User Response Driven Content Understanding with Causal Inference

Fei Tan[†⋆], Zhi Wei[‡*], Abhishek Pani[§], and Zhenyu Yan[§*]
[†]Yahoo Research
[‡]New Jersey Institute of Technology
[§]Adobe Inc.
fei.tan@verizonmedia.com, zhiwei@njit.edu, {apani, wyan}@adobe.com

***Abstract*—Content understanding with many potential industrial applications, is spurring interest by researchers in many areas in artificial intelligence. We propose to revisit the content understanding problem in digital marketing from three novel perspectives. First, our problem is to explore the way how user experience is delivered with divergent key multimedia elements. Second, we treat understanding as to elucidate their causal implications in driving user responses. Third, we propose to understand content based on observational audience visit logs. To approach this problem, we measure and generate heterogeneous content features and model them as binary, multivalued or continuous genres. Multiple key performance indicators (KPIs) are introduced to quantify user responses. We then develop a flexible and adaptive doubly robust estimator to identify the causality between these features and user responses from observational data. The comprehensive experiments are performed on real-world data sets. We show that the further analysis of the experimental results can shed actionable insights on how to improve KPIs. Our work will benefit content distribution and optimization in digital marketing.**

***Index Terms*—Content Understanding, User Engagement, Digital Marketing, Causal Inference**

I. Introduction

In digital marketing, content can be delivered to the customers through various channels (e.g., websites). Different from the well-defined content (e.g., image), contents (e.g., webpage) in digital marketing may contain heterogeneous multimedia elements [1]. It is intriguing to understand the manner how different parts of contents jointly deliver a user experience. We regard the perceived integration form to be content, which shapes audience experience [2] and further influences their willingness-to-purchase [3]. To this end, we propose a novel KPI-driven content understanding problem. Understanding content well will be a crucial step towards optimizing marketing performance. It's noted that well-defined content understanding conventionally refers to interpreting their literal meaning (e.g., object recognition in an image). Such tagging tasks can be done using machine learning, with labels produced through crowd-sourcing efforts [4]. We emphasize that, for our problem, understanding is driven by user responses, which are instead generated by audience.

A product usually walks through three major phases of interactions with audiences: user reach, engagement, and conversion. For example, a user is probably exposed to an advertisement when he/she submits a query in search engine. Audiences can click them for more details or just dismiss them. If users are directed to channels (e.g., a website) successfully, they will make conversion decisions (e.g., purchase) based on their evolving experiences of the channel. Content design is essential for capturing their fleeting attention, which is the premise of audience retention and the ultimate conversion [5]. The focus of this work is on online user engagement. User response and engagement will be used interchangeably in the sequel. In addition, we are dedicated to elucidating causality, which will make the human intervention more effective than statistical correlations.

There are three major issues. First, content here refers to the overall presentation of key multimedia elements in channels, which is more difficult than the well-studied text and images and relatively understudied. Second, different from objective and literal labels, user engagement is the perceived quality of user experience. Reasonable quantification of user responses is critical for the inference of useful implications. Third, the causality needs to be estimated based on observational user logs. Observational data possess many features including confounders and treatments, which are much challenging compared to randomized experimental data. Furthermore, treatments are multiple and heterogeneous.

To approach this task, we posit that user engagement comes from both usability and presentation of multimedia elements. The latter is also termed aesthetics [6]. Specifically, we first curate two heterogeneous sources of a channel: screenshot images and key elements. Then we extract low-level statistics of screenshot images, which form the primitive and heterogeneous visual features (e.g., visual complexity). Meta information is retrieved from key elements (e.g., hyperlinks). We group these features into usability and aesthetics. Then we model these content features as binary, multivalued, and continuous treatments. We introduce three representative metrics to quantify user engagement. Previous studies showed that aesthetics is correlated with perceived usability and the existence of the halo effect from aesthetic judgment [6]. To address the issue of non-randomness existing in observational data, we develop *Adaptive Doubly Robust* (AdaDR) analysis

⋆This work is done when Fei Tan was a PhD student at New Jersey Institute of Technology and interned at Adobe Inc.
*Corresponding author.

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00168

based on classical Rubin Causal Model [7] to control for confounders and other covariates. We further develop a naive rotation mechanism for the multivariable analysis.

II. Related Work

Prior studies related to several threads of research in this work are content understanding, user engagement, and causal inference.

A. Content Understanding

The primary research goal of content understanding is to identify the involved relationship among contents and interactions between contents and target labels of interest. It has advanced the subareas of artificial intelligence recently such as computer vision [4], language understanding [8] and speech recognition [9]. Our work can be regarded as an extension of the above content understanding. There are also some explorations on the aesthetics as another genre of content such as the visual quality rating of natural images [10] and visual complexity and preference prediction of web pages [2], [11]. Different from these pioneering works in content understanding, user responses in log data instead of crowdsourcing-based labels are utilized to calibrate understanding procedure. Furthermore, our work focuses on causation inference instead of correlation discovery.

B. User Engagement

User engagement often refers to the quality of user experience [12]. There are two broad types of measurement, including subjective and objective. Although the user's subjective (e.g., self-reported and cognitive engagement) experiences are central to engagement measurement, it is unfeasible to assess them (e.g., post-experience questionnaires) at large scale. Fortunately, it is possible to have objectively observable consequences [13] (e.g., online behavior metrics) of subjective experiences, which might serve as indicators of user engagement [12], [14]. Engaging user experiences have been comprehensively explored and the corresponding metrics are defined [5], [14]. In this work, we focus on objective engagement metrics.

Previous studies focus on the visual appeal of aesthetics [10], [11], the correlation between aesthetics and perceived usability [6], and interaction between aesthetics and user experience [15]. However, the principal methodology of some works rests on the correlation-based analysis, where the estimated effects are not actual causal effects. Others are based on experimental data and rely on A/B testing. Our work concentrates on the large-scale causality of aesthetics on user responses from observational data.

C. Causal Inference

Understanding the causality of a treatment or intervention on an individual is a central problem across an array of fields such as health and medicine [16], online advertising [17] and so on. Common frameworks for causal inference are Structural Equation Modeling (SEM) and Rubin Causal Model (RCM). One significant difference rests on the assuming existence of latent confounders. The latter presumes that all confounders can be measured and observed, which is subsumed by SEM [18]. There is a large volume of works under frameworks of either SEM [19] or RCM [7], [20], [21]. The doubly robust estimators have been widely used in causal inference and missing data models [22], [23]. These are asymptotically unbiased when either one of two component models is correctly specified. However, neither one of two working models could be adequately specified in reality.

In this work, we follow RCM to analyze the causality between treatments and user engagement metrics. We then introduce multilayer neural networks to two working models of the doubly robust framework to make model specifications more adaptive and flexible. Furthermore, different from previous works on single or homogeneous treatments with one outcome, we explore heterogeneous treatments with multiple outcomes.

III. Problem Formulation

The proposed problem concerns what would happen to a user response as a result of a hypothesized treatment or intervention on contents. The focus here is on the causal effects instead of pure prediction. We thus formulate it as causal inference in observational data. There are three major parties involved in causality: confounders, treatments and outcomes as illustrated in Fig. 1. T refers to all treatments, and Y covers user response metrics. T can have multiple types, including binary, multivalued, and continuous treatments. $\boldsymbol{X}$ as confounders impact both treatments T and outcomes Y.

Our problem is to estimate the causal effects of each treatment (e.g., T_i) on user responses by controlling for different variables, including confounders $\boldsymbol{X}$ and other treatments excluding T_i from non-randomized observational data.

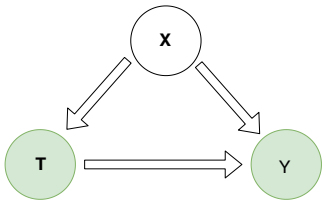

Fig. 1. The relationship illustration for treatments T, confounding factors $\boldsymbol{X}$ and outcomes Y.

IV. Methodology

A. Content Feature

Given a delivering channel, we have two content sources: screenshot images and key elements. For each screenshot at a 1024×768 pixel resolution, we calculate low-level image metrics (e.g., W3C Color[1]) based on a set of algorithms in human-machine interaction [11]. Regarding key elements, we extract meta information accordingly, such as the presence and the

[1]aqua, olive, blue, purple, silver, maroon, gray, yellow, red, black, teal, lime, green, navy, fuchsia, white

number of multimedia compositions. In our causal inference framework, they are all called treatments. For example, what is the average effect of the presence of image on user dwell time conditioned on the same other covariates? We model features as types of binary, multivalued, and continuous treatments. In addition to treatments, we also extract channel categories at different levels as confounding factors such as product details, and support services. Lastly, we gather auxiliary covariates, which only have impacts on outcomes. They are helpful to correct for the potential post-treatment bias. For example, users' network connection conditions could influence the dwell time.

B. User Response

To measure users' depth of engagement with channels, we employ online behavior metrics as a proxy [12]. The rationale is that online behavior metrics can collect data from large-scale users without reliance on user subjectivity. The metrics are *dwell time* (DT), *return frequency* (RF) and *bounce rate* (BR). All metrics are measured at the level of channel unit. They are detailed as follows.

Dwell time measures the average time of a visit to a channel unit, which is defined as $\text{DT} = \frac{\text{total time spent on channel unit}}{\text{total hits on channel unit}}$. It is an activity indicator of page content satisfying the intent of users and reflects the sustaining attention of users. Return frequency captures how frequently users return to a channel unit, which is formally calculated as $\text{RF} = \frac{\text{total hits on channel unit}}{\text{total unique channel unit visitors}}$. It is an important metric concerned with user loyalty. A bounce is a single-unit session[2]. Bounce rate of a channel unit is calculated specifically as $\text{BR} = \frac{\text{total hits in a single-unit session}}{\text{total hits as landing unit}}$. It indicates the impact of landing units on the user experience.

Altogether, different user engagement metrics are introduced to capture different aspects of user responses.

C. Adaptive Doubly Robust

We propose a simple and generic estimator based on the classical doubly robust method. It combines regression model and propensity score approaches such that only one of them needs to be correctly specified to obtain an unbiased effect estimator [20], [22], [23]. However, neither one can be specified correctly for most cases in reality. Thus, our novel estimator needs to address or at least alleviate the model misspecification issue. To facilitate the explanation, we only explore one treatment T here. For a regression model, the treatment-outcome and confounder-outcome association are estimated simultaneously as

$$E(Y|\boldsymbol{X}, T) = \boldsymbol{\beta}\boldsymbol{X} + \gamma T \tag{1}$$

The average treatment effects (ATE) can be estimated as

$$\text{ATE} = E_{x\sim p(x)}[E(Y|x,1) - E(Y|x,0)] = \gamma \tag{2}$$

[2]https://www.optimizesmart.com/two-powerful-ways-to-reduce-bounce-rate/

If this model is misspecified, the estimated ATE will be misleading. Regarding the propensity score, the formula is

$$\begin{aligned}\text{ATE} &= E_{x\sim p(x)}[E(Y|x,1) - E(Y|x,0)] \\ &= \frac{1}{n}\sum_{is.t.t_i=1}\frac{y_i}{\hat{p}(t_i=1|x_i)} - \frac{1}{n}\sum_{is.t.t_i=0}\frac{y_i}{\hat{p}(t_i=0|x_i)}\end{aligned} \tag{3}$$

Propensity score, however, is contingent upon another assumption of the substantial overlap between two groups [24].

To alleviate these issues, we replace component models with adaptive and flexible multilayer neural networks, which are powerful to create mathematical models to approximate complex functions [25]. To handle the limited population overlap, we adopt a simple rule of thumb to discard all samples with estimated propensity scores outside the range [0.1, 0.9], which is a good approximation to the optimal selection rule [26]. We will discuss two scenarios: discrete and continuous treatments.

1) Discrete: We mainly study binary and multivalued cases.

For binary treatments, the propensity score of sample $i(i = 1, \cdots, N)$ is modeled as binary-class neural networks $f_{bps}(\cdot)$ and denoted as $r_i = p(T_i = 1|X_i) = f_{bps}(X_i; \theta_{bps})$. Parameters θ_{bps} can be estimated by minimizing the binary cross-entropy loss $\arg\min_{\theta_{bps}} \mathcal{J}_1 = -\frac{1}{N}\sum_{i=1}^{N} T_i\log(r_i) + (1-T_i)\log(r_i)$. Then, the general forms of treatment and non-treatment responses are derived [20] as $\hat{R}_i(1) = \frac{Y_i(T_i=1)\times T_i}{r_i} - \frac{\hat{Y}_i(T_i=1)\times(T_i-r_i)}{r_i}$ and $\hat{R}_i(0) = \frac{Y_i(T_i=0)\times(1-T_i)}{1-r_i} + \frac{\hat{Y}_i(T_i=0)\times(T_i-r_i)}{1-r_i}$. Here $Y_i(T_i = t)$ is the factually observed outcome with treatment $t = 0, 1$ and $\hat{Y}_i(T_i = t)$ is the potential outcome based on the outcome model $f^t_{bo}(\cdot)$ as $\hat{Y}_i(T_i = t) = \text{E}(Y|T_i = t, X_i) = f^t_{bo}(X_i; \theta^t_{bo})$. Parameters θ^t_{bo} can be estimated by minimizing the mean squared error loss $\arg\min_{\theta^t_{bo}} \mathcal{J}_2 = \frac{1}{|A|}\sum_{i\in A=\{i:T_i=t\}}\left(Y_i - f^t_{bo}(X_i; \theta^t_{bo})\right)^2$. ATE can be derived accordingly as

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}\hat{R}_i(1) - \hat{R}_i(0) \tag{4}$$

For multivalued treatments, the generalized propensity score (GPS) is introduced to extend the binary case [21]. GPS is the conditional probability of assigning subject i to one of M treatment groups $(t = 0, \cdots, M-1)$ given a vector of observed covariates, which is denoted as $r_i(t) = p(T_i = t|X_i) = f_{mps}(t|X_i; \theta_{mps})$, where r_i is modeled as multi-class neural networks $f_{mps}(\cdot)$. Parameters θ_{mps} can be estimated by minimizing the multinomial cross-entropy loss $\arg\min_{\theta_{mps}} \mathcal{J}_3 = -\frac{1}{N}\sum_{t=0}^{M-1}\sum_{i=1}^{N} 1\{T_i = t\}\log\big(f_{mps}\big(t|X_i; \theta_{mps}\big)\big)$. The estimated outcome of subjects under treatment t is $\hat{Y}_i(T_i = t) = \text{E}(Y|T_i = t, X_i) = f^t_{mo}(X_i; \theta^t_{mo})$, where $\hat{Y}_i(T_i = t)$ is the potential outcome based on the outcome model $f^t_{mo}(\cdot)$. Parameters θ^t_{mo} can be estimated by minimizing the mean squared error loss $\arg\min_{\theta^t_{mo}} \mathcal{J}_4 = \frac{1}{|A|}\sum_{i\in A=\{i:T_i=t\}}\left(Y_i - f^t_{mo}(X_i; \theta^t_{mo})\right)^2$. The response of treatment can then be derived as $\hat{R}_i(t) = \frac{Y_i(T_i=t)I(T_i,t) - [I(T_i,t) - r_i(t)]\hat{Y}_i(T_i=t)}{r_i(t)}$, where

TABLE I
USER VISIT LOGS OF A WEBSITE

Time Span	Days	# Observations	# Web pages
7/1/2018-7/31/2018	31	131,579,443	458,999

$\mathrm{I}(T_i, t) = 1$ if $T_i = t$, it is 0 otherwise. ATE between different treatments j and k is

$$\mathrm{ATE}(j,k) = \frac{1}{N}\sum_{i=1}^{N} \hat{R}_i(j) - \hat{R}_i(k) \tag{5}$$

When $M = 2$, the binary treatment effect estimation is a special case of multivalued treatments and Eq. 5 is equivalent to Eq. 4.

2) Continuous: Regarding continuous treatments, we use a flexible parametric approach and assume a normal distribution for the treatment given the covariates [27] $T_i|X_i \sim \mathcal{N}(f_{cps}(X_i;\theta_{cps}), \sigma^2)$, where parameters θ_{cps} and σ can be estimated by $\arg\min_{\theta_{cps}} \mathcal{J}_5 = \frac{1}{N}\sum_{i=1}^{N}\big(T_i - f_{cps}(X_i;\theta_{cps})\big)^2$, $\sigma = \sqrt{\frac{\sum_{i=1}^{N}\big(T_i - f_{cps}(X_i;\theta_{cps})\big)^2}{N-1}}$. The estimation of GPS can be derived as $r_i(T_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{1}{2\sigma^2}\big(T_i - f_{cps}(X_i;\theta_{cps})\big)^2\Big)$. With T_i and r_i, we model the conditional expectation of Y_i as a flexible function of its two arguments $\hat{Y}_i(T_i) = E[Y|T_i, r_i(T_i)] = f_{co}(T_i, r_i(T_i);\theta_{co})$. Parameters θ_{co} can be estimated by ordinary least squares $\arg\min_{\theta_{co}} \mathcal{J}_6 = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i - f_{co}(X_i;\theta_{co})\big)^2$. We then estimate the average potential outcome at treatment level $t \in [t_{min}, t_{max}]$ as

$$\mathrm{ATE}(t) = \frac{1}{N}\sum_{i=1}^{N} f_{co}(t, r_i(t);\theta_{co}) \tag{6}$$

The estimation of the entire dose-response curve can be obtained for each level of treatments of interest.

The overall inference procedure is called *Adaptive Doubly Robust* or AdaDR for short.

D. Treatment Multiplicity

We assume only one treatment of interest in the above causal models for different types. In reality, multiple treatments need to be explored due to the nature of the problem. To this end, we propose a rotation mechanism. Specifically, each treatment of interest is studied by regarding the remaining treatments as part of covariates alternately. This is conditioned on the presumption that no explicit temporal dependency among different treatments exists. This assumption is reasonable for our problem as all treatments are placed simultaneously once a web page is created. The correlation between different treatments does exist, which have common confounding factors. However, they are actually independent conditioned on confounders.

V. Experiments

In this section, we report and analyze experimental results on a private real-world data set of user visit logs.

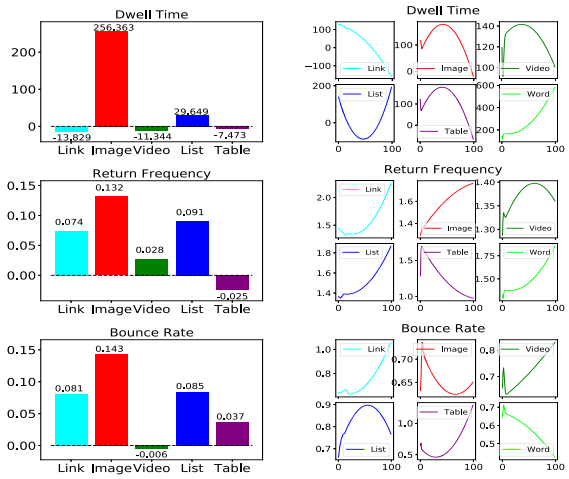

Fig. 2. Average treatment effects of the presence of hyperlink, image, video, list, table on engagement metrics. Responses of number of hyperlink, image, video, list, table, word on engagement metrics.

TABLE II
CONTENT FEATURES EXTRACTED FROM WEB PAGES

Feature	Type	Description
Color	Continuous	The percentage of pixels that are close to one of sixteen colors defined by the W3C.
Leave	Multivalued	Number of leaves calculated by the space-based decomposition (modified from [28]). Binning the number into low, medium and high levels.
Image area	Continuous	Number of leaves that the algorithm identifies as separate images. Several adjacent images are counted as one image area.
Text group	Continuous	Number of horizontal groups of text characters. Each group may represent a word, one-line text, multiple lines of text or a paragraph.
Symmetry	Continuous	Evaluates the symmetrical arrangement of the leaves along the horizontal and vertical axes.
Balance	Continuous	Measures whether the top and bottom, as well as the right and left part of an image have an equal number of leaves, independent of their spatial distribution.
Equilibrium	Continuous	Evaluates whether the quadtree's leaves mainly center around an image's midpoint.
Hyperlink, Image, Video, List, Table	Binary, Continuous	Number of hyperlinks, image, video, list, table and their corresponding binary presence indicator.
Page category	Confounders	Hierarchical categories of web pages.
Continent	Covariate	North America (NA), Europe (EU), Asia (AS), None, South America (SA), Oceania (OC), Africa (AF).
Preferred language	Covariate	English, None, Chinese, German, Portuguese, Spanish, French, Russian, Arabic.
Browser	Covariate	Google, Microsoft, Mozilla, Apple, Yandex, None, QQ Browser.
Carrier	Covariate	Non-mobile, Mobile.
Connect type	Covariate	None, Lan, Modem, Offline.
Operating system	Covariate	Windows, Android, Linux, OS X, iOS, None.

A. Real-world Dataset

The dataset used in this work is user visit logs of a commercial website through the whole of July 2018. A web page is a channel unit here. The basic statistics are detailed in Table I. One page hit is an observation with details such as page hit and session visit identifiers and user responses, which are collected and generated by our data analytics server. To maintain business confidentiality, we cannot divulge the volume of audiences. Both treatments and confounders are generated based on Section IV-A and reported in Table II.

B. Primary Results

1) HTML Elements: We first analyze the causal impacts of basic HTML elements including hyperlinks, images, videos, lists, and tables as shown in the left side of Fig. 2. The results clearly show that the presence of different elements has varied impacts on an array of metrics. To facilitate the understanding of feature effects on different measures, we

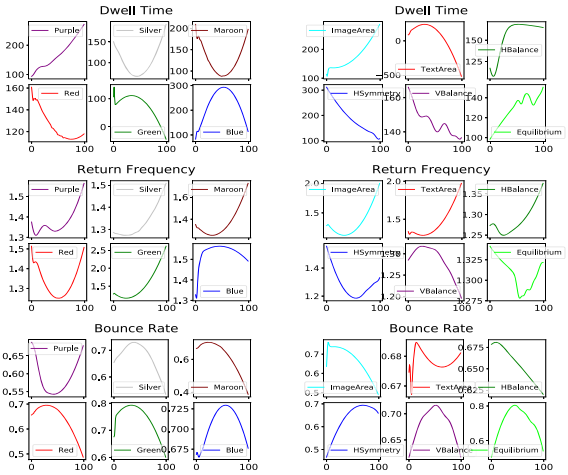

Fig. 3. Responses of color distribution on engagement metrics. Responses of number of image area, text area, balance, symmetry, equilibrium on engagement metrics.

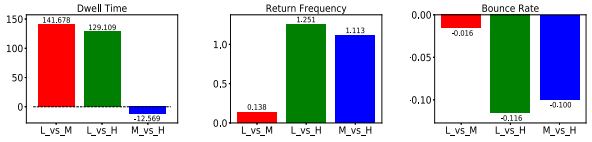

Fig. 4. Average treatment effect of number of leaves on engagement metrics. L, M, H are low, medium and high levels, respectively.

interpret them in detail. Specifically, more hyperlinks in a page don't necessarily increase dwell time, as users are more likely to navigate to other pages with little dwell time or even bounce to other websites (Bounce Rate) through these hyperlinks. Users are more likely to revisit the page (Return Frequency) after hopping as the content of hyperlinked pages are usually related to the original page. For example, users probably dig into details or social media through other related pages (e.g., Facebook) and navigate back to them. These are further verified by the analysis from the perspective of continuous numerical treatments as detailed in the right side of Fig. 2. Regarding images, they increase both dwell time and return frequency as they can help to improve users' visual appeal and thus retain their attention [11], [29]. However, too many images have opposite effects on user engagement measures as the visual complexity is more likely to boost accordingly [11]. Thus, a moderate number of images are preferred. For videos, they exhibit complicated patterns as video elements consist of external hyperlinks and embedded clips (the right side). Their coarse-grained binary existence increases the return frequency and reduces bounce rate, which helps to catch online users' attention and to enhance their web-surfing experiences [30]. It seems to not increase dwell time. However, the right side further shows that it benefits dwell time initially but deteriorates later. The moderate number of videos does good dwell time and other metrics. Overall, the placement of videos benefits users' engagement. For lists and tables, they shape metrics in the opposite manner. It is recommended that tables should be placed moderately while lists are placed as few as possible. Lastly, textual words are demonstrated to benefit user engagement (in the right side of Fig. 2) as they can provide detailed descriptions of products or services.

Previous studies have also shown that the color will impact users' perceived visual complexity and appeal [2], [11] as well as purchasing willingness [3]. We here thus analyze the average treatment effects of color distribution, as reported in the left side of Fig. 3. For the color of silver and maroon, the higher their proportion is, the more users are engaging with web pages. Regarding purple, the moderate composition does good for this goal. For three additive primary colors (i.e., red, green, and blue), their causal impacts are somewhat sophisticated across different metrics. Overall, their proportions are suggested to be restricted, improved, and moderated jointly for maintaining engagement quality. As a rough point of reference to color theory[3], warm colors evoke warmth because they remind us of things like the sun or fire, which include red. Cool colors evoke a cool feeling because they remind us of things like water or grass, which include blue, green and purple [31]. Previous studies on the relationship between color and emotion show that it is preferable to use low contrast levels between colors. Cold colors are perceived as more suitable than warm ones in this regard [32]. These studies are in line with our findings.

In the right side of Fig. 3, we report results for the visual composition. The areas of images and text are further analyzed from a high-level visual perspective. For image area, scattering images is better than clumping them together. Regarding text area, the distribution of textual elements has quite divergent effects on different metrics. The corresponding intervention should be developed on a case-by-case basis. The experiments on balance show that elements should be placed with high and low balance in horizontal and vertical dimensions, respectively. The overall equilibrium contributes to the improvement of engagement metrics. Low horizontal symmetry is beneficial as well.

In addition, we use the binning level of leaves to study the visual complexity of the overall web page, as reported in the right side of Fig. 4. The less complex page screenshots are, the better they are suited for the engagement improvement. A high number of leaves are more likely to be perceived as unprofessional and thus influence the ratings of visual complexity and appeal [11].

In summary, these results are derived by controlling for confounders and will provide general insights into the content understanding for user engagement.

VI. Discussion and Future Work

The candidate treatments are generated by customized algorithms. The treatment generation procedure is separated from

[3]https://cios233.community.uaf.edu/design-theory-lectures/color-theory/

the causal modeling in this work. It facilitates the human-level intervention by turning experimental results into actionable insights. That being said, there could be so many candidate features that this two-stage approach is not efficient. Furthermore, feature-handcrafting time is wasted if the generated features make no significant influence as determined by the KPI models later on. Thus, automatic content generation for shaping KPIs is a promising direction to explore in the future.

Actually, user response consists of different phases, and we focus on user engagement. Our work, however, can be readily applied to other user responses. Our AdaDR is based on RCM, where all confounders are assumed to be measurable and observable. However, this assumption of ignobility is usually untestable, and some unmeasured or latent confounders probably do exist. We will thus extend our study by inferring the latent confounders, which might remove a possible bias in the current settings due to hidden confounders.

VII. Conclusions

In this article, we have proposed to understand contents from a user engagement perspective. We refer to understanding as causality rather than correlation here. The adaptive doubly robust analysis is then developed to identify the causal effects of key elements and visual compositions. The gained insights based on the in-depth analysis of a private real-world dataset of user visit logs help to identify attribution and optimize content distribution in digital marketing. Our work is an attempt to advance content understanding in user experience.

References

[1] C. Cheng, F. Tan, X. Hou, and Z. Wei, "Success prediction on crowdfunding with multimodal deep learning," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2019, pp. 2158–2164.

[2] K. Reinecke and K. Z. Gajos, "Quantifying visual preferences around the world," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 11–20.

[3] R. Bagchi and A. Cheema, "The effect of red background color on willingness-to-pay: The moderating role of selling mechanism," *Journal of Consumer Research*, vol. 39, no. 5, pp. 947–960, 2012.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[6] J. Hartmann, A. Sutcliffe, and A. De Angeli, "Investigating attractiveness in web user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 387–396.

[7] D. B. Rubin, *Matched sampling for causal effects*. Cambridge University Press, 2006.

[8] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[9] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[10] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 457–466.

[11] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos, "Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2049–2058.

[12] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret, "Models of user engagement," in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2012, pp. 164–175.

[13] F. Tan, Z. Wei, J. He, X. Wu, B. Peng, H. Liu, and Z. Yan, "A blended deep learning approach for predicting user intended actions," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 487–496.

[14] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski, "Towards a science of user engagement (position paper)," in *WSDM workshop on user modelling for Web applications*, 2011, pp. 9–12.

[15] A. Sutcliffe, "Designing for user engagement: Aesthetic and attractive user interfaces," *Synthesis lectures on human-centered informatics*, vol. 2, no. 1, pp. 1–55, 2009.

[16] T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet, "Causal inference in public health," *Annual review of public health*, vol. 34, pp. 61–75, 2013.

[17] L. Li, S. Chen, J. Kleban, and A. Gupta, "Counterfactual estimation and optimization of click metrics in search engines: A case study," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 929–934.

[18] J. Pearl, "Causality: models, reasoning, and inference," *Econometric Theory*, vol. 19, no. 675-685, p. 46, 2003.

[19] J. Pearl, "Graphs, causality, and structural equation models," *Sociological Methods & Research*, vol. 27, no. 2, pp. 226–284, 1998.

[20] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian, "Doubly robust estimation of causal effects," *American journal of epidemiology*, vol. 173, no. 7, pp. 761–767, 2011.

[21] C. Tu, W. Y. Koh, and S. Jiao, "Using generalized doubly robust estimator to estimate average treatment effects of multiple treatments in observational studies," *Journal of Statistical Computation and Simulation*, vol. 83, no. 8, pp. 1518–1526, 2013.

[22] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[23] K. Vermeulen and S. Vansteelandt, "Bias-reduced doubly robust estimation," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 1024–1036, 2015.

[24] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, vol. 46, no. 3, pp. 399–424, 2011.

[25] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[26] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, vol. 96, no. 1, pp. 187–199, 2009.

[27] K. Hirano and G. W. Imbens, "The propensity score with continuous treatments," *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, vol. 226164, pp. 73–84, 2004.

[28] J. Ha, R. M. Haralick, and I. T. Phillips, "Recursive xy cut using bounding boxes of connected components," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 2. IEEE, 1995, pp. 952–955.

[29] P. B. Lowry, D. W. Wilson, and W. L. Haig, "A picture is worth a thousand words: Source credibility theory applied to logo and website design for heightened credibility and consumer trust," *International Journal of Human-Computer Interaction*, vol. 30, no. 1, pp. 63–93, 2014.

[30] S. Kim and A. S. Mattila, "An examination of electronic video clips in the context of hotel websites," *International Journal of Hospitality Management*, vol. 30, no. 3, pp. 612–618, 2011.

[31] C. K. Coursaris, S. J. Swierenga, and E. Watrall, "An empirical investigation of color temperature and gender effects on web aesthetics," *Journal of usability studies*, vol. 3, no. 3, pp. 103–117, 2008.

[32] E. Papachristos, N. Tselios, and N. Avouris, "Inferring relations between color and emotional dimensions of a web site using bayesian networks," in *IFIP Conference on Human-Computer Interaction*. Springer, 2005, pp. 1075–1078.