Helpfulness Prediction for VR Application Reviews: Exploring Topic Signals for Causal Inference

Meng Zhang*
School of Management,
Hefei University of Technology

Yang Qian†
School of Management,
Hefei University of Technology

Yuanchun Jiang‡
School of Management,
Hefei University of Technology

Yuyang Wang§
Hong Kong University of
Science and Technology (Guangzhou)

Yezheng Liu¶
School of Management,
Hefei University of Technology

Abstract

Recently, with the development of stereo display and 3D graphics technology, virtual environments and applications are growing rapidly, e.g., video conferencing, and virtual reality (VR) applications. These are several E-commerce platforms that are designed for the transactions of VR applications. Consumers on these platforms can comment on a VR application after purchase. In this paper, we attempt to predict the helpfulness of these VR application reviews. To this end, we propose a topic-causal regression model that explores the influence of topic features in VR application reviews and numerical information on helpfulness. Specifically, we first apply the most classic topic model, Latent Dirichlet Allocation, to extract the topic signals in VR application reviews. Then, we construct a topic regression for causal inference. We perform extensive experiments on a real-world dataset collected from Oculus. The experimental results demonstrate that our model can estimate regression weights for topic factors and analyze their influence on the helpfulness of VR application reviews.

Keywords: Topic Model; Causal Inference; Two-stage; VR Application; Review Helpfulness

Index Terms: Human computer interaction (HCI)-HCI design and evaluation methods-User studies;Human-centered computing-Collaborative and social computing-Collaborative and social computing design and evaluation methods-Social network analysis

1 Introduction

Virtual Reality (VR) is an important interaction technology in virtual 3D worlds. Due to its entertainment and immersion for consumers, more and more enterprises participate in the development of VR applications, e.g., Facebook. According to the latest data from Harvard Business Review[1], the VR applications like Fortnite and Roblox have attracted almost 400 million consumers, and others like Decentraland and the Sandbox are increasing rapidly. In this context, some platforms are designed for the transactions of VR applications, e.g., Oculus[2]. these e-commerce platforms allow consumers to choose VR applications according to their interests. In addition, consumers will post product-related reviews based on their personal experiences. These reviews play an important role for other consumers to purchase VR applications. Due to the growing amount of reviews, it is critical to find the helpfulness of reviews and recommend them to consumers who intend to purchase the VR applications. This paper aims to identify helpful reviews by exploring textual factors that contribute to the helpfulness of online reviews in VR scenarios.

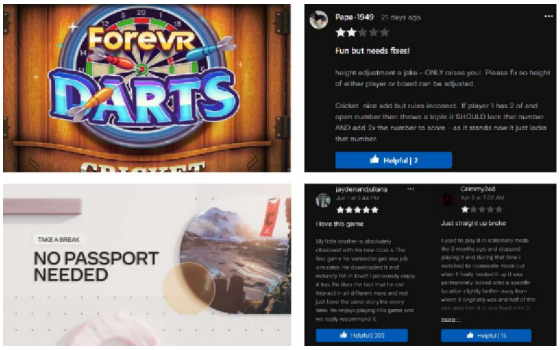

Figure 1: Example of Reviews of VR applications

Recent studies on review helpfulness prediction can be divided into two categories: empirical analysis and machine learning methods. The empirical analysis explores what new factors will affect the usefulness of comments. For example, Kashyap et al. [1] proposed a review usefulness measurement model based on review features, e.g., review valence, review length, and feature-wise information. They used orthogonal design to build questionnaires that contained example reviews with different features. Karimi and Wang [2] investigated the potential effects of the reviewer profile image, and the reviewer name on review helpfulness. Choi and Leon [3] adopted factors from three dimensions: source factors, review factors, and context factors to examine the causes of online review helpfulness. However, these empirical-based studies mostly focus on the acquisition of review usefulness and analysis of the influencing variables from different perspectives. They depend on a large number of observations, experiments, and investigations of existing data, then to summarize the latent variables that may affect the usefulness of reviews. Machine learning methods for helpfulness prediction focus on how to construct features or how to capture better representations. Chen et al. [4] proposed a multi-domain Gated Convolutional Neural Network (CNN) to enrich the word representations by integrating multi-granularity information. In addition, a set of handcrafted features are often constructed by machine learning methods, e.g., semantic features [5], aspect-based [6], and argument-based features [7]. Although these machine learning methods yield better

*e-mail: zhangzm981@mail.hfut.edu.cn
†e-mail: soberqian@hfut.edu.cn (Corresponding author)
‡e-mail: ycjiang@hfut.edu.cn
§e-mail: yuyangwang@ust.hk (Corresponding author)
¶e-mail: liuyezheng@hfut.edu.cn

[1]https://hbr.org/2022/07/exploring-the-metaverse
[2]https://www.oculus.com/

2771-1110/22/$31.00 ©2022 IEEE
DOI 10.1109/ISMAR-Adjunct57072.2022.00014

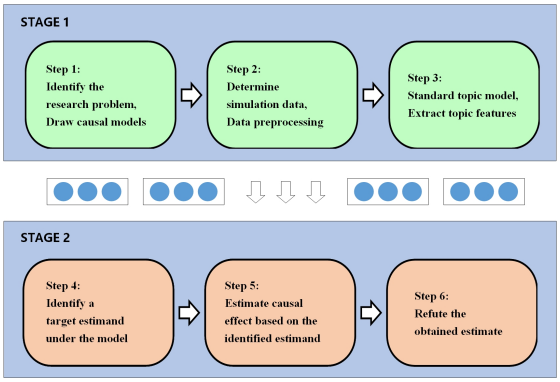

Figure 2: An inference step of the topic-causal regression model

results, they ignore the causal relationship among latent variables (or features).

VR applications are different from products on traditional e-commerce websites, e.g., electronic goods and clothing. The characteristics of traditional e-commerce products are more reflected in the quality and price, while VR applications are experiential virtual products with several unique characteristics, i.e., immersion and sickness. Therefore, to better explore the impact of potential unique features of VR applications on the usefulness of reviews, this paper aims to use online consumer reviews to identify potential and unique features for VR applications and take into account these variables such as textual topic features and digital confounding to predict the review usefulness. This has not been explored yet in the previous studies. Fig. 1 shows a sample of the VR applications. In the context of this research, consumers share their reviews on different aspects of the VR applications they purchased on the platform. This allows us to capture potential textual features in VR application reviews.

To obtain the interpretable features in online reviews, we use a topic model technique to identify the potential signals that may affect the usefulness of reviews in the VR context. With the rapid growth of user-generated content (UGC), topic models have become prevailing tools to explore and extract insights from textual data [8]. This paper applies the most widely used topic modeling technique, Latent Dirichlet Allocation (LDA) [9], to discover specific features for VR applications. Compared with other text analysis methods, LDA can analyze the text without human intervention. To assess the impact of different topic features on the usefulness of reviews, we select a DoWhy framework [10] for causal inference. This framework can separate the identification of causality from the estimation of correlation, which makes it capable of handling complex causality. Fig. 2 shows the model framework. In the first stage, the proposed framework uses a standard topic model for topic feature extraction. In the second stage, we use causal inference methods to predict review usefulness by combining the estimated topic features information of each VR application review with confounders.

The main contributions of this paper can be summarized in the following three aspects:

- To the best of our knowledge, we are the first to propose a model framework that fuses topic features, numeric features, and numeric confounders to predict review usefulness in the VR context. The proposed method is a two-stage framework, namely the topic-causal regression model.
- We combine the topic model's advantage of extracting topics with causal inference to evaluate the causal relationship between causal variables.
- We collect the dataset about the VR application from Oculus.

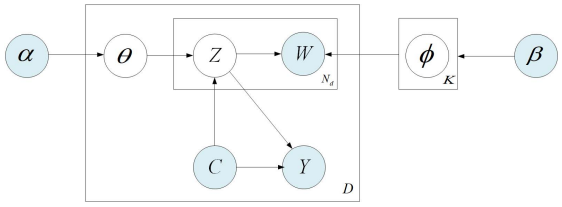

Figure 3: Topic-causal regression model

The experimental results help people understand the extent to which different topics affect the usefulness of reviews when controlling for intervention variables. And it also deepens VR product managers to judge the usefulness of online reviews.

2 RELATED WORK

We organize the related literature into the following two aspects:

User behavior in VR scenarios: At present, due to the wide application of VR in the field of marketing, a large number of studies focus on consumer behavior in VR scenarios. Xi and Hamari [11] evaluated the impact of virtual reality and related stimuli on consumer psychology and behavior in a shopping context. Han et al. [12] argued that the consumer flow experiences (e.g., e. telepresence, challenge, and control) positively affect consumers' technology acceptance. Yang et al. [13] proposed a topic model to identify several meaningful features for VR reviews, e.g., challenge, immersive, and sickness. Alzayat and Lee [14] found that a VR retail environment (vs. an online retail website)) positively affects hedonic consumption.

Review helpfulness: Customers often depend on different kinds of online reviews to make the decision when they are going to purchase. The assessment of review helpfulness can help online platforms to present consumers with more useful information. Both academia and industry have paid close attention to the helpfulness prediction for reviews. For example, Du et al. [15] proposed an end-to-end deep learning method via introducing the context clues inferred from reviews' neighbors. Fan et al. [16] proposed an end-to-end deep neural architecture that is fed by both the raw text of its reviews and the metadata of a product to obtain product-aware review representations for helpfulness prediction. By comparison, this paper focuses on helpfulness prediction in the VR context. We attempt to use the topic model to extract the specific topic features related to VR products and then use the causal inference to predict the review usefulness fed by topic features, numerical features, and other confounder factors.

3 MODEL

The topic-causal regression model proposed in this paper is shown in Fig. 3. To describe the model in detail, we give the research problem in Section 3.1. In Section 3.2, the modeling and inference processes for the model are introduced.

3.1 Problem Formulation

Suppose we have a corpus $D = (d_1, d_2, \ldots, d_M)$ of user reviews about VR applications and M is the number of reviews. Each review d_m in the corpus is represented by a word tuple $w_m = (w_{m1}, w_{m2}, \ldots, w_{mN_m})$, where each word $w_{mn} \in \{1, 2, 3, \ldots, V\}$ and V are sets of vocabularies for reviews. Let $z \in [1, K]$ be a topic indicator variable. For each review, we use θ_m to describe the review-topic distribution, and ϕ_k to describe the topic-word distribution. C denotes the confounders and $Y = (y_1, y_2, \cdots, y_M)$ denotes the helpfulness of all reviews. α and β are the Dirichlet priors. T denotes the treatment variables for causal inference. U denotes the unobserved confounders.

3.2 Learning and inference

3.2.1 Topic discovery

We first use the LDA model to discover the hidden semantic information in VR application reviews. The LDA model is a probabilistic generative model that assumes each document is a mixture of topics and each topic is multinomial over a fixed word vocabulary. In the LDA model, the topic distribution θ_{m} is sampled from a Dirichlet distribution with K-dimensional parameters. Based on the topic distribution θ_{m} , we then draw the topic assignment z_{mn} for the word w_{mn} in review d_m using a multinomial distribution. Using the topic assignment z_{mn} , we can draw the word w_{mn} from a multinomial distribution $\phi_{z_{mn}}$. For the topic-word distribution ϕ_k , it is sampled from Dirichlet distribution with V-dimensional parameters. By repeating the above process, we can generate all reviews related to VR applications. The generation process of the LDA can be summarized as follows:

- For each topic k , draw a topic-word distribution $\phi_k \sim Dirichlet(\beta)$
- For each review m ,
  - Draw the topic distribution $\theta_m \sim Dirichlet(\alpha)$,
  - For each review n ,
    - Draw a topic $z_{mn} \sim Multinomial(\theta_m)$,
    - Draw a word $w_{mn} \sim Multinomial(\phi_{z_{mn}})$,

We used the Gibbs sampling method for model inference, and the sampling process is as follows:

$$p(z_{mn} = k \mid w_{mn} = v, w_{m,\neg n}, z_{m,\neg n}) \propto \frac{N_{mk,\neg n} + \alpha}{\sum_{k=1}^{K} N_{mk,\neg n} + \alpha} \frac{N_{kv,\neg n} + \beta}{\sum_{v=1}^{V} N_{kv,\neg n} + \beta} \tag{1}$$

After converging the above sampling steps, the latent parameters and can be estimated using the following formula:

$$\theta_{mk} = \frac{N_{mk} + \alpha}{\sum_{k=1}^{K} N_{mk} + \alpha} \tag{2}$$

$$\phi_{kv} = \frac{N_{kv} + \beta}{\sum_{v=1}^{V} N_{kv} + \beta} \tag{3}$$

N_{mk} represents the number of words that are assigned topic k in review m. N_{kv} represents the number of word v that are assigned topic k. $\neg n$ means to remove the word with subscript n.

3.2.2 Causal Estimation

Causal inference is the process of concluding the causal relationships using the conditions under the results occurring [17]. Several frameworks have been designed for causal inference, including the potential outcome framework (POF) and the Structural Causal Model (SCM) [17]. In this paper, we choose a classical causal model, namely DoWhy, to explore the relationships between topic features and helpfulness for VR applications. DoWhy combines the advantages of the POF and SCM. Specifically, DoWhy contains four key steps for causal analysis: model, identify, estimate, and refute.

- Model: We first identify our research problem which is to explore the relationships between topic features and helpfulness for VR applications. Then we construct an underlying causal graphical model to deal with this problem. In Figure 4, we give the causal graphical model. Specifically, c_1 and c_2 indicates the review rating and the review length, respectively. c_3 denotes the holistic ratings for VR applications. A solid arrow means that changing the value of a variable (e.g., z) may change the distribution of helpfulness y. U denotes the random variables. Thus, based on Figure 4, we can make each causal assumption explicitly.

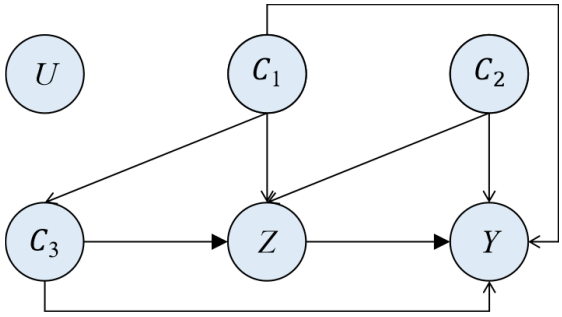

Figure 4: The causal model

- Identify: Based on the causal graph in Fig. 4, the DoWhy can find all possible ways to test a causal effect between the topic features and helpfulness. Specifically, in this step, it applies graph-based criteria and do-calculus to identify expressions that can test the causal effect.
- Estimate: In this step, we use backdoor adjustment to block backdoor paths in the DoWhy framework. Due to the existence of counterfactuals, we can never observe the treatment effect and the untreated effect of the VR review, simultaneously. Therefore, we adopt the EconML and CausalML packages to estimate the average treatment effect (ATE) to measure the causal effect instead. In this process, we choose a linear regression:

$$ATE = E[Y|do(T = 1) - Y|do(T = 0)] \tag{4}$$

- Refute: To verify an effect estimate from a causal estimator, we refute the obtained estimate by implementing three robustness checks: adding random common cause as a confounder, Placebo treatment, and data subset validation.

4 EXPERIMENT

In this section, we present experimental results on a VR application dataset from Oculus. We first describe the dataset in detail. We then use the method described in Section 3.2.1 for topic extraction and the method described in Section 3.2.2 for causal inference. Finally, we give the influence weights of different topics on the helpfulness of VR application reviews.

4.1 Data Descriptions

VR applications, as the representative virtual products, have more unique features than those of traditional e-commerce products, e.g., clothing. We select one of the most representative E-commerce platforms for the transaction of VR applications, namely Oculus. Consumers who have purchased a VR application on Oculus, can make comments on this application. This guarantees the credibility and authenticity of the data sources. We collect nearly 200,000 reviews, review ratings, and other related data from the Oculus platform. We use the following steps to preprocess the data set. First, we convert all reviews to lowercase. Second, to avoid the influence of irrelevant factors, we remove punctuations and frequent words in the reviews, as well as other non-English characters (e.g., URL, Chinese). Finally, we retain reviews containing at least 5 words. To reduce computing overhead, we only select 50,000 reviews in a randomized form as the experimental dataset.
Table 1 shows the descriptive statistics of our final dataset. There are 18.98 words per review on average. We regard the helpfulness prediction task as a classification task that discretizes the number of total helpfulness to intervals to represent the helpfulness levels

Table 1: Data description and statistics

Statistic	Number
Number of reviews	50,000
Mean	18.98
Standard deviation	9.60
Maximum	336

of online reviews. We divide the overall helpfulness of reviews into four categories based on the quantitative distribution: [0, 3], [4, 7], [8, 11], [12, ∞].

4.2 Topic Results

For topic models, there are no fixed rules for the setting of parameter values. However, previous studies [18] have proved that the empirical method can help us to better determine the parameter value. Therefore, we set hyperparameters to $\alpha = 50/K$ and $\beta = 0.1$. We differ in the number of topics to infer the topic model. We use the perplexity score [9] to assess the predictive performance of the topic model. According to the results of the perplexity score, we set the topic model with 30 topics.

Table 2: Topics discovered by LDA

Topic1	Topic6	Topic13	Topic19	Topic22
download	room	ball	world	scare
install	world	table	beautiful	scary
issue	chat	tennis	art	watch
account	talk	hit	show	jump
problem	community	physics	life	door
error	social	throw	visual	wait
version	meet	pinball	environment	fear
purchase	join	bowling	explore	horror
load	multiplayer	sport	interesting	guy
connect	vrchat	racket	scene	face

Due to space constraints, this paper only lists five representative topics in Table 2. From the topic results, we note the LDA model can identify a set of interpretive topics. For example, topic 1 is related to the installation problem of VR applications, and the related semantic words contain ”download,” ”install,” ”issue”, and ”version.” Topic 6 is related to the social functions provided in VR applications, as evidenced by the semantic words ”room,” ”world,” ”chat,” ”talk,” and ”community.” Topic 13 is related to the category of VR applications, including “ball,” “tennis,” “hit” and “pinball.” Topic 19 is related to the VR visual environment. Topic 22 is related to the virtual scene experience of VR applications, as evidenced by the semantic words “scare,” “scary,” “watch,” and “fear.”

4.3 Results of Causal Inference

We use the method introduced in Section 3.2.2 to estimate the causal effect by exploiting the DoWhy framework for causal inference. DoWhy combines the graph model of causal inference with the potential outcome model. Specifically, the expression of a causal effect is identified by the criterion of the graph. Then the hypothesis is modeled by do-integral. Finally, the non-parametric causal effect can be calculated. The experimental results are shown in Table 3.
As shown in Table 3, we obtain an average estimate of -0.27, which denotes that the expected value of review helpfulness decreases by about 27%. At the same time, we test the prediction performance of the model on three evaluation metrics: precision, recall, and f1-score. The results show that our model achieves good prediction.

Table 3: Performance of the causal inference

Estimated:	-0.27	**Accuracy:**	%45.12
category	**precision**	**recall**	**f1-score**
0	0.46	0.44	0.45
1	0.41	0.48	0.44
2	0.60	0.43	0.50
3	0.47	0.63	0.53

In addition, we design three additional experiments to verify the topic-causal regression model's robustness. First, We randomly add a confounder to the model and set its coefficient to zero. If the assumption is correct, the causal effect after adding the random confounder is similar to that of the initial model. And its p-value should be greater than 0.05, revealing the null hypothesis with a coefficient of zero cannot be rejected. For the second experiment, we replace the true treatment variable T with an independent random variable. And the causal effect should be close to zero if the model assumptions are correct. For the third experiment, we randomly select the subset of the dataset. If the assumptions of the model are correct, there should be no significant change in causal effects. Table 4 gives the results of robustness checks. From Table 4, we note that the p-value for the first experiment is 0.46, revealing that adding an independent random variable has no significant impact on the helpfulness prediction for VR application. From the results of the second experiment, we find that its estimated effect is 2*10-3 that close to zero, revealing that we cannot effectively predict the helpfulness of VR applications after replacing with random variables. For the results of the third experiment, we note that the estimated effect does not change significantly when we use a randomly selected subset. In summary, the proposed model helps us to estimate the effect between topic features and the helpfulness of VR applications.

Table 4: Results of Robustness Checks

Method	Estimated effect	p-value
Adding a confounder	-0.27	0.46
Placebo Treatment	2*10-3	0.48
Data Subset Validation	-0.27	0.49

We give the weight of text topic features for the helpfulness of VR application reviews. As shown in Fig. 5, the top five topics have a significant impact on the helpfulness of VR application reviews. These topics include installation (topic 1), price (topic 2), immersion (topic 10), control (topic 16), and visual environment (topic 19). We find that the importance of topic 1 is greater than that of other topics. It reveals many consumers pay more attention to the download and installation of VR applications. The underlying reason behind this is that the installation is the first step for users to experience VR applications.

5 Conclusions

In this paper, we propose a framework that combines topic modeling with causal inference for helpfulness prediction in the VR context. We focus on the joint analysis of topic features, numeric features, and several confounders. To better illustrate the predictive performance of the model, we first use the three evaluation metrics to evaluate the model. To test the robustness of the model, we also design three

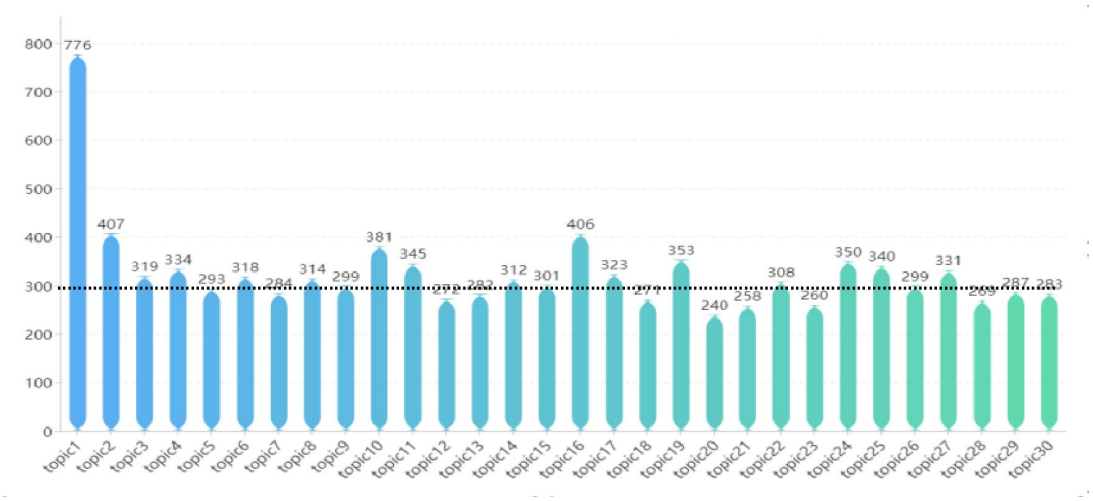

Figure 5: The topic features importance

additional experiments. We construct these experiments on a new dataset from the Oculus platform. The experimental results show that the model can identify a set of meaningful topic features, and the identified topic features have a certain effect on the helpfulness of reviews. Therefore, predicting the helpfulness of reviews should consider the topic features related to VR context, e.g., immersive, control, and visual environment. In the future, we can design a unified modeling framework that incorporates both textual data and numerical data for modeling the helpfulness of VR applications, and jointly infer all model parameters.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (72101072, 91846201, 72171071, 71722010), the Postdoctoral Research Foundation of China (2021M690852), the Fundamental Research Funds for the Central Universities (JZ2022HGTB0282, JZ2021HGQB0272) and the National Engineering Laboratory for Big Data Distribution and Exchange Technologies.

References

[1] Rachita Kashyap, Ankit Kesharwani, and Abhilash Ponnam. Measurement of online review helpfulness: A formative measure development and validation. *Electronic Commerce Research*, pages 1–34, 2022.

[2] Sahar Karimi and Fang Wang. Online review helpfulness: Impact of reviewer profile image. *Decision Support Systems*, 96:39–48, 2017.

[3] Hoon S Choi and Steven Leon. An empirical investigation of online review helpfulness: A big data perspective. *Decision Support Systems*, 139:113403, 2020.

[4] Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Sheng Bao. Multi-domain gated cnn for review helpfulness prediction. In *The world wide web conference*, pages 2630–2636, 2019.

[5] Lionel Martin and Pearl Pu. Prediction of helpful reviews using emotions extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[6] Yinfei Yang, Cen Chen, and Forrest Sheng Bao. Aspect-based helpfulness prediction for online product reviews. In *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)*, pages 836–843. IEEE, 2016.

[7] Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. Using argument-based features to predict and analyse review helpfulness. *arXiv preprint arXiv:1707.07279*, 2017.

[8] Yi Yang, Kunpeng Zhang, and Yangyang Fan. sdtm: A supervised bayesian deep topic model for text analytics. *Information Systems Research*, 2022.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[10] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[11] Nannan Xi and Juho Hamari. Shopping in virtual reality: A literature review and future agenda. *Journal of Business Research*, 134:37–58, 2021.

[12] Sang-Lin Han, Myounga An, Jerry J Han, and Jiyoung Lee. Telepresence, time distortion, and consumer traits of virtual reality shopping. *Journal of Business Research*, 118:311–320, 2020.

[13] Yang Qian, YingQiu Xiong, Yuyang Wang, Yuanchun Jiang, Yezheng Liu, and Yidong Chai. Identification of key features for vr applications with vreview: A topic model approach. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 183–188. IEEE, 2022.

[14] Ayman Alzayat and Seung Hwan Mark Lee. Virtual products as an extension of my body: Exploring hedonic and utilitarian shopping value in a virtual reality retail environment. *Journal of Business Research*, 130:348–363, 2021.

[15] Jiahua Du, Jia Rong, Hua Wang, and Yanchun Zhang. Neighbor-aware review helpfulness prediction. *Decision Support Systems*, 148:113581, 2021.

[16] Miao Fan, Chao Feng, Lin Guo, Mingming Sun, and Ping Li. Product-aware helpfulness prediction of online reviews. In *The world wide web conference*, pages 2715–2721, 2019.

[17] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

[18] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235, 2004.