LLMs as Emotion Analyzers for Causal Models: Partial Identification with Fuzzy Interval Data

Huidi Ma
Cockrell School of Engineering
The University of Texas at Austin
Austin, TX 78705
mhd0601@utexas.edu

Wendao Xue
McCombs School of Business
The University of Texas at Austin
Austin, TX 78705
wendao.xue@mccombs.utexas.edu

Yifan Yu [*]
McCombs School of Business
The University of Texas at Austin
Austin, TX 78705
yifan.yu@mccombs.utexas.edu

Abstract

Information systems (IS) researchers often use machine learning algorithms to recognize emotions in massive and unstructured online content and then study the causal effect of emotions on various business outcomes. Large language models (LLMs) as automatic emotion analyzers offer new opportunities for IS researchers to advance the causal understanding of emotions. Nevertheless, we show that directly plugging LLM-generated emotional variables into econometric models can induce bias in causal estimation because LLMs are imperfect in emotion recognition, which adds noise to the causal models. We propose a novel algorithm to correct such bias. A key feature is that the algorithm considers predictions generated by different LLMs to form fuzzy interval data. Then, partial identification of causal parameters is achieved. The algorithm meets three design requirements. First, it is unsupervised, meaning it does not need additional labeled data to correct the causal estimation. Second, it is flexible in incorporating the predictions of different LLMs (leveraging the "wisdom of the AI crowd") and can be easily adapted to various causal models. Third, the causal estimators are theoretically guaranteed to achieve consistency. This work provides important implications for causal inference with LLMs, the causal study of emotions, and prescriptive analytics for fuzzy interval data.

Keywords: LLM, causal inference, partial identification, fuzzy interval data.

1 Introduction

Emotions in online communications are of substantial managerial value because they are influential to various business outcomes, including sales performance [Song et al., 2019, Yu et al., 2023], online review helpfulness [Yin et al., 2014, 2016], and customer service effectiveness [Han et al., 2023]. With the proliferation of social media, valuable emotional information is embedded in massive and unstructured online content, which calls for the use of automatic emotion analyzers [Yu et al., 2023]. Emotion analyzers are algorithms or models designed to detect, interpret, and classify human emotions from unstructured data such as text and images.

[*] Corresponding author. Contact at yifan.yu@mccombs.utexas.edu.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Information systems (IS) research has shown that the imperfect emotion analyzers built on traditional machine learning (ML) models, such as the random forest and XGBoost, can produce measurement errors in emotional variables and induce bias in causal estimation [Yang et al., 2018, Qiao and Huang, 2021, Zhang et al., 2023a]. Recently, Large Language Models (LLMs) as emotion analyzers have become increasingly popular due to their state-of-the-art (SOTA) performance and ease of use [Wang et al., 2023]. An important question is how the capabilities of the LLMs as emotion analyzers could improve our understanding of emotions and business outcomes. This paper empirically illustrates that LLMs suffer from the same measurement error issue as the traditional ML algorithms do. Further, we show that bagging the predictions of different LLMs[2] does not significantly mitigate the bias. The results indicate that, while LLMs are SOTA emotion analyzers, they are still imperfect in recognizing human emotions [Zhang et al., 2023b] and can bias our understanding of emotions and business outcomes. The literature and our empirical illustration motivate the need for an algorithm that accurately draws a causal understanding of emotions from the massive unstructured data with (imperfect) LLMs.

We propose three key design requirements: unsupervised, flexible, and accurate. First, the unsupervised algorithm should avoid using labeled datasets, aligning with the recent AI revolution that aims to develop general-purpose pre-trained models to avoid expensive and time-consuming processes of data labeling and supervised model training. Second, flexibility means that the algorithm should incorporate different versions of LLMs to leverage the "wisdom of the AI crowd" and its performance should not rely on a specific LLM version to ensure its generalizability [Abbasi et al., 2024]. Flexibility also means the algorithm can be easily adapted in various causal models [Zhang et al., 2023a]. Third, accuracy means that, technically, the algorithm should produce consistent causal estimators [Qiao and Huang, 2021, Zhang et al., 2023a], which offer estimates sufficiently close to the true causal parameters asymptotically.

This paper proposes a novel algorithm to meet the three design requirements. An important distinction of this algorithm compared to those in the IS literature [Yang et al., 2018, Qiao and Huang, 2021, Zhang et al., 2023a] is that it treats LLM-predicted variables as *fuzzy interval data*. We expect that any point prediction generated by a specific LLM may be biased. When directly plugging the predicted value into the causal model, the model would falsely take the prediction error as information, thereby biasing the causal parameters. Multiple LLMs will likely produce different predicted values (e.g., different valence levels) for the same unstructured data point (e.g., an online review). The minimal and maximum predicted values form an *interval* for the true emotional value. It is not plausible to expect that all LLMs would constantly over- or under-estimate the true emotional value for every data point, and thus, the interval should contain the true emotional value in a probabilistic way. Considering the probabilistic nature, we refer to these intervals as fuzzy interval data.

The rest of the paper is organized as follows. We first review the related literature and design theories and identify key design requirements and principles. Then, we elaborate on our model setup and the algorithm design. Next, we present our empirical analysis of the emotional capabilities of LLMs and conduct numerical experiments to evaluate the algorithm with multiple benchmarks. Finally, we conclude the work and discuss its implications.

2 Literature Review and Design Theories

2.1 Debiasing ML-generated Emotional Variables in Causal Models

A large body of IS literature adopts ML to generate important variables from unstructured data and then plug the variables into econometric models to study their business impacts [Yang et al., 2018, Shin et al., 2020, Qiao and Huang, 2021, Zhang et al., 2023a]. Emotion is a leading example of ML-generated variables in the IS literature on unstructured data analytics [Yin et al., 2014, Song et al., 2019, Hou et al., 2023, Yu et al., 2023]. A possible reason is that emotions significantly impact a wide range of business outcomes and have deep theoretical implications for understanding individual behaviors [Yin et al., 2014, Yu et al., 2023].

Building on this literature, Liu et al. [2024] offers a related approach using LLMs to annotate unstructured data and derive causal insights, but their methodology differs from ours in key aspects. While they take a data-driven approach to uncover causal factors, we focus on theory-driven features

[2] This is a common strategy to reduce prediction errors in ensemble learning.

like emotions. Additionally, Liu et al. [2024] uses the FCI algorithm aligned with Pearl's causal diagrams, whereas we adopt the causal regressuib framework from econometrics. Finally, while they validate their methods empirically, we extend this by providing theoretical guarantees for our approach.

Because measurement errors in ML-generated variables can cause endogeneity and bias the causal estimation, it is critical for empirical IS researchers to account for such errors [Yang et al., 2018]. The IS literature on debiasing ML-generated variables in causal models also focuses on empirical applications with emotional variables [Yang et al., 2018, Qiao and Huang, 2021, Yang et al., 2022, Zhang et al., 2023a]. We focus on emotions in this work because ML algorithms are almost always imperfect in recognizing human emotions [Zhang et al., 2023b].

To construct consistent estimators for the effects of emotions, the literature relies on labeled datasets [Qiao and Huang, 2021, Zhang et al., 2023a]. By comparing labeled and ML-predicted values, researchers can estimate the error distributions conditional on predicted values and leverage this information to design more accurate econometric models. This design idea is smart for traditional ML contexts because researchers have to construct labeled datasets to train ML-based emotion analyzers before causal studies. To correct biases, researchers only need to randomly split the on-hand labeled datasets respectively for training and correcting [Fong and Tyler, 2021]. However, when the available labeled dataset is small in scale, the variance of the causal estimator is often large, leading to poor finite-sample property [Zhang et al., 2023a]. It means that researchers can still get biased or insignificant results. More importantly, using labeled datasets is incompatible with the recent AI revolution, which aims to use pre-trained general-purpose models to avoid expensive and time-consuming processes of data labeling and model training. It remains unclear if LLMs have sufficient capabilities to offer accurate emotional information needed by causal models and how we can debias LLM-generated emotional variables without the costly labeling processes.

2.2 Partial Identification with Interval Data

Generalized Method of Moments (GMM) is a core methodological framework for causal identification [Greene, 2018]. If the identification assumption relies on moment equations, the framework offers point identification. Point identification is unavailable and partial identification is needed when moment conditions are inequalities. Partial identification produces an identified set (e.g., an interval) for a causal parameter.

When a data point is an interval instead of a single value, Manski and Tamer [2002] showed that moment conditions can be formulated as inequalities, and partial identification may be achieved. However, their method by is not applicable when the interval data is fuzzy. Because we cannot ensure that the true emotional values are always between the minimum and maximum values generated by a series of LLMs, we have to account for the fuzzy interval data issue in our method, which extends their work.

2.3 Design Requirements and Principles

Based on the following design theories and literature, we propose three key design requirements, i.e., unsupervised, flexible, and accurate. First, the algorithm performs unsupervised learning, i.e., automatically learning the causal effects of emotions from unlabeled and unstructured data with LLMs.

Second, the flexibility requirement is both for LLMs and causal models. On the one hand, the algorithm should allow for incorporating different versions of LLMs. Due to the subtle and subjective nature of emotions [Yu et al., 2023], different humans can interpret the same emotional information distinctly. We expect that the different LLMs usually cannot achieve an agreement in emotion recognition. However, the current proliferation of LLMs may enable us to leverage the “wisdom of the AI crowd" to achieve a better causal understanding of emotions. Further, the rapid iteration of LLM versions poses challenges for the design of LLM-based applications. To make sustainable impacts, the performance of the designed algorithm should not rely on the performance of any specific LLM version [Abbasi et al., 2024]. On the other hand, management researchers leverage a wide range of causal models, and therefore, the designed algorithm should be easily adaptable to various model specifications [Zhang et al., 2023a].

Third, the accuracy requirement indicates that using LLM-predicted variables should not introduce bias into the causal estimation. Technically, the algorithm should provide a consistent estimator for the causal relationship, robust to imperfect emotion recognition [Qiao and Huang, 2021, Zhang et al., 2023a]. A consistent estimator is also essential for valid statistical inference.

3 Model and Algorithm

Suppose a researcher is interested in estimating the following model:

$$\mathbb{E}[y|x,v] = F(x'\beta_o + \alpha_o v), \tag{1}$$

where x is a $d-$dimensional vector and v is a scalar. The functional form of $F(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is known to the researcher. The parameter vector $\gamma_o := (\beta_o', \alpha_o)'$ is known to lie in a finite-dimensional vector space, i.e., $\gamma_o \in \mathcal{C} \subset \mathbb{R}^{d+1}$. Let $f(x, v, \gamma_o) := F(x'\beta_o + \alpha_o v)$. For example, in the linear regression, the researcher assumes the underlying model as follows:

$$y = f(x, v, \gamma_o) + \epsilon = x'\beta_o + \alpha_o v + \epsilon \quad \text{s.t. } \mathbb{E}[\epsilon|x, v] = 0, \tag{2}$$

and therefore $F(\cdot)$ being the identity function. Another example is the Logistic regression, where

$$\mathbb{E}[y|x,v] = P\left(y = 1|x, v\right) = \frac{1}{1 + e^{-(x'\beta_o + \alpha_o v)}}. \tag{3}$$

Here, $F(x) = 1/(1 + e^{-x})$.

Suppose the researcher can observe a sample of the random variable y, x, but they cannot observe a sample of the random variable v. For example, it is difficult for the firm to discern the true emotion in a movie review. But LLMs can generate a series of predictions of v. Suppose the minimum and maximum predictions are v_0 and v_1, respectively. More formally, let the y, x, v, v_0, v_1 denote the random variable, and let $Y_i, X_i, V_i, V_{1i}, V_{0i}$ denote one realization of the random variable. To summarize, we have a data sample $\{Y_i, X_i, V_{0i}, V_{1i}\}$, where V_{0i} and V_{1i} construct a fuzzy interval which contains V_i with probability p. We are interested in estimating the γ_o in (1). In Section 3.1, we describe the identification assumptions and the basic finding on the identification of γ_o.

3.1 Identification and Estimation

We begin with the identification assumptions detailed in Appendix A. These assumptions guide the relationship between the latent variable and the outcome variable, providing a theoretical foundation for the model.

The assumptions collectively allow us to form the basis for identifying the latent structure of the model. Assumption 1 imposes a monotonic relationship between the latent variable v and the outcome variable, ensuring that we can infer the latent effects through their observable manifestations. Assumption 2 incorporates the uncertainty in the latent variable estimation, accounting for the possibility that the latent variable may fall within an interval, rather than being observed exactly. Finally, Assumption 3 ensures that the interval boundaries do not provide additional predictive power beyond the latent variable itself, simplifying the estimation.

With these assumptions in place, we now proceed to establish the identification result, which allows us to determine the parameters of interest in the model based on the observable data.

Proposition 1. Let Assumptions 1, 2 and 3 hold. Let $c \in \mathcal{C}$ and

$$V(c) := \left[(x, v_0, v_1) : \begin{array}{c} \left\{ \begin{array}{c} f(x, v_0, c)p_2(x, v_0, v_1) + f(x, v_1, c)p_1(x, v_0, v_1) \\ +f(x, v_1 + b, c)p_3(x, v_0, v_1) < \eta_o(x, v_0, v_1) \end{array} \right\} \\ \cup \left\{ \begin{array}{c} \eta_o(x, v_0, v_1) < f(x, v_0 - b, c)p_2(x, v_0, v_1) \\ +f(x, v_0, c)p_1(x, v_0, v_1) + f(x, v_1, c)p_3(x, v_0, v_1) \end{array} \right\} \end{array} \right]. \tag{4}$$

Then γ_o is identified relative to c if and only if $P(V(c)) > 0$. Therefore, the identified set C^* solves:

$$C^* = \{c : P(V(c)) = 0\}. \tag{5}$$

Proposition 1 discusses the identification of the set C^*. For C^* to be identified, it is necessary that $P(V(c)) > 0$. Consider, for example, an uninformative observed interval where $v_0 \to -\infty$ and $v_1 \to \infty$. Under Assumption 1, it follows that $f(x, v_0, c) < \eta_o(x, v_0, v_1) < f(x, v_1, c)$ for all $c \in \mathcal{C}$, resulting in $P(V(c)) = 0$ and C^* is not identified.

Lemma 1. Let $g_1(c, x, v_0, v_1) = \mathbf{1}\big[f(x, v_0, c)p_2(x, v_0, v_1) + f(x, v_1, c)p_1(x, v_0, v_1) + f(x, v_1 + b, c)p_3(x, v_0, v_1) < \eta_o(x, v_0, v_1)\big]$, and $g_0(c, x, v_0, v_1) = \mathbf{1}\big[\eta_o(x, v_0, v_1) < f(x, v_0 - b, c)p_2(x, v_0, v_1) + f(x, v_0, c)p_1(x, v_0, v_1) + f(x, v_1, c)p_3(x, v_0, v_1)\big]$. Then every $c \in C^*$ solves the problem:

$$\min_{c \in C} Q(c, \eta) \tag{6}$$

where

$$Q(c, \eta) = \int \left[g_1(c, x, v_0, v_1)w_1(c, x, v_0, v_1) + g_0(c, x, v_0, v_1)w_0(c, x, v_0, v_1)\right] dP(x, v_0, v_1), \tag{7}$$

and the two weighting $w_1(\cdot)$ and $w_0(\cdot)$ are given by:

$$\begin{aligned}
&w_1(c, x, v_1, v_0)\\
&= w\left(f(x, v_0, c)p_2(x, v_0, v_1) + f(x, v_1, c)p_1(x, v_0, v_1) + f(x, v_1 + b, c)p_3(x, v_0, v_1), \eta_o(x, v_0, v_1)\right),\\
&w_0(c, x, v_1, v_0)\\
&= w\left(f(x, v_0 - b, c)p_2(x, v_0, v_1) + f(x, v_0, c)p_1(x, v_0, v_1) + f(x, v_1, c)p_3(x, v_0, v_1), \eta_o(x, v_0, v_1)\right).
\end{aligned}$$

No $c \in C^*$ solves this problem.

Lemma 1 suggests an estimator for C^* by solving the sample analog. Let $\widehat{\eta}(x, v_0, v_1)$ be an consistent estimator for $\eta_o(x, v_0, v_1)$. Let $\widehat{g}_1(c, x, v_0, v_1) = \mathbf{1}\big[f(x, v_0, c)p_2(x, v_0, v_1) + f(x, v_1, c)p_1(x, v_0, v_1) + f(x, v_1 + b, c)p_3(x, v_0, v_1) < \widehat{\eta}(x, v_0, v_1)\big]$, and $\widehat{g}_0(c, x, v_0, v_1) = \mathbf{1}\big[\widehat{\eta}(x, v_0, v_1) < f(x, v_0 - b, c)p_2(x, v_0, v_1) + f(x, v_0, c)p_1(x, v_0, v_1) + f(x, v_1, c)p_3(x, v_0, v_1)\big]$. Every $c \in \widehat{C}^*$ solves the following analog problem:

$$\widehat{C}^* := \left[c \in \mathcal{C} : \widehat{Q}(c, \widehat{\eta}) \leq \min_{\widetilde{c} \in \mathcal{C}} \widehat{Q}(\widetilde{c}, \widehat{\eta}) + \varepsilon_n\right] \tag{8}$$

where $\varepsilon_n = o_p(1)$,

$$\widehat{Q}(c, \widehat{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{g}_1(c, X_i, V_{0i}, V_{1i})w_1(c, X_i, V_{0i}, V_{1i}) + \widehat{g}_0(c, X_i, V_{0i}, V_{1i})w_0(c, X_i, V_{0i}, V_{1i})\right], \tag{9}$$

and

$$\begin{aligned}
&w_1(c, x, v_1, v_0)\\
&= w\left(f(x, v_0, c)p_2(x, v_0, v_1) + f(x, v_1, c)p_1(x, v_0, v_1) + f(x, v_1 + b, c)p_3(x, v_0, v_1), \widehat{\eta}(x, v_0, v_1)\right),\\
&w_0(c, x, v_1, v_0)\\
&= w\left(f(x, v_0 - b, c)p_2(x, v_0, v_1) + f(x, v_0, c)p_1(x, v_0, v_1) + f(x, v_1, c)p_3(x, v_0, v_1), \widehat{\eta}(x, v_0, v_1)\right).
\end{aligned}$$

For practical use, we choose a quadratic loss as the weighting function, i.e., $w(a, b) = (a - b)^2$. We propose the following algorithm:

In the following subsection, we provide the asymptotic property for the given estimator.

3.2 Asymptotic Property

Given that the parameter of interest and the estimator could be sets, we first need to define the distance between the two sets before presenting the convergence result. Let $\rho(\widehat{C}^*, C^*)$ measure the distance between $\widehat{C}^*$ and C^*, such that:

$$\rho(\widehat{C}^*, C^*) := \sup_{\widehat{c} \in \widehat{C}^*} \inf_{c^* \in C^*} |\widehat{c} - c^*|, \tag{10}$$

$$\rho(C^*, \widehat{C}^*) := \sup_{c^* \in C^*} \inf_{\widehat{c} \in \widehat{C}^*} |\widehat{c} - c^*|. \tag{11}$$

Then we define $\widehat{C}^* \xrightarrow{p} C^*$ if $\rho(\widehat{C}^*, C^*) \xrightarrow{p} 0$ and $\rho(C^*, \widehat{C}^*) \xrightarrow{p} 0$.

```
Algorithm 1 Proposed algorithm
Input: Unstructured data
       Structured covariates and output data $\{X_i, Y_i\}_{i=1}^n$
  Step 1: Interval Construction
  Step 1.1: Use $m$ LLMs to generate predicted variables: $\{M_{1i}, \ldots, M_{mi}\}_{i=1}^n$;
  Step 1.2: Construct interval data $\{V_{0i}, V_{1i}\}_{i=1}^n$, where $V_{0i} = \min\{M_{1i}, \ldots, M_{mi}\}$ and $V_{1i} = \max\{M_{1i}, \ldots, M_{mi}\}$.
  Step 2: Consistently estimate $\eta_o(\cdot) = E[Y|X, V_0, V_1]$ with the data $\{Y_i, X_i, V_{0i}, V_{1i}\}_{i=1}^n$.
      ▷ Comment: One can use ML algorithms (e.g., random forest) or nonparametric methods (e.g., sieve estimator) to consistently estimate $\eta_o(\cdot)$. Use cross-validation to avoid over-fitting.
  Step 3: Criterion Value Calculation. Use the full data and the estimated $\eta_o$ to solve for the optimization problem in (8).
      ▷ Comment: Feasible optimization methods include discretization Hütter and Rigollet [2021] and simulated annealing Manski and Tamer [2002]. Hyperparameters $b$, $p_1$, $p_2$, and $p_3$ are chosen based on contextual information.
Output: Estimator for $\gamma_o$ using $\widehat{C}^*$.
```

Proposition 2. Let Assumptions 1, 2 and 3 hold. Let the parameter space $\mathcal{C}$ be compact with $\gamma_o \in \mathcal{C}$. Assume there exists $\phi(x, v_0, v_1)$ such that

$$|g_1(c, x, v_0, v_1)h_1(c, x, v_0, v_1) + g_0(c, x, v_0, v_1)h_0(c, x, v_0, v_1)| \leq \phi(x, v_0, v_1) \tag{12}$$

where

$$h_1(c, x, v_0, v_1) = \begin{pmatrix} f(x, v_0, c)p_2(x, v_0, v_1) + f(x, v_1, c)p_1(x, v_0, v_1) \\ +f(x, v_1 + b, c)p_3(x, v_0, v_1) - \eta_o(x, v_0, v_1) \end{pmatrix}^2, \tag{13}$$

$$h_0(c, x, v_0, v_1) = \begin{pmatrix} f(x, v_0 - b, c)p_2(x, v_0, v_1) + f(x, v_0, c)p_1(x, v_0, v_1) \\ +f(x, v_1, c)p_3(x, v_0, v_1) - \eta_o(x, v_0, v_1) \end{pmatrix}^2, \tag{14}$$

for all (c, x, v_0, v_0) and $\int \phi(x, v_0, v_1) dF_{X,V_0,V_1}$ is finite. Assume that $\widehat{\eta}(x, v_0, v_1) \xrightarrow{p} \eta_o(x, v_0, v_1)$, a.e. (x, v_0, v_1). Let $\{\sup_{c \in \mathcal{C}} |\widehat{Q}(c, \widehat{\eta}) - Q(c, \eta_o)|\}/\varepsilon_n \xrightarrow{p} 0$, then $\widehat{C}^* \xrightarrow{p} C^*$.

Proposition 2 established the theoretical foundation for the proposed estimator, demonstrating that the estimated set converges to the true set as the sample size approaches infinity. This proposition is crucial for ensuring the reliability of the proposed estimator. To ensure that $\widehat{\eta}(x, v_0, v_1) \xrightarrow{p} \eta_o(x, v_0, v_1)$ a.e. (x, v_0, v_1), we suggest using machine learning algorithms such as random forests or neural networks to estimate the nuisance function η_o. The consistency results for $\widehat{\eta}$ can be established under certain regularity conditions [Athey et al., 2019, Chen, 2007].

4 Empirical Analysis and Numerical Experiments

In this section, we first provide an empirical analysis of the emotional capabilities of LLMs. We then conduct numerical experiments to evaluate our algorithm. We utilize the SST-5 dataset, which consists of 11,855 sentences extracted from movie reviews, to perform sentiment analysis [Socher et al., 2013]. This dataset is chosen for several reasons. First, SST-5 provides a wide range of valence labels, from very negative emotion (labeled as 0) to very positive emotion (labeled as 4), which can be seen as an ordinal or continuous variable that fits our framework. Second, each sentence in this dataset has been carefully annotated by three human judges, offering solid ground truth. Third, the emotions contained in movie reviews have been studied in the literature and shown to be meaningful for the movie industry [Song et al., 2019, Yu et al., 2023]. We randomly select 5,000 sentences from the SST-5 dataset and utilize gpt-3.5-turbo to generate the valence for each sentence. The prompts used in this paper strictly follow the practice suggested by Zhang et al. [2023b].

To incorporate different versions of LLMs, we use two different prompts (zero-shot vs. few-shot) and temperature parameters (randomness in LLM responses, ranging from 0 as the most deterministic to 1 as the most random). First, we adopted a zero-shot prompt with three distinct temperature settings: 0, 0.5, and 1. The corresponding models are referred to as Model 1 (M1), Model 3 (M3), and Model

5 (M5), respectively. Second, we adopted a few-shots prompt with the three temperature settings: 0, 0.5, and 1. The corresponding models are referred to as Model 2 (M2), Model 4 (M4), and Model 6 (M6), respectively. For each sentence, we collect LLMs-generated valence from each model.

4.1 Descriptive Analysis for the LLMs-generated Values

We find that Fleiss' Kappa statistic for the six models is 0.791, which implies that (1) there is some agreement among the models, likely because all models may capture some common emotional signals from the sentences, and (2) the predictions are not fully aligned across models, likely induced by different prompts, temperature parameters, and/or the random nature of generative AI models.

Table 1 summarizes the true versus predicted valence from different models. We observe that the LLMs-generated variable exhibits different error patterns across varying valence levels. For positive valence (3 and 4), the LLMs tend to underestimate the true values, whereas, for non-positive valence (0, 1, and 2), the LLMs tend to overestimate the true values. To provide a more thorough analysis of how these differing error patterns affect our algorithm's performance, we divided the dataset into two subsets by the true valence. The positive valence dataset consists of data with true valence ranging from positive (3) to very positive (4), while the non-positive valence dataset consists of data with true valence ranging from very negative to neutral (0 to 2).

For non-positive valence, the few-shot models outperform the zero-shot models, with mean values being slightly closer to that of the true valence, and overall accuracy being higher. For positive valence, although the means of few-shot models are closer to the mean of true valence on average, their accuracy has decreased compared to the zero-shot models. This indicates an intricate error pattern regarding LLMs-generated variables within the dataset. The last column, Bagging, presents the average generated valence from M1-M6. Interestingly, despite the general belief that bagging predictors in ensemble learning enhances accuracy [Breiman, 1996], this does not hold true for LLMs-generated variables in our case. This discrepancy calls for algorithms to account for inaccuracies in generated variables and correct the subsequent bias in downstream model estimation. In the last row, we demonstrate that when we use the minimum and maximum values from the generated valence to form an interval, only 63% of these intervals cover the true valence. This makes the method proposed by Manski and Tamer [2002] inapplicable in our case, as it assumes the true value lies within the interval.

4.2 Performance Analysis and Results

The previous subsection examined the efficacy of LLMs-generated valence. However, a more pertinent question for businesses is whether such inaccuracies significantly impact causal estimation. If they do, does our algorithm offer a viable solution? In this subsection, we employ simulated experiments to demonstrate the efficacy of our algorithm. We adopt the following data-generating process (DGP): The latent variable of interest, V, represents the true valence in the SST-5 dataset. The covariate, X, serves as the structured control variable. For instance, in movie sales, the structured control variable could be the number of competing movies released. In this experiment, we assume that X follows a normal distribution with a mean of 1 and a variance of 4. The outcome variable (e.g., movie sales) is simulated using the equation $Y = 1 + 2V - X + e$, where e follows a normal distribution with a mean of 0 and a variance of 0.1. Here, we are interested in estimating the causal effect of a latent variable (e.g., valence in movie reviews) on the outcome variable (e.g., movie sales). The DGP provides a ground truth value of 2.

However, without annotation, we cannot obtain the true valence from the text data (e.g., movie reviews). Utilizing the LLMs, we generate six valence estimates from models M1 to M6. Using the minimum and maximum values from these generated valences, we form an interval $[V_0, V_1]$ for each sentence from SST-5. During the implementation of Algorithm 1, we choose b to be 1 because LLMs may struggle to distinguish “very negative" valence from “negative" valence; however, it is less common for them to confuse “negative" valence with “positive" valence. We choose the coverage rate as $p_1 = 0.63$ and set $p_2 = p_3$ accordingly, ensuring $\sum_i p_i = 1$. The discretization method solves the optimization problem [Hütter and Rigollet, 2021].

For a detailed visualization of the experimental setup and intermediary results, refer to Figure 1 in the Appendix. Figure 1 presents the main results from our proposed Algorithm 1 alongside results from various models. Initially, we conducted an OLS regression of the outcome variable on the covariate and the generated valence, utilizing results from M1-M6. The 95% confidence intervals of the estimated coefficients are illustrated as the first six intervals in each subfigure. The results implied that directly using the generated valence from LLMs tends to introduce bias in the downstream regression estimation. Moreover, bagging on the generated valences cannot mitigate this bias.

To correct for bias, we propose using the LLMs-generated variables as interval data. Instead of our proposed Algorithm 1, one might consider a naive method to estimate the coefficients by using the upper and lower endpoints to conduct two separate OLS, yielding two coefficients to form the interval. However, this interval remains biased. The intuition is that when using the upper and lower endpoints to perform two separate regressions, the variance of the latent variable and its covariance with the outcome may be underestimated, resulting in biased outcomes.

In the last two columns, we present the results from Manski and Tamer [2002] (MT method) and our Algorithm 1. Our results outperform theirs for non-positive valence by covering the true value. The intuition behind this is that their underlying assumptions are not satisfied in our dataset (i.e., latent variable must be contained in the interval), whereas our proposed method relaxes their assumptions by allowing for a fuzzy design for the interval data, as previously described in Section 3.1.

One would observe a wider confidence interval in both the MT method and our proposed method compared to methods M1 through M6. The intuition behind this is that methods M1 to M6 provide precise but inaccurate estimates. This is because they use predicted labels that are precise but inaccurate. In contrast, the MT method and our proposed method use less precise information (inputs are intervals instead of points). As a result, the MT method and our proposed method yield interval estimates, which can be seen as 'looser' versions of point estimates. In our empirical application, we demonstrate that although the MT method also relies on imprecise information, they made wrong assumptions (i.e., the interval does not necessarily cover the true label)[3]. Therefore, the interval estimates produced by the MT method do not encompass the true value as one might expect.

5 Concluding Remarks

We propose a novel algorithm that can estimate the causal effects of emotions from unstructured data using LLMs as emotion analyzers. The algorithm meets three key design requirements, i.e., unsupervised, flexible, and accurate. We empirically explored the emotional capabilities of LLMs and demonstrated the performance of the algorithm through numerical experiments. A key feature of this algorithm is to view LLM-generated data as fuzzy intervals and to achieve partial identification. This work not only contributes to the IS literature on the causal studies of emotions but also offers implications for LLM-enabled causal inference of other features in unstructured data. The paper advances the econometric method of partial identification with interval data by allowing the intervals to be fuzzy.

Building on current findings, future research could explore the following areas: (1) understanding how a multi-LLM solution can scale in practical settings and (2) exploring more in specific contexts beyond text data in business. While our empirical focus has been on analyzing unstructured data in business, we acknowledge the importance of domain adaptation. To apply this approach successfully in other domains, further adaptation may be required to improve performance.

[3] In fact, the coverage rate is only 63% in the real world dataset in our empirical application.

References

A. Abbasi, J. Parsons, G. Pant, O. R. L. Sheng, and S. Sarker. Pathways for design research on artificial intelligence. *Information Systems Research*, 2024.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. 2019.

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

L. Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

X. Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

C. Fong and M. Tyler. Machine learning predictions as regression covariates. *Political Analysis*, 29 (4):467–484, 2021.

W. H. Greene. Econometric analysis. *Retrieved on July*, 15:2022, 2018.

E. Han, D. Yin, and H. Zhang. Bots with feelings: Should ai agents express positive emotion in customer service? *Information Systems Research*, 34(3):1296–1311, 2023.

J.-R. Hou, J. Zhang, and K. Zhang. Pictures that are worth a thousand donations: How emotions in project images drive the success of online charity fundraising campaigns? an image design perspective. *MIS Quarterly*, 47(2):535–584, 2023.

J.-C. Hütter and P. Rigollet. Minimax estimation of smooth optimal transport maps. 2021.

C. Liu, Y. Chen, T. Liu, M. Gong, J. Cheng, B. Han, and K. Zhang. Discovery of the hidden world with large language models. *arXiv preprint arXiv:2402.03941*, 2024.

C. F. Manski and E. Tamer. Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70(2):519–546, 2002.

M. Qiao and K.-W. Huang. Correcting misclassification bias in regression models with variables generated via data mining. *Information Systems Research*, 32(2):462–480, 2021.

D. Shin, S. He, G. M. Lee, A. B. Whinston, S. Cetintas, and K.-C. Lee. Enhancing social media analysis with visual data analytics: A deep learning approach. *MIS Quarterly*, 44(4):1459–1492, 2020.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

T. Song, J. Huang, Y. Tan, and Y. Yu. Using user-and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms. *Information Systems Research*, 30(1):191–203, 2019.

Z. Wang, Q. Xie, Y. Feng, Z. Ding, Z. Yang, and R. Xia. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*, 2023.

M. Yang, G. Adomavicius, G. Burtch, and Y. Ren. Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, 29(1): 4–24, 2018.

M. Yang, E. McFowland III, G. Burtch, and G. Adomavicius. Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science*, 1(2):138–155, 2022.

D. Yin, S. D. Bond, and H. Zhang. Anxious or angry? effects of discrete emotions on the perceived helpfulness of online reviews. *MIS quarterly*, 38(2):539–560, 2014.

D. Yin, S. Mitra, and H. Zhang. Research note—when do consumers value positive vs. negative reviews? an empirical investigation of confirmation bias in online word of mouth. *Information Systems Research*, 27(1):131–144, 2016.

Y. Yu, Y. Yang, J. Huang, and Y. Tan. Unifying algorithmic and theoretical perspectives: Emotions in online reviews and sales. *MIS Quarterly*, 47(1), 2023.

J. Zhang, W. Xue, Y. Yu, and Y. Tan. Debiasing machine-learning-or ai-generated regressors in partial linear models. *Available at SSRN*, 2023a.

W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023b.

A Assumptions and Theoretical Foundations

Assumption 1. Assume the known function $F(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is a strictly increasing function with $\beta_o \in \mathbb{R}^d$ and $\alpha_o \geq 0$.

Assumption 1 posits a qualitative relationship between the latent variable v and the outcome variable. This can be generalized to $E[y|x, v]$ weakly increasing in v. While this assumption may not hold in all contexts, it is plausible in certain scenarios when the researcher has some theoretical foundation on the causal relationship between the latent and outcome variables. For example, within reasonable data ranges, the experience of pleasantness (valence) is positively associated with review helpfulness based on the confirmation bias theory [Yin et al., 2016, Yu et al., 2023]; emotional intensity is positively associated with content virality based on the theory of arousal and social activation [Berger and Milkman, 2012]. Additionally, following Manski and Tamer [2002], this assumption can be tested operationally.

Assumption 2. Assume that v belongs to the interval $[v_0, v_1]$ with some known probability, i.e., $P(v \in [v_0, v_1]) = p_1(x, v_0, v_1)$. Further, assume that $P(v \in [v_0 - b, v_0]) = p_2(x, v_0, v_1)$ and $P(v \in [v_1, v_1 + b]) = p_3(x, v_0, v_1)$ with $p_1(x, v_0, v_1) + p_2(x, v_0, v_1) + p_3(x, v_0, v_1) = 1$.

Assumption 2 extends Assumption I in Manski and Tamer [2002] by allowing the latent variable to belong to the interval in a fuzzy design. This assumption is grounded in our empirical evidence (detailed later) that it is unlikely that LLM-generated intervals always encompass the true emotional variable. As a result, the estimation method proposed by Manski and Tamer [2002] becomes infeasible. It is possible that one may have prior knowledge about the performance of the LLM-generated variables, e.g., as stated in the literature [Wang et al., 2023, Zhang et al., 2023b].

Assumption 3. The following equation always holds:

$$\mathbb{E}[y|x, v, v_0, v_1] = \mathbb{E}[y|x, v]. \tag{15}$$

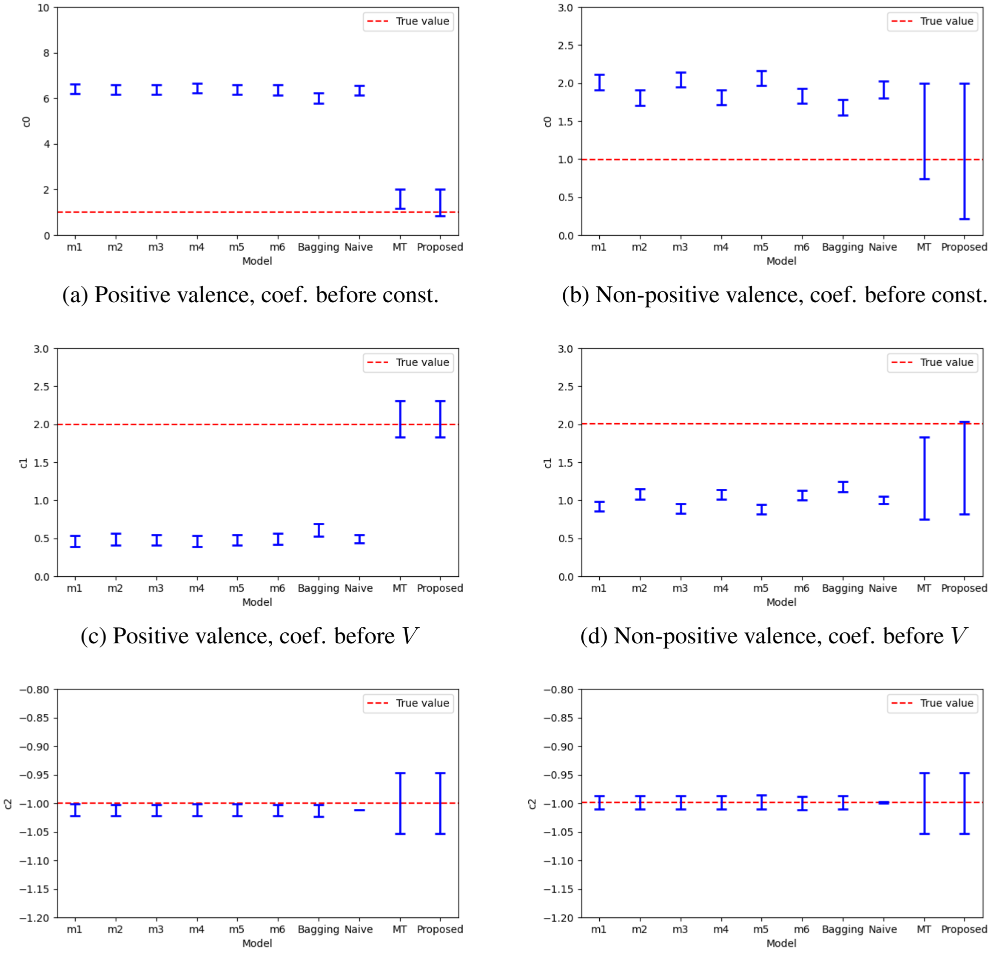
Assumption 3 is identical to Assumption MI in Manski and Tamer [2002]. This assumption indicates that if one can observe v, then observing the interval $[v_0, v_1]$ would be superfluous in predicting y.

B Summary statistics for LLMs-generated valence

	True Valence	M1	M2	M3	M4	M5	M6	Bagging
				Positive Valence				
mean	3.37	2.89	2.83	2.88	2.83	2.89	2.83	2.86
std	0.48	0.54	0.52	0.55	0.52	0.55	0.53	0.47
Accuracy		0.56	0.50	0.56	0.51	0.56	0.51	0.51
N				2,138				
				Non-positive Valence				
mean	1.12	1.32	1.31	1.32	1.31	1.31	1.31	1.31
std	0.73	0.74	0.69	0.74	0.69	0.75	0.70	0.66
Accuracy		0.49	0.52	0.48	0.52	0.49	0.52	0.53
N				2,862				
Overall Coverage				63%				

Table 1: Summary statistics for LLMs-generated valence

C Results for coefficient estimators



(a) Positive valence, coef. before const.

(b) Non-positive valence, coef. before const.

(c) Positive valence, coef. before V

(d) Non-positive valence, coef. before V

(e) Positive valence, coef. before X

(f) Non-positive valence, coef. before X

Figure 1: Results for coefficient estimators