, warm, dance, recently, trips, middle, approachable, excited, 8th grade, drugs, beautiful, sizes, secretary, new friends, incredible, tour, community parents, happening, music, 2008, emails, strides, ranking, house, oldest, facility, reviews, feel comfortable *Top Categories: overall quality (.14); school culture (.1); resources (.08)*
Math	charter, club, honors, gone, looking, didn, graduated, performing, teachers don, honors classes, forced, safe, parent coordinator, best teachers, courses, faculty, really good, tell, absolutely, french, teaches, joke, kids teachers, age, drama, alot, students involved, understand, athletic, says, lake, guidance, used, games, rules, played, turnover, los, control, love students, like children, cliques, athletics, needs students, bullied, knowledgeable, charter schools, center, specials, la *Top Categories: resources (.1); instruction & learning (.06); (overall quality (.06)*
RLA	computer, grade level, 6th, charter, dress, safety, children teachers, best teachers, problems, clean, 7th, 6th grade, away, hour, lake, awesome, code, rude, learning environment, allowed, discipline, grade teachers, won, bullies, weekly, volunteers, preschool, wont, children attend, suggest, punished, strict, public, updated, saw, art, advanced, ago, great teachers, gone, level, lacking, grandson, rules, chorus, hoping, reason, learns, safe, answer *Top Categories: resources (.08); school culture (.08); physical environment (.06)*