Using Text-Based Causal Inference to Disentangle Factors Influencing Online Review Ratings

Linsen Li and **Aron Culotta** and **Nicholas Mattei**
Department of Computer Science
Tulane University, New Orleans, LA, USA

Abstract

Online reviews provide valuable insights into the perceived quality of facets of a product or service. While aspect-based sentiment analysis has focused on extracting these facets from reviews, there is less work understanding the impact of each aspect on overall perception. This is particularly challenging given correlations among aspects, making it difficult to isolate the effects of each. This paper introduces a methodology based on recent advances in text-based causal analysis, specifically CausalBERT, to disentangle the effect of each factor on overall review ratings. We enhance CausalBERT with three key improvements: temperature scaling for better calibrated treatment assignment estimates; hyperparameter optimization to reduce confound overadjustment; and interpretability methods to characterize discovered confounds. In this work, we treat the textual mentions in reviews as proxies for real-world attributes. We validate our approach on real and semi-synthetic data from over 600K reviews of U.S. K-12 schools. We find that the proposed enhancements result in more reliable estimates, and that perception of school administration and performance on benchmarks are significant drivers of overall school ratings.

1 Introduction

Understanding the influence of specific aspects mentioned in text reviews on the overall ratings of products or services is a complex yet important endeavor in many industries. For example, in education, how does feedback on academic performance or facility quality influence a school's overall ratings? Precisely quantifying this influence can help businesses identify key areas for improvement.

Traditional approaches to this problem include aspect-based sentiment analysis, which extracts sentiments tied to predefined aspects (Zhang et al., 2022; Kandhro et al., 2024), and exploratory analyses that measure the correlation between certain terms and overall rating (Geetha et al., 2017). However, these methods generally fail to account for confounding variables, leading to biased results. For example, consider a school often praised for its academic performance, with reviews about its excellent programs and great teachers. Traditional analysis might directly link these positive attributes to the school's high ratings. However, if these reviews also frequently mention extensive extracurricular opportunities or high parental involvement, it could imply that the school's high ratings reflect its socioeconomic advantages, not educational quality alone. Overlooking such factors could lead to incorrect assessments of the true influence of educational quality on overall perceptions.

Our goal is to estimate the impact of aspects mentioned in reviews on overall ratings. To achieve this, we have developed a causal inference framework to control for confounding variables within text. Our framework starts by identifying an aspect of interest from the text related to an entity, e.g., if the school reviews frequently praise facility quality. We then utilize the remaining text of the entity, excluding the aspect-related text, as covariates to analyze the overall rating for the entity. Here, the textual content is treated as a proxy for real-world factors that may influence an entity's rating. This approach helps us control for confounders associated with the entity and isolate the treatment effect of the specific aspect on the overall rating.

Contribution We apply CausalBERT (Veitch et al., 2020) to estimate the effects of specific topics on overall review ratings, isolating genuine influences from other textual elements. We enhance CausalBERT by (i) integrating Temperature Scaling to calibrate propensity scores; (ii) optimizing a key hyperparameter that balances treatment and outcome prediction, reducing overadjustment for confounds; (iii) employing interpretability methods to characterize discovered confounds. We validate

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11259–11277

our approach using 600K U.S. K-12 school reviews from GreatSchools.org, finding that issues of administration personnel and academic performance are significant drivers of perceived school quality.

2 Methods

We apply the potential outcomes framework (Neyman, 1923), observing for each subject (school) i a tuple (X_i, Y_i, T_i), where $X_i \in \mathbb{R}^p$ denotes text covariates, $Y_i \in \mathbb{R}$ is the continuous outcome (average review rating), and $T_i \in \{0, 1\}$ is the treatment assignment (presence of topic in reviews). The potential outcomes $Y_i(0)$ and $Y_i(1)$ represent the outcomes under control (no treatment) and treatment scenarios, respectively. The outcome Y_i is defined as $Y_i = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$. The goal is to estimate the Average Treatment Effect (ATE), τ, which quantifies the expected difference in outcomes due to the treatment:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \quad (1)$$

To estimate the ATE, we consider several core assumptions and estimators: *Ignorability* assumes that the treatment assignment T_i is independent of the potential outcomes, a critical condition that allows the use of a naive unbiased estimator ($\hat{\tau}_{unadjust}$) directly:

$$\hat{\tau}_{unadjust} = \mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0] \quad (2)$$

However, this assumption is often unrealistic when treatment assignment correlates with confounds. Thus, we further assume *Conditional Ignorability*, which posits that the treatment assignment T_i is independent of the potential outcomes given the covariates X_i. If we denote $\mathbb{E}[Y_i|X_i = x, T_i = 1]$ as $Q(1, x)$ and $\mathbb{E}[Y_i|X_i = x, T_i = 0]$ as $Q(0, x)$, then the ATE can be estimated by:

$$\hat{\tau}_Q = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Q}(1, X_i) - \hat{Q}(0, X_i)\right) \quad (3)$$

Here, $\hat{Q}(T_i, X_i)$ is the estimated response given treatment status and covariates. *Positivity* assumes that every subject has non-zero probability of receiving the treatment ($0 < P(T_i = 1|X_i = x) < 1$ for all x). With the propensity score $g(x) = P(T = 1|X = x)$, let $\hat{g}(x)$ denote the estimate of the true propensity score $g(x)$. Then the Inverse Probability Weighting (IPW) estimator is:

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\hat{g}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{g}(X_i)}\right) \quad (4)$$

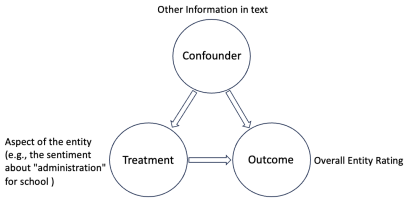

Figure 1: The causal graph of the framework

To mitigate the instability in IPW estimates due to extreme propensity scores, we also use augmented inverse propensity weighted (AIPW) estimator (Robins et al., 1995):

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{\hat{g}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{g}(X_i)}\right] - \left[\frac{T_i - \hat{g}(X_i)}{\hat{g}(X_i)}\hat{Q}(1, X_i) - \frac{T_i - \hat{g}(X_i)}{1-\hat{g}(X_i)}\hat{Q}(0, X_i)\right] \quad (5)$$

2.1 Estimating Framework

We explore how specific features reflected in reviews impact the aggregate rating of the entity. To frame this as a causal inference task, we define the treatment to be any identifiable feature associated with the entity being reviewed, such as sentiments about specific aspects (e.g., 'administration' for schools) or mentions of particular topics (e.g., 'bullying'). For simplicity, we use a keyword-based treatment: We define a set of keywords related to the topic of interest, then categorize each entity into treatment or control groups based on the presence of these keywords in their reviews. The outcome variable is the entity's average review rating across all reviews.

To address the challenges noted by Pryzant et al. (2021) in assessing conditional ignorability – where the treatment label may itself be influenced by other text properties – we separate reviews that determine treatment status from those that do not. That is, all reviews that are not used to determine treatment are concatenated together to serve as covariate variables X_i, while the remaining reviews are discarded after being used to determine treatment. Thus, in this causal inference task (Figure 1), average review ratings are the outcome (Y_i), and the treatment (T_i) is some real world effect determined by the presence of keywords in the reviews. We use the text from reviews without the treatment keywords as covariates (X_i)—serving as a proxy for real-world factors that may influence the entity's rating—to estimate the ATE according to Equation 1.

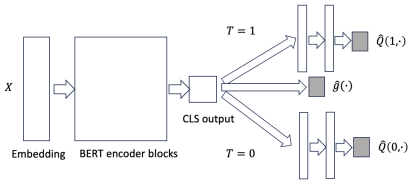

Figure 2: CausalBERT architecture.

2.2 CausalBERT

A key challenge in performing causal inference with text is adjusting for confounding effects within the text. CausalBERT (Veitch et al., 2020), an extension of BERT (Devlin et al., 2018), addresses this challenge by learning text representations that predict both the propensity score $g(.)$ and the conditional expected outcomes $Q(t_i, .)$, thereby learning causally sufficient text representations.

The architecture of CausalBERT (Figure 2), inspired by Dragonnet (Shi et al., 2019), processes textual data into causally relevant embeddings. Initially, CausalBERT utilizes BERT[1] to transform input text x_i into a dense representation, $h_i = \mathrm{B}(x_i; \theta_\mathrm{B})$, captured at the CLS token output. This representation is essential as it captures the critical textual information needed for causal analysis. Following the initial embedding process, CausalBERT extends into three predictive branches: (1) $g_{nn}(h_i; \theta_g)$ for binary treatment assignment using a sigmoid-activated linear map, and (2) $Q^0_{nn}(h_i; \theta_0)$ and $Q^1_{nn}(h_i; \theta_1)$ for potential outcomes under non-treatment and treatment scenarios, respectively, modeled through fully connected layers (each with two hidden layers). CausalBERT optimizes a multi-objective loss function combining mean squared error for outcome predictions and cross-entropy for treatment assignment:

$$(\theta^*, \theta^*_\mathrm{B}) = \arg\min_{\theta,\theta_\mathrm{B}} \frac{1}{n} \sum_i \Big[\left(Q^0_{nn}(h_i;\theta_0) - y_i\right)^2 (1 - t_i) + \left(Q^1_{nn}(h_i;\theta_1) - y_i\right)^2 t_i \Big] + \alpha \, \mathrm{CE}(g_{nn}(h_i;\theta_g), t_i)$$

Here, θ^* represents the parameters of the downstream predictive models, including those for treatment assignment (θ_g) and conditional expected outcome prediction (θ_0, θ_1). The parameter θ_B corresponds to the underlying BERT parameters that are fine-tuned to optimize text representations for the specific causal inference tasks. α is a hyperparameter balancing the prediction accuracies of treatment assignments and outcomes. For scalability, we do not include the 'next sentence prediction' and 'masked language model' tasks typically found in BERT's training regime.

Inference involves passing data through the model to obtain propensity scores $\hat{g}(.)$ and potential outcomes $\hat{Q}(0, .)$, $\hat{Q}(1, .)$, which are then used in the ATE estimators from Section 2. We next describe several enhancements to refine the performance of CausalBERT.

2.3 Temperature Scaling in CausalBERT

In observational studies, treatment and control groups often exhibit a lack of complete overlap in confounder distributions, which hinders the derivation of empirical counterfactuals (Gelman et al., 2021). In CausalBERT, this lack of overlap can arise due to poorly calibrated propensity scores, where extreme $\hat{g}(.)$ values make adjustments unstable. To address this issue, we enhance CausalBERT with Temperature scaling (Guo et al., 2017), which introduces a temperature parameter $M > 0$ to adjust the confidence levels of the propensity score predictions. This parameter helps align the predicted probabilities with their actual confidence levels, mitigating the risk of extreme propensity scores that could negatively affect subsequent estimations such as IPW or AIPW.

In CausalBERT, for a given logit vector z from the treatment prediction branch, the adjusted confidence prediction is: $\hat{g}_{scaled} = \max_k \sigma_\mathrm{SM}\left(\frac{z_k}{M}\right)$, where σ_SM denotes the softmax function and k denotes the class index. The temperature M (where $M > 1$) "softens" the probabilities, increasing output entropy and leading to more uniform class probabilities. Conversely, as M approaches zero, the softmax probabilities converge to a point mass, favoring more confident predictions.

To determine M, we minimize the Negative Log-Likelihood (NLL) of the propensity score predictions on a heldout validation set: $M^* = \arg\min_M \mathrm{NLL}\left(\frac{z}{M}; t\right)$, where t is the true treatment label. Adjusting M does not change the predicted class — it only refines the softmax probabilities to better represent the underlying uncertainties in treatment assignment predictions.

2.4 Mitigating Overadjustment

Overadjustment for potential confounds can lead to biased effect estimates (VanderWeele, 2009). In the CausalBERT objective, α determines the importance placed on the treatment prediction head, which in turn can influence the amount of confounder adjustment. We propose setting α based

[1] In our project, we use pre-trained DistilBERT (Sanh et al., 2019), a smaller, faster, and lighter version of BERT base.

on estimates of the amount of confounding in the data. To estimate the amount of confounding, we use the accuracy of treatment prediction as a signal. This assumes that lower treatment classification accuracies generally indicate weaker confounding. In such cases, we can increase α, thereby intensifying the model's focus on treatment classification without the risk of substantial bias from confounds. By correlating α with observed treatment accuracy, we employ an empirical approach to adjust α, enhancing causal effect estimation across confounding scenarios. We explore this further in §4.

2.5 Interpreting CausalBERT

As true causal effects are rarely known, it is important to have qualitative methods to assess the validity of CausalBERT. We explore two qualitative methods to do so: **CLS Comparative Analysis** and **Integrated Gradients**.

First, building on interpretability methods for deep learning, our **CLS Comparative Analysis** quantifies the aggregate attention for the CLS token and compares the fine-tuned CausalBERT with baseline DistilBERT to determine how fine-tuning affects term importance. We analyze tokens that significantly influence the CLS token – excluding stopwords and punctuation – using two strategies. **General Top Contributing Tokens** ranks tokens by attention score, selecting the most influential for each document, while **Max Subarray for Continuous Contribution** identifies contiguous token subarrays with maximum influence on the CLS token. By aggregating these tokens across all documents, we compare the top influential tokens between CausalBERT (A), which controls for confounding effects through its design, and DistilBERT (B), which does not, using $A \setminus B$ to assess changes in attention due to fine-tuning. Additionally, in semi-synthetic experiments below, we assess the proportion of $A \setminus B$ tokens that correspond to the confounding variable we inject into the data, providing an additional check that CausalBERT is discovering confounders appropriately.

Second, we employ **Integrated Gradients (IG)** (Sundararajan et al., 2017), an interpretability technique that attributes the prediction of a deep learning model to its input features. For each output component of CausalBERT (treatment and outcomes), IG identifies the tokens that significantly increase or decrease the model's predictions. By aggregating across instances, we compile the most influential tokens for each prediction task. We denote g^+ and g^- as the top terms that respectively increase and decrease the propensity score prediction, while Q_0^+ and Q_1^+ describe the top terms that enhance the outcome predictions for the control and treated groups, and Q_0^- and Q_1^- for those that diminish these predictions. Each token in these categories is associated with a contribution weight that quantifies its impact on the model's output. These weights are normalized within each respective list to highlight the relative importance of each term.

3 Experiments

We empirically examine the capabilities of CausalBERT and the proposed enhancements, using both real and semi-synthetic review data, with a focus on the following questions. **RQ1:** How does CausalBERT performance vary with confounder strength? **RQ2:** What effect does Temperature Scaling have on treatment effect estimation? **RQ3:** How does hyperparameter α in the loss function influence overadjustment and how does its optimal value relate to treatment prediction accuracy? **RQ4:** How effective are interpretability methods at surfacing the confounders discovered by CausalBERT? **RQ5:** To what extent do educational aspects, such as 'bullying' and 'administration,' impact overall school ratings in real-world data?

Dataset We analyze 677,210 reviews from GreatSchools.org [2] , covering 83,795 public, private, and charter schools in the United States between 2002-2019. We investigate the impact of school-related topics such as 'bullying,' 'academic performance,' 'administration,' 'extracurricular activities,' and 'curriculum,' each defined by a keyword list established by prior work (Harris et al., 2022; Gillani et al., 2021) (Appendix A.6). For each topic, we first separate reviews into those that discuss the topic and those that do not. For outcome, we normalize review ratings by computing state-specific z-scores and average these scores for each school over the selected period. Thus, outcome values are in standard units. For treatment assignment, we adopt different methods based on the nature of the topics discussed in the reviews. For neutral topics such as 'administration,' treatment is determined by sentiment about that topic. In this case, an entity is considered treated ($T = 1$) if all relevant reviews express positive sentiment and untreated ($T = 0$) if all are negative. Schools

[2]The dataset was provided by our collaborator GreatSchools and is not publicly available.

with a mix of positive and negative reviews on administration are excluded from the analysis to maintain a clear treatment distinction. For the negative topic of 'bullying,' treatment is 1 if the topic is mentioned, 0 otherwise.

In our framework, the 'bullying' task aims to assess how the presence of bullying affects the school's overall ratings. For other topics like 'administration,' we treat the sentiment of text referencing administration as a proxy for real-world administrative quality and measure how that factor's sentiment influences the school's overall rating. Our causal pathway assumes multiple school attributes can affect overall ratings, so we isolate the effect of a specific aspect by controlling for other factors. For instance, to examine bullying, we separate it from other negative conditions—like poor administration—that might also lower ratings. We have between 3,900 and 13,300 schools for each topic; more detailed statistics (e.g., average total reviews per school) are provided in §A.1.

Semi-Synthetic Data Setup First, we employ a semi-synthetic evaluation framework (Weld et al., 2022) to evaluate CausalBERT's treatment effect estimation capabilities using the bullying topic. We simulate a binary confound $C_i \in \{1, 2\}$ by inserting text related to an academic challenges topic into certain reviews (see §A.2 for terms). Schools in Class 1 receive these injected sentences, while those in Class 2 do not.

To manipulate the ATE, we vary the true ATE by defining two outcome models based on treatment status: for Class 1, outcomes are modeled as $Y \sim \mathcal{N}(u_2, 0.3)$ when treatment $T = 1$ and $Y \sim \mathcal{N}(u_1, 0.3)$ when $T = 0$. Class 2 maintains uniform effects with $Y \sim \mathcal{N}(u_2, 0.3)$, indicating no treatment effect from textual confounding. Here, u_2 is set at -0.3, and u_1 varies within $\{0.3, 0.4, 0.5\}$, creating corresponding true ATE u values of $\{-0.3, -0.35, -0.4\}$.

The confounder strength is controlled by adjusting the probability p, which defines the treatment assignment probabilities within each class. We vary p from 0.9 to 0.5, where for Class 1: $P(T = 1|C = 1) = 1 - p$ and $P(T = 0|C = 1) = p$, and inversely for Class 2: $P(T = 1|C = 2) = p$ and $P(T = 0|C = 2) = 1 - p$.

In our semi-synthetic experiments, we construct data according to specific values for u and p by sampling schools to satisfy these constraints. Thus, the only synthetic part of these experiments is the sampling procedure and the injection of confounding sentences. Each experiment samples 5,000 instances from a population of 13,361.

Evaluation Metrics We perform 5-fold cross-validation for the semi-synthetic data experiments and bootstrap aggregation for real-world data analysis (§A.7). Estimators $\hat{\tau}_Q$, $\hat{\tau}_{IPW}$, and $\hat{\tau}_{AIPW}$, and their calibrated versions, are used to estimate the ATE, with the naive estimator $\hat{\tau}_{unadjust}$ as a baseline. Since we have different true causal treatment effect designs, we use the error ratio to evaluate the estimation results, defined as $|\hat{\tau}_{\text{est}} - \tau_{\text{true}}|/\tau_{\text{true}}$, where $\hat{\tau}_{\text{est}}$ is the estimated treatment effect and τ_{true} is the true treatment effect. We also report the accuracy of treatment prediction and the mean squared error (MSE) of outcome prediction.

4 Results

4.1 Evaluation on Semi-Synthetic Data

Treatment Assignment Prediction Table 1 reports average treatment prediction accuracy (α=0.33) for different confounder strengths p and true ATE u. For every fixed u, the accuracy of treatment assignment prediction increases linearly as the confounder strength increases. These results provide evidence that CausalBERT can capture the relationship between treatment and confounder variables. Furthermore, the results suggest that treatment prediction accuracy can serve as a reliable indicator of confounder strength in real-world data, which we will return to below.

	Treatment Accuracy			MSE		
p	u = -.40	u = -.35	u = -.30	u = -.40	u = -.35	u = -.30
.5	.59±.01	.58±.03	.57±.02	.077±.01	.069±.00	.080±.01
.6	.62±.02	.62±.01	.61±.02	.088±.01	.074±.00	.085±.01
.7	.67±.02	.67±.02	.63±.01	.067±.00	.077±.01	.085±.02
.8	.77±.01	.74±.01	.73±.02	.066±.01	.071±.01	.070±.00
.9	.83±.02	.82±.02	.83±.01	.068±.01	.069±.01	.075±.01

Table 1: Average treatment classification accuracy and outcome MSE using CausalBERT(α=0.33) for different confounder strength p and true ATE u.

Outcome Prediction Table 1 also shows the average MSE of outcome prediction. Unlike treatment accuracy, MSE does not exhibit sensitivity to confounder strength. This suggests that the regression task of predicting outcomes is less challenging for CausalBERT compared to the classification task of treatment prediction, indicating different levels of complexity and sensitivity in these tasks.

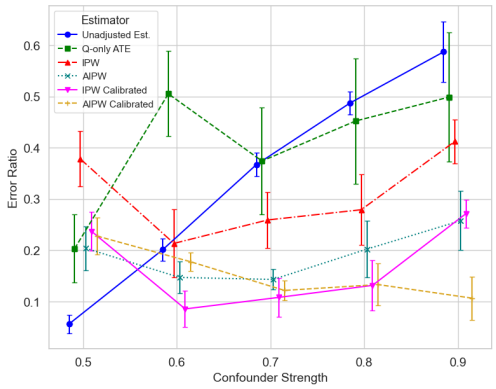

Figure 3: Average error ratio (with standard error) of treatment effect estimation by confounder strength.

Effects of Confounder Strength Figure 3 shows the performance of different estimators across various confounder strengths for semi-synthetic datasets with a fixed true ATE of -0.3 (see Appendix for similar results with true ATE values -0.35 and -0.4). The baseline unadjusted ATE estimator performs better as confounder strength decreases, achieving optimal results at a confounder strength of 0.5. This observation aligns with expectations for a randomized trial scenario, where treatment assignment is completely random and devoid of confounding biases, thus rendering this naive estimator unbiased.

In contrast, the Q-only, IPW, and AIPW estimators significantly outperform the baseline when confounder strength is high ($p > 0.7$). Both IPW and AIPW outperform Q-only in scenarios with strong confounders, suggesting that the propensity scores calculated by CausalBERT are particularly beneficial in aiding robust ATE estimation.

Calibrated IPW and AIPW estimators display the most consistent performance across various settings, suggesting that temperature scaling, which aligns predicted probabilities more closely with their true confidence levels, enhances reliability. However, under conditions of weak confounder strength ($p = 0.5$), CausalBERT tends to underperform compared to the unadjusted estimator. This could be due to the model capturing irrelevant information perceived as confounding. Next we will further explore these observations on temperature scaling and overadjustment.

Effect of Temperature Scaling Figure 4 (Left) presents the average error reduction provided by temperature scaling on estimates from IPW ATE ($\alpha = 0.4$) across various confounder strengths and true ATEs. Temperature scaling consistently enhances IPW estimates by aligning predicted probabilities more closely with their actual confidence levels, mitigating the issue of extreme propensity scores. This is evident as scaled IPW consistently shows reduced error ratios compared to non-scaled IPW across all configurations. For AIPW (see Appendix), the benefit of scaling is less pronounced and primarily observed in high confounder strength scenarios. AIPW's inherent double robustness mechanism may make it less susceptible to the pitfalls of extreme propensity scores.

Both IPW and AIPW show significant improvement from scaling at a confounder strength of 0.9. This may be attributed to the propensity model's increased risk of producing extremely confident predictions (too close to 0 or 1) at high treatment probabilities, where the calibration can have a substantial impact. For a further analysis, see §A.5.

Mitigating overadjustment Figure 4 (Center) shows the linear relationship between confounder strength and treatment prediction accuracy, trained with $\alpha = 0.5$. This analysis suggests that treatment prediction accuracy can serve as an indicator of the amount of confounding within a dataset. Building upon this relationship, Figure 4 (Right) explores the impact of varying $\alpha \in \{0.2, 0.5, 0.8\}$ on model performance for calibrated AIPW on a fixed true ATE of $u = -0.3$. (See Appendix for IPW results.) Taken together, these results suggest that we can guide our selection of α based on treatment prediction accuracy. Notably, at lower treatment accuracies, indicative of weaker confounding, $\alpha = 0.8$ yields the best ATE estimates. Conversely, in regions of higher treatment accuracy, $\alpha = 0.5$ performs best. This analysis can guide the selection of an appropriate α value: a higher α when treatment accuracy is around or below 0.6 and a moderate α when treatment accuracy exceeds 0.65.

A possible explanation for the effectiveness of a higher α in low-confounding scenarios is that it shifts the model's focus towards treatment classification, enhancing its sensitivity to treatment signals. In these settings, clearer signals emerge because the primary challenge is not adjusting for confounds but accurately identifying treatment presence. Thus, by prioritizing treatment prediction, the model more effectively captures and learns from these direct treatment effects, improving its performance in estimating treatment impacts.

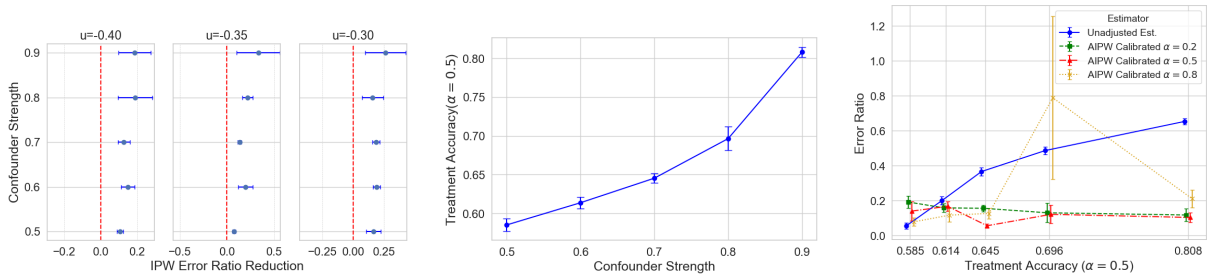

Figure 4: (Left) Error ratio decrease by temperature scaling on IPW. (Center) Treatment accuracy (standard error) for CausalBERT (α=0.5) by confounder strength. (Right) Error ratio by treatment accuracy and α (trained with $\alpha = 0.5$) on a semi-synthetic dataset with a fixed true ATE $u = -0.3$, for the AIPW Calibrated model.

CLS Comparative Analysis We now turn to a qualitative analysis to investigate which words and phrases drive the predictive signal in CausalBERT. The analyses in this section are conducted across semi-synthetic datasets, fixing the true ATE $u = -0.3$, setting $\alpha = 0.4$, and varying confounder strengths from 0.9 to 0.5.

To understand which factors are potential confounds discovered by CausalBERT, Table 2, displays results from applying the CLS comparison method while varying confounder strength. We observe that when the confounder strength is strong ($p = .9$ or $.8$), the tokens in $A \setminus B$ predominantly derive from our inserted confounder text, indicating that supervised fine-tuning enhances the model's emphasis on confounder information within the text representation. When the confounder strength is weak, however, the model shifts its focus towards the treatment itself, with tokens such as 'bull,' 'horrible,' 'rude,' and 'bad' emerging prominently.

p	Top terms	Prop.
.9	**##pf**, **exams**, sweet, special, involvement, involved, **handling**, **##istic**, **course**, make, **structured**, **study**, want, kind, **thinking**, **transferring**, **##hel**, **sy**, even, **face**, **understand**, **challenges**, **##ul**, **##bus**, **believe**, know, **semester**, environment, **advice**, **professor**	.667
.8	simply, **##pf**, **exams**, deeply, **##real**, **handling**, **course**, **deal**, **sizes**, **materials**, level, **feedback**, caring, **transferring**, **assignments**, focused, elementary, oldest, ##ted, **sy**, gifted, **face**, **understand**, **poorly**, **challenges**, **assistants**, **##ul**, awesome, growing, job, **large**, **believe**, **semester**, amazing, **advice**, smile, **professor**	.622
.5	told, needs, ##ing, experience, middle, last, people, bad, need, never, ##t, **un**, nothing, even, horrible, rude, go, bull, ##ied	.053

Table 2: Terms with high attention in CausalBERT but not DistilBERT by confounder strengths p. Bold terms were inserted by the synthetic confounder. Prop. is fraction of top tokens from the confounder template.

These tokens are closely associated with the manifestation of bullying, suggesting that CausalBERT is sensitive to direct treatment signals when the confounder-treatment correlation is minimal. These results indicate that CausalBERT effectively identifies confounding topics when the confounding strength is high, but has more difficulty doing so when confounding strength is low.

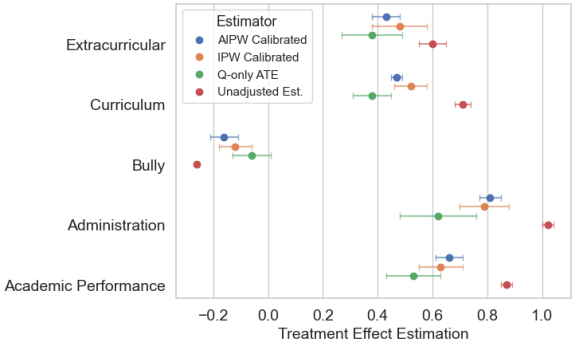

Figure 5: ATEs by topic (α=0.5).

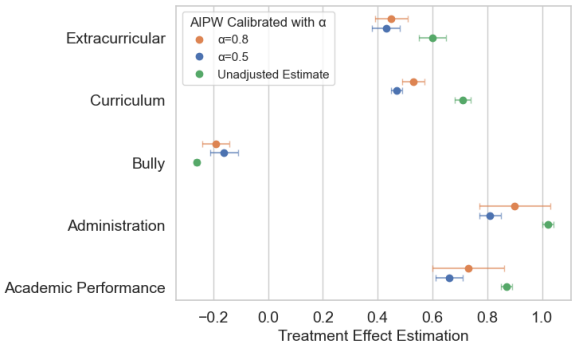

Figure 6: AIPW Calibrated estimates by topic and α .

4.2 Application to Original Data

We next apply CausalBERT to the original school review data, without any synthetic data injection. Figure 5 presents the bootstrapped treatment effect estimates by topic with $\alpha = 0.5$. We observe that our adjustment for confounders consistently reduces the magnitude of the effect estimates

when compared to the unadjusted estimates. For instance, in the case of 'bullying', AIPW-Calibrated produces an effect of -0.16, in contrast to a more substantial decrease of -0.26 in the unadjusted estimate. This difference suggests the presence of confounding factors, such as poor administration, that often co-occur with bullying incidents, lead to a compounded negative impact on school ratings. Similarly, in 'academic performance', where AIPW-Calibrated shows a treatment effect of +0.66 compared to the unadjusted effect of +0.87, positive academic performance may coexist with other favorable conditions, such as educational programs and community involvement.

Comparing the overall effect sizes, 'administration' has the highest effect, followed by 'academic performance,' 'curriculum', and 'extracurricular activities.' In contrast, 'bullying' has a negative impact, as expected. We note that it may be difficult to directly compare the magnitudes of the bullying topic with the others, due to the difference in how the treatment categories were determined (§3).

4.2.1 Integrated Gradients

g^+	**Bullying**	g^-
horrible, terrible, rude, worst, ##ing, un, bad, lack, bull, seems, office, needs, nothing, ##t, negative, different, disappointed, many, ##ied, problem		great, love, best, wonderful, amazing, excellent, loves, every, caring, awesome, much, top, say, happy, truly, family, dedicated, community, support, pleased
g^+	**Academic Performance**	g^-
great, love, school, excellent, amazing, best, program, staff, happy, community, know, always, wonderful, make, learn, really, highly, part, awesome, dedicated		principal, get, administration, needs, horrible, education, new, want, rude, worst, care, leadership, bad, nothing, terrible, bullying, feel, reviews, di, bull

Table 3: Top 20 terms for treatment assignment prediction (g^+ and g^-) by Integrated Gradients.

Table 3 presents the top 20 terms for g^+ and g^- by applying Integrated Gradients for 'bullying' and 'academic performance' topics, highlighting how specific terms influence treatment prediction. For 'bullying,' terms like 'horrible', 'terrible', and 'rude' significantly increased the propensity score, indicating that bullying often co-occurs in schools with many other negative reviews. Conversely, positive terms such as 'great', 'happy', 'love', 'support', 'family', and 'community' decreased the propensity score, reflecting environments where bullying is likely absent and the school atmosphere is perceived positively. For 'academic performance', positive sentiment terms are associated with schools perceived favorably academically, while administrative terms like 'principal' and 'administration' correlate with negative perceptions of academic performance. This pattern indicates that perceptions of academic quality are correlated not only with leadership and administrative factors but also with community engagement and the quality of educational programs.

Selecting α To apply CausalBERT to real data, we assess the confounder strength by first applying CausalBERT with a moderate α to estimate treatment prediction accuracy. The relationship between this initial accuracy and the estimation performances, as in Figures 4 (Center), informs the selection of an appropriate α for subsequent analysis. Once an optimal α is determined, we recommend re-running CausalBERT with this adjusted α and employing temperature scaling to enhance the robustness of the estimate.

Applying this approach, we find that treatment prediction accuracies by topic are: bullying=.53, extracurricular=.59, curriculum=.62, academic performance=.63, administration=.65, suggesting low to moderate confounder strength across all topics. According to Figure 4 (Right), we repeat the experiment with a higher $\alpha = 0.8$, using AIPW Calibrated estimator, since it is the most robust in our semi-synthetic experiments. The results in Figure 6 show that the estimated treatment effects are consistently larger than those at $\alpha = 0.5$ and closer to the unadjusted estimates. This new result with $\alpha = 0.8$ may offer more accurate estimates, as prior experiments with semi-synthetic data showed that, under weak confounder conditions, $\alpha = 0.8$ tended to be the most precise.

5 Related Work

Aspect-based sentiment analysis analyzes sentiment toward specific topics (Hu and Liu, 2004; Tang et al., 2015; Ruder et al., 2016), yet generally does not address the causal impact of these aspects on overall ratings, potentially conflating correlation with causation. There is growing research in causal inference with text (Keith et al., 2020; Feder et al., 2021; Keith et al., 2023; Veljanovski and Wood-Doughty, 2024). A common approach is to use NLP to identify confounds and adjust for them in estimation (Sridhar and Getoor, 2019; Roberts et al., 2020; Mozer et al., 2020). Other methods method adjust for confounders from text with supervised NLP. Veitch et al. (2020) introduced CausalBERT, leveraging pre-trained BERT

models (Devlin et al., 2018) to derive “sufficient" embeddings that capture confounding properties within texts, optimizing predictions for both treatment and counterfactual outcomes. While they highlighted the potential for deep learning methods to improve estimation accuracy, they also outlined several future challenges: refining deep learning approaches to enhance estimation accuracy; developing visualization and sensitivity analysis tools to clarify the “black box” nature of embeddings; and expanding semi-synthetic simulations into a comprehensive benchmarking strategy. These challenges have inspired the work in this paper.

6 Conclusions

In this study, we have extended CausalBERT to understand how factors mentioned in school reviews affect overall ratings. Through semi-synthetic experiments, we verified the effectiveness of CausalBERT in a controlled setting, which then guided our application to real-world data. Our analysis indicates that Temperature Scaling and Integrated Gradients can refine causal estimates and enhance interpretability. Analysis of U.S. K-12 school reviews found that educational aspects like ‘Administration’ and ‘Academic Performance’ have significant influence on school ratings.

7 Limitations

Like most studies involving causal inference, true effects are unknown, and thus there is unavoidable uncertainty in the results. Despite rigorous validation using semi-synthetic data that demonstrates the model’s effectiveness in controlled scenarios, the extrapolation of these results to real-world data must be treated with caution.

Additionally, our reliance on keyword-based treatment identification introduces another layer of potential noise. This method assumes that the presence of predefined keywords, such as ’bullying’, sufficiently identifies relevant reviews.

8 Ethics Statement

Our analysis focuses on publicly available online school reviews. While we are primarily interested in understanding specific school-related topics—such as "bullying," "academic performance," "administration," "extracurricular activities," and "curriculum" that influence overall school ratings, our work could also be utilized by administrators or parents, potentially leading to unintended consequences for certain schools. We caution against over-reliance on our results and emphasize considering each school’s unique context. All data were provided by our collaborator, GreatSchools, in an anonymized form, containing no personally identifiable information. We only use aggregated, school-level data; any individual-level identifiers, such as reviewer names or addresses, were removed prior to our access.

Acknowledgments

This work was supported in part by the Harold L. and Heather E. Jurist Center of Excellence for Artificial Intelligence at Tulane University and the Tulane University Center for Community-Engaged Artificial Intelligence. Linsen Li was supported by NSF Award IIS-III-2107505. Aron Culotta was supported by NSF Awards IIS-HCI-2333537 and SCC-IRG-2427237. Nicholas Mattei was supported by NSF Awards IIS-RI-2007955, IIS-III-2107505, IIS-RI-2134857, IIS-RI-2339880 and CNS-SCC-2427237.

We would like to thank Douglas N. Harris and Jamie M. Carroll in their help with this work. Portions of this research were 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 R305C100025 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, Great Schools, or any other organization.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

M Geetha, Pratap Singha, and Sumedha Sinha. 2017. Relationship between customer sentiment and online customer ratings for hotels-an empirical analysis. *Tourism Management*, 61:43–54.

Andrew Gelman, Jennifer Hill, and Aki Vehtari. 2021. *Regression and other stories*. Cambridge University Press.

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*, 7:2332858421992344.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

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)*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Irfan Ali Kandhro, Fayyaz Ali, Mueen Uddin, Asadullah Kehar, and Selvakumar Manickam. 2024. Exploring aspect-based sentiment analysis: an in-depth review of current methods and prospects for advancement. *Knowledge and Information Systems*, pages 1–31.

Katherine A Keith, Sergey Feldman, David Jurgens, Jonathan Bragg, and Rohit Bhattacharya. 2023. Rct rejection sampling for causal estimation evaluation. *arXiv preprint arXiv:2307.15176*.

Katherine A Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*.

M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L Jason Anastasopoulos. 2020. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, 28(4):445–468.

Jerzy Neyman. 1923. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2021. Causal effects of linguistic properties. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121.

Donald B Rubin. 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2:169–188.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, Austin, Texas. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. *arXiv preprint arXiv:1906.04177*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal. Association for Computational Linguistics.

Tyler J VanderWeele. 2009. On the relative nature of overadjustment and unnecessary adjustment. *Epidemiology*, 20(4):496–499.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on uncertainty in artificial intelligence*, pages 919–928. PMLR.

Marko Veljanovski and Zach Wood-Doughty. 2024. Doublelingo: Causal estimation with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 799–807.

Galen Weld, Peter West, Maria Glenski, David Arbour, Ryan A Rossi, and Tim Althoff. 2022. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1109–1120.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

A Technical Appendix

A.1 Schools Per Topic Table

topic	$T = 1$	$T = 0$	total
bullying	4,688	8,673	13,361
administration	2,940	1,049	3,989
academic performance	4,111	1,670	5,781
extracurricular	5,406	1,385	6,791
curriculum	4,431	2,063	6,494

Table 4: Schools in treatment and control groups by topic.

Topic	Total Reviews		Keyword-Containing Reviews	
	$T = 1$	$T = 0$	$T = 1$	$T = 0$
Bullying	11.08 ± 9.06	8.34 ± 5.23	1.60 ± 1.10	0.00 ± 0.00
Administration	8.09 ± 5.34	6.62 ± 2.14	3.41 ± 2.87	2.35 ± 1.39
Academic Performance	8.01 ± 4.76	6.79 ± 2.37	2.39 ± 1.74	1.77 ± 0.98
Extracurricular	9.64 ± 6.55	8.22 ± 4.12	1.82 ± 1.35	1.27 ± 0.58
Curriculum	8.61 ± 5.74	7.37 ± 2.90	2.05 ± 1.69	1.48 ± 0.80

Table 5: Average ($\pm$ standard deviation) of total and keyword-containing reviews for treated ($T = 1$) and untreated ($T = 0$) per schools by topic.

A.2 Academic Challenge Post Templates

This subsection outlines the templates and words used to generate academic challenge posts for our semi-synthetic data generation process. We use the following sentence templates to simulate academic challenges faced by students, with placeholders indicated by {}:

- "I can't believe I have to deal with {} in this course."
- "Every semester, I face {} in my classes."
- "The professor doesn't understand the challenges of {}."
- "Does anyone have advice on handling {} in school?"
- "I'm thinking of transferring because of {}."

The placeholders are filled with words representing academic challenges: "unrealistic assignments", "difficult exams", "lack of study materials", "unhelpful teaching assistants", "large class sizes", "lack of feedback", "poorly structured syllabus".

Data Generation Procedure: Each synthetic data point is generated by randomly selecting one of the above words and inserting it into a randomly chosen template. This process introduces variability and simulates real-world academic challenges, facilitating the evaluation of CausalBERT's performance in handling text-based confounders.

A.3 Estimator Comparison

In this section, we provide additional insights into the comparison of various treatment effect estimators used in our analysis.

Naive Estimator ($\hat{\tau}_{unadjust}$): The simplest approach, assumes ignorability, where the treatment is independent of potential outcomes. While straightforward, this estimator is often unreliable in practice due to potential confounding variables that correlate with treatment assignment.

Q-only Estimator ($\hat{\tau}_Q$): It refines the Naive Estimator by assuming conditional ignorability, where treatment assignment is independent of potential outcomes given covariates X_i. This method adjusts for confounders but depends heavily on accurately modeling the relationship between covariates and outcomes, which can be challenging with varying covariate distributions across treatment and control groups(Rubin, 2001).

Inverse Probability Weighting (IPW): To address the limitations of the Q-only Estimator, IPW introduces the concept of propensity scores, $g(x)$, estimating the probability of treatment given covariates. This method reweights observations to balance covariate distributions between treated and untreated groups. However, IPW is sensitive to extreme propensity scores ($\hat{g}(x)$ close to 0 or 1), which can lead to unstable estimates.

Augmented IPW (AIPW): AIPW enhances IPW by combining the propensity score with outcome modeling. This dual adjustment stabilizes the estimation process, particularly when extreme propensity scores are present, offering a more robust approach to treatment effect estimation.

A.4 CausalBERT Parameter Detail

The CausalBERT model, similar to DistilBERT, utilizes a model dimension of 768 with an input maximum length of 512 tokens. The embeddings layer consists of word embeddings with a size of $[30\,522, 768]$ and position embeddings with a size of $[512, 768]$. The transformer layers, comprising six blocks, have attention and feedforward networks, each with parameter vectors of size $[768, 768]$ and $[768, 3072]$ respectively. For the downstream tasks, such as treatment assignment prediction and outcome regression, the model uses fully connected layers with a hidden dimension of 200. In total, the model contains approximately 66 million parameters, with key components distributed as follows: embeddings layer (24 million),

transformer layers (43 million), and downstream layers (0.3 million).

A.5 Propensity score distribution

To investigate why temperature scaling most significantly enhances IPW and AIPW at higher confounder strengths (as shown in Figure 4 (Left) and Figure 11), we begin by defining the overlap metric as follows:

$$\text{Overlap} = \frac{1}{N}\left(\sum_{i:T_i=0} \hat{g}_i + \sum_{i:T_i=1} (1-\hat{g}_i)\right),$$

where N is the total number of observations, $\hat{g}_i$ is the predicted propensity score for observation i, and T_i is the treatment label for observation i. This Overlap metric assesses the similarity of propensity score distributions between treatment/control groups. A higher Overlap indicates a better balance of observed covariates between groups, which enhances the reliability of causal inference.

Figure 7 shows the Q–Q Plot of Overlap Differences (the overlap increment after temperature scaling) vs. Temperature Scaling Boost on AIPW across different confounder strengths p from 0.9 to 0.5 and true ATE $u = -0.4$, computed from an overall validation set in cross-validation with CausalBERT(α=0.4). We observe that at higher confounder strengths, particularly $p = 0.9$, there is a marked increase in the overlap increment, which correlates with significant improvements in AIPW estimation accuracy and vice versa for lower confounder strength. This trend suggests that the temperature scaling interventions effectively mitigate the distortions in the propensity score distributions that are more pronounced at higher confounder levels. Consequently, the propensity score estimate is rectified towards a more moderate range, preventing the extreme values that typically skew the analysis at high confounder strengths. This improved covariate balance directly enhances AIPW estimation, demonstrating the benefits of scaling interventions in strongly confounded scenarios.

To enhance the robustness of our findings in Figure 7, we extend our analysis by examining the Pearson Correlation Coefficients (Schober et al., 2018) across multiple true ATE values ($u = -0.4, -0.35, -0.3$). This involves two specific correlations: first, between the confounder strength and the increment in overlap, and second, between the overlap increment and the boost in AIPW estimation accuracy. Across these different true ATE

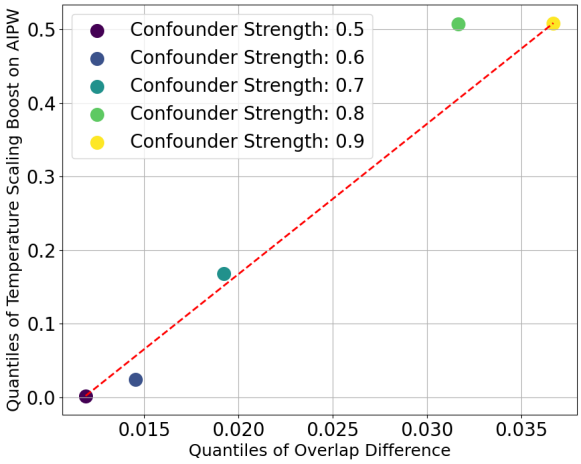

Figure 7: Q–Q Plot of Overlap Differences (the overlap increment after temperature scaling) vs. Temperature Scaling Boost on AIPW (the AIPW error ratio reduction after the temperature scaling), with α=0.4.

scenarios, we compute and average the Pearson correlation coefficients, with the first correlation yielding a value of 0.974 ± 0.022, and the second yielding a value of 0.503 ± 0.061. Both correlations are consistently positive, indicating that a larger confounder strength is always associated with a more pronounced increase in overlap after temperature scaling, and that a larger increase in overlap is associated with a more significant boost in AIPW estimation accuracy. This consistent positivity affirms the observations from the Q–Q Plot and underscores the efficacy of temperature scaling in adjusting for confounder-related distortions within propensity score distributions, particularly in settings with high confounder strength.

A.6 Keywords for School-Related Topics

The following lists detail the keywords used to identify mentions of various school-related topics in reviews, adapted from the work by (Harris et al., 2022). Refer to Table 6 for the specific keywords categorized by topic.

A.7 Data Process Details

A.7.1 Selected Period Define

To determine the most relevant time period for analyzing the influence of specific topics in school reviews, we employ a systematic approach to identify the four consecutive years with the highest frequency of topic mentions across schools. This method involves aggregating the review data annually for each school and isolating those reviews that contain keywords associated with the topic

Table 6: Keywords used to identify school-related topics in reviews.

Topic	Keywords List
Bullying	bully, bullying, harassment, intimidation, teasing, taunting, tormenting, victimization, abuse, threatening, coercion, humiliation, exclusion, cyberbullying, aggression, peer pressure, verbal abuse, physical abuse, emotional abuse, marginalization, ostracization, discrimination, hazing, stalking
Administration	administrator, board, counselor, counselors, coaches, handle, governance, head, advisor, superintendent, headmaster, vice, administration, policymaker, dean, policies, policy, staff, policymakers, coach, director, faculty, educators, educator, admin, assistant, headmistress, professional, administrative, leadership, principle, principal, administrators, principals, management
Academic Performance	exams, grade, score, standards, benchmark, ratings, evaluations, exam, reputation, performance, review, achievement, assessment, tests, rated, results, test scores, scores, grades, evaluation, star, rate, standard, academic, test, performing, benchmarks, assessments, success, rating
Extracurricular	athletic, club, extracurricular, music, teams, basketball, team, band, sports, drama, art, football, clubs, dance, athletics, soccer, choir, orchestra, volleyball, cheerleading, theater, debate, speech, track, swimming, tennis, robotics, student council, volunteer
Curriculum	language, reading, spanish, books, projects, courses, pe, writing, course, arts, science, assignments, book, subjects, math, english, social, history, lessons, write, homework, curricular, ap, curriculum, subject, physical

of interest. We then count the number of unique schools discussing these topics each year. By examining these annual counts, we pinpoint the time span where the conversation around each topic is most concentrated. This focused analysis ensures that our causal inference study is grounded in the period of maximal relevance for each educational aspect under consideration.

For each topic, once we define the time period, we then aggregate the review data for each school in this period to get school-level data. The data is further refined by excluding schools with fewer than five reviews and ensuring that concatenated non-keyword comments comprise at least 100 tokens. In this process, we define the treatment, outcome, and covariate for each school accordingly, as we discuss in Section 3. The summary of the school-level data is shown in Appendix A.1.

A.7.2 Bootstrap Detail

For the real-world data experiment, to ensure statistical robustness, we utilize a bootstrap sampling method without replacement, selecting equal numbers of treated and control schools for each topic to create balanced datasets. Specifically, based on the school-level data summary shown in Appendix A.1, we sample 9,000 schools for 'bullying,' 3,200 schools for 'academic performance,' 2,000 schools for 'administration,' 2,600 schools for 'extracurricular,' and 4,000 schools for 'curriculum.' This process is repeated six independent times for each topic to mitigate any sampling bias and provide a comprehensive overview of the causal impacts. The experimental results are then averaged across these six iterations to present a consolidated finding on how specific aspects influence school ratings.

A.7.3 Reviews Concatenate

When concatenating non-keyword reviews for each school, we insert the '[SEP]' special token between individual reviews. This token is preserved during training to serve as a delimiter, indicating the transition between separate reviews from the same school. To accommodate the DistilBERT model's token limit of 512, we employ a two-step truncation method for concatenated reviews exceeding 450 tokens. Initially, up to four sentences are included in each component review unless adding another sentence would exceed 450 tokens. If the token limit has not been reached after this initial pass, additional sentences are sequentially added from each review in rounds, continuing until reaching the token limit or exhausting all sentences. This approach is based on the assumption that the first several sentences of a review typically contain the most relevant information. The aim is to balance the depth of content detail within each review and the breadth of including multiple reviews, ensuring

comprehensive coverage without exceeding token constraints.

A.7.4 Confounder Insert

In our semi-synthetic data experiments, confounder text is randomly inserted at different positions within the reviews, specifically at locations marked by '[SEP]' tokens, at the beginning or at the end of the review texts. This insertion ensures that the confounder text is separated from the existing content by a '[SEP]' token, maintaining the structure of the original reviews. This approach simulates the addition of an authentic single review, seamlessly integrating the confounder text to reflect realistic review scenarios while preserving the integrity of individual reviews.

A.8 Semi-Synthetic Data Detail

In this section, we introduce the probability distribution framework we employ for the semi-synthetic data. Building on the definitions in Section 2.1, we expand our probability model to include a confounder class for each subject i, represented as (X_i, Y_i, T_i, C_i). Here, $C_i \in \{1, 2\}$ categorizes the confounder.

Outcome distributions are modeled to assess the effect of treatment across confounder classes. When $T = 1$ and $C = 1$, outcomes follow a target distribution $Y \sim \mathcal{N}(u_2, 0.3)$, contrasting with $Y \sim \mathcal{N}(u_1, 0.3)$ when $T = 0$ and $C = 1$. For Class 2, irrespective of the treatment, outcomes are both modeled as $Y \sim \mathcal{N}(u_2, 0.3)$, indicating uniform effects in the absence of confounding text. In this example, the conditional ATE for Class 1 is $(u_2 - u_1)$, and for Class 2 is 0, then the true ATE $u = (u_2 - u_1)/2$ for the whole dataset.

The probability of confounding class assignment is evenly distributed with $P(C = 1) = P(C = 2) = 0.5$. Class 1 subjects receive academic challenge posts inserted into their observed text, whereas Class 2 subjects' text remains unaltered, simulating environments with varying levels of textual confounding. Within these classes, the treatment assignment probabilities are designed to reflect their respective confounding impacts: $P(T = 1|C = 1) = 1 - p$ and $P(T = 0|C = 1) = p$, contrasting with $P(T = 1|C = 2) = p$ and $P(T = 0|C = 2) = 1 - p$. The magnitude of p directly modulates the strength of the confounding effect in our model. As p approaches 1, it signifies a strong correlation between the treatment assignment T and the confounder class C, indicating robust confounding. Conversely, when p nears 0.5, treatment assignment becomes effectively random within each class C, implying minimal or no confounding influence. Thus, p serves as a key parameter to adjust the intensity of confounding in our study. In our case, we vary p from 0.9 to 0.5.

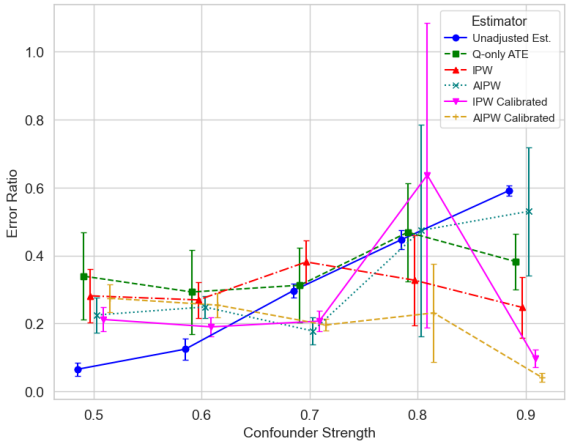

Figure 8: Estimator performance at a true ATE of $u = -0.35$, consistent with the settings in Figure 3.

The selection of u_1 and u_2 in our experiment is informed by the statistical properties of our 13,361 population dataset. The primary objective is to ensure that our sample not only includes a sufficient number of instances but also reflects a high sampling quality that aligns closely with the target distribution. Specifically, we set $u_2 = -0.3$, as the median and mean of $Y|T = 1$ in our dataset hover around -0.3. For u_1, we explore values within the range [-0.2, -0.1, 0.0, 0.1, 0.2, 0.3, 0.4, 0.5], and ultimately select 0.3, 0.4, and 0.5, where the resulting sample distribution closely matches our target. Ultimately, we choose a sample size of 5,000 for our analysis, ensuring both robustness and relevance in our evaluation of CausalBERT's performance.

A.9 Additional Results

Estimation for others true ATE Figures 8 and 9 show the performance of different estimators across various confounder strengths for semi-synthetic datasets where the true ATEs are -0.35 and -0.4, respectively.

Effect of α Figure 10 shows the effect of α on the IPW Calibrated model, mirroring the results found on AIPW in Figure 4 (Right).

Temperature Scaling Figure 11 shows the average error reduction provided by temperature scaling on the AIPW ATE estimates, mirroring the

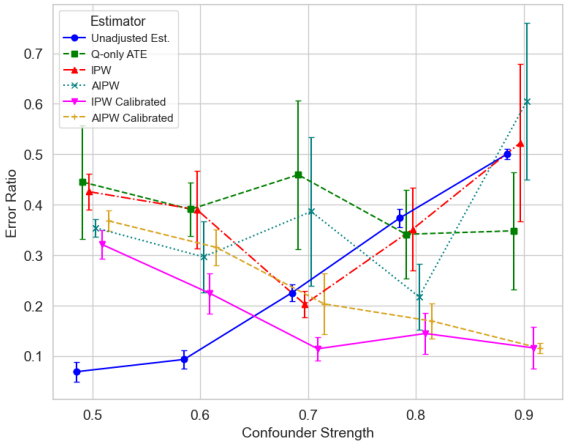

Figure 9: Estimator performance at a true ATE of $u = -0.4$, consistent with the settings in Figure 3.

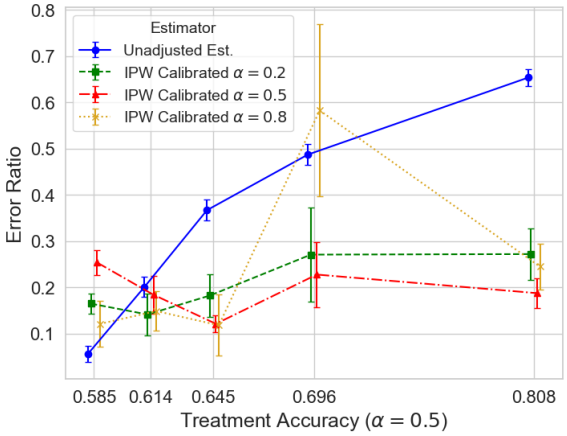

Figure 10: Error ratio comparison for CausalBERT under varying α values (0.2, 0.5, 0.8) indicated by treatment accuracy (trained with $\alpha = 0.5$) on a semi-synthetic dataset with a fixed true ATE $u = -0.3$, for the IPW Calibrated model.

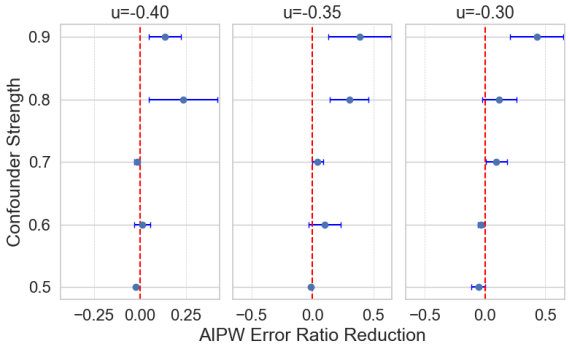

Figure 11: Error ratio decrease by temperature scaling on AIPW estimations

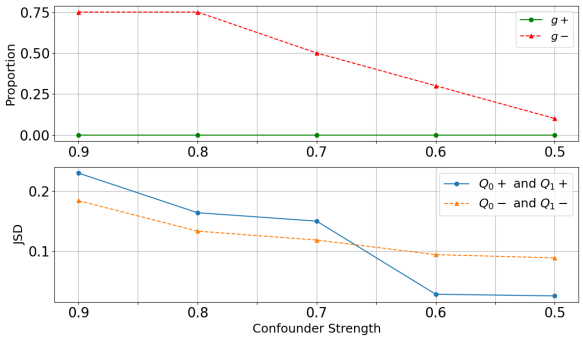

Figure 12: Application of Integrated Gradients on CausalBERT models trained across semi-synthetic datasets with a fixed true ATE $u = -0.3$ across a varying of confounder strengths from 0.9 to 0.5. The analysis identifies the top 20 tokens that influence the model's predictions across different output components. The top graph displays the proportion of these tokens from '$g+$' and '$g-$' originating from the inserted confounder template, and the bottom graph compares the Jensen–Shannon divergence(JSD) between the weighted tokens from 'Q_0^+' and 'Q_1^+' as well as 'Q_0^-' and 'Q_1^-'. Both graphs are presented across varying confounder strengths.

results found on IPW in Figure 4 (Left).

Integrated Gradients In this section, we extend our qualitative analysis on semi-synthetic datasets with a fixed true ATE $u = -0.3$ and varying confounder strengths from 0.9 to 0.5 using Integrated Gradients. Figure 12 reveals how top tokens(top 20) from 'g^+', 'g^-', 'Q_0^+', 'Q_1^+', 'Q_0^-' and 'Q_1^-' respond to changes in confounder strengths within CausalBERT. The top graph tracks the proportion of tokens from 'g^+' and 'g^-' that derive from the confounder template, showing how the model's reliance on confounder-driven features for treatment prediction varies with confounder strength. The bottom graph employs Jensen–Shannon divergence(JSD)[3] to measure the similarity between weighted tokens from 'Q_0^+' and 'Q_1^+' as well as 'Q_0^-' and 'Q_1^-'.

We observe from the top graph that the proportion of words in 'g^-' that originate from the confounder template decreases linearly as confounder strength is reduced from 0.9 to 0.5. This aligns with our expectations from the semi-synthetic dataset design. Under strong confounder conditions, a substantial portion of the non-treatment text incorporates inserted confounder text, making these features strong predictors of non-treatment. As confounder strength weakens, and the distribu-

[3]The Jensen–Shannon divergence (JSD) is a symmetric measure of the similarity between two probability distributions (Menéndez et al., 1997)

tion of confounder text between treatment and non-treatment groups becomes more balanced, the model's reliance on these features for predicting treatment diminishes accordingly.

In the bottom graph, we observe a decrease in the JSD between the weighted token lists Q_0^+ and Q_1^+, as well as between Q_0^- and Q_1^-, as confounder strength diminishes. This trend indicates that not only are the same terms present in both lists, but their relative weights also converge, suggesting a decreasing distinction in the outcome prediction signals for treated and untreated groups with weaker confounder influence. This result aligns with the design of our semi-synthetic data, which expects diminishing differences in reviews between treated and control groups as confounder strength decreases. Additionally, the range of change in JSD is notably greater for Q_0^+ and Q_1^+ than for Q_0^- and Q_1^-. This pattern aligns with the structural design of our semi-synthetic data, where a higher target mean outcome ($u_1 = 0.3$) is associated with the presence of inserted confounder text, in contrast to a lower target mean ($u_2 = -0.3$), which is mostly derived from Class 2 (without inserted confounder text). This design implies that the inserted confounder text serves as an indicator of higher overall ratings. Thus, the prediction signals for increasing outcomes exhibit greater sensitivity to shifts in confounder strength, as reflected by the larger divergence observed in Q_0^+ and Q_1^+. These findings collectively highlight CausalBERT's nuanced ability to detect, adapt to, and accurately represent subtle changes and biases in treatment effects, confirming its effectiveness in evaluating the impacts on outcomes under varying conditions of confounder strength.

Weighted Token Comparison In this section, we focus on the Integrated Gradients analysis applied to real data, emphasizing the weighted token differences between top terms (top 20) Q_1^+ and Q_0^+, as well as Q_1^- and Q_0^-. Figure 13 shows the Jensen–Shannon divergence(JSD) between the weighted tokens from 'Q_0^+' and 'Q_1^+' as well as 'Q_0^-' and 'Q_1^-' across all topics. we observe a distinct pattern across all topics: the JSD for the outcome-decreasing predictions top terms ('Q_0^-' and 'Q_1^-') consistently exhibits greater divergence compared to the outcome-increasing predictions top terms('Q_0^+' and 'Q_1^+'), particularly notable in the 'bully' topic. This trend suggests that, compared to positive comments, negative comments

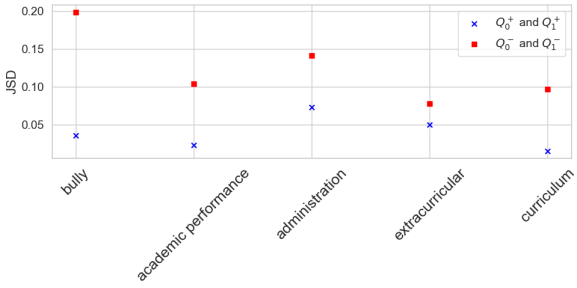

Figure 13: Comparison of the Jensen–Shannon divergence(JSD) between the weighted tokens from 'Q_0^+' and 'Q_1^+' as well as 'Q_0^-' and 'Q_1^-' across all topics. All weighted tokens are derived from the Integrated Gradients analysis applied to each topic from a single bootstrap sample.

about schools often cover a broader and more varied range of concerns, based on specific school attributes.

To provide additional context for Figure 13, we present the detailed weight distribution of the top terms used in the analysis along with their corresponding JSD values for several topics ('bullying' and 'curriculum'), as shown in Figures 14, 15, 16, and 17. Generally, the observed variations between control and treatment groups in outcome predictions are subtle, especially for outcome-increasing predictions. Universally positive terms such as 'great,' 'love,' and 'amazing' are prevalent in both groups (Figures 15&16), indicating a consensus on the attributes that positively impact school perceptions. For outcome-decreasing predictions, while negatively charged terms like 'bad' and 'horrible' are common in both scenarios (Figure 17), there remain slight distinctions, highlighting subtler variations in concerns that negatively impact school ratings. For example, in the 'bullying' topic (Figure 14), we notice that the term 'administration' is predominantly featured in the control group's negative predictors, highlighting that in the absence of bullying, issues like poor administration significantly impact school ratings. Conversely, this term is less pronounced in the treated group, suggesting that the presence of bullying tends to overshadow other administrative problems in influencing school ratings.

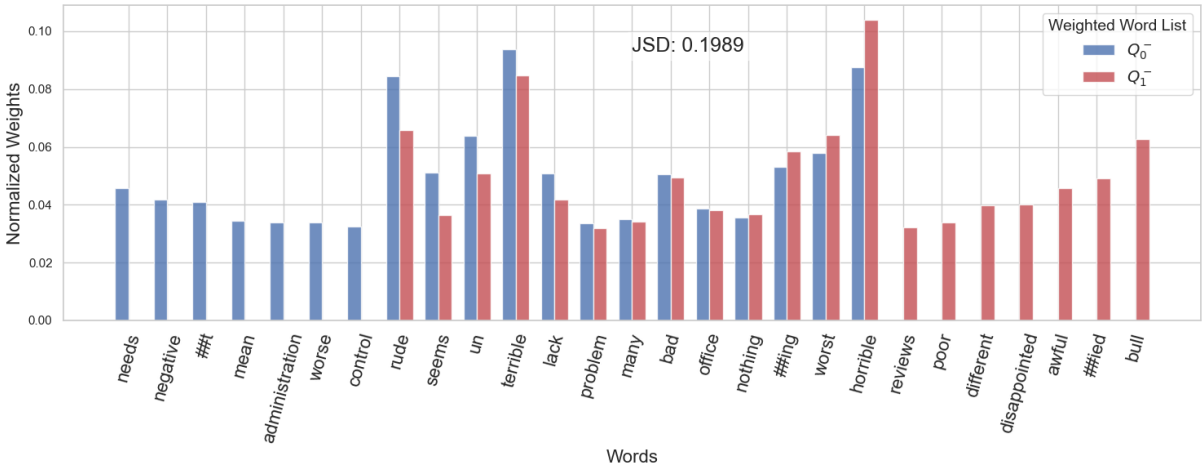

Figure 14: The weight distribution for top 20 weighted tokens from 'Q_0^-' and 'Q_1^-' for 'bullying' topic. We also include the JSD between these two weighted tokens.

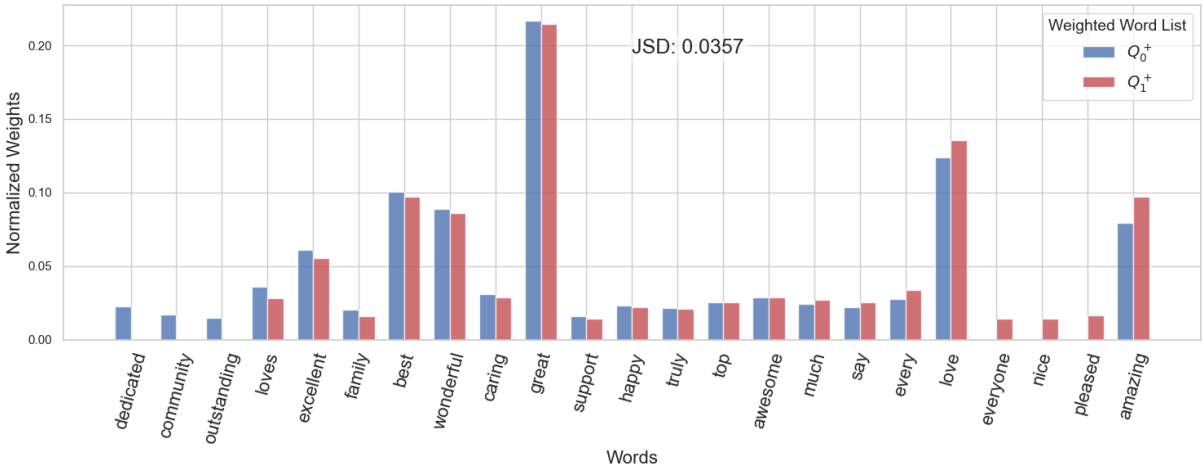

Figure 15: The weight distribution for top 20 weighted tokens from 'Q_0^+' and 'Q_1^+' for 'bullying' topic with JSD between these weighted tokens.

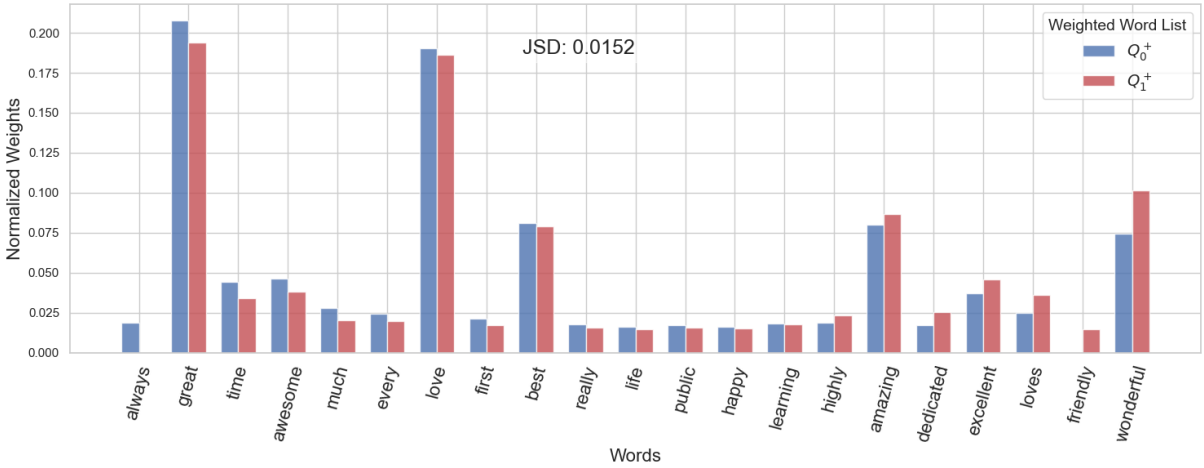

Figure 16: The weight distribution for top 20 weighted tokens from 'Q_0^+' and 'Q_1^+' for 'curriculum' topic with JSD between these weighted tokens.

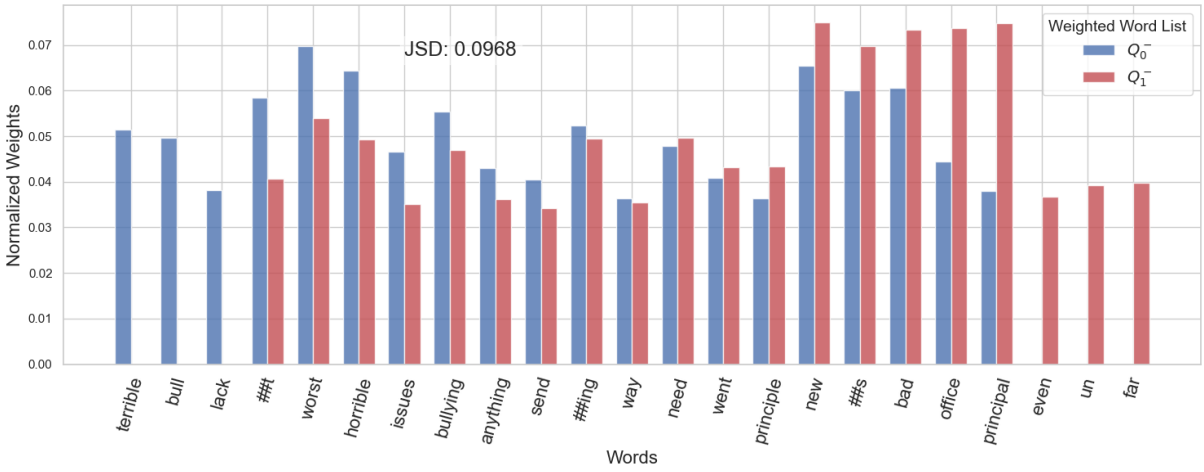

Figure 17: The weight distribution for top 20 weighted tokens from 'Q_0^-' and 'Q_1^-' for 'curriculum' topic with JSD between these weighted tokens.