Marketing Science Institute Working Paper Series 2025

Report No. 25-112

Words That Matter: Analyzing the Causal Effect of Words

Alain Lemaire, Mingzhang Yin and Oded Netzer

Words That Matter: Analyzing the Causal Effect of Words

Alain Lemaire[*] Mingzhang Yin[†] Oded Netzer[‡]

Abstract

Language plays a crucial role in marketing, influencing outcomes such as consumer engagement and decision-making. Although prior research has extensively analyzed the relationship between linguistic features and business outcomes, most approaches have been descriptive or predictive, limiting their value for crafting more effective content. Understanding the causal effects of specific linguistic features is essential but challenging because, in real-world settings, the focal textual feature often changes simultaneously with other confounding factors. This paper builds on recent advances in causal text analysis and introduces an embedding-based causal inference framework that isolates the impact of specific linguistic elements while controlling for both textual confounders and nontextual controls. The approach leverages foundational language models to create text representations optimized for causal inference, enhancing the accuracy of causal estimates. We rigorously validate our methodology using both semi-synthetic experiments and experimental data from large-scale A/B tests of news headlines. The experimental data allow us to offer a first-of-its-kind validation of the causal text approach using a marketing-relevant application. Applying the causal text approach to online donation and crowdfunding applications, we find that, for example, pre-thanking and second-person pronouns have a strong positive causal effect on success rates. However, these effects can be weakened or reversed if textual confounding is not properly controlled.

Keywords:
Causal Inference, Natural Language Processing, Text Analysis, Foundation Model, Deep Learning

[*]McCombs School of Business, University of Texas, USA. alain.lemaire@mccombs.utexas.edu
[†]Warrington College of Business, University of Florida, USA. mingzhang.yin@warrington.ufl.edu
[‡]Columbia Business School, Columbia University, USA. onetzer@gsb.columbia.edu

[*]The authors thank Kathy Li, Andy Gershoff, Hortense Fong, and Sanjana Rosario for their valuable comments. In writing the manuscript, the authors used ChatGPT to correct grammatical mistakes and to refine (prompted) sentences or paragraphs, but never to create new content.

INTRODUCTION

"il n'y a pas de hors-texte." — "There is nothing outside the text."

— Jacques Derrida

Much of the marketing and communication world revolves around the use of words, whether in crafting news headlines (Banerjee and Urminsky 2023) or social media posts that drive engagement (Berger and Milkman 2012) or in constructing the most persuasive sales pitch or donation appeals (Hong and Hoban 2022).

Identifying the right words to use requires an understanding of the causal effect of each textual element on relevant outcomes. Although text analysis has gained significant traction in the marketing literature, the majority of the academic and practical applications of text analysis in marketing and related fields has been correlational or predictive in nature. For example, marketing scholars have used text analysis to describe market structures (Lee and Bradlow 2011; Netzer et al. 2012), to predict loan defaults (Netzer, Lemaire, and Herzenstein 2019), or to predict stock returns (Tirunillai and Tellis 2012). Using a descriptive or predictive approach for text analysis works well if the goal of the research is to use text as a window into consumers' preferences or to predict their behavior; however, if the goal is more prescriptive, such as to create better textual content in social media posts to engage consumers or a more persuasive marketing pitch, descriptive or predictive approaches are not suitable. Such research objectives call for a more causal inference approach.

The challenge in assessing the causal effect of words is that "holding all else constant" is nearly impossible when it comes to textual data because the language space is very large, and naturally occurring sentences often vary by more than one word at a time. Table 1 shows two examples of news headlines used in our empirical study. The two headlines differ in whether they contain angry words ("outrage"), which may be the topic of interest for the researcher to explore, but they also differ in other ways, such as the inclusion of words like "threatens," "security," or "support," as well as variations in the length of the headline. If these additional

differences are related to the inclusion of the word "outrage" and also influence the outcome of interest, then a simple comparison of click-through rates of the headlines would result in a biased causal estimate of the effects of the angry words on the click-through rates (CTRs). Even when the data come from field A/B experiments (*e.g.*, Banerjee and Urminsky (2023)), different text corpora (*e.g.,* article titles) often vary with respect to more than one word, making word-level attribution difficult. A few studies have focused on very specific textual elements in lab settings that allowed them to carefully assess the causal effect of specific word changes (*e.g.,* Packard and Berger (2020); Packard, Li, and Berger (2024)). However, such studies often are limited in scope and generalizability.

In this paper, we build on recent work in computer science on estimating causal effects in textual data (Feder et al. 2022; Keith, Jensen, and O'Connor 2020; Veitch, Sridhar, and Blei 2019). We present an approach that allows marketing academics and practitioners to infer the causal effect of a group of words in complex marketing settings- settings in which multiple observed textual and nontextual variations may naturally co-occur along with the variation of the focal textual variable, affecting the outcome and hence acting as the confounders. The approach builds on the causal Bidirectional Encoder Representations from Transformers (BERT) to control the textual confounder (Veitch, Sridhar, and Blei 2019). Specifically, the approach starts with a pre-trained foundation model to summarize the entire document as an embedding vector, which captures linguistic information while accounting for the rich dimensionality of the textual confounder space. To control for the textual features' acting as the confounders and to remove redundant noise, this foundation model is further optimized to learn the causally-sufficient text embedding that relates to both the treatment and the outcome. We then extend the model to also account for nontextual controls that commonly occurs in marketing (*e.g.,* the date the document was created). Additionally, we improve the estimation accuracy by using augmented inverse probability weighting (AIPW) as a doubly robust estimator.

To demonstrate the value of the model in capturing the causal effect of words, we start by estimating the proposed model on semi-synthetic data. The textual corpora are actual news

Table 1: Two examples of headlines with and without "angry" words used by Upworthy (Matias et al. 2021)

	HEADLINE DATA	CTR
Treatment	**Headline with angry word**	
	OUTRAGE as New Policy Threatens Jobs and Families	6%
Control	**Headline without angry word**	
	New Policy Aims to Improve Job Security and Support Families.	3%

Note: Capitalization and blue font for the word "outrage" is not in the original text.

titles from the news website Upworthy.com, but the dependent variable is simulated. Using semi-synthetic data, we can assess the ability of the proposed approach to recover the true causal effect of changes in words between news headlines, and compare the proposed model with competing methods that do not fully account for possible confounding effects.

After estimating the proposed model, we apply it to three different marketing applications used in the academic literature to assess the effect of words on meaningful marketing outcomes. The first empirical application leverages data from the same Upworthy news website used to generate the semi-synthetic data (Banerjee and Urminsky 2023; Matias et al. 2021). A unique feature of the Upworthy data is that the platform ran many A/B tests by exposing random groups of website visitors to different headlines and measuring engagement through CTRs. Such A/B tests would seemingly allow us to observe the causal effect of words directly. However, because the two headlines often vary with respect to multiple words (see, e.g., the headlines in Table 1), even a well-conducted experiment cannot causally attribute the different outcomes between the headlines to any particular word or writing style.

However, the Upworthy dataset provides a unique opportunity to validate the proposed causal text approach in a marketing-relevant context. We demonstrate that when only a single word varies between the two headlines, the estimated causal effect (i.e., CTRs) for the two headlines based on our model and the raw difference in CTRs between the conditions converges; the A/B test indeed captures the causal effect of changing a specific word. However, as the level of confounding factors increases – measured by the number of word changes between the pair of headlines in the A/B test – the more our model and the raw causal effect diverge

due to increased confounders. Importantly, our model's textual causal estimates for the focal textual feature remained robust to increasing the number of words that changed between the headlines, as our model is capable of accounting for such confounds We demonstrate this effect across 43 linguistic treatment effects that were commonly investigated in the marketing literature, such as anger, aesthetics, gender words, and deliberation words. We identify meaningful linguistic drivers of headline engagement, such as the use of words related to "morality" and "instructional" words, which are often related to the call to action. We also demonstrate that linguistic treatment effects estimated using the proposed causal model align more closely with established findings in marketing, psychology, and linguistics literature compared to both simple mean comparisons from A/B tests and regression models that cannot properly account for confounders. Furthermore, we illustrate that discrepancies between our causal estimates and simple mean differences can be attributed to the degree of confound (Keith, Jensen, and O'Connor 2020; Mickey and Greenland 1989). Collectively, these results highlight the effectiveness of the proposed approach in accurately estimating the causal impact of linguistic features while appropriately controlling for textual confounders. This analysis also contributes to the causal effect of textual data literature (*e.g.,* Keith, Jensen, and O'Connor (2020); Veitch, Sridhar, and Blei (2019)) by leveraging the unique experimental nature of the Upworthy data to empirically demonstrate the ability of the embedding-based causal inference approach to truly capture the causal effect of words.

Following this analysis, we apply the model to two additional applications previously used in the marketing literature in the context of online donations and in loan request settings. The data and variation across textual documents in these applications do not come from experimental data. We compare the treatment effect found from the proposed causal model with approaches that do not fully account for textual confounders. We specifically focus on two commonly investigated textual features in this literature (often with mixed results): "prethanking," and "second-person pronouns."

Using the proposed text-based causal model, we find that in both donation and investment

settings, pre-thanking has a significant positive effect on the likelihood of funding success. This result aligns with prior experimental research results Rind and Bordia (1995); Merchant, Ford, and Sargeant (2013) on pre-thanking's persuasive effects, reinforcing the idea that expressions of gratitude can enhance appeal. Similarly, we find that in both donation and investment settings, the use of second-person pronouns ("you" words) increases the likelihood of securing funding. This result is consistent with extant experimental work Packard and Berger (2020) demonstrating that "you" words facilitate persuasion by prompting readers to associate the appeal with a specific person they know, thereby enhancing engagement and relatability. While these effects are consistent with prior literature, we show that methods that do not account for possible confounds often lead to the opposite effects.

Using both synthetic data and three empirical applications from the marketing literature, we demonstrate how causal methods for textual analysis can help estimate the causal effect of a group of words on engagement using secondary data in settings where randomized controlled trials are not feasible. To the best of our knowledge, this paper is the first to assess the causal impact of words in marketing. The approach we propose can be used by marketing academics to study the causal effect of linguistics on relevant marketing outcomes and by practitioners, such as journalists, editors, and copywriters, to choose words and a linguistic style that leads to the preferred marketing outcomes. We contribute to the textual causal inference literature by empirically testing the validity of the embedding-based causal inference approach using a unique dataset that involves A/B tests on textual data (news headlines), demonstrating the ability of the model to recover the causal effect of textual data in the face of textual confounders.

The rest of this paper is organized as follows. In the next section, we review the literature on text analysis, causal inference, and text-based causal inference in marketing and related fields, providing the foundation for our study. Following this, we illustrate the issue of textual confounding and demonstrate how our proposed approach can mitigate such confounds. Next, we introduce our model aimed at accounting text-based confounders. We then validate the

model using both synthetic data and real-world data from an A/B test of news headlines. With the model validated, we apply it to examine the effects of two key linguistic constructs—"pre-thanking" and "you" words—on engagement, providing empirical evidence for their effects. Finally, we conclude with a discussion of our findings and outline potential directions for future research.

TEXT ANALYSIS AND CAUSAL INFERENCE IN MARKETING

The availability of unstructured data and advances in methods to analyze such data have led to a proliferation of research in the marketing literature that uses textual data (Berger et al. 2020). Given the nature and scale of unstructured data, much of this work has been descriptive or predictive. However, one of the main goals of marketing studies is to identify effective marketing interventions, including optimizing textual content, which requires causal inference. The main approaches to obtaining such causal inference are through experiments or by identifying exogenous variation in the data. The problem with applying such methods to textual data is that, unless the goal is to explore the effect of a specific word or a specific construct, the dimensionality of variation in textual data is often too large to run experiments or to leverage exogenous variation. However, recent advances in causal machine learning enable us to close the gap in assessing causal inference from observational textual data.

Text Analysis in Marketing

Marketing researchers have shown that textual user-generated content, such as blog posts and consumer reviews, can serve as a substitute for expensive survey data collection for purposes like market structure analysis (Lee and Bradlow 2011; Netzer et al. 2012; Tirunillai and Tellis 2014) or understanding consumer preferences (Timoshenko and Hauser 2019). Marketers have also leveraged textual data to predict outcomes related to stock prices (Tirunillai and Tellis 2012), loan default (Netzer, Lemaire, and Herzenstein 2019), donations (Hong and Hoban 2022), and consumers' movie choices (Toubia et al. 2019).

Recent studies quantify the effect of text elements on business outcomes using predictive models. For example, Banerjee and Urminsky (2023) leveraged a large dataset of experimental data to identify which linguistic features make for an appealing news headline. Similarly, Hong and Hoban (2022) leveraged secondary data and a deep learning attention model to find sentences that are either detrimental or helpful in collecting donations.

These studies, however, often fall short of being able to make causal claims. For example, identifying how to modify the text of a donation appeal or loan request to generate a higher likelihood of success would be more useful than merely predicting donation success or loan default. Answering these interventional questions requires the application of a causal inference method (Berger et al. 2020; Packard, Li, and Berger 2024; Packard, Moore, and McFerran 2018).

Causal Inference with Textual Data

The prevalence of textual data in marketing provides a valuable resource for conducting causal inference. However, traditional causal inference methods are often restricted to low-dimensional structured data. The outcome, treatment, and confounders must be recorded as tabulated data for computational analysis, making these methods unsuitable for unstructured text. In contrast, deep learning methods excel at predicting and generating texts, but they are not designed to satisfy the necessary identification assumptions underlying causal inference.

Studies attempting to make causal inferences with textual data often reduce the problem to a specific word or concept. For example, Packard and Berger (2020) shows that using second-person pronouns can determine whether people will like cultural items like music. Others have used textual data as the outcome variable and used traditional methods, such as difference-in-differences and synthetic control, to analyze the policy effects of exogenous variation (Puranam, Kadiyali, and Narayan 2021; Puranam, Narayan, and Kadiyali 2017). However, such studies often examine a single textual feature using lab experiments. When studying multiple textual features, even well-executed experiments are limited in their ability to randomize text treatment without changing other text elements (*e.g.,* Table 1).

Recent advances at the intersection of NLP and causal inference provide promising avenues for estimating the causal effect of words and controlling the textual confounders. For example, Olteanu, Varol, and Kiciman (2017) used n-grams to encode textual confounders. However, n-grams cannot capture long-range dependencies or effectively encode textual corpora into meaningful summary statistics. Roberts, Stewart, and Nielsen (2020) analyzed the causal effect of perceived author gender on scientific paper's citation frequency, controlling for textual confounders using topic modeling. However, this method overlooks potential textual confounders other than the topics, such as sentiment, tone, and writing style (Mozer et al. 2018). Researchers recently have projected the text confounders onto an embedding space (*e.g.,* Gui and Veitch (2022a); Veitch, Sridhar, and Blei (2019); Veljanovski and Wood-Doughty (2024)). This approach is particularly useful because the embedding function leverages the effective language models to capture multifaceted linguistic confounders, and the compact embedding space ensures an overlap between the treatment and control texts. The embedding space is then optimized to focus on representing the confounding words that are correlated with both the treatment and the outcome.

Our work applies and extends this line of research in a couple of ways. First, substantively, we explore the opportunity of these models in a marketing context. Second, methodologically, we extend these models to account for both textual confounders and nontextual controls commonly found in marketing applications, as well as using the augmented inverse probability weighting (AIPW) adjustment to the estimator. Third, and more importantly, we leverage a dataset of A/B tests of news headlines to empirically validate the ability of such models to truly capture causal effects with meaningful confounders.

ADJUSTING TEXT CONFOUNDERS FOR CAUSAL INFERENCE

Estimating causal effects from observational data is challenging due to confounders—variables influencing both treatment and outcome. For instance, when evaluating how an author's gender affects a news article's click-through rate (CTR), the article's content

itself acts as a confounder since it likely impacts CTR while potentially correlating with the author's gender. Similarly, as demonstrated in Table 1, determining the causal impact of specific words or concepts in a headline (e.g., anger-related words) on CTR requires controlling for other headline words that may serve as confounders.

The relationships between treatment and outcome in the presence of confounders are summarized in Figure 1; here, the treatment is denoted by $T \in \{0,1\}$, and the outcome variables are denoted by $Y \in \mathbb{R}$, and both treatment and outcome are affected by a confounder $C \in \mathscr{C}$.

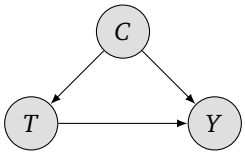

Figure 1: Graphical model showing a confounding variable C affecting both treatment T and outcome Y. Shaded cells indicate observed variables.

We adopt the potential outcome framework (Neyman 1990; Rosenbaum and Rubin 1983). For the treatment $T = t$, $Y_i(t)$ is the potential real-valued response if the unit i receives treatment t. In this paper, our focus is on estimating the average treatment effect (ATE) $\tau := \mathbb{E}[Y_i(1) - Y_i(0)]$.

The following set of assumptions is sufficient for the identification of the ATE (Imbens and Rubin 2015):

Assumption 1 (Ignorability): Given a set of covariates C, the treatment assignment and potential outcomes are independent; that is, $(Y(0), Y(1)) \perp T \mid C$

Assumption 2 (Overlap): Every observation has a positive probability of receiving each treatment level; that is, $\eta < P(T_i = t \mid C_i = c) < 1 - \eta$ for some $\eta > 0$.

Assumption 3 (Stable Unit Treatment Value Assumption (SUTVA)): The treatment applied to one unit does not affect the outcomes of other units (no interference), and each unit's outcome depends solely on its assigned treatment without hidden variations or inconsistencies.

Ignorability is satisfied when the observed confounders C contain all the information that may affect the outcome except the treatment. For example, in one of our empirical applica-

tions, consumers decide whether to click on a headline based only on the text in the headline. Thus, apart from the treatment (*e.g.,* whether the headline includes an anger word), all possible confounders are observed in the headline text. In situations where there exist nontextual controls, these variables need to be adjusted in addition to the textual confounders to satisfy ignorability.

Regarding the overlap assumption, it is likely to be violated when words are used directly as confounders, as the treatment and control texts rarely share the same set of words, and the high dimensionality of text can lead to extreme propensity scores (D'Amour et al. 2021; Gui and Veitch 2022b). Our approach achieves overlapping by encoding the text into a more compact embedding space and eliminating redundant information that does not pertain to confounders. For the SUTVA assumption, we assume no interference between units. This assumption is commonly made in marketing applications that do not involve network effects, which is the case in our empirical applications.

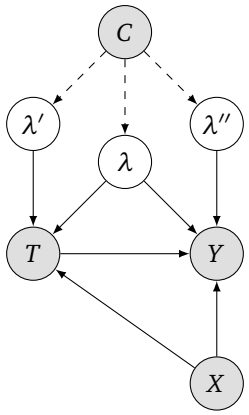

Figure 2: Graphical model showing confounding variables C and X affecting both treatment T and outcome Y, and the representation of C through λ.

Figure 2 provides a visualization of the problem and sketches a solution. We want to estimate the effect of the observed treatment of a group of words denoted by the shaded T on the outcome Y. But the context – comprising the surrounding words – is a confounder, denoted by C. Given that we are dealing with secondary data, we must also account for other nontextual covariates that may affect Y and T, denoted by X. In the donation application,

which we discuss in a later section, nontextual possible controls that we need to account for are the number of competing donation requests. Thus, one goal of the proposed approach is to find an accurate representation of the textual variable C. This representation is denoted by the unshaded λ, λ', and λ'' in Figure 2 because they are unobserved. Recall that for a text element to be a confounder, it must be related to *both* the treatment and the outcome. Hence, in controlling for confounders, we need to focus on the embedding space of λ, rather than λ' or λ''. That is, to capture C in a causally sufficient way, we construct embeddings that focus on minimally sufficient confounding information, removing textual content (λ' and λ'') that does not affect the treatment and outcome simultaneously, and hence does not serve as confound.

In addition to finding the representation (λ) to control for textual confounders, we extend Veitch, Sridhar, and Blei (2019) by capturing additional nontextual covariates (denoted by X in Figure 2), which are commonly found in marketing applications. Accounting for both textual confounders and possible nontextual controls is important to satisfy the ignorability assumption. We propose to integrate the textual representation (λ) in a doubly robust estimator to achieve accurate causal estimation.

MODEL DESCRIPTION

Our model brings together both language modeling and causal inference. It represents the words in the document in a more compact and causally-aware manner using embedding methods.

From a causal inference perspective, the proposed approach recognizes that not all textual information is equally relevant to estimating causal effects. For example, the appearance of a popular brand name in a headline might be related to the outcome (*e.g.,* CTR), but it may have little effect on whether the treatment (*e.g.,* "you" words) appears. Thus, to increase efficiency and improve statistical overlapping, the proposed approach strategically focuses on possibly confounding words and their corresponding embeddings that are relevant to both the treatment and the outcome.

More formally, we observe a set of documents W_i, $i = 1, \cdots, N$ where each document W_i consists of words and sentences and is associated with an observed outcome y_i (e.g., the CTR for news headlines). The key to the proposed approach is the application of text representation techniques that enable causal inference (Veitch, Sridhar, and Blei 2019). We use a BERT model $f(\cdot)$ with parameters γ^W to learn the text embedding $\lambda_i = f(W_i; \gamma^W)$. Specifically, we use BERT (Devlin 2018), a large language model designed to capture contextual relationships among words, sentences, and paragraphs. This capability allows BERT embeddings to effectively encode complex linguistic nuances, thus adequately controlling for textual confounding. We adapt the BERT model by optimizing an objective function designed specifically for causal inference.

Consider the following causal inference problem: Researchers are interested in studying the effect that including a specific word or dictionary of words in a document i (t_i) has on outcome y_i; possible textual confounders are represented by the embedding function λ_i, and possible other nontextual controls are represented by X_i (e.g., the date the article was written).

The proposed approach simultaneously fits a treatment model $\hat{e}$, an outcome model $\hat{Q}$, and learns the embedding function of the textual confounders. Specifically, we use the following objective:

$$\min_{\gamma^{t_i}, \gamma^e, \gamma^W} \mathbb{E}[L_o(y_i, \hat{Q}(t_i, X_i, \lambda_i; \gamma^{t_i})) - t_i \log \hat{e}(X_i, \lambda_i; \gamma^e)) - (1 - t_i) \log(1 - \hat{e}(X_i, \lambda_i; \gamma^e)) + L_u(W_i, \lambda_i; \gamma^W)], \quad (1)$$

where the expectation is over all the documents. A simple neural net is used to model the outcome given the text embedding λ_i:

$$\hat{Q}(t_i, X_i, \lambda_i; \gamma^{t_i}) = \omega_2^{t_i} \cdot \text{ReLU}(\omega_1^{t_i} \cdot [t_i, X_i, \lambda_i] + b_1^{t_i}) + b_2^{t_i}, \quad (2)$$

where $[t_i, X_i, \lambda_i]$ is the concatenated inputs, $\text{ReLU}(z) = \max(0, z)$, and the parameters $\gamma^{t_i} =$

$(\omega_2^{t_i}, b_2^{t_i}, \omega_1^{t_i}, b_1^{t_i})$ are the weights and bias. The propensity score is modeled as the logistic regression $\hat{e}(X_i, \lambda_i; \gamma^e) = 1/(1 + \exp(-[X_i, \lambda_i]^\top \gamma^e))$.

The model in Equation (1) has three components. The first component $L_o(y_i, \hat{Q}(t_i, X_i, \lambda_i; \gamma^{t_i}))$ is the standard estimation of the outcome, where L_o is a square loss function for a continuous outcome variable such that $L_o = (y_i - \hat{Q}(t_i, X_i, \lambda_i; \gamma^{t_i}))^2$. $\hat{Q}(t_i, X_i, \lambda_i; \gamma^{t_i})$ is the estimated outcome under treatment t_i, and γ^{t_i} are the parameters of this outcome function. This component ensures that the embedding λ_i is related to the outcome.

The second component, $-t_i \log \hat{e}(X_i, \lambda_i; \gamma^e)) - (1 - t_i) \log(1 - \hat{e}(X_i, \lambda_i; \gamma^e))$, represents the propensity estimation. It is modeled as a binary cross-entropy, which is equivalent to the negative log-likelihood of a logistic regression model. This component links the embedding vector, λ_i, to the treatment, t_i, via the parameters, γ^e. Thus, this component ensures that when we estimate the embedding vector, λ_i, we are focusing on confounders that are correlated with the treatment.

The third component, $L_u(W_i, \lambda_i; \gamma^W)$, is the standard unsupervised BERT likelihood function that generates the embedding vector, λ_i, to represent document W_i, which optimizes the set of transformer neural network parameters γ^W.

Note that the unique aspect of Equation (1) is its ability to capture textual information as confounders. Including the embedding vector, λ_i, in all three components facilitates this ability. Specifically, λ_i is generated similar to the traditional transformer model in the third component, but it is adapted to the causal inference problem by making sure that the inferred embedding is related to both the outcome (first component) and the treatment (second component).

The causal effect to be estimated is the ATE, which measures the mean difference in outcomes when all units receive the treatment, compared to when none of them do. Under the assumptions of SUTVA, positivity, and ignorability, the ATE can be identified as

$$\begin{aligned}\tau_{\text{ATE}} =& \mathbb{E}[Y_i(1) - Y_i(0)] \\ =& \mathbb{E}_{W_i, X_i}[\mathbb{E}[Y_i \mid W_i, X_i, T_i = 1] - \mathbb{E}[Y_i \mid W_i, X_i, T_i = 0]].\end{aligned}$$

Given the trained language model embedding by Equation (1), the conditional expectation $\mathbb{E}[Y_i \mid W_i, X_i, T_i = t_i]$ is estimated by $\hat{Q}(t_i, X_i, \lambda_i(W_i); \gamma^{t_i})$, which can be used in the outcome-only estimator of the ATE:

$$\hat{\tau}_{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} [\hat{Q}(1, X_i, \lambda_i) - \hat{Q}(0, X_i, \lambda_i)]. \tag{3}$$

The outcome-only estimator might be sensitive to model misspecification. To address this issue, we propose to use the augmented inverse probability weighting (AIPW) estimator as a doubly robust estimator that incorporates both the outcome model and propensity scores (Robins, Rotnitzky, and Zhao 1994):

$$\begin{aligned} \hat{\tau}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \Bigg(& \hat{Q}(1, X_i, \lambda_i) - \hat{Q}(0, X_i, \lambda_i) \\ & + \frac{T_i (Y_i - \hat{Q}(1, X_i, \lambda_i))}{\hat{e}(X_i, \lambda_i)} - \frac{(1 - T_i)(Y_i - \hat{Q}(0, X_i, \lambda_i))}{1 - \hat{e}(X_i, \lambda_i)} \Bigg). \end{aligned} \tag{4}$$

Here, $\hat{e}$ is the propensity score function estimated from the second component in Equation (1). The doubly robust property of AIPW means it is a consistent estimator for the ATE as long as either the outcome or the propensity model is accurate (Bang and Robins 2005). In simulation studies, we will show that the doubly robust estimator provides an ATE estimate closer to the ground truth causal effect than the outcome-only estimator.

EMPIRICAL VALIDATION STUDIES

Validating a causal inference method, such as the one presented in this paper, requires knowing the actual treatment effect, which is rarely observed. One of the main objectives of our paper is to empirically validate the embedding-based causal inference approach, specifically with respect to relevant textual data and outcomes for marketing purposes. To this end, we leverage experimental data from the news website Upworthy.com (Matias et al. 2021), which includes a large set of field experiments comparing the CTRs of different news head-

lines. However, as we discussed earlier, because different headlines vary with respect to more than one textual unit, even these experiments do not allow us to directly test the causal effect of a single textual unit.

To overcome this challenge, we first apply the proposed model to semi-synthetic data, where the treatment and confounders are from the Upworthy data but the outcome is simulated. The semi-synthetic data provide us with a clear ground truth of the treatment effect and allow us to test the model robustness under different levels of confounding strength. Next, we use the actual experiments in the Upworthy data and the proposed approach to estimate various treatments that the literature has shown could increase engagement. We leverage the variation in the number of words changed between different headlines in the experiment to assess the degree of possible textual confounding. Specifically, when only one word changes between the two headlines, the difference-in-mean estimate should serve as the ground truth and should be similar to the estimate of the BERT-based causal model. However, we expect that the raw difference between the outcomes of headlines (RAW-ATE) would deviate from our model estimate as more words change between a pair of headlines.

Upworthy Data

Upworthy is a news website created in 2012 with the goal of focusing on positive storytelling. For two years, Upworthy ran randomized experiments on their homepage, where they assigned different readers to see different headlines for the same story. The Upworthy dataset comprises 32,488 A/B tests conducted over two years, from 2013 to 2015 (Matias et al. 2021). These experiments resulted in more than 538 million impressions and more than 8 million clicks.[1] These experiments aimed to identify the most engaging combinations of headlines, subheadings, and images, collectively called "packages." Each package, which resembles the content presentation of news on the Upworthy website, was designed to maximize viewer engagement. Upworthy conducted the A/B tests by randomly assigning different viewers to vari-

[1] Upworthy made the data available for academic research through their OSF repository (https://OSF.io/jd64p/).

ous packages of the same news story, and it recorded the number of impressions and clicks each package received. To ensure the integrity of the experiments, only one experiment was conducted on the Upworthy main page, minimizing potential interference. The dataset provides comprehensive details for each package, including the experiment ID, creation time, headline, subhead, social media preview text, preview image, and the number of impressions and clicks.

Table 2: Descriptive Statistics for the Upworthy Dataset

	Mean	Min	Max	N
Packages	4.81	2	12	1,558
Impressions	3,576.50	758	11,225	7,498
Clicks	37.95	0.0	811	7,498
CTR	0.01	0.0	0.13	7,498

To focus on changes in textual features in the headline, and to satisfy the ignorability assumption, we analyze the 1,588 A/B tests that varied only with respect to the headline, keeping the subhead and image constant across conditions in the same package. In this subset of the data, experiments had a mean of 4.81 packages per experiment. The average number of views for each package was 3,577, with a maximum of 11,225 views. The average number of clicks for the packages was 37.95, while some packages registered 0 clicks, and the maximum clicks received was 811. In total, the dataset contained 7,498 packages. See Table 2 for related summary statistics.

To fully assess the causal effect of the words on outcomes, we start by analyzing semi-synthetic data generated from the Upworthy data and then follow up with an empirical analysis of the Upworthy experiments.

Semi-synthetic Outcome Validation Analysis

In this analysis, we use the headlines from Upworthy but generate the outcome variable y_i based on:

$$
\begin{aligned}
y_i &= \alpha t_i + \beta(\pi(z_i) - 0.5) + \epsilon_i \\
\pi(z_i) &= Pr(t_i = 1 | z_i) \\
P(t_i = 1 | z_i = 1) &= \frac{\sum 1_{(t_i=1 \cap z_i=1)}}{\sum 1_{(z_i=1)}} \\
P(t_i = 1 | z_i = 0) &= \frac{\sum 1_{(t_i=1 \cap z_i=0)}}{\sum 1_{(z_i=0)}} \\
\epsilon_i &\sim N(0, \sigma_\epsilon),
\end{aligned}
\tag{5}
$$

where t_i is a binary variable to represent treatment, taking a value of 1 if a word from the linguistic dictionary chosen as treatment (*e.g.*, "time words") appears in the headline, and 0 otherwise. α represents the causal effect we are trying to recover. We set α to 1. In this analysis, we consider a single confounder z_i, captured as a binary variable and indicating the appearance of words from a dictionary other than the treatment dictionary (*e.g.*, "adverb words").

Since the propensity score is a sufficient statistic to control for the confounder (Rosenbaum and Rubin 1983), the confounder z_i influences both the treatment t_i and the outcome y_i through the propensity score $\pi(z_i)$. We subtracted 0.5 from $\pi(z_i)$ to center it. The scalar $\beta \in \mathbb{R}$ represents the strength of the confounders' impact on the outcome y_i. The larger β is, the greater the estimation bias becomes if the confounder is not properly controlled. In addition, ϵ_i is a random error in the outcome that is generated from a standard normal distribution. In our analysis, we do not inform the model of the identity of the confounder z_i; instead, we let the model infer it from the document W_i. In this data-generating process, the text W_i, the dictionaries z_i, and treatment t_i are from actual Upworthy data, and the outcome y_i is synthetic.

We use the Linguistic Inquiry and Word Count (LIWC) 2015 dictionaries to represent the data. LIWC is a group of dictionaries that group words into a coherent theme. It contains 125

dictionaries that cover context-independent linguistic features, such as positive emotion (*e.g.,* beauty, brilliant, carefree), time words (*e.g.,* afterward ago, begin, bedtime), and adverbs (*e.g.*, about, anyways, finally, fortunately) (Pennebaker et al. 2015). In this analysis, we use *time words* as the treatment and *adverb words* as the confounders. In the actual Upworthy data, the probability of time words in the headlines, given the existence of adverb words ($\pi(z_i = 1)$), is 0.71, and the probability of time words given no adverb words ($\pi(z_i = 0)$) in the headline is 0.33. Thus, adverbs and time words are positively correlated in our data. Finally, for our analysis, we vary the confounding strength parameter (β) from 1 to 20 by increments of 1. We do this to simulate the recovery powers of the proposed approach under increasing confounding strength.

We compare the estimated treatment effect based on the BERT-based model with two common baseline causal inference measures. The first is the raw ATE (RAW ATE; see Equation (6)); this measure is the difference between the mean outcome when the treatment is applied and when the treatment is not applied:

$$\text{RAW ATE} = \frac{1}{n_1} \sum_{i:T_i=1} y_i - \frac{1}{n_0} \sum_{i:T_i=0} y_i. \tag{6}$$

n_1 is the number of observations/packages in the treatment condition (headlines with time words) and n_0 is the number of observations in the control (headlines without time words). We have $n_1 = 3,235$ and $n_0 = 4,263$.

The second baseline measure is the inverse probability weighting (IPW ATE). IPW ATE augments the RAW ATE by accounting for the probability of the package's receiving the treatment. The idea behind IPW is to create a "pseudo-population" by reweighting the data so that the treatment becomes independent from the confounders. This reweighting is done through the propensity score $\hat{e}$, computed by

$$\text{IPW ATE} = \frac{1}{n_1} \sum_{i:T_i=1} \frac{y_i}{\hat{e}(B_i)} - \frac{1}{n_0} \sum_{i:T_i=0} \frac{y_i}{1-\hat{e}(B_i)}. \tag{7}$$

In Equation (7), we first convert the text W_i in each headline to the bag-of-words representation $B_i \in \mathbb{N}^V$, where the v-th element of B_i is the count of the v-th unique word that appears in the headline i. We use a binary logistic regression to estimate $\hat{e}(B_i)$, where the independent variables are the words in headline i that are not in the treatment dictionary. The propensity score $\hat{e}$ captures the likelihood of the treatment (time word), based on the other words in the headline.

Using the 7,498 packages in the Upworthy dataset, we fit our proposed model by optimizing Equation (1) using Pytorch to update the parameters $(\gamma^{t_i}, \gamma^e, \gamma^W)$ of the outcome model, treatment model, and BERT language model. After fitting the model to Upworthy data, the causal effect is estimated using AIPW estimator in Equation (4) with the learned parameters.

Semi-synthetic Data Results

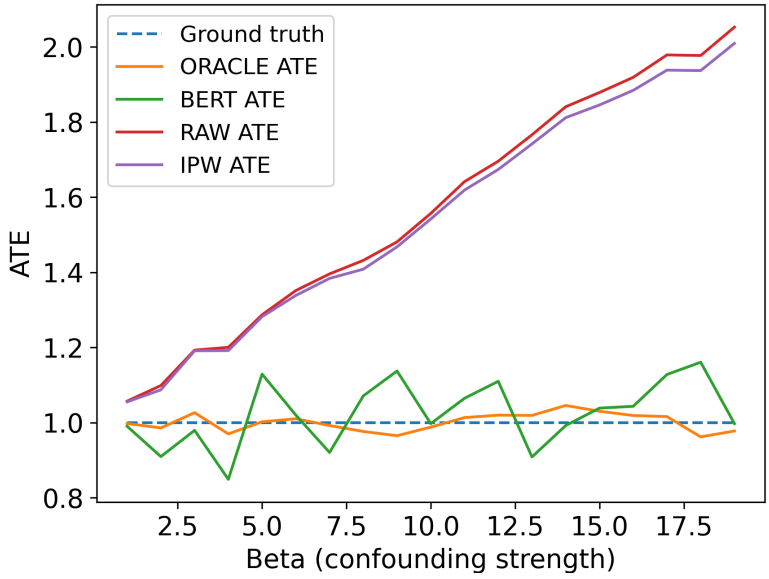

Figure 3: Estimated ATEs for different levels of confounding strength across different models

Using the synthetic data, we can test the ability of the proposed model and the benchmark ATE measures to capture the observed causal treatment effect of time words under different levels of confounding strength. The results of this analysis can be seen in Figure 3. The blue

(dashed) line represents the ground truth ($\alpha = 1$), which is constant across various levels of confounding strength. The ORACLE model assumes the confounder (adverb words) is known a-priori, and includes the indicator of adverb words as a covariate in regression, which represents the ideal estimation achievable by the data.

From the figure, we can see that as the level of confounding strength level increases, the baseline estimates increase and deviate from the true treatment effect. When $\beta = 1$, the RAW ATE and IPW ATE are close to the true treatment effect, but at $\beta = 20$, the two ATE estimates are above 2. On the other hand, the causal BERT ATE (green line) is consistently between 0.9 and 1.1 across all levels of confounding strength. The fact that the BERT ATE tracks closely the ORACLE ATE suggests that the the causal BERT embedding model is capable of capturing the true confounder. In the Figure W1 of the Web Appendix II, we also show the causal BERT estimate with the outcome-only estimator in Equation (3), which is significantly more accurate than the baselines but slightly less accurate than the AIPW estimator.

Overall, this analysis using semi-synthetic data, in which we can observe the actual treatment effect and varying degrees of confounding that arise from text, shows that the proposed causal BERT model is capable of recovering the treatment effect in the face of different levels of textual confounding, whereas the traditional ATE measures suffer from bias. Having validated the BERT-based causal model using the semi-synthetic data, our next step is to validate the proposed model using real observed outcomes.

Upworthy Data Validation Analysis

After demonstrating that the BERT-based causal model is capable of capturing causal effects in the face of textual data, we explore its ability to do so using secondary data. Although the Upworthy experiments causally indicate which headlines perform better, the proposed approach can shed light on the specific linguistic features that cause an increase in CTRs. Such inference is necessary to generalize the results from the current experiment to the crafting of new headlines. We leverage the same data described in the *Upworthy Data* section. A unique

aspect of this dataset is that it is based on hundreds of field experiments. This dataset allows us to assess the causal effect of different writing style elements (*e.g.,* Banerjee and Urminsky (2023)) on CTRs, while controlling for varying degrees of confounding, which arises from changes in other words in the headline in addition to the change in the focal word. In our semi-synthetic data analysis, we focused on a single treatment word dictionary (time words) and a single confounder (adverb words), but the 1,558 A/B tests in the Upworthy data allow us to test the causal effect of multiple writing style dictionaries under varying degrees of confounding, based on how many words change between the different headlines in the package.

Using the BERT-based causal model, we causally estimate the effect of the presence of specific dictionaries on engagement, as measured by CTRs. We then compare these estimates to the baseline ATE estimates (RAW ATE and IPW ATE), as well alternative estimates that control for possible confounds.Moreover, similar to the semi-synthetic analysis, we show that as the strength of confounding increases (i.e., more words vary between any pair of headlines), the traditional baseline measures overstate the magnitude of the treatment effect, whereas the BERT ATE estimates remain stable. We also test for the presence of confounders when the BERT ATE estimates deviate from the RAW ATE estimates. Finally, we compare the estimated causal effects based on the proposed model and the benchmark models with the effects found in the literature, which are often based on randomized controlled lab settings.

Linguistic Style Effects on Headline Engagement on Upworthy.com

To explore the causal effect of language on headline performance, we use the validated engagement dictionaries created by Banerjee and Urminsky (2023) as treatments. We test the effect of each of these linguistic variables, one at a time.[2] To estimate the causal effect of these linguistic features on engagement (i.e., CTRs), and to control for variation between experiments, we add experimental fixed effects as the nontextual X_i in Equation (1), We use

[2]We use Table 1a from Banerjee and Urminsky (2023) and the word list on Banerjee and Urminsky (2023) Open Science Framework (OSF) directory to recreate the linguistic dictionaries. Given that the model takes in a binary variable for treatment variables, we did not estimate the effect of continuous dictionaries, such as reading ease, or dictionaries to which we did not have access, such as intensity of emotion and positive and negative intensity.

the average CTR *multiplied by 100* for each package as dependent variables.

We ran the BERT-based causal model for each of the 43 dictionaries, one at a time, to estimate the causal effect of each linguistic feature. The results of this analysis are in Table 3. We compare the BERT ATE estimates with the RAW ATE based on Equation (6) and the IPW ATE based on Equation (7).

We also compare our causal estimates with the results of a model that controls for all the linguistic dictionaries using a regression. This model is similar to the model of Banerjee and Urminsky (2023). For the linguistic feature regression adjustment (REG ATE), we use the following regression equation:

$$\mathbb{E}[y_i \mid t_{ij}, \boldsymbol{x}_i] = \alpha_i + \tau_j t_{ij} + \sum_{j' \neq j} \beta_{j'} x_{ij'}, \tag{8}$$

where α_i controls for the experiment fixed effects. $x_{ij'}$ is the indicator of the existence of a word from dictionary j' in headline i, t_{ij} indicates whether the treatment dictionary exists, τ_j estimates ATE of dictionary j, and $\beta_{j'}$ is the regression coefficient that controls the other dictionaries.

The proposed model leverages a causal-oriented text embedding. Thus, a strong baseline for the proposed model would be a model that controls for the standard text embedding using a regression approach. To create this baseline model, we use a pre-trained sentence embedding to represent the textual confounders in the headline (Reimers and Gurevych 2019). Specifically, we first remove any words in the treatment dictionary from the headline. Next, we use the MiniLM model (Wang et al. 2020) as a state-of-the-art sentence encoder to transform the modified headline W_i into a 384-dimensional dense vector $\boldsymbol{v}_i$ for confounding control. Note that, unlike the causal BERT model, the sentence encoder is an off-the-shelf model that is not adjusted for causal inference. This EMBED ATE model is estimated for each treatment dictio-

nary j by fitting

$$\mathbb{E}[y_i \mid t_{ij}, \boldsymbol{v}_i] = \alpha_i + \tau_j t_{ij} + \sum_{j'=1}^{384} \tilde{\beta}_{j'} v_{ij'}, \tag{9}$$

where α_i controls the experiment fixed effect; $\tilde{\beta}_{j'}$ are regression coefficients that relate the sentence embeddings to the outcome variable, and τ_j estimates the treatment effect of the appearance of dictionary j in the headline after controlling for the sentence embeddings.

Comparing the causal estimates based on the different methods (i.e., the proposed BERT ATE, RAW ATE, IPW ATE, REG ATE, and EMBED ATE), we see a positive correlation between the proposed approach and the benchmark causal effect measures. Specifically, the correlations between BERT ATE and RAW ATE, IPW ATE, REG ATE, and EMBED ATE are 0.11, 0.12, 0.21, and 0.30, respectively. The greater similarity between the BERT ATE and the EMBED ATE, compared to other baselines, suggests that the vanilla embeddings capture a richer set of confounding variables in the text than the dictionary indicators. The moderate correlation of 0.30 also suggests meaningful differences between the causally adjusted embedding and unadjusted embedding. We also find some other meaningful differences. The BERT-based causal model shows that of the 43 dictionaries tested, 17 positively affected engagement, while 26 had a negative effect. Of the 17 that had a positive effect, BERT ATE differed in sign from the EMBED ATE in four cases (instructional, men, social plural family, threat). However, BERT ATE disagreed 60% of the time with EMBED ATE and 80% of the time with RAW ATE in dictionaries that negatively affected engagement. Similar patterns hold for the IPW and REG baselines.

Now that we have estimated a causal linguistic model on the Upworthy data, we can use the BERT ATE estimates to learn which linguistic features lead to high CTRs. Table 3 shows that terms related to “visual languages,” with words such as “warm,” “view,” and “sweet,” had the biggest positive effect on engagement, with an increase in the CTR of 0.1065%. This result is supported by extant work in the literature showing that "visual language” increases engagement in certain product categories (Nelson and Hitchon 1999; Elder and Krishna 2010). The

Table 3: Average Treatment Effects (ATEs) for Different Linguistic Features Across ATE Estimation Methods

Feature	BERT ATE	RAW ATE	IPW ATE	REG ATE	EMBED ATE
Visual language	0.1065	0.0750	0.0629	-0.0092	0.0033
Conflicting	0.0982	0.0462	0.0350	0.0166	0.0167
Women	0.0903	0.1328	0.0967	0.0309	0.0220
Goals	0.0864	-0.0265	-0.0189	0.0039	0.0014
Authority	0.0787	-0.0156	-0.0114	0.0101	0.0063
Instructional	0.0752	-0.0158	-0.0052	-0.0429	-0.0249
Morality	0.0739	0.0142	0.0154	0.0525	0.0426
Negative emotion	0.0622	0.0444	0.0336	-0.0098	0.0020
Positive emotion	0.0562	-0.0247	-0.0173	0.0071	0.0272
Cute	0.0561	0.0525	0.0448	-0.0107	0.0046
Numeric	0.0467	0.0578	0.0528	-0.0213	0.0077
Forward reference	0.0456	0.0844	0.0754	0.0333	0.0639
Men	0.0237	0.1513	0.1064	-0.0185	-0.0144
Social plus family	0.0143	-0.0372	-0.0244	0.0072	-0.0068
Threat	0.0069	0.0086	0.0021	0.0027	-0.0151
Auditory language	0.0045	-0.0003	-0.0072	0.0074	0.0041
Aesthetic	0.0007	0.0999	0.0831	-0.0000	0.0103
Interrogation	-0.0006	0.0153	0.0121	0.0028	-0.0032
Informal	-0.0014	0.0203	0.0105	0.0250	0.0208
Safekeeping	-0.0069	-0.1070	-0.0991	0.0505	0.0069
Fearful	-0.0308	0.0065	0.0106	0.0106	0.0200
Compare	-0.0311	0.0168	0.0181	-0.0092	-0.0118
Harm	-0.0469	0.0284	0.0299	0.0218	0.0255
First Person	-0.0477	0.0058	0.0036	0.0185	-0.0120
Verbs categories	-0.0556	0.0565	0.0671	0.0007	0.0125
Past	-0.0558	0.0695	0.0496	0.0510	0.0610
Location	-0.0566	0.0377	0.0414	0.0613	0.0210
Hedges	-0.0582	0.0100	0.0093	-0.0070	0.0032
Verbs	-0.0612	0.0350	0.0419	0.0084	0.0220
Future	-0.0656	0.0213	0.0195	0.0278	0.0240
Swear	-0.0703	0.0408	0.0343	-0.0215	-0.0258
Negation	-0.0707	0.0061	0.0106	0.0093	0.0177
Second Person	-0.0711	0.0001	0.0003	-0.0021	-0.0118
Disgust	-0.0864	-0.0048	0.0077	0.0137	0.0110
Social plus other	-0.0946	-0.0030	-0.0007	0.0033	0.0065
Anger	-0.0949	0.0342	0.0298	0.0309	0.0108
Time	-0.1066	0.0362	0.0254	-0.0544	-0.0393
Fairness	-0.1106	-0.1081	-0.0946	-0.0698	-0.0194
Secret	-0.1187	0.1074	0.0979	-0.0006	0.0035
Conjunction	-0.1202	0.0200	0.0149	-0.0035	-0.0116
Deliberation	-0.1356	0.0456	0.0436	-0.0001	-0.0038
Present	-0.1518	-0.0237	-0.0155	-0.0650	-0.0535

textual feature with the next highest lift in engagement is "conflicting words" (*e.g., "against," "destroy," "eviscerate," "tirade"*). Indeed, research has shown that online news articles high in conflict language tend to receive more comments Tenenboim and Cohen (2015). We also find that language associated with instruction, which contain clickbait words (*e.g.,* "here is why," "click here," "this is why") and tutorial words (*e.g.,* "learn how," "teach you how") has a positive effect on engagement. Note that the average treatment obtained from the BERT-based causal model for the Instructional words differs in sign from the ATE of the other methods. Clickbait language ("click here") has been found to increase engagement (Matias and Munger 2019). Aesthetic terms (*e.g.,* "look nice," "breathtaking," "splendid") and terms associated with "cuteness" (*e.g.,"adorable," "cuddly," "munchkin"*) also increase engagements. (Wagner, Baccarella, and Voigt 2017) show that language associated with aesthetics improves social media engagement in the automobile category. As expected, we found that words related to negative emotion (*e.g.,* "kill,""killer," "wars," "violent") increase engagements by 0.0622%. This result confirms that violence and conflict tend to get top billing in newspapers because of their effectiveness.[3]

In terms of groups of words that negatively affect engagement, we found that "present" terms (i.e., terms that use the present tense and words that refer to the present) cause the largest decline in engagement (-0.1518%). Such terms include words like "do," "seem," "has," "feel," "follow," and "exclude." This category was followed by safekeeping words, such as "deliberation" and "conjunction." Deliberation words include cognitive words (*e.g.,* "accept," "affect," exclude"); insight words (*e.g.,* "become," "learn,""know"), and conjunction words (*e.g.,*"also,""although,""unless"), which also decreased engagements. This result accords with the findings of Rennekamp (2012), which show that fewer conjunction words improve fluency in the reader's processing of texts. Using hedging language in headlines (*e.g.,* "could,""potentially," "varies") decreases engagements. This finding is supported by prior studies showing that the use of hedging words decreases the likelihood of receiving funds in peer-

[3]https://pepperdine-graphic.com/opinion-if-it-bleeds-it-leads-the-modern-implications-of-an-outdated-phrase/

to-peer lending contexts (Larrimore et al. 2011). It also indicates that readers prefer articles that have definitive headlines. Similarly, we found that asking questions has a negative effect on clicking behavior. In addition, using social-plus-other words (*e.g., "them," "they," "crowd," "team"*) decreased engagement (-0.0946%), while using social-plus-family words (*e.g.,* "buddies," "fiance," "girlfriends," "brother") had the opposite effect (0.0143%).

Given the causal nature of our analysis, these results could provide editors with guidance on how to improve the effectiveness of the headlines they create.

Table 3 highlights some meaningful discrepancies in the various causal estimates. Hence, in the following subsections, we use different methods and unique aspects of the experimental Upworthy data to examine which estimate aligns more closely with the true treatment effect and to identify potential confounding factors.

Detecting the Degree of Confounding

Although the exact effects of confounders cannot be directly observed, we can assess the magnitude of confounding through a *change-in-estimate approach* (Lee 2014; Maldonado and Greenland 1993; Vander Weele and Shpitser 2011). The idea behind this approach is to compare the treatment effect estimates from models that include adjustments for potential confounders with the estimates from models that do not include them. If a confounder exists, we would expect the treatment effect to change significantly after accounting for the confounder.

To assess the degree of confounding for the LIWC dictionaries in the Upworthy data, we conducted two linear regression analyses using CTRs as the dependent variable. The first regression includes only the focal treatment dictionary, and the second includes the focal treatment dictionary and a possible confounder dictionary (another LIWC dictionary other than the treatment). By comparing the treatment effect between the two regressions, we can assess the strength of confounding. Specifically, we first run a regression of treatment (captured as the existence of a word from the focal dictionary in the headline) on CTRs while controlling for experiment fixed effects. Following the literature, we call this measure of treatment "crude"

because it excludes adjustments for possible textual confounders:

$$\mathbb{E}[y_i|t_{ij}] = \alpha_i + \tau_{j,crude} t_{ij}. \tag{10}$$

Here, α_i represents fixed effects for the Upworthy experiment, $\tau_{j,crude}$ is the unadjusted treatment effect coefficient, and t_{ij} is the indicator for the presence (t_{ij}=1) or absence (t_{ij}=0) of dictionary j in headline i.

The second regression extends this model by adjusting for the presence of a possible confounding dictionary ($v_{ij'}$):

$$\mathbb{E}[y_i|t_{ij}, v_{ij'}] = \alpha_i + \tau_{j,j',adjusted} t_{ij} + \beta_{j'} v_{ij'} \qquad j' \neq j. \tag{11}$$

Here, j' represents a dictionary other than the focal dictionary j, and $v_{ij'}$ is an indicator for the presence of a word from dictionary j' in headline i.

To determine whether a variable is a confounder, we compute the percentage change between the unadjusted and adjusted treatment effect coefficients:

$$\tau_{j,j',\text{percentage}} = \left| \frac{\tau_{j,j',adjusted} - \tau_{j,crude}}{\tau_{j,crude}} \right|. \tag{12}$$

A commonly used cutoff for identifying confounding is $\tau_{j,j',\text{percentage}} \geq 10\%$ (Maldonado and Greenland 1993; Lee 2014). If $\tau_{j,j',\text{percentage}} \geq 0.1$, the variable j' is considered a potential confounder for j.

Application and Results We tested for confounding strength for each of the 43 dictionaries used in the analysis of (Table 3), testing them as both treatments (j) and potential confounders (j'). For each dictionary pair, we ran regressions to evaluate confounding effects. That is, for each dictionary used as treatment j, we estimated 42 regressions following Equation 11, with each of the other 42 dictionaries acting as a possible confounder(j'). We also obtained a crude estimate for the treatment (j) following Equation 10. We then calculated the

strength of the possible confounding of each dictionary (j') for each focal treatment dictionary j following Equation 12. For instance, we examined the possible confounding strength that the "Women" dictionary has for the presence of words from the "Aesthetic" dictionary in the headline ($\tau_{\text{"Aesthetic,""Women," percentage}}$). If this measure exceeded 10%, we flagged the "Women" dictionary as a potential confounder for "Aesthetic." This process resulted in 1,806 regressions (43 treatments × 42 potential confounders per treatment).

Our findings reveal that *all* dictionaries had at least 15 confounder dictionaries in our data. For example, the "Cute" and "Aesthetic" dictionaries were confounded by the "Women" dictionary, indicating that the RAW ATE and IPW ATE estimates for "Aesthetic" are likely to be biased. Furthermore, 12 dictionaries had more than 40 confounders. Across our 43 treatment estimates, the average number of confounders with $\tau_{j,j',\text{percentage}} \geq 0.1$ is 33. These results suggest that RAW estimates are likely to be biased when they do not account for correlations among texts.

To determine whether our BERT-based causal model accounts for such possible confounders, we related the degree of confounding for each treatment dictionary to the discrepancy between the BERT ATE and the RAW ATE. Specifically, we measure this discrepancy using:

$$\text{ATE-DIFF}_{j,\text{percentage}} = \left| \frac{\text{ATE}_{j,\text{BERT}} - \text{ATE}_{j,\text{RAW}}}{\text{ATE}_{j,\text{RAW}}} \right|. \tag{13}$$

We find that the correlation between ATE-DIFF and the average $\tau_{j,j',\text{percentage}}$ across all j' for each treatment j ($\overline{\tau}_{j,\cdot,\text{percentage}}$) is positive and significant ($\rho = 0.62$, p-values < 0.01), indicating that the divergence of BERT estimates from RAW ATE estimates increases as the number of confounders increases. We find a similar correlation when correlating ATE-DIFF with the number of confounders of each treatment dictionary (n_j), defined as $n_j = \sum_{j'} \mathbb{1}_{\tau_{j,j',\text{percentage}}>0.1}$, (ρ=0.28, p-value= 0.06).

This analysis demonstrates that confounders significantly influence treatment effect estimates. The presence of these confounders can explain why the BERT ATE estimates differ from

the RAW ATE estimates. The divergence between the BERT ATE and RAW ATE estimates increases with the strength of the confounding, providing indirect evidence that the BERT model is able to control for such confounding effects. In the next section, we further explore when the BERT ATE and RAW ATE estimates diverge and the role of the confounders' strength.

Differences between Headlines as a Proxy for Confounders' Strength

In the *Semi-synthetic Outcome Validation Analysis* section, we varied the strength of the confounder to demonstrate the ability of the proposed BERT-based causal model to recover the treatment effect, even in the face of such confounders and the inability of traditional measures to correct for confounders. A unique aspect of the Upworthy data allows us to estimate the strength of the confounder effect. Specifically, the number of words (apart from the treatment word) that vary between each pair of headlines can serve as a proxy for the degree of possible confounding. Consider the extreme case in which the only difference between the treatment headline and the control headline is the treatment word itself. In this situation, given that the data comes from an A/B test, the RAW ATE is unbiased estimate of the true treatment effect because there is no other confounder in the headline. Thus, the BERT ATE and RAW ATE should be similar. As more of the words in the two headlines deviate, we would expect the RAW ATE and the BERT ATE to diverge because of possible confounders introduced by the difference in words (other than the treatment words) between the pair of headlines. Moreover, if the proposed model indeed controls for possible confounders, we would expect the BERT ATEs to be robust to the increase the number of words that vary between the headlines.

To study this effect, we performed a pairwise headline analysis. In each experiment, we labeled a package as treatment if it included a word from the focal dictionary and labeled it as control if it did not. For each pair, we computed the number of words that changed as the sum of the "words added" – that is, words that are present in the treatment headline and absent in the control headline – and the "words removed" – that is, words that are in the control headline but not in the treated headline. Overall, we analyzed 128,390 pairs of headlines. The

Table 4: Example of pairwise analysis with visual related words as treatment

	Headline	Words Changed	RAW ATE	BERT ATE	RAW - BERT ATE	Experiment ID
Treatment	"**watch** an eccentric animation of an idea that could change the way we eat in cities forever"	1	-0.0325	-0.043	0.0113	1
Control	"an eccentric animation of an idea that could change the way we eat in cities forever"	1	-	-	-	1
Treatment	"the most simple argument against raising the minimum wage might not be as simple as **it sounds**"	4	-0.5149	-0.0942	-0.4206	2
Control	"the most simple argument against raising the minimum wage might not be as simple as **you think**"	4	-	-	-	2

pairs of headlines in our data differed with respect to 1 to 37 words. The pairs of headlines differed on average by 17.67 words. We then computed the BERT ATE and RAW ATE for each pair of headlines. To analyze the degree of confounding strength, we averaged the ATE estimates based on the number of words that changed between each pair of headlines. Similar to the results of the semi-synthetic data analysis, we find that as the number of word changes between headlines increases, the RAW ATE and BERT ATE estimates diverge. Moreover, the BERT ATE estimate is robust to the change in the number of words.

Table 4 illustrates the pairwise analysis using visual words (e.g., "watch," "video," "sounds") as the treatment. The first two rows of the table report a pair in which the only word that varied between the two headlines is the treatment word: The treatment headline reads "**watch** an eccentric animation of an idea that could change the way we eat in cities forever," and the control headline removes "watch." In this case, the RAW ATE in CTRs should estimate the true causal effect of the negation word. In the second pair of headlines, the two headlines differ by a total of four words: The treatment headline reads "the most simple argument against raising the minimum wage might not be as simple as **you think**," and the control headline reads "the most simple argument against raising the minimum wage might not be as simple as

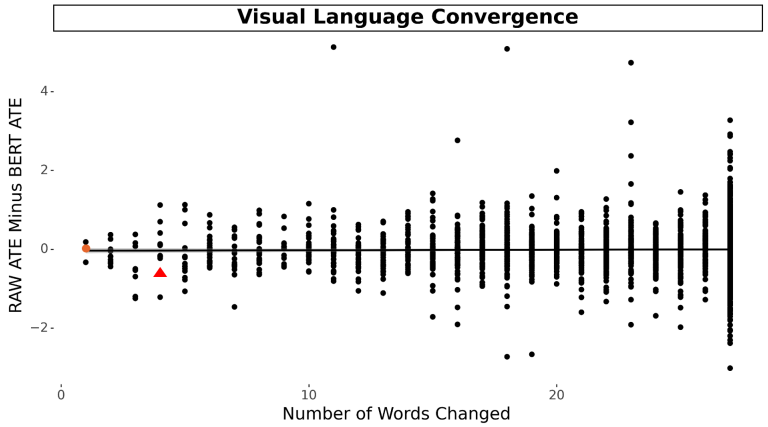

Figure 4: Difference between RAW ATE and BERT ATE estimates for the visual language dictionary as a function of number of words changed, ranging from 1-word change to 27 or more word changes. The orange dot represents the outcome for Experiment ID 1, and the red triangle represents the outcome for Experiment ID 2 in Table 4.

it sounds." The four differences are the addition of two words ("you," "think") and the removal of two words ("it," "sounds"). The treatment word ("sounds") was replaced with "think." The two additional word changes are possible confounders between these two headlines. Thus, this difference may introduce a bias if we do not account for the confounders. Indeed, we can see in the fifth column of Table 4 that in the first experiment, where no confounder exists, the difference between the RAW ATE and BERT ATE estimates is smaller. However, when a confounder exists, as in the second experiment, the two estimates diverge.

To go beyond the anecdotal example in Table 4, Figure 4 depicts the difference between the RAW ATE and the BERT ATE for the 128,390 pairs of headlines that vary with respect to a visual word, assessed according to how many words differed between the two headlines. The difference ranges from 1 word (i.e., only the treatment word differs) to 27 or more words. Each dot in the plot represents the difference between the RAW ATE and BERT ATE estimates. For example, the orange dot indicates the difference in the first experiment (0.0113) in Table 4, whereas the red triangle represents the difference in the second experiment (-0.4206). The plot in Figure 4 has a funnel shape, which means that when a small number of words change between the two headlines, the difference between the RAW ATE estimate and the BERT ATE estimate is small. However, as more words differ between the two headlines, creating a higher

potential for confounding, the two estimates diverge.

We recreated the graph shown in Figure 4 for all the treatment variations.[4] We can see in Figure 5 that across the treatment dictionaries, we observe the same type of funnel. That is, the spread between the differences in the BERT ATE estimates, and the RAW ATE estimates are negligible at one-word changes and increase as the number of word changes increases. To statically confirm that these two estimates indeed are diverging as the number of words changes, we ran a regression in which the dependent variable is the variance of the difference between the RAW ATE and BERT ATE estimates, and the independent variable is the number of words that differed between the pair of headlines. The idea is to find out whether, as the number of word changes increases, the difference between the RAW ATE and the BERT ATE estimates is more volatile. Thus, we would expect the regression coefficient for the number of word changes to be positive and significant. Indeed, the results of our regression analysis show that the coefficient for the number of words changed is 0.028 (C.I. = [0.01, 0.035]), with a t-value of 8.372. The R-squared for this regression was 0.834.

We postulated that the results in Figures 4 and 5 are driven by the confounding bias of the RAW ATE estimates. Thus, we would expect that as more words change between a pair of headlines, the treatment RAW ATE would differ across brackets of number of other words that have changed between headlines because the confound created by the additional change in words is likely to differ between headlines. On the other hand, the BERT ATE, which controls for possible confounders, is likely to lead to a robust ATE estimate a brackets of word changed. We demonstrate this pattern by plotting the mean square deviation ($\sum_i \frac{(ATE_i - \bar{ATE})^2}{n}$) of the ATEs within a number of the word-changed brackets for the BERT ATE and the RAW ATE as a function of the number of words changed. In Figure 6, the red triangles represent the mean square deviation for the RAW ATE estimates, and the orange dots represent the mean square deviation for the BERT ATE estimates. From this plot, we can see that the variation clearly comes from the RAW ATE estimates. Importantly, it shows that the BERT ATE estimates are robust to the

[4] We removed from this figure dictionaries that did not have sufficient observations for all 27 brackets of word changes between pairs of headlines. Specifically, we removed the dictionaries for Disgust and Fairness.

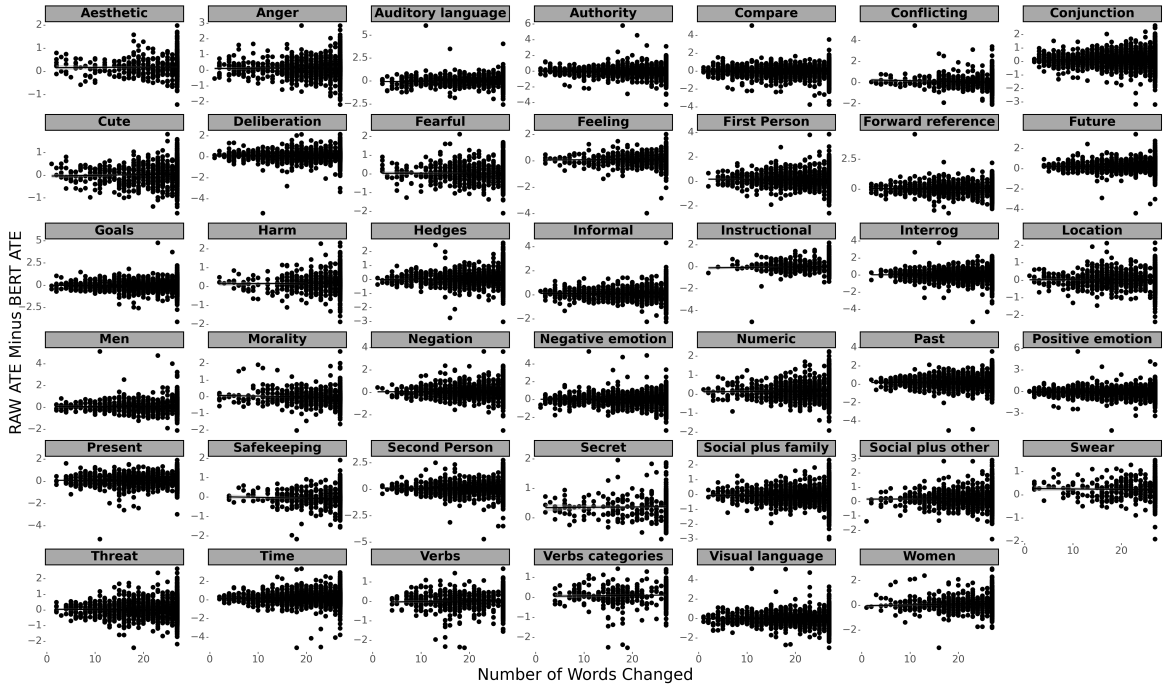

Figure 5: Difference between RAW ATE and BERT ATE estimates for different treatment dictionaries as a function of the number of words that differ between a pair of headlines

increase in the number of words that changed between the pair of headlines.

In the next subsection, we test whether accounting for text confounders using the BERT-based causal model helps to obtain estimates that are more consistent with estimates in the extant literature, which often estimated linguistic effects in lab settings.

Comparing Our Linguistic Treatment Effects with Findings from the Literature

Numerous studies have investigated the effect of linguistic features on people's reactions to written text. Most of these studies involve testing one or a few of these textual features at a time using lab studies (*e.g.,* Packard and Berger (2020)). Because randomized controlled trials are considered the gold standard for causal inference, we assume that findings from prior experimental literature provide a reliable benchmark for evaluating treatment effects. Thus, another way to measure whether our proposed approach is capable of capturing the true treatment effect is to compare whether the proposed BERT-based causal approach correlates better with previous findings in the literature than approaches that do not control for confounders.

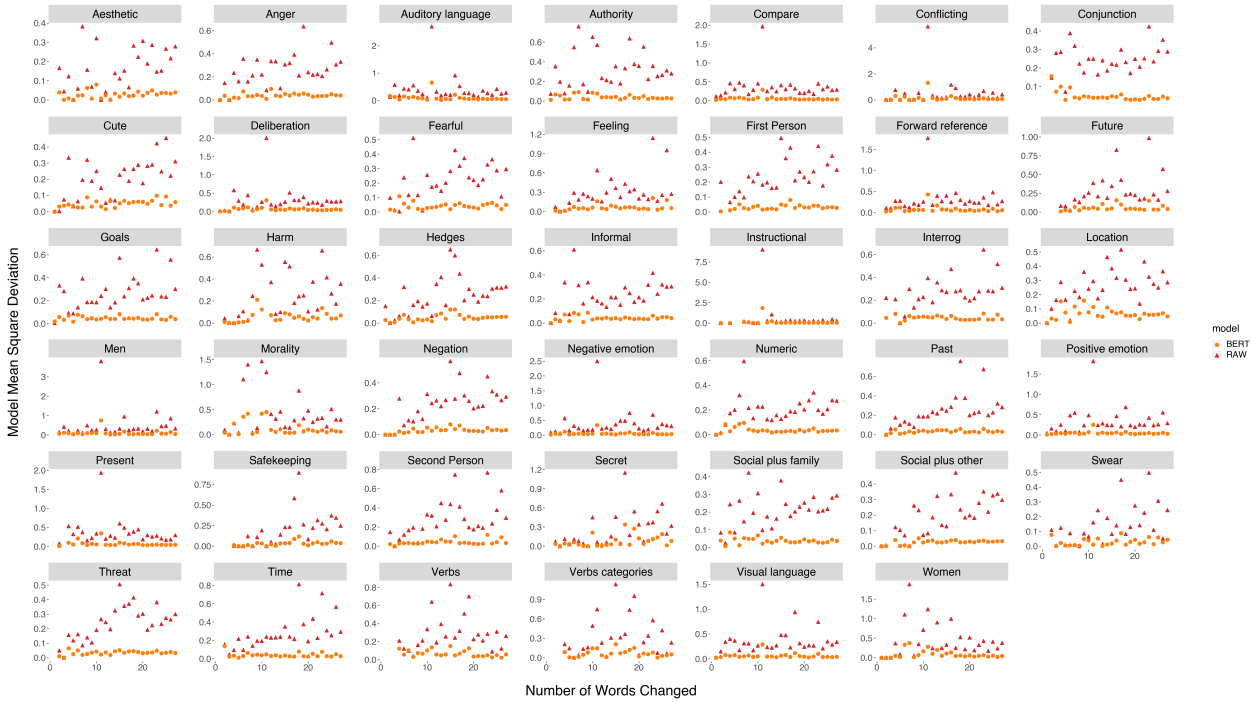

Figure 6: Plot of mean square deviation between RAW ATE (red triangles) and BERT ATE (orange circle) estimates for all treatments as a function of the number of words that vary between pairs of headlines

Banerjee and Urminsky (2023) provides an excellent overview of the findings in the literature. To code these findings, we use Table 2 in Banerjee and Urminsky (2023). For each linguistic construct j, we define the effect of that construct on engagement, as found in previous studies, as:

$$\text{Linguistic Construct Effect}_j = \frac{\#\text{of studies with positive effect} - \#\text{of studies with negative effect}}{\text{Total number of studies}}. \tag{14}$$

For example, if a linguistic construct had seven studies, two with positive effects, three with negative effects, and two with no effects, the linguistic construct effect in the literature would be coded as $-1/7$.

We then compute the correlation between the linguistic construct effects in the literature using Equation (14) and the causal estimates from different methods in Table 3. In addition

Table 5: Correlation between the Findings in the Literature and Different ATE Estimates

Models	Correlation with Literature	Correlation with RAW ART
BERT ATE	0.3297	0.1114
LASSO	0.2470	0.4786
REG ATE	0.0633	0.6735
IPW ATE	0.0719	0.9885
EMBED ATE	0.1995	0.2748
RAW ATE	0.0832	1

to the ATE estimates we have discussed thus far, we follow Banerjee and Urminsky (2023) and estimate a LASSO regression that includes all 43 dictionaries and experiment fixed effects. Such a model can reduce possible collinearity across dictionaries.

As Table 5 shows, the BERT ATE estimates have the highest correlation with the literature (0.3297), followed by the LASSO regression (0.2470), the EMBED ATE (0.1995), the RAW ATE (0.0832), and the IPW ATE (0.0719). The full linear regression had the lowest correlation with the literature (0.0633), possibly because of collinearity across the dictionaries.

Consistent with our prior results, we found that the BERT ATEs are less correlated with the RAW ATEs (0.1114) compared to the LASSO estimates (0.4786), whereas the REG ATE estimates are highly correlated with the RAW ATEs (0.6735). These results are expected because the primary strength of the proposed BERT-based causal model lies in its ability to account for confounders. In contrast, estimates from linear or LASSO regression models depend heavily on the context defined by the covariates that are included in the regression.

Discussion of Empirical Validation

We validated the performance of the BERT-based causal model using two distinct approaches. First, we applied a semi-synthetic data approach, where real data were combined with generated dependent variables to create a known ground truth. This setup allowed us to evaluate the model's ability to successfully recover the actual treatment effects, as well as the inability of simple RAW causal estimates to capture the effect of confounders.

Second, we applied the model to a secondary dataset from the Upworthy news platform. Comparing the BERT ATE with several alternative ATE estimates reveals significant differences between the different approaches. Further analysis shows that the RAW ATE, as well as regression and basic embedding approaches, fail to fully account for the confounding effect of other words in the headlines. We also demonstrate that as the confounding strength increases, as measured by the change in the treatment effect when accounting for confounders, the RAW ATE estimates deviate further from the BERT ATE estimates. A subset analysis confirms that as the confounding strength—measured by the number of word changes surrounding the treatment words—increased, the RAW ATE estimates became increasingly distorted. In contrast, the BERT-based causal model's estimates remained stable. This stability is consistent with patterns observed in the semi-synthetic analysis, indicating that the proposed model effectively mitigates the effect of confounding.

Third, we compared the BERT ATE estimates with the alternative measures in terms of their correlation with results previously found in lab settings and found that our model estimates are more closely related to the findings in the previous literature.

We wish to highlight that, to the best of our knowledge, this is the first empirical validation of the causal textual analysis framework (Veitch, Sridhar, and Blei (2019)), demonstrating empirically using secondary data that the BERT-based causal inference approach is capable of capturing a consistent treatment effect even as the number of possible confounders increases.

LINGUISTIC DRIVERS OF ENGAGEMENT IN CROWDFUNDING: A CAUSAL TEXT ANALYSIS

With the BERT-based causal model validated, in this section, we illustrate the types of research questions that the proposed model can help marketing researchers explore. Specifically, the proposed approach can serve as a confirmatory tool to evaluate whether specific words or groups of words *causally* affect desired outcomes. Our approach is well-suited for analyzing secondary data, especially when both textual confounders and nontextual controls are present.

We use two recent examples from the marketing literature that explore the effect of words on meaningful marketing outcomes to demonstrate how the BERT-based causal model could be particularly valuable in assessing the effect of writing style: (1) online donation behavior on DonorsChoose (Hong and Hoban 2022); and (2) crowdfunding investments on Prosper (Netzer, Lemaire, and Herzenstein 2019). The proposed approach can be used to assess the causal effect of textual features (words or phrases), providing deeper insights into the linguistic drivers of donor and investor behaviors.

For this analysis, we focus on examining the effect of *pre-thanking* and of *second person pronouns* on donation and funding decisions. Pre-thanking refers to expressing gratitude in advance, regardless of whether the reader ultimately complies with the request. Extant research has demonstrated that pre-thanking can promote pro-social behaviors and encourage acts of helping others (Clark, Northrop, and Barkshire 1988; Carey et al. 1976; McGovern, Ditzian, and Taylor 1975; Clark III 1975). Rind and Bordia (1995) showed experimentally that writing "thank you" on a check increases customer tips. Similarly, Merchant, Ford, and Sargeant (2013) found that pre-thanking improves donor retention for nonprofits. In our contexts, Netzer, Lemaire, and Herzenstein (2019) identified a complex link between pre-thanking and loan outcomes, associating gratitude expressions with successful funding but at the same time with higher likelihood of default. Likewise, Hong and Hoban (2022) observed appreciation phrases like "thank you" improved the likelihood of receiving a donation.

"Second-person pronoun" includes words such as "you," "yours," "yourself," and "you are." Past research has shown that "you" words can function as a persuasion technique because they not only engage the audience but also prompt them to think of a specific person they know. Incorporating "you" words, readers are more likely to project themselves into the narrative, making the argument more persuasive by evoking personal connections and emotions. This effect arises because second-person pronouns stimulate mental simulation, helping individuals vividly relate to their own experiences (Brunyé et al. 2009; Green and Brock 2000; Hartung et al. 2016). Packard and Berger (2020) demonstrated that using "you" in song lyrics signifi-

cantly increases downloads and purchase intent. These findings suggest strategic use of "you" words can enhance communication effectiveness such as in storytelling, advertising, and persuasive writing—key elements in crowdfunding contexts. We, therefore, investigate whether donation and crowdfunding data reflect similar patterns, given literature predictions of a strong positive effect on engagement (Banerjee and Urminsky 2023).

The conflicting evidence in the literature on the effect of pre-thanking makes these textual features an ideal test case for our BERT-based causal model. To evaluate its effectiveness, we compare BERT ATE with the RAW ATE, which measures the mean difference in outcome between textual units that include, versus those that do not include, the treated text ("pre-thanking" and "you" words). In addition, we benchmark against two alternative models that control for different confounders. REG NOTEXT estimates the effect of the focal words using a regression model that includes nontextual covariates but excludes textual features apart from the treatment. EMBED ATE extends this approach by incorporating text embeddings from a pre-trained sentence encoder to control for other linguistic variations within the textual unit.

DonorsChoose

Founded in 2000, DonorsChoose is a crowdfunding website enabling U.S. public school teachers (K-12) to request funds for classroom projects, such as supplies, electronic tablets, or musical instruments. Each project includes descriptors like grade level and teacher location. Donors may fully or partially support projects within a four-month period. Projects reaching their funding goal receive pledged amounts; otherwise, donations are returned.

Textual and Nontextual Variables For our analysis, we focus on the donation requests made in 2013. The *dependent variable* was a binary indicator, with one indicating that a classroom project was fully funded and zero indicating that the project was canceled. To control for nontextual variable, we included several relevant features:

1. *Project Cost*: The amount requested by the teacher.
2. *Requested Resource Types*: Categories such as books, supplies, technology, or trips.

3. *Grade Level*: Categorical variables indicating which grade level the project targets.
4. *Metro Type*: Whether the school is in a rural, urban, or suburban area.
5. *Teachers' Experience*: a dummy variable indicating whether the request is the first time the teacher has sought funding from DonorsChoose.
6. *School Location*: categorical variables for the state in which the school is located.
7. *Poverty Level*: The percentage of students in the school who are eligible for free and reduced-cost lunch through the National School Lunch Program
8. *Competition Variables:* These variables fall into four categories:
    - *Platform Competition:* The average number of active projects on the platform during the focal project's funding period.
    - *School District Competition:* The average number of competing donation requests from teachers in the same school district.
    - *Resource Type Competition:*The average number of requests for similar resources as the focal project.
    - *School Competition:* The average number of donation requests from the same school while the focal request was active.

We provide detailed descriptive statistics for these variables in the Web Appendix (Table W1). Regarding textual confounders, we analyzed the full text of the project needs statements using both causal BERT embeddings and a pre-trained embedding (Wang et al. 2020) for the EMBED ATE model. We preprocessed the text by removing HTML tags.[5]

Regarding treatment variables, we defined pre-thanking in the context of DonorsChoose as using words such as "thank(s)," "thank you," and different tenses of the word "appreciate." We used the "you" words dictionary ÷- you, your, you're, yourself etc – from the LIWC 2015 (Pennebaker et al. 2015) for the second-person pronoun categorization. We defined treatment as a binary variable. This construct was defined as 1 if at least one of the words related to it (Pre-Thanking/Second-Person Pronoun) existed in the donation request and was 0 otherwise. Table 6 summarizes these words and the estimates of the considered methods.

Causal Estimates for Pre-thanking and You Words: We estimated the average treatment effect for pre-thanking and you words, based on the four different measures: our proposed BERT ATE, RAW ATE, REG NOTEXT, and EMBED ATE. We find that pre-thanking has a positive

[5]We did not correct any grammatical errors to preserve the authenticity of the text.

Table 6: Words associated with thanking and second person and their causal impact on donations likelihood: DonorChoose

	WORDS LIST	BERT ATE	RAW ATE	REG NOTEXT	EMBED ATE
Pre-thanking	Thank(s), thank you, Appreciate(s)(d)	0.0691	0.0304	0.0121	0.0075
You	You, you're, your, yours, yourself	0.0339	0.0068	-0.0022	-0.0021

effect on funding (ATE:0.0691). This finding is consistent with prior results based on experiments (Rind and Bordia 1995; Merchant, Ford, and Sargeant 2013). However, the magnitude of the treatment effect is much stronger based on our model, relative to the benchmarks that use fewer controls.

Consistent with the experimental results from Packard, Moore, and McFerran (2018) and others, we found that using "you" words has a positive effect on donation requests (ATE:0.0339). These results have the opposite sign relative to the REG NOTEXT and EMBED ATE estimates. When calculating the degree of confounding following the same analysis as in the *Detecting the Degree of Confounding* section, we found that, for pre-thanking, only one LIWC dictionary had a confounding strength of more than 10%: the "you" words dictionary; meanwhile for "you" words, five different dictionaries indicate confounding at more than 10%: biology, body, perception, negation, and hear words. These findings suggest that when estimating the effect of textual features the presence of additional confounders that are correlated with the treatment and dependent variables can lead to biased estimates; thus, properly accounting for such confounding is crucial.

Prosper

Founded in 2005, Prosper is a crowdfunding platform connecting borrowers requesting personal loans (1,000˘25,000) with lenders. Like DonorsChoose, it initially followed an all-or-nothing funding model. Borrowers specify the loan amount and maximum interest rate they are willing to pay, originally determined via a Dutch-like auction.[6] Borrowers must provide

[6]The auction mechanism was replaced by pre-defined rates in 2009.

personal and financial details, including debt-to-income ratio and loan amount. Additionally, borrowers may include a project image and write a loan description explaining their needs, though Prosper does not verify this textual information.

Textual and nontextual Variables We limited our analysis to loans originating between April 2007 and October 2008. The dependent variable in our analysis is the loan status, coded as a binary outcome: 1 indicates that the loan was successfully funded, and 0 indicates that the loan expired. We excluded all loans that were canceled or withdrawn by borrowers during the bidding process.

Regarding nontextual variables, we controlled for the important information a lender would use to decide whether to bid on a loan:

1. *Amount Requested:* The total loan amount requested by the borrower.
2. *Debt-to-Income Ratio:* A financial indicator used by lenders to assess borrowers' repayment ability, as well as a dummy variable indicating whether the debt-to-income ratio was missing.
3. *Maximum Rate Willing to Pay*: The maximum interest rate the borrower is willing to accept.
4. *Credit Score (Risk Grade):* A categorical variable represented as a letter grade indicating credit risk, ranging from AA (lowest risk) to HR (highest risk). The complete scores are AA, A, B, C, D, E, and HR.
5. *Loan Category:* The intended use of the loan (*e.g.,* debt consolidation, business, medical, wedding).
6. *Group Membership:* A dummy variable indicating whether the borrower is part of a Prosper group. Group membership can be a positive signal because group leaders may perform additional borrower vetting.
7. *Loan Request with Images:* A dummy variable indicating whether the borrower included an image with their loan request.

For textual confounders, we focused on the loan description field, where borrowers describe themselves and explain why they need the loan. Similar to the DonorsChoose analysis, we used words such as "thank(s)," "thank you," and "appreciate(s)(d)" for the pre-thanking treatment and the "you" words dictionary for the second-person pronouns.

Table 7: Words associated with thanking and second person, and their impact on loan funding: Prosper

	WORDS LIST	BERT ATE	RAW ATE	REG-NOTEXT	EMBED ATE
Pre thanking	Thank(s), thank you, appreciate(s)(d)	0.0950	-0.0244	0.0141	-0.0035
You	You, you're, your, yours, yourself	0.2165	-0.0101	0.0148	0.0077

Causal Estimates for Pre-thanking and You Words: We applied the BERT-based causal model to analyze the binary outcome of loan funding. The treatment variable is binary, indicating the presence or absence of at least one word from the respective dictionaries in the loan description field. As shown in Table 7, consistent with our findings from DonorsChoose and prior experimental research on pre-thanking, we find that the presence of pre-thanking words positively affects the likelihood of securing funding (BERT ATE: 0.095). This result suggests that even in investment settings, where lenders expect a financial return, politeness still holds value. It aligns with the results of Netzer, Lemaire, and Herzenstein (2019), who found that pre-thanking is correlated with funding and suggested that expressions of gratitude may signal an attempt to build goodwill with lenders that evidently succeeds. In addition, in line with Packard and Berger (2020) and our DonorsChoose analysis, second-person pronouns ("you" words) significantly increased the likelihood of loan approval (BERT ATE: 0.2165).

Notably, failing to account for confounders in Prosper loan applications leads to even bigger discrepancies and misleading conclusions. Specifically, if confounding is ignored, pre-thanking and "you" words would erroneously appear to decrease the likelihood of funding (RAW ATE: -0.0244 and -0.0101, respectively). In fact, we find a strong degree of confounding between the focal dictionaries and the rest of the text in the loan application. For pre-thanking words, all LIWC dictionaries serve as confounders at the 10% level, with "Insight words" and "Differ words" exerting the most significant confounding levels: a 95.6% and 95.5% change, respectively, in the parameters of pre-thanking when these dictionaries are included versus not included). Likewise, second-person pronouns were confounded by 66 of 72 LIWC dictionaries, with "Affiliation words" and "Ingest words" causing the most substantial changes in estimates.

These findings from the change-in-estimate method suggest that textual confounding is significantly stronger in the Prosper data compared to the DonorsChoose data. This relationship indicates that, when writing their loan request, borrowers often include words that together are used to increase the likelihood of loan funding; hence, studying each of the words in isolation without accounting for their potential confounding can lead to misleading inferences.

CONCLUSION

In this research, we introduce to marketing contexts a novel BERT-based causal model designed to estimate the causal effect of specific words or groups of words. Text analysis in marketing research has traditionally been correlational, relying on machine learning and deep learning models for prediction. A limitation of this approach is that these models often lack interpretability and causal insights. Recent advances in marketing research have increasingly focused on causal inference, but most studies rely on experimental or quasi-experimental settings, such as policy changes or marketing interventions, to infer causality. Few tools exist for estimating causal effects from secondary data, particularly when text is involved.

To address this gap, we build on recent causal modeling work outside of marketing (Veitch, Sridhar, and Blei 2019) and extend it to handle the complex nature of marketing data, incorporating both textual confounders and nontextual controls. This approach allows researchers to estimate causal effects from naturally occurring text, providing deeper insights into engagement drivers.

To validate the proposed model, we conducted two analyses. In the first, we used semi-synthetic data, combining real textual data with generated dependent variables. This approach allowed us to establish a causal effect ground truth against which we could benchmark our empirical estimates. We demonstrated that the BERT-based causal model effectively accounts for textual confounders, outperforming traditional causal inference methods, such as IPW.

In the second analysis, we used real A/B headline testing data from the news website Upworthy.com (Matias et al. 2021). A/B testing provides a unique validation opportunity because

the experimental design inherently eliminates potential unobserved confounders. This analysis enabled a direct comparison of the BERT-based causal model with prior approaches using traditional machine learning (LASSO) and econometric models to estimate engagement effects measured through CTRs. Using findings from the marketing literature as a proxy for ground truth and the naive difference in CTR as a biased baseline, we demonstrated that the BERT ATE estimates were more strongly correlated with the literature and less correlated with the raw CTR differences. In contrast, models based on LASSO regression and econometric methods were more correlated with the biased raw CTR differences and less consistent with established findings in the literature.

In addition, a post hoc analysis revealed where and why the BERT ATE estimates deviated from the raw CTR differences. For control and treatment headline pairs differing by only one word, our BERT ATE estimates closely matched the RAW ATE estimates. However, as the number of word changes between control and treatment headlines increased, the raw estimates diverged significantly from the BERT ATE estimates. This result indicates that contextual complexity—measured by the number of word changes—drives the difference, highlighting the BERT-based model's ability to account for contextual nuances that traditional models miss. To the best of our knowledge, this paper is the first to validate the BERT-based textual causal framework using empirical data.

With the proposed model validated, we detailed a few potential applications of the proposed approach based on recent marketing publications. We investigated two long-standing research questions in marketing: the impact of second-person pronouns and pre-thanking (expressing gratitude before receiving a service or donation) on engagement. Previous findings in the literature on these topics have been contradictory. Our goal was to clarify these relationships through a causal lens.

In the context of donations, where funds are not expected to be returned, we found that pre-thanking and "you" words both had a positive effect. These results align with prior experimental findings showing that pre-thanking can enhance pro-social behavior and that using

second-person pronouns causes the reader to be invested by encouraging the reader to put on the hat of the protagonist.

In investment settings, where funds are expected to be returned with interest. We found that pre-thanking had a positive impact on obtaining funds. This outcome suggests that acts of gratitude do sway investors. Similarly, we found that using "you words" had a positive effect on loan funding, consistent with experiment-based results from the literature. These results highlight that the effectiveness of certain language strategies depends on the specific context and the expectations surrounding the interaction.

There several possible extensions to our work. First, our paper primarily focuses on identifying the causal effect of text. One could extend our work to account for both text and image data as confounders. Advances in image-to-text (Alayrac et al. 2022) may permit combining these two sources of unstructured data in a single model. Second, we focus on estimating the average treatment effect (ATE), one could extend our work to explore heterogeneity in treatment effect and conditional average treatment estimation (CATE).

The BERT-based causal model provides a robust tool for estimating the causal effects of words or dictionaries in contexts with both textual confounders and nontextual controls. By bridging the gap between predictive and causal modeling in marketing research, our research advances text-based causal inference, enabling researchers to explore more complex linguistic effects on relevant marketing outcomes. We encourage future research to use such causal textual model to further explore important marketing questions that involve like textual data like what makes for an engaging ad, social media post or the causal effect of specific word used in product reviews on sales.

REFERENCES

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds et al. (2022), "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, 35, 23716–23736.

Banerjee, Akshina and Oleg Urminsky (2023), "The Language That Drives Engagement: A Systematic Large-scale Analysis of Headline Experiments.," *Available at SSRN 3770366*.

Bang, Heejung and James M Robins (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61 (4), 962–973.

Berger, Jonah, Ashlee Humphreys, Stephan Ludwig, Wendy W Moe, Oded Netzer, and David A Schweidel (2020), "Uniting the tribes: Using text for marketing insight," *Journal of marketing*, 84 (1), 1–25.

Berger, Jonah and Katherine L Milkman (2012), "What makes online content viral?," *Journal of marketing research*, 49 (2), 192–205.

Brunyé, Tad T, Tali Ditman, Caroline R Mahoney, Jason S Augustyn, and Holly A Taylor (2009), "When you and I share perspectives: Pronouns modulate perspective taking during narrative comprehension," *Psychological Science*, 20 (1), 27–32.

Carey, J Ronald, Steven H Clicque, Barbara A Leighton, and Frank Milton (1976), "A test of positive reinforcement of customers," *Journal of Marketing*, 40 (4), 98–100.

Clark, Hewitt B, James T Northrop, and Charles T Barkshire (1988), "The effects of contingent thank-you notes on case managers'visiting residential clients," *Education and Treatment of Children*, pages 45–51.

Clark III, Russell D (1975), "The effects of reinforcement, punishment and dependency on helping behavior," *Personality and Social Psychology Bulletin*, 1 (4), 596–599.

Devlin, Jacob (2018), "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon (2021), "Overlap in observational studies with high-dimensional covariates," *Journal of Econometrics*, 221 (2), 644–654.

Elder, Ryan S and Aradhna Krishna (2010), "The effects of advertising copy on sensory thoughts and perceived taste," *Journal of consumer research*, 36 (5), 748–756.

Feder, Amir, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts et al. (2022), "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond," *Transactions of the Association for Computational Linguistics*, 10, 1138–1158.

Green, Melanie C and Timothy C Brock (2000), "The role of transportation in the persuasiveness of public narratives.," *Journal of personality and social psychology*, 79 (5), 701.

Gui, Lin and Victor Veitch (2022a), "Causal Estimation for Text Data with (Apparent) Overlap Violations,".

Gui, Lin and Victor Veitch (2022b), "Causal estimation for text data with (apparent) overlap violations," *arXiv preprint arXiv:2210.00079*.

Hartung, Franziska, Michael Burke, Peter Hagoort, and Roel M Willems (2016), "Taking perspective: Personal pronouns affect experiential aspects of literary reading," *PloS one*, 11 (5), e0154732.

Hong, Jiyeon and Paul R Hoban (2022), "Writing more compelling creative appeals: A deep learning-based approach," *Marketing Science*, 41 (5), 941–965.

Imbens, Guido W and Donald B Rubin (2015), *Causal inference in statistics, social, and biomedical sciences* Cambridge university press.

Keith, Katherine, David Jensen, and Brendan O'Connor "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates," "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics," pages 5332–5344, Online: Association for Computational Linguistics (2020).

Larrimore, Laura, Li Jiang, Jeff Larrimore, David Markowitz, and Scott Gorski (2011), "Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success," *Journal of Applied Communication Research*, 39 (1), 19–37.

Lee, Paul H (2014), "Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification?," *Journal of epidemiology*.

Lee, Thomas Y and Eric T Bradlow (2011), "Automated marketing research using online customer reviews," *Journal of Marketing Research*, 48 (5), 881–894.

Maldonado, George and Sander Greenland (1993), "Simulation study of confounder-selection strategies," *American journal of epidemiology*, 138 (11), 923–936.

Matias, J Nathan and Kevin Munger (2019), "The Upworthy Research Archive: A Time Series of 32,488 Experiments in US Advocacy," *Preprint*.

Matias, J Nathan, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole (2021), "The Upworthy Research Archive, a time series of 32,487 experiments in US media," *Scientific Data*, 8 (1), 195.

McGovern, Leslie P, Jan L Ditzian, and Stuart P Taylor (1975), "The effect of one positive reinforcement on helping with cost," *Bulletin of the Psychonomic Society*, 5 (5), 421–423.

Merchant, Altaf, John B Ford, and Adrian Sargeant "'Don't forget to say thank you': The effect of an acknowledgement on donor relationships," "New Horizons in Arts, Heritage, Nonprofit and Social Marketing," pages 5–22, Routledge (2013).

Mickey, Ruth M and Sander Greenland (1989), "The impact of confounder selection criteria on effect estimation," *American journal of epidemiology*, 129 (1), 125–137.

Mozer, Reagan, Luke Miratrix, Aaron Russell Kaufman, and L Jason Anastasopoulos (2018), "Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality,".

Nelson, Michelle R and Jacqueline C Hitchon (1999), "Loud tastes, colored fragrances, and scented sounds: How and when to mix the senses in persuasive communications," *Journalism & Mass Communication Quarterly*, 76 (2), 354–372.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine your own business: Market-structure surveillance through text mining," *Marketing Science*, 31 (3), 521–543.

Netzer, Oded, Alain Lemaire, and Michal Herzenstein (2019), "When words sweat: Identifying signals for loan default in the text of loan applications," *Journal of Marketing Research*, 56 (6), 960–980.

Neyman, Jerzy (1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 5 (4), 465–472 Originally published in Polish in 1923. Translated by Dorota M. Dabrowska and Terence P. Speed.

Olteanu, Alexandra, Onur Varol, and Emre Kiciman "Distilling the outcomes of personal experiences: A propensity-scored analysis of social media," "Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing," pages 370–386 (2017).

Packard, Grant and Jonah Berger (2020), "Thinking of you: How second-person pronouns shape cultural success," *Psychological Science*, 31 (4), 397–407.

Packard, Grant, Yang Li, and Jonah Berger (2024), "When language matters," *Journal of Consumer Research*, 51 (3), 634–653.

Packard, Grant, Sarah G Moore, and Brent McFerran (2018), "(I'm) happy to help (you): The impact of personal pronoun use in customer–firm interactions," *Journal of Marketing Research*, 55 (4), 541–555.

Pennebaker, James W, Ryan L Boyd, Kayla Jordan, and Kate Blackburn (2015), "The development and psychometric properties of LIWC 2015,".

Puranam, Dinesh, Vrinda Kadiyali, and Vishal Narayan (2021), "The impact of increase in minimum wages on consumer perceptions of service: A transformer model of online restaurant reviews," *Marketing Science*, 40 (5), 985–1004.

Puranam, Dinesh, Vishal Narayan, and Vrinda Kadiyali (2017), "The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors," *Marketing Science*, 36 (5), 726–746.

Reimers, Nils and Iryna Gurevych "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing," (2019).

Rennekamp, Kristina (2012), "Processing fluency and investors' reactions to disclosure readability," *Journal of accounting research*, 50 (5), 1319–1354.

Rind, Bruce and Prashant Bordia (1995), "Effect of server's "thank you" and personalization on restaurant tipping 1," *Journal of Applied Social Psychology*, 25 (9), 745–751.

Roberts, Margaret E, Brandon M Stewart, and Richard A Nielsen (2020), "Adjusting for confounding with text matching," *American journal of political science*, (ajps.12526).

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American statistical Association*, 89 (427), 846–866.

Rosenbaum, Paul R and Donald B Rubin (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70 (1), 41–55.

Tenenboim, Ori and Akiba A Cohen (2015), "What prompts users to click and comment: A longitudinal study of online news," *Journalism*, 16 (2), 198–217.

Timoshenko, Artem and John R Hauser (2019), "Identifying customer needs from user-generated content," *Marketing Science*, 38 (1), 1–20.

Tirunillai, Seshadri and Gerard J Tellis (2012), "Does chatter really matter? Dynamics of user-generated content and stock performance," *Marketing science*, 31 (2), 198–215.

Tirunillai, Seshadri and Gerard J Tellis (2014), "Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation," *Journal of marketing research*, 51 (4), 463–479.

Toubia, Olivier, Garud Iyengar, Renée Bunnell, and Alain Lemaire (2019), "Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption," *Journal of Marketing Research*, 56 (1), 18–36.

Vander Weele, Tyler J and Ilya Shpitser (2011), "A new criterion for confounder selection," *Biometrics*, 67 (4), 1406–1413.

Veitch, Victor, Dhanya Sridhar, and David M Blei (2019), "Adapting Text Embeddings for Causal Inference,".

Veljanovski, Marko and Zach Wood-Doughty "DoubleLingo: Causal Estimation with Large Language Models," "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 (2024).

Wagner, Timm F, Christian V Baccarella, and Kai-Ingo Voigt (2017), "Framing social media communication: Investigating the effects of brand post appeals on user interaction," *European Management Journal*, 35 (5), 606–616.

Wang, Wenhui, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei (2020), "MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers," *arXiv preprint arXiv:2012.15828*.

WEB APPENDIX

WEB APPENDIX I: SUMMARY STATISTICS OF THE DONORCHOOSE AND PROSPER DATASET

Table W1 provide detailed descriptive statistics for the nontextual confounding variables of DonorsChoose, and Table W2 provide the statistics for Prosper.

Table W1: DonorsChoose Descriptive Statistics

	Min	Max	50%	Mean	Std.
Project Cost	92.00	111596.67	491.47	601.72	986.07
compete	177.00	6005.00	5014.00	4552.07	1368.77
compete school	0.00	50.00	1.00	1.81	2.71
compete district	0.00	444.00	6.00	36.91	85.30
compete type	0.00	2017.00	471.00	736.33	621.43
Books	0.00	1.00	0.00	0.19	0.39
Other	0.00	1.00	0.00	0.11	0.31
Supplies	0.00	1.00	0.00	0.35	0.48
Technology	0.00	1.00	0.00	0.35	0.48
Trips	0.00	1.00	0.00	0.01	0.08
Visitors	0.00	1.00	0.00	0.00	0.04
Grades 3-5	0.00	1.00	0.00	0.31	0.46
Grades 6-8	0.00	1.00	0.00	0.17	0.37
Grades 9-12	0.00	1.00	0.00	0.12	0.33
Grades PreK-2	0.00	1.00	0.00	0.40	0.49
rural	0.00	1.00	0.00	0.09	0.29
suburban	0.00	1.00	0.00	0.30	0.46
town	0.00	1.00	0.00	0.05	0.21
unknown	0.00	1.00	0.00	0.06	0.24
urban	0.00	1.00	0.00	0.49	0.50
Alabama	0.00	1.00	0.00	0.01	0.10
Alaska	0.00	1.00	0.00	0.00	0.05
Arizona	0.00	1.00	0.00	0.02	0.15
Arkansas	0.00	1.00	0.00	0.01	0.11
California	0.00	1.00	0.00	0.18	0.38
Colorado	0.00	1.00	0.00	0.02	0.12
Connecticut	0.00	1.00	0.00	0.02	0.12
Delaware	0.00	1.00	0.00	0.00	0.07
District of Columbia	0.00	1.00	0.00	0.01	0.08
Florida	0.00	1.00	0.00	0.06	0.24

Georgia	0.00	1.00	0.00	0.03	0.18
Hawaii	0.00	1.00	0.00	0.00	0.07
Idaho	0.00	1.00	0.00	0.00	0.06
Illinois	0.00	1.00	0.00	0.05	0.23
Indiana	0.00	1.00	0.00	0.03	0.17
Iowa	0.00	1.00	0.00	0.00	0.07
Kansas	0.00	1.00	0.00	0.00	0.06
Kentucky	0.00	1.00	0.00	0.01	0.10
Louisiana	0.00	1.00	0.00	0.01	0.11
Maine	0.00	1.00	0.00	0.01	0.08
Maryland	0.00	1.00	0.00	0.01	0.11
Massachusetts	0.00	1.00	0.00	0.02	0.14
Michigan	0.00	1.00	0.00	0.03	0.16
Minnesota	0.00	1.00	0.00	0.01	0.10
Mississippi	0.00	1.00	0.00	0.01	0.09
Missouri	0.00	1.00	0.00	0.02	0.12
Montana	0.00	1.00	0.00	0.00	0.05
Nebraska	0.00	1.00	0.00	0.00	0.05
Nevada	0.00	1.00	0.00	0.02	0.13
New Hampshire	0.00	1.00	0.00	0.00	0.05
New Jersey	0.00	1.00	0.00	0.02	0.15
New Mexico	0.00	1.00	0.00	0.00	0.07
New York	0.00	1.00	0.00	0.09	0.28
North Carolina	0.00	1.00	0.00	0.05	0.21
North Dakota	0.00	1.00	0.00	0.00	0.03
Ohio	0.00	1.00	0.00	0.02	0.13
Oklahoma	0.00	1.00	0.00	0.02	0.13
Oregon	0.00	1.00	0.00	0.01	0.09
Pennsylvania	0.00	1.00	0.00	0.02	0.15
Rhode Island	0.00	1.00	0.00	0.00	0.07
South Carolina	0.00	1.00	0.00	0.02	0.15
South Dakota	0.00	1.00	0.00	0.00	0.05
Tennessee	0.00	1.00	0.00	0.02	0.14
Texas	0.00	1.00	0.00	0.05	0.22
Utah	0.00	1.00	0.00	0.01	0.12
Vermont	0.00	1.00	0.00	0.00	0.04
Virginia	0.00	1.00	0.00	0.02	0.13
Washington	0.00	1.00	0.00	0.02	0.13
West Virginia	0.00	1.00	0.00	0.00	0.06
Wisconsin	0.00	1.00	0.00	0.01	0.09
Wyoming	0.00	1.00	0.00	0.00	0.03
Teacher First Trial	0.00	1.00	0.00	0.34	0.47
School Percentage Free Lunch	0.00	100.00	72.00	65.97	24.93

Table W2: Prosper Descriptive Statistics

	Min	Max	50%	Mean	Std.
Amount Requested	1000	25000	5000	7517	6337
Debt To Income Ratio	0.000000	10.010000	0.290000	0.514809	1.220235
Lender Rate	0.000000	0.350000	0.190000	0.205256	0.092441
Missing DTI	0.000000	1.000000	0.000000	0.104790	0.306284
Is Borrower Homeowner	0.000000	1.000000	0.000000	0.366771	0.481925
Group Membership	0.000000	1.000000	0.000000	0.163092	0.369451
Has Images	0.000000	1.000000	1.000000	0.518332	0.499665
A	0.000000	1.000000	0.000000	0.044134	0.205394
AA	0.000000	1.000000	0.000000	0.034097	0.181480
B	0.000000	1.000000	0.000000	0.068370	0.252381
C	0.000000	1.000000	0.000000	0.120583	0.325643
D	0.000000	1.000000	0.000000	0.167655	0.373561
E	0.000000	1.000000	0.000000	0.173663	0.378820
HR	0.000000	1.000000	0.000000	0.391491	0.488085
category 0	0.000000	1.000000	0.000000	0.442473	0.496681
category 1	0.000000	1.000000	0.000000	0.253928	0.435258
category 2	0.000000	1.000000	0.000000	0.021192	0.144023
category 3	0.000000	1.000000	0.000000	0.087610	0.282727
category 4	0.000000	1.000000	0.000000	0.110019	0.312914
category 5	0.000000	1.000000	0.000000	0.021637	0.145496
category 6	0.000000	1.000000	0.000000	0.013458	0.115226
category 7	0.000000	1.000000	0.000000	0.049683	0.217291
Number of Prior Listing	0.000000	46.000000	0.000000	0.915673	2.015160

WEB APPENDIX II: BERT AIPW VERSUS BERT OUTCOME ONLY ESTIMATORS

Figure W1 presents results for the baseline methods, and the causal BERT approach using the outcome-only estimator from Equation (3) and AIPW estimator from Equation (4). The BERT ATE by either outcome-only estimator or doubly robust AIPW estimator is closer to the ground truth than baseline methods. However, the BERT ATE using AIPW estimator achieves the most accurate estimation. Accordingly in the result of the paper we use the AIPW estimaor for teh BERT ATE approach.

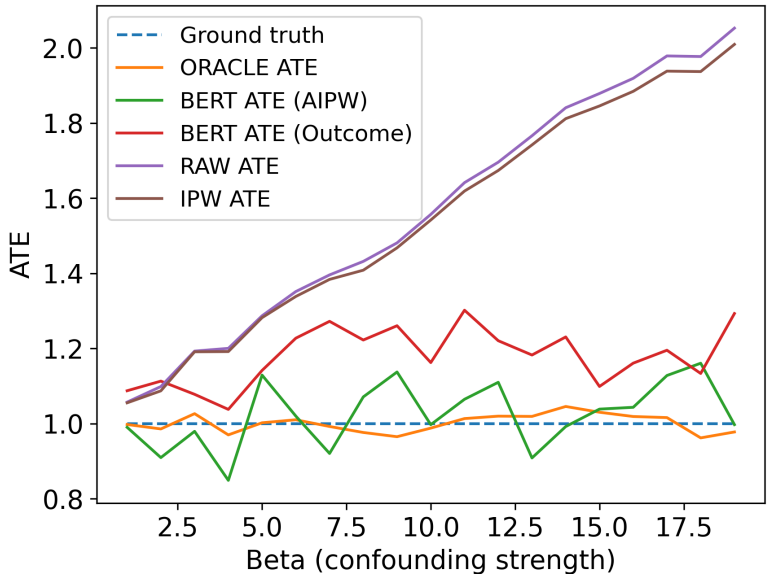

Figure W1: Estimated ATEs for different levels of confound across different models, including two estimation methods using causal BERT embeddings.