Advances in Reliability, Safety and Security, Part 3 – Kołowrocki, Dąbrowska (eds.)
 ISBN 978-83-68136-15-9 (printed), ISBN 978-83-68136-02-9 (electronic)

Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

When Are Customers Willing To Pay For Better Reliability And How Much: Empirical Study On E-Commerce Observational Data Through Causal Inference And Natural Language Processing

Adam Younsi, Jean Meunier-Pion, Zhiguo Zeng

Chair of Risk and Resilience of Complex Systems, Laboratoire Genie Industriel, Centralesupélec, Université Paris-Saclay,Gif-sur-Yvette, France

Abstract

Product reliability is widely acknowledged as a crucial aspect for both enterprises and customers. However, solely investing in improving product reliability does not guarantee enhanced sales performance. Various factors, including reliability, price, brand, and customer sentiment, contribute to a product's success, with intricate interactions among them. This paper introduces a statistical model leveraging observational data from e-commerce websites to explore these interactions and their impact on sales. Our study employs explanatory variables such as reliability, price, brand, and customer sentiment, with product sales as the dependent variable. Popularity and sales figures are approximated using the number of reviews per unit of time in the product life cycle. Reliability and customer sentiment are extracted from customer reviews through two natural language processing (NLP) models. Recognizing the limitations of observational data, where statistical association may differ from causality, we apply the propensity score matching approach to estimate the average treatment effect of each explanatory variable. Causality is crucial for decision-makers seeking to understand why a product succeeds or fails. Our approach addresses this by offering insights into the factors influencing product popularity. The analysis reveals that 'customer sentiment' is the most significant factor, followed by 'reliability' and 'price.' In contrast, a comparable study, neglecting causal effects, prioritizes 'reliability,' followed by 'customer sentiment' and 'price.' The results underscore the importance of considering causal inference, as our model corrects biases introduced by confounders in observational data, aiding decision-makers in prioritizing features to enhance a product's popularity.

Keywords: causal inference, reliability, sentimental analysis, Natural Language Processing, Average Treatment Effect, propensity score matching

1. Introduction

Being able to tell what a company has to change in order to make its products more popular can be very interesting for the company. There are usually many contributing factors, and they interact with one another to impact the popularity of a product or a service (Cheng et al. 2022). There are two major challenges when we try to investigate how these factors impact the product popularity. First, how to collect the data necessary to support estimating the treatment effect of an influencing factor on the product popularity. Ideally, one should design randomized controlled trials (RCT), in which samples are purely randomly selected to be subject to the treatment and observe the response. However, RCTs can be very expensive and sometimes unethical to implement. To solve this problem, we propose a new framework to investigate the causal effects of different contributing factors to product popularity through observational data from e-commerce websites. A web scraping scheme is established to collect the needed data. As some explanatory variables we are interested in are contained in customer reviews (e.g., reliability, customer sentiment), natural language processing (NLP) and machine learning models are developed to extract this information.

The second challenge is, since we are using observational data, there are inevitably confounders in the data. A confounding variable is a variable that influences the variables studied, thereby distorting the observed

relationship and preventing an accurate representation of the relationships between the variables studied (Pourhoseingholi et al., 2012).

We then established causal relationships between the features (for example reliability is linked by a causal relation with customer's sentiment). We finally estimated the ATE (Average Treatment Effect) thanks to propensity score matching, which makes possible for us to estimate the true importance of a given factor to improving the popularity of the product. When confounders are present, statistical association does not equal to causality. To solve this problem, in this paper, we introduce a causal inference model based on propensity score matching to estimate the average treatment effect of each explanatory variable.

The rest of this paper is organized as follows. In Section 2, we present the methods for data collection and the models developed for extracting information through NLP and estimating the ATE based on causal inference. Then, in Section 3, we present the results obtained and compare the features. Finally, in Section 4, we discuss problems related to our method before concluding our work in Section 5.

2. Methods

To investigate the impacts of the main contributing factors on the popularity of a product, we follow the procedures defined in Figure 1. First, we collect the needed data through web scraping (Section 2.1). Since some of the explanatory variables are only implicitly expressed in the reviews from customers, natural language processing and machine learning models are developed to extract this information and estimate the values of the related explanatory variables (see Section 2.2 for details). A propensity score matching based method is presented in Section 2.4 for estimating the average treatment effect (ATE) from the collected observational data.

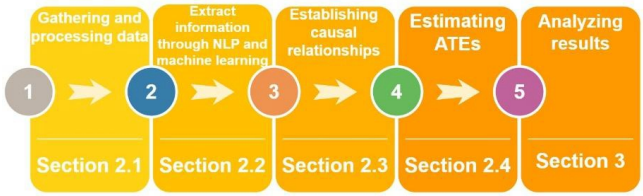

Fig. 1. Flowchart of our approach.

2.1. Data collection using web scraping

The objective of this study is to collect data and investigate the main contributing factors and their contributions to the popularity of a product. As we do not directly have access to the data from the manufacturing company, it is in general difficult to collect data by designing RCTs. Alternatively, we collect open data from e-commence websites through web scraping and use this observational data for our research purposes.

The explanatory variables considered in our model are listed in Table 1, along with their collection method. The response variable is the popularity of the product. In particular, the response variable, the popularity of a product, is approximated through the number of reviews received per day, as shown in Table 1. To do this, we calculated the number of days since the product went on sale. We then divided the number of reviews received by the product by the number of days, finally giving us the number of reviews received per day, which we believe represents the product's popularity. The reason for making this approximation is that the e-commerce website we used for data collection does not directly share the sales number for each product.

In order to extract these variables from a e-commerce website (Amazon), we first developed a scrapping algorithm based on the BeautifulSoup module (https://www.crummy.com/software/BeautifulSoup/) in Python. It made possible for us to fill in a csv file all the reviews, the brand, the operating system, the technical features (such as processor, ram, the size of the screen), the price, the number of stars and the release date of the laptop. A part of the dataset is shown in Figure 2. For 23 laptops, we managed to obtain a total of 1543 lines like those shown in Figure 2.

Text	Stars	Price	Number_reviews	Popularity	Brand	Size	Processor	Color	OS	RAM
This laptop isn't perfect, it's less than two hundred quid with full fat windows.It is however pretty damned good at the price point, it's a MacBook	5.0 out	179	549	0.3553398058	iSTYLE	10.1 In	Atom Z8350	White	Window	2 GB
This laptop was a waste of money! It wouldn't switch on and when I arranged for it to be collected for return nobody turned up on the	1.0 out	179	549	0.3553398058	iSTYLE	10.1 In	Atom Z8350	White	Window	2 GB
With the school being closed and lessons moving on line we needed a lap top quick so that we could continue to work at home	5.0 out	179	549	0.3553398058	iSTYLE	10.1 In	Atom Z8350	White	Window	2 GB

Fig. 2. A part of our dataset.

Table 1. Important variables and the data collection scheme

Variable name	Type	Meaning	Collection method
Reliability	Explanatory	Reliability of the product	Develop NLP and machine learning model to extract from customer reviews.
Price	Explanatory	Price of the product	Directly scrape from e-commerce website.
Customer sentiment	Explanatory	How customers feel about the product	Develop NLP and machine learning model to extract from customer reviews.
Brand	Explanatory	Brand recognition	Scrape from an e-commerce site and then compare it to a data set of well-known brands.
Popularity	Response	Popularity of the product	Approximate it by the number of reviews over the product life cycle.

2.2. Information extraction through NLP and machine learning

As shown in Table 1, the variables reliability and customer sentiment cannot be directly extracted from the scraped data. These two variables need to be inferred from the information conveyed in customer reviews. In this section, we show how to develop machine learning models based on NLP to extract information from customer review data and estimate these two variables.

Once we collected all the reviews in csv files (each laptop has its own csv file with all the features we extracted). We noticed that the number of stars people gave a product did not really represent how they felt about it, whether they were satisfied or not.

Therefore, we needed to know how consumers felt about the product. This is why we used the BERT language model (which stands for Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). Unlike previous language models, BERT considers bidirectional context during training, meaning it analyzes the context of words in a sentence by taking into account both left and right surrounding words. The BERT model is pre-trained on a large corpus of text (from Wikipedia and Bookcorpus). We then have to fine-tune the model (Sun et al., 2019). Fine-tuning a pre-trained model requires adjusting or adapting the structure/parameters of the latter for a specific task or dataset. The fine-tuning is done by adding an extra layer to the pre-trained layers of the BERT model, as shown in the Figure 3.

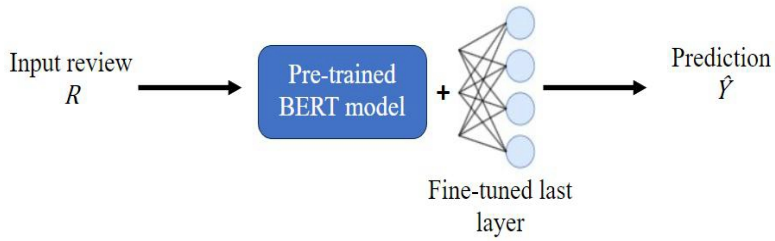

Fig. 3. Fine-tuning BERT.

We then train the whole model to perform a particular task, which will lead the layers from the pre-trained network to slightly modify their hyperparameters. In contrast, the layer we have added will radically alter its parameters since it has been trained for the first time (Merchant et al., 2020). In other words, the last layers are the most task-specific and, therefore are the most modified during the fine-tuning process, whereas the first layers change only slightly.

For example, in our case, we want to classify reviews into two classes, i.e., positive sentiment or negative sentiment. Therefore, we provided the model with a dataset containing reviews labeled as positive/negative (Kotzias et. al., 2015), this dataset contains sentences related to electronics labeled with a positive or negative sentiment, for sentiment analysis. The model then adjusts its weights and parameters to better fit this task. This enabled us to find out, for each review, whether the consumer had positive or negative feelings about the product.

We worked in the same way to find out whether or not a consumer complains about a failure of the product. In fact, we fine-tuned another BERT language model with a dataset that we labeled manually (Meunier-Pion et al., 2021), in which a total number of 2415 reviews were labelled as containing failure complaints or not. Here is a sample in Figure 4 of the dataset.

	comment	stars	Failure class
3	Just what I wanted	5	
7	Great quality	5	
8	Bought this computer because my Mac laptop is a bit erratic. Works well, lightweight. Can fit in purse.	5	
11	The computer stopped working. Will not power up. I am returning it.	1	IF

Fig. 4. A sample of the data set, IF means that the consumer has reported a failure.

Finally, the trained model is used to identify if a review contains a failure description. Then, the reliability of the product is estimated by the fraction of reviews that do not report failure over the total number of reviews, as shown in (1).

$$Reliability = \frac{N_{reliability}}{T_n} \times 100\,, \tag{1}$$

where $N_{reliability}$ is the number of reviews without a report of failure and T_n is the total number of reviews.

2.3. Establishing causal relationships

Now that we have put together all the variables we need, it is time to establish how one relates to the other. To do this, we have used our prior knowledge to define the potential interactions among the variables (Constantinou et al., 2023), as a causal graph shown in Figure 5. For example, price, brand directly affect popularity and that reliability affects how people feel about a product and popularity. Reliability is therefore a confounding factor.

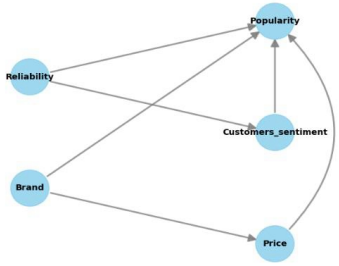

Fig. 5. One of our causality graphs.

In the following analysis, we chose one of the explanatory variables as treatment, and the others as covariates. Covariates are variables that are measured and included in the analysis to account for potential sources of variability or to control for their effects, as shown in Figure 6. For example, in a study investigating the effect of a new drug (the primary treatment) on blood pressure (the outcome), age, gender, and baseline blood pressure may be considered as covariates. Including these covariates in the analysis allows for a more accurate estimation of the drug's causal effect by controlling for potential confounding factors (Morgan, 2018), which occurs when a third variable (confounder) is associated with both the exposure and the outcome, leading to a spurious association between the treatment and the outcome. Confounding can introduce bias and distort the estimation of the true causal effect.

The treatment refers to an intervention, condition, or variable deliberately manipulated or observed to assess its impact on an outcome. In other words, to determine whether or not there is a causal relationship between the variable referred to as treatment and the outcome. In our case, the treatment will be price, reliability or consumer sentiments, and we will observe how these "treatments" affect the product's popularity referred to as the outcome.

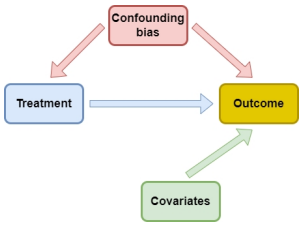

Fig. 6. Example.

2.4. Estimating the ATE through propensity score matching

The term Average Treatment Effect (ATE) is a statistical concept used in the context of causal inference. It quantifies the average difference in outcomes between a group of individuals receiving a particular treatment or intervention and a comparable group that does not. ATE helps to measure the causal effect of a treatment or intervention by comparing the average outcomes between these two groups. It is calculated as follows for an individual (Kuang et al., 2020):

$$ATE = \mathbb{E}(Y^{Treated}) - \mathbb{E}(Y^{Control}), \quad (2)$$

where Y^{Treated} is the value of the outcome if the individual received the treatment whereas Y^{Control} is the value of the outcome if the individual did not receive the treatment and $\mathbb{E}$ the expectancy. In this study, the outcome is defined as the popularity of the product and in each trail, we chose the treatment to be one of the explanatory variables in Table 1 that want to know its effects on the outcome.

In order to adjust for the effect of confounder and estimate the ATE, we decided to use the Dowhy module (https://www.pywhy.org/dowhy/v0.11.1/) in Python. The package has built-in supports for most state-of-the-art approaches for ATE estimation, including:

- Linear regression (Gomila, 2021);
- Propensity score stratification (Austin, 2011);
- Propensity score matching (Austin, 2011);

- Inverse propensity score weighting (Austin, 2011);
- Instrumental variable (Aronow et al., 2017);
- Regression discontinuity (Melly et al., 2020).

Here, we chose propensity score matching because it helps address confounding and reducing possible bias while estimating the ATE by creating comparable groups with similar distributions of covariates. In addition, this method applies in the case of an observational study that we are carrying out.

Propensity score matching (Austin, 2011) involves estimating the probability of receiving the treatment (this is the propensity score) and then matching treated individuals with similar untreated individuals based on this score, as illustrated in Figure 7 The two samples with similar propensity score are regarded as similar in terms of other covariates and only different in terms of the treatment. Therefore, the ATE is then estimated by calculating the difference in the outcomes from a matched pair of two experiments. The main advantage is that it reduces the selection bias inherent in observational studies (we are conducting an observational study here).

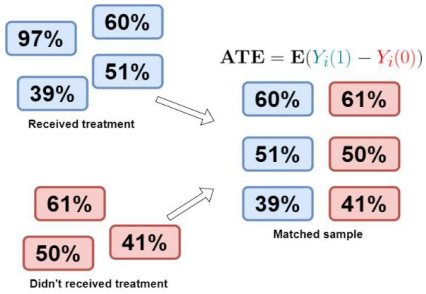

Fig. 7. Propensity Score Matching.

The propensity score, which is defined as the probability of receiving the treatment given the covariates (Abdia et al., 2017):

$$\boldsymbol{e_i} = P(T_i = 1|X_i), \tag{3}$$

where $\boldsymbol{e_i}$ is the propensity score for an individual i, T_i is the treatment variable and X_i is the covariate variables.

Conditioning on this probability can objectively estimate the average treatment effect (Valojerdi et al., 2018).

The propensity score can be estimated based on the following steps (Harris et al., 2019):

- *Define Covariates X*: Identify a set of observed covariates X associated with treatment assignment and outcome. It can be the brand, the reliability, the price and so on. We could include all measured covariates in the propensity score model; this is the simplest approach, and it may enhance the precision of the estimates (Emsley et al., 2008). However, other authors have performed simulations to illustrate that covariates related to the outcome are required for obtaining the least biased estimates of treatment effect (Brookhart et al., 2006). We decided to include all the covariates we have measured.
- *Train a logistic regression model to predict the propensity score*:
  - Formulate a logistic regression model with treatment T as the dependent variable and covariates X as independent variables: $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$
  - Estimate coefficients $(\beta_0, \beta_1, \ldots, \beta_k)$ using maximum likelihood estimation.

- For each observation, calculate the predicted propensity score of receiving the treatment $\widehat{\pi_i}$:

$$\widehat{\pi_i} = \frac{1}{1 + e^{-(\widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \cdots + \widehat{\beta_k} X_{ki})}}$$

- *Check overlap*: Ensure sufficient overlap in the distribution of estimated propensity scores between treatment and control groups.

Once the propensity scores are calculated, the ATE (Kuang et al., 2020) can be estimated by

$$ATE = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left(Y_i^{Treated} - Y_i^{Control}\right), \tag{4}$$

where:

- N is the number of matched pairs;
- $Y_i^{Treated}$ denotes the outcome for the treated unit i;
- $Y_i^{Control}$ denotes the outcome for the matched control unit i.

This formula shows that the ATE estimate only considers matched pairs.

3. Results

In this section, we present and discuss the main results from the analysis. We decided to choose a laptop as the product to study. We collected data on 23 laptops on Amazon. We then drew up a ranking of the features that a seller should consider to improve laptop sales.

3.1. Results for several features

We chose different variables of interest for which we wanted to measure their importance on the popularity of the product. We considered price, product reliability and customer sentiment as treatments (the last two variables were obtained from comments on the product's website). For simplicity of analysis, we assume that the treatment is binary, i.e., either receiving or not receiving the treatment. We use a threshold-based approach to define each treatment. For example, if more than 70% of the comments did not mention a system failure, we consider the system to be reliable.

Table 2. Decision threshold.

Variable	Threshold	Criteria
Consumer's sentiment	70%	1 if value ≥ 70%
Reliability	70%	1 if value ≥ 70%
Price	€350	1 if value ≤ €350

The result of the analysis is presented in Table 3 which ranks each of these variables of interest (or treatment) according to the impact they have on a product's popularity. For example, the model predicts that making the product reliable will increase its popularity by 76%.

Table 3. ATE calculated from the developed model.

Variable	Increase of popularity
Consumer's sentiment	88%
Reliability	76%
Price	14%

From the results, if a seller wants his product to be more popular, then he has to make sure that people's opinions of his product are positive. Improving reliability can contribute to product popularity but the product popularity is less sensitive to the reduced price.

3.2. Without considering causality

In this section, we compare the results from causal inference to a benchmark model that does not consider causality. In this simple model, the impact on product popularity can be evaluated directly from data, using the following formula:

$$\text{Increase in popularity} = \frac{\left|Avg_{reliable} - Avg_{not\_reliable}\right|}{Avg_{reliable}} \times 100\%. \quad (5)$$

where $\text{Avg}_{\text{reliable}}$ and $\text{Avg}_{\text{not\_reliable}}$ are respectively the average popularity score of reliable products (the ones that receive treatment) and the average popularity score of the not-reliable products (the ones that do not receive treatment). The effect on popularity calculated from this simple model is summarized in Table 4.

Table 4. Impact on popularity without considering causality.

Variable	Increase of popularity
Reliability	97%
Consumer's sentiment	23%
Price	5%

It can be seen from Table 4 that the treatment effect of reliability on product popularity will increase to 97%, which is much higher than the 76% previously estimated. This can be explained by the fact that when we estimated the increase in popularity, which ultimately led to 76%, we considered covariates such as price. The price can be higher when the product is more reliable, since it should use more robust materials. Therefore, if the increase in reliability also leads to an increase in price, the increase in popularity may be less than the naive result in which price is not considered.

It should be noted that not only has the treatment effect on product popularity of each treatment changed, but also the general order of importance of the variables has changed. As decision-makers rely on the ranking to decide the most important way to improve the popularity, it is clear that without properly considering confounding effect in the observational data through the causal inference model, the decision-makers might be misled by the results of the analysis.

4. Discussions

4.1. Issues with propensity score

In order to estimate the ATE, we estimated the propensity score. However, using the propensity score raises some issues. A first problem may arise when the sample size is small. Indeed, propensity score matching requires a sufficient number of individuals to be matched in order to produce accurate estimates of treatment effects. If the sample size is very small, it may not be possible to find enough matched individuals, which can lead to bias in the estimates of treatment effects.

Another problem is that applying propensity score matching relies on several assumptions. One of these assumptions is that all covariates that are related to both the outcome and the treatment are observed and included in the propensity score model. Many authors (Austin, 2011) highlighted that this is a strong assumption that is difficult to validate. Another major assumption of propensity score is the Stable Unit Treatment Value Assumption (SUTVA). This assumption says that the treatment effect for one individual is not affected by the treatment status of another.

We could also ask ourselves, which covariates should we include in a logistic regression model for estimating the propensity scores? Many authors have explored this question of variable selection. A few authors say that including all measured covariates in the propensity score model is the simplest approach and enhances the precision of the estimates. Other authors have performed simulations to illustrate that covariates related to the outcome is required for obtaining the least biased estimates of treatment effect (Brookhart et al., 2006).

5. Summary and conclusions

In this paper, we meticulously gathered and processed data using web scraping techniques. NLP and machine learning models are developed to extract useful information regarding product features like reliability and customer sentiment. Then, a causal inference model is developed based on propensity score matching to estimate the true treatment effect of different influencing factors on improving the popularity and sales of the product. We developed robust causal inference models by employing causal graphs and leveraging the Dowhy module in Python. The culmination of our efforts resulted in a systematic ranking of extracted features, offering sellers strategic insights into enhancing their product's popularity.

Remarkably, our study yielded distinct rankings contingent on considering causal relationships. The significance of incorporating causal inference became evident as it impacted the prioritization of features crucial for a product's success. Notably, 'customer sentiment' emerged as the most influential feature for enhancing product popularity.

The findings underscore the importance of considering confounding bias and covariates in understanding outcomes. This study provides a practical framework for sellers to optimize their products and emphasizes the

necessity of nuanced analyses that account for causal relationships in unravelling the intricacies of product popularity.

Acknowledgements

The resarch of Jean Meunier-Pion and Zhiguo Zeng are partially supported by Chaire on Risk and Resilience of Complex Systems (chaire EDF, Orange and SNCF). The research of Zhiguo Zeng is supported by ANR under grant number ANR-22-CE10-0004.

References

Abdia, Y., Kulasekera, K.B., Datta, S., Boakye, M., Kong, M. 2017, Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. Biom. J. 59, 967-985.

Altman, N., Krzywinski, M. 2015. Association, correlation and causation. Nat Methods 12, 899–900.

Aronow PM, Carnegie A. 2017. Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable. Political Analysis.

Austin PC. 2011. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, St ̈Urmer T. 2006. Variable selection for propensity score models.

Cheng, L., Guo, R., Candan, K., Liu, H. 2022. Effects of Multi-Aspect Online Reviews with Unobserved Confounders: Estimation and Implication. Proceedings of the International AAAI Conference on Web and Social Media 16(1), 67-78.

Constantinou, A.C., Guo, Z., Kitson, N.K. 2023. The impact of prior knowledge on causal structure learning. Knowl Inf Syst 65, 3385–3434

Devlin J. , Chang M-W, Lee K. , Toutanova K. N. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Emsley R, Lunt M, Pickles A, Dunn G. 2008. Implementing double-robust estimators of causal effects.

Gomila, R. 2021. Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. Journal of Experimental Psychology: General, 150(4), 700–709.

Harris, H., Horst, S. J., 2019. A Brief Guide to Decisions at Each Step of the Propensity Score Matching Process. Practical Assessment, Research, and Evaluation: Vol. 21, Article 4.

Kotzias D., Denil M., de Freitas N., and Smyth P. 2015. From Group to Individual Labels Using Deep Features. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). Association for Computing Machinery, New York, NY, USA, 597–606.

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., Jiang, Z., 2020. Causal Inference. Engineering, Volume 6, Issue 3, 253-263.

Meunier-Pion, J., Zeng, Z. and Liu, J., 2021. Big Data Analytics for Reputational Reliability Assessment Using Customer Review Data. In 31st European Safety and Reliability Conference, ESREL 2021, 2336-2343.

Melly, B., Lalive, R. 2020. Regression Discontinuity Design. Zimmermann, K.F. (eds) Handbook of Labor, Human Resources and Population Economics.

Merchant, A., Rahimtoroghi, E., Pavlick, E., Tenney, I. 2020. What happens to BERT embeddings during fine-tuning?

Morgan, C.J. 2018. Reducing bias using propensity score matching. J. Nucl. Cardiol. 25, 404–406

Pourhoseingholi MA, Baghestani AR, Vahedi M. 2012. How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench. Spring 5(2), 79-83.

Sun, C., Qiu, X., Xu, Y., Huang, X. 2019. How to fine-tune bert for text classification? Chinese computational linguistics: 18th China national conference, 194-206. Springer International Publishing.

Valojerdi, E., Janani L. 2018. A brief guide to propensity score analysis.