Persuade to Click: Context-aware Persuasion Model for Online Textual Advertisement

Yuan Yuan, Fengli Xu, Hancheng Cao, Guozhen Zhang, Hui Pan, *Fellow, IEEE,* Yong Li, *Senior Member, IEEE,* and Depeng Jin, *Member, IEEE*

Abstract—In recent years, due to the prevalence of online textual advertisements, increasing businesses recognize their huge potential in product promotion. The high-quality textual content has been empirically shown to have a substantial impact on consumers' attitudes and decisions. As a result, persuasive tactics play an essential role in online textual advertisements, which are employed to increase the attractiveness, and sequentially increase the conversion rate and sales volume. As the context of persuasion, product attributes, e.g., category and price, also greatly influence the persuasion outcomes. However, they are largely overlooked by existing works. In this paper, we propose a novel framework to study context-aware persuasion by designing a multi-tasking learning model and performing extensive causal analysis. First, the prediction model recognizes the persuasive tactics employed in an advertising text and predicts their promotion effectiveness. Specifically, we design a disentangled representation learning algorithm to capture the persuasive tactics, and then develop a novel context-aware attention module to model the relationships between persuasive tactics and product attributes. Experiments on a large-scale real-world dataset demonstrate the superior performance of our proposed model over state-of-the-art baselines. Then we show its great practical value by conducting an in-depth causal analysis of context-aware results that our model learns, which offers insightful interpretations and guidelines for marketers to employ persuasive tactics in textual advertisements.

Index Terms—Persuasive tactics, disentangled representation learning, context-aware, causal analysis

1 Introduction

Web applications provide important services for online marketing, such as online retail, online advertisements. Recently, more businesses have recognized the tremendous potential value of content on e-commerce websites, which also inspires the interests of researchers [1]–[4]. On e-commerce websites, online textual advertisements are becoming increasingly important, as customers can choose to interact with the advertisement, e.g., clicking the product link, that interests them. Undoubtedly, high-quality advertising text is of great significance, which employs various persuasive tactics to improve its promotion effectiveness. Therefore, obtaining a better understanding of persuasion in the textual advertisement is especially important. It not only contributes to promotion effectiveness prediction, but also leads practitioners to know why their persuasive tactics work. Figure 1 shows a typical example of online textual advertisement, which includes textual content with a promoted product (along with product info, e.g., price).

It is acknowledged that persuasive content contributes to a good advertisement, while persuasion in advertisement with computational methods has not been well-explored.

Fig. 1. An example of textual advertisement, which contains two components: advertising text and related product. The sentences are highlighted according to the used persuasive tactics.

The hypothesis that online textual advertisements affect consumer behavior has been supported by prior empirical studies [5]–[7]. Besides, [8] and [9] extracted predictive textual features of economic outcomes from product reviews. However, these studies only used simple summary statistics of text like length, readability. In addition, related studies are concerned with lexicon discovery by using odds ratios to select features and hierarchical regression to determine their importance [10], [11]. Despite these explorations, we propose that embedded in these advertisements are persuasion theories that affect consumers, which receive less attention. In the meantime, persuasive communication has been widely studied from various perspectives, including persuasive essays [12], [13], crowdfunding [14], web argu-

- *Y. Yuan, G. Zhang, Y. Li and D. Jin are with Beijing National Research Center for Information Science and Technology (BNRist), and Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: y-yuan20@mails.tsinghua.edu.cn, liyong07@tsinghua.edu.cn*
- *F. Xu and H. Pan are with Department of Computer Science and Engineering, Hong Kong University of Science and Technology.*
- *H. Cao is with Department of Computer Science, Stanford University, CA, 94305.*

ments [15]. However, we still know little about the way how to model the persuasion in advertisements computationally and it also remains hard to analyze how it functions in affecting the advertisement effectiveness with the corresponding product.

In this paper, we propose a framework to study persuasion in online advertisement, where we jointly consider multiple persuasive tactics and diverse contexts. The task has not been well explored and is faced with the following challenges. (1) it is difficult to operationalize persuasive tactics, which are defined by the classic psychological theory, in real-world advertisements practically. Since classic theories are quite abstract and general, there exists a natural gap between theories and real-world advertisements. Some studies predict or assess persuasiveness with linguistic features [15]–[17]. However, they cannot integrate explicit knowledge of persuasion theory. (2) The relationships between persuasive tactics and contexts are complicated, and the context, such as the product's price and popularity, can affect the effectiveness of persuasive tactics. Due to the mixed employment of persuasive tactics and diverse contexts, it is challenging to design effective modules to capture their relationships. To the best of our knowledge, there is no prior research taking the persuasion contexts into consideration, which, however, is important to improve the advertisement persuasiveness. (3) It is also essential to understand the causal relationships between persuasive tactics and contexts in addition to accurate prediction results, which facilitates applications of our study in real-world applications. Existing related works investigate factors that are related to persuasiveness with various methods [16], [18], such as important lexicon induction and examining the effect of length scales. However, they only focus on basic linguistic features and fail to offer high-level instructions based on theories to improve persuasiveness.

To address the above challenges, we propose a framework to model context-aware persuasion in online advertisements. Specifically, we address challenge 1 by two steps. We first operationalize multiple persuasive tactics based on classic persuasion theories with rich expert knowledge. Then, we design a GRU-based persuasion encoder to generate tactic-related representations, which is supervised by a tactic classification task. The innovative utilization of well-studied neural nets maps advertisement sentences to persuasive tactics computationally. To address challenge 2, we design a novel context-aware attention module, which effectively model the relationships between multiple tactics and diverse contexts. It is distinguished from the vanilla attention mechanism by capturing the interactions. In this way, we are able to model persuasion contexts more realistically as employed persuasive tactics are often related to promoted products. Last, we address challenge 3 with two elaborate designs: learning disentangled tactic representations and performing causal analysis to offer practical guidelines. Specifically, to improve the stability and interpretability of tactic representations, we model distinct tactics independently by reducing the spurious correlations between different tactic embeddings. In addition, we design a causal analysis, which use CEM to control for the confounders, to further reveal the causal relationships between multiple persuasive tactics and diverse contexts. The novel design of causal analysis provides insights to improve the interpretability of DL models.

We evaluate our proposed model on a real-world dataset, and the results demonstrate that modeling persuasion in a context-aware manner is significantly useful in identifying persuasive tactics and measuring the promotion effectiveness. By learning disentangled representations of persuasive tactics and leveraging the context-aware mechanism, it outperforms state-of-the-art baselines significantly, and achieves a 14.0% improvement in accuracy of persuasive tactics identification, and reduces the MAE of promotion effectiveness prediction by 11.3%. In addition, we validate the utility of learning disentangled representations. The overall performance indicates that our method can effectively model context-aware persuasion with high-quality disentangled representations of persuasive tactics, which provides a more accurate quantitative framework to study persuasion in online advertisements.

In the causal analysis, we uncover the context-aware mechanism, via the difference of attention weights between the treated and control groups with a method for improving causal inferences called "Coarsened Exact Matching" (CEM) [19]. The detailed analysis provides insights into applying persuasive tactics in practice. For example, the social proof tactic should be utilized with best-selling products but not for niche products, and the price sensitivity could be changed by using persuasive tactics: pricing tactic increases the price sensitivity while nonprice tactics reduce it. The findings of causal analysis offer insightful suggestions to marketers for better persuasion in the textual advertisement. Furthermore, we compare the causal analysis results between dataset with accurate annotation on persuasive tactics and unlabeled dataset, whose persuasive tactics labels are predicted by our model. The high consistency between them demonstrates the models' robustness, which has the potential to be applied to datasets without human labeling. In summary, our contributions are as follows:

- We investigate persuasive tactics in online textual advertisement and present a new paradigm for persuasion modeling. To our best knowledge, it is the first attempt to model context-aware persuasion in the commercial web.
- We propose a multi-task learning model CAPE, which jointly identifies persuasive tactics and predicts their promotion effectiveness. It outperforms existing state-of-the-art baselines significantly on a real-world dataset, demonstrating the essential roles of the proposed disentangled representation learning and context-aware mechanism.
- We perform a causal analysis to understand the context-aware mechanism, which provides valuable insights and suggestions to marketers for generating more effective advertisements in real scenarios. Further comparison results validate that our model facilitates the causal analysis of unlabeled data, demonstrating the model's generalization ability and robustness.

2 Preliminary

2.1 Research Context

We collect the dataset from Beidian[1], one of the largest social e-commerce platforms in China [20], [21]. After reading the

1. https://www.beidian.com

advertisement posted on the website, customers can choose to click the product link to get more detailed information about the product. Figure 1 is a typical example of the advertising text. The user, Della, posted an advertisement to promote facial masks, during which she used several persuasive tactics, such as *commitment and consistency*, *product description* and *social identity*, to increase the attractiveness. We use the click-through rate (CTR) to represent the promotion effectiveness score, as it is the number of reads divided by the number of clicks, which reflects the effectiveness of an advertising text. Briefly, CTR is the probability of a click on the attached product link, given a consumer sees the advertising text.

2.2 Persuasive Tactics

Researchers identify a set of persuasive tactics that affect people's reactions toward persuasion [22], including *social proof*, *social identity*, *commitment and consistency*, *source credibility* and *scarcity*. Based on persuasion theories, we operationalize persuasive tactics, including *commitment and consistency*, *social identity*, *social proof* and *credibility*. We also propose tactics considering our scenario with product promotion: *pricing*, *product description* and *motive*. The selected persuasive tactics cover four essential factors of product promotion: product, reader, marketer, and other customers. Then we examine that these persuasive tactics are commonly used and comprehensive in product promotion. We introduce them as follows:

Social identity: Identity refers to people's adoption of particular groups and traits [23]. *Social identity* is defined as peoples' self-perception and self-concept of an existing membership of social groups [24]. In the textual advertisement, marketers describe their characteristics and traits to increase customers' sense of identity, so that it is easier to persuade them. For example, marketers may express and share their feelings about becoming new moms, and then guide potential consumers to select diapers.

Social proof: People tend to act in conformity to the majority of others and use others' attitudes and behaviors as cues, which is called *social proof* [25]. When promoting products, marketers only show many others' purchasing behaviors, which is convincing enough to attract customers to act. For example, many advertisers always inform consumers when the products are *best-selling* or *fastest-growing*.

Credibility: Prior studies find the effect of source credibility on the recipients' responses, indicating that higher credibility gives rise to more behavioral compliance. [26]. In the textual advertisement, not only the authority of marketers, but also the worthiness of promoted products, enhance the credibility of promotion. For example, marketers emphasize their expertise in the related field of the promoted product. Besides, they stress the product's famous brand, the renowned spokesperson, or some expert certifications.

Commitment and consistency: It states that we desire to be consistent with what we have done or decided before to justify our earlier behaviors and decisions. A positive relationship is reported between *commitment* and attitude change in [27]. In the textual advertisement, this tactic focuses on convincing others of their correct choices. For example, practitioners commit that: *I have bought it*, or show their consistency with previous behaviors: *I will repurchase it after running out*.

Pricing: In many cases, *price*, serving as a stimulus to attract customers to e-commerce, has a significant influence on online purchases [28]. According to [29], lower prices contribute to increasing consumers' probability of purchasing from e-commerce. Thus, whether mentioning the information of a better price in the text may have an impact on customers' behaviors. Expressions like *low price*, *cheap* are regarded as the *pricing* tactic.

Product description: Marketers usually use detailed descriptions to introduce products [30]. Therefore, we leverage this expression as one of the persuasive tactics in the textual advertisement. If the content of the text refers to the necessary information of the product, such as appearance, use, and function, we classify it as the *product description* tactic.

Motive: The tactic encourages consumers to purchase the product by reasoning, which provides buying motivation. Prior studies examine how the motive perceptions influence customers' likelihood of purchasing the product, and motives are thought to have a direct and motivating impact on people's behaviors [31], [32]. When promoting products, marketers improve customers' motivations to choose the product by offering reasons for purchasing. For example, alternation of seasons and climate changes are often mentioned to suggest purchasing season-related products.

Note that other persuasive tactics, such as the *scarcity* tactic, which leads people to value the rare products, and the *liking* tactic, which leverages good relationships to persuade others, are also widely applied in marketing. However, either there are not enough observed examples, e.g., *scarcity*, or it is impractical to use them in text, e.g., *licking*, so we only consider the operationalized persuasive tactics mentioned above.

To present the framework of context-aware persuasion study, we first introduce the prediction model in Section 3, and then we examine the model's performance of identifying persuasive tactics and predicting advertisement effectiveness in Section 4. Subsequently, we perform extensive causal analysis with large-scale observational data to uncover the context-aware attention mechanism of modeling interactions between persuasive tactics and product attributes in Section 5.

3 MODEL

3.1 Overview

Problem definition: Let an advertising text $\mathbf{T} = \{S_0, S_1, ..., S_L\}$ consists of L sentences to promote a product, and $\mathbf{A} = \{a_1, a_2, ..., a_q\}$ consists of q attributes to describe the product. The goal of the proposed model f is to identify the sentence-level label p_i of employed persuasive tactics and predict the promotion effectiveness $\hat{y}_i$ of the advertising text, i.e., $[\mathbf{p}, \hat{y}_i] = f(\mathbf{T}, \mathbf{A})$.

As a solution, we propose a deep multi-task learning model: CAPE, which models the **C**ontext-**A**ware **P**ersuasion in online textual advertisement and measure the promotion **E**ffectiveness. We first design an attentional gated recurrent neural network as a persuasion encoder to identify

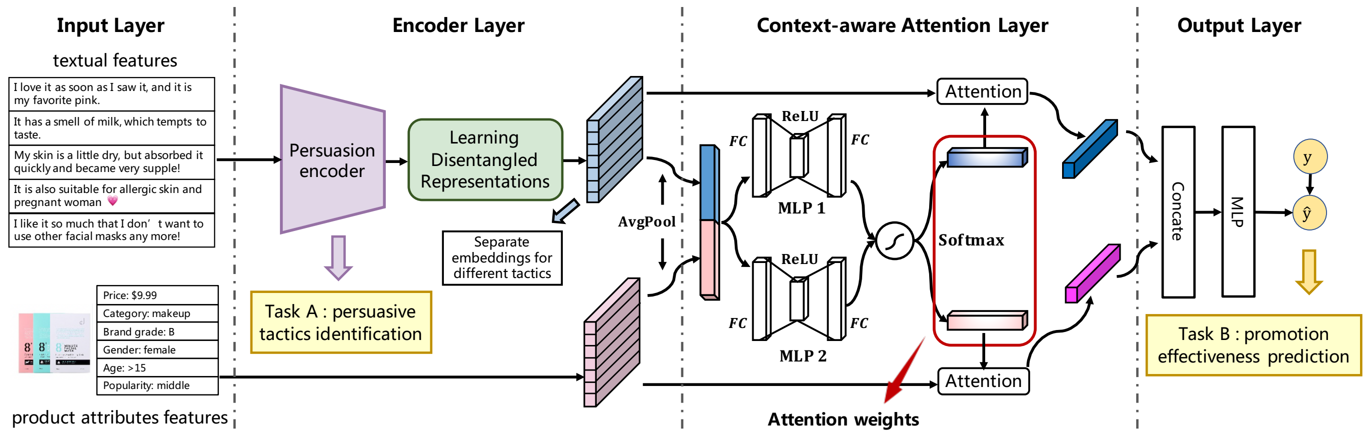


Fig. 2. The overall architecture of our proposed model CAPE. After the input layer takes in the product attributes and advertising text, in the encoder layer, the persuasion encoder generates the text-level persuasion representation and the one-hot vector of product attributes is transformed into embeddings. Then the context-aware attention layer models the relationships by simultaneously generating attention weights on persuasive tactics and product attributes, and develops new embeddings. Lastly, the output layer concatenates the embeddings and predicts promotion effectiveness.

sentence-level persuasive tactics. Then we learn disentangled text-level representation for persuasive tactics with a covariance regularization to improve the interpretability and robustness of representations. Furthermore, we propose a novel context-aware attention module to capture the relationships between persuasive tactics and product attributes. Moreover, it allows us to better understand the relationships by analyzing the learned attention weights. The overall architecture of the proposed model is illustrated in Figure 2. It is composed of four layers: (1) Input layer, which takes in the textual features and representations of the product attribute. (2) Encoder layer, where each advertising text is encoded into a representation of persuasive tactics, and product attributes are transformed to embeddings. (3) Context-aware attention layer, which models the relationship between persuasive tactics and product attributes. (4) Output layer, which is used for promotion effectiveness prediction. Overall, it is an end-to-end model with multi-task learning. For easy understanding, we elaborate on the technical details of the key modules of the model in the following subsections.

3.2 Persuasion Encoder

First of all, our goal is to represent the textual advertisement as persuasive tactics. Considering that an advertising text consists of several sentences, which may use different persuasive tactics, we design a persuasion encoder to generate each sentence's embedding and classify it as persuasive tactics. Gated Recurrent Unit RNN (GRU) [33] could be seen as an effective tool to keep the long-term information of the sequential input data, and we utilize it to encode the semantic features of each sentence into a representation of persuasive tactics. The architecture is shown in Figure 3. To associate words with persuasive tactics, we present an attention mechanism [34] to aggregate the hidden states generated from GRU.

As pre-trained word vectors are utilized, we directly adopt input embeddings of words from Tencent-AI lab [35]. For each element $\mathbf{x}_t$ in the sequence, GRU computes a hidden state $\mathbf{h}_t$ at each time-step using gated mechanism to

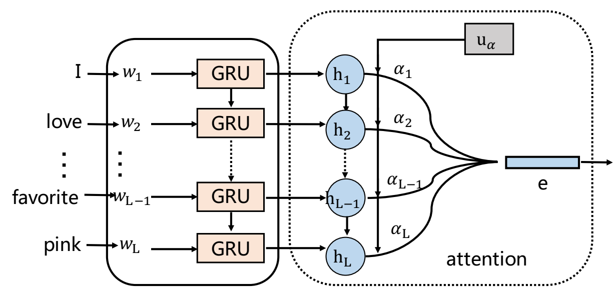


Fig. 3. The design of persuasion encoder. GRU net with an attention mechanism encodes each sentence into persuasion embedding.

gather information of current memory $\mathbf{h}'_t$ and the previous hidden state $\mathbf{h}_{t-1}$. Given the sentence $\mathbf{S}_i$, word vectors from $\mathbf{w}_{i,1}$ to $\mathbf{w}_{i,l}$ are encoded into hidden state vectors from $\mathbf{h}_{i,1}$ to $\mathbf{h}_{i,l}$:

$$\begin{aligned} \mathbf{r}_t &= \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1}) \ , \\ z_t &= \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1}) \ , \\ \mathbf{h}_t &= z_t \odot \mathbf{h}_{t-1} + (1 - z_t) \odot \sigma(\mathbf{W}_s\mathbf{x}_t + \mathbf{r}_t \odot \mathbf{U}_s\mathbf{h}_t) \ , \end{aligned} \tag{1}$$

where $\mathbf{W}_r, \mathbf{W}_z, \mathbf{W}_s \in \mathbb{R}^{h\times n}, \mathbf{U}_r, \mathbf{U}_z, \mathbf{U}_s \in \mathbb{R}^{x\times h}$ are parameters of the network and $\mathbf{x}_t$ is the input at time t. σ is the sigmoid function, and $\odot$ is the Hadamard product.

After we obtain the hidden state vectors of each word in the sentence, we leverage an attention mechanism [33] to capture the tactic-related features and aggregate the word embeddings into sentence-level representations. The process can be formulated as follows:

$$s_{i,j} = \mathbf{u}_\alpha^T \tanh(\mathbf{W}_\alpha \mathbf{h}_{i,j} + \mathbf{b}_\alpha) \ , \tag{2}$$

$$\alpha_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=1}^{l} \exp(s_{i,k})} \ , \tag{3}$$

$$\mathbf{e}_i = \sum_j \alpha_{i,j} \mathbf{h}_{i,j} \ , \tag{4}$$

where $\mathbf{W}_\alpha$ and $\mathbf{b}_\alpha$ are parameters of the fully-connected layer, and $\mathbf{u}_\alpha$ is a context vector of parameters serving as a query, which is learnable and randomly initialized.

To assign concrete tactic meanings to the sentence representations, we design a tactic classification task with supervised learning. Here human labels based on expert knowledge serve as ground truth, which provide information of associations between sentences and persuasive tactics. With such supervision, the hidden representation E can be optimized to extract effective tactic-related features. Specifically, we feed the embedding $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_m]^T$ into a fully-connected layer and then to a SoftMax layer, where m is the the number of sentences in a text. The probability distribution matrix $\mathbf{P} \in \mathbb{R}^{m\times k}$ is calculated as follows:

$$\mathbf{P} = \mathrm{softmax}(\mathbf{W}_p E + \mathbf{b}_p) \tag{5}$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are parameters of the fully-connected layer. The obtained matrix describes the probability distribution of using different persuasive tactics, which is used in the tactic classification task. Then we calculate the text-level tactic representation by aggregating the sentence embeddings according to the tactic probability as follows:

$$\mathbf{T}_p = \mathbf{P}^T \mathbf{E} \tag{6}$$

where $\mathbf{T}_p = [\mathbf{p}^1, \mathbf{p}^2, ..., \mathbf{p}^k]^T$ denotes the employment of persuasive tactics, h is the hidden size, and k is the number of persuasive tactics.

3.3 Learning Disentangled Representations

Section 3.3: "Although we have obtained representations of persuasive tactics with the supervision of expert labels, there are probably multiple persuasive tactics in an advertisement. As a result, different tactic embeddings can be affected by each other and give rise to low expressive ability. For example, if representations of tactic A and tactic B are highly dependent, the model may perform well when both tactics are employed but give unsatisfactory results when only one tactic is used. Thus, it is important to avoid the modelling of spurious correlations embedded in the employment of persuasive tactics. At this point, it is essential to model distinct tactics independently to improve robustness, which means high stability under distribution shift which is mainly caused by spurious correlations hidden in the data. Meanwhile, flexible modelling is also important as it is easier to transfer and adapt when adding or removing a tactic. To achieve this goal, we propose to learn disentangled representations of persuasive tactics. Here the disentanglement means that representations model different persuasive tactics independently, and each channel only captures patterns of one desired persuasive tactic. By disentangled learning, we obtain more robust representations, which means high stability under distribution shift which is mainly caused by spurious correlations hidden in the data. Figure 4 illustrates the details of the disentangled learning.

Specifically, by utilizing the covariance regularization on $\mathbf{T}_p$, we reduce the mutual effect between different tactic embeddings. First, we define the covariance matrix of the tactic representations by:

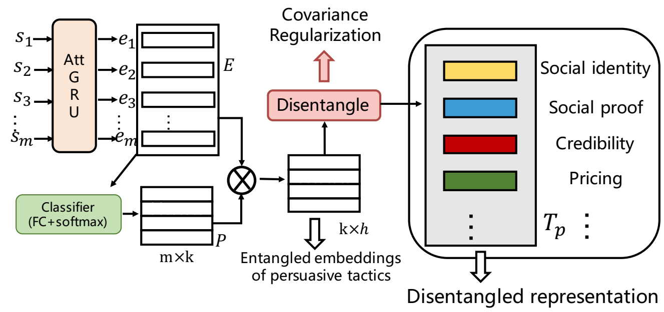


Fig. 4. The process of learning disentangled representations for persuasive tactics. AttGRU refers to the persuasion encoder.

$$\mathbf{C} = \frac{1}{h}(\mathbf{T}_p - \bar{\mathbf{T}}_p)(\mathbf{T}_p - \bar{\mathbf{T}}_p)^T, \tag{7}$$

where $\bar{\mathbf{T}}_p$ is the mean matrix with regard to each row and it has the same shape with matrix $\mathbf{T}_p$. Then we add a l_2 regularization term on all off-diagonal elements of the covariance matrix, and the auxiliary loss is as follows:

$$\mathcal{L}_C = \frac{1}{2}(\|\mathbf{C}\|_F^2 - \|\mathrm{diag}(\mathbf{C})\|_F^2), \tag{8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Note that this objective is jointly optimized in the training process via multi-task learning.

3.4 Context-aware Attention Module

As we discuss before, modeling context-aware persuasion is essential, which, however, is not resolved by existing works. In order to effectively model the persuasion in a context-aware manner to improve our model's generalization ability and explainability, we design a novel context-aware attention module, which can further explore the relatedness between persuasive tactics and product attributes. It learns the contributing weights from different channels to exploit the interactions between features for prediction, and serves as a bridge to link persuasive tactics and the context. Specifically, it emphasizes crucial feature channels in predicting the promotion effectiveness and aggregates feature channels with corresponding attention weights. The design of this module is shown in the context-aware attention layer of Figure 2.

Given the persuasion matrix $\mathbf{T}_p \in \mathbb{R}^{k\times h}$ and product matrix $\mathbf{A} \in \mathbb{R}^{q\times d}$ from two aspects, we summarize each feature map with avg-pooling and obtain the embedding $\mathbf{a}$ and $\mathbf{p}$ with q and k channels, respectively. The goal is to generate attention weights for each channel. We concatenate the two embeddings into $\mathbf{x}$ and utilize two FC/MLP layers connected with the softmax function parameterize the context-aware process. We depict the context-aware attention operation as follows:

$$a_i = \mathrm{mean}(\mathbf{A}_{ij}) \ , j \in [0, d-1], \tag{9}$$
$$p_i = \mathrm{mean}(\mathbf{T}_{ij}) \ , j \in [0, h-1] \ ,$$
$$\mathbf{x} = [\mathbf{a}, \mathbf{p}] \ , \tag{10}$$
$$\alpha = \mathrm{F}_{\mathrm{ca-ex}}(\mathbf{x}, \mathbf{W}_a) = \mathrm{softmax}(\mathbf{W}_{a2}\delta(\mathbf{W}_{a1}\mathbf{x})) \ , \tag{11}$$
$$\beta = \mathrm{F}_{\mathrm{ca-ex}}(\mathbf{x}, \mathbf{W}_p) = \mathrm{softmax}(\mathbf{W}_{p2}\delta(\mathbf{W}_{p1}\mathbf{x})) \ ,$$

where α, β are attentional weights over the feature channels respectively, and δ is the ReLU function, and $\mathbf{W}_{c1}$,$\mathbf{W}_{c2}$,$\mathbf{W}_{p1}$,$\mathbf{W}_{p2}$ are learnable parameterized matrixes of the MLP network.

The final output embeddings of product feature and persuasion feature, which consider each other, are obtained by aggregating the original feature map with channel-wise attention weights as follows:

$$\mathbf{a}' = \alpha^T \mathbf{A} \ , \tag{12}$$
$$\mathbf{p}' = \beta^T \mathbf{T_p} \tag{13}$$

3.5 Output and Loss Function

In order to predict the promotion effectiveness, we feed the concatenated vector into the output layer. Two fully connected layers are used for the task:

$$\hat{y} = \mathbf{W_2}(\sigma(\mathbf{W}_1[\mathbf{a}', \mathbf{p}'] + \mathbf{b}_1)) + \mathbf{b_2} \ , \tag{14}$$

where $\mathbf{W_1}, \mathbf{W_2}, \mathbf{b}_1, \mathbf{b}_2$ are parameters, and $\mathbf{x}$ is the concatenated vector from the context-aware attention layer. Here we apply ReLU as the non-linear activation function σ.

The model is trained end-to-end with multi-task learning. The loss function consists of four parts: (i) The cross entropy loss for persuasive tactics classification:

$$\mathcal{L}_p = -\frac{1}{M} \sum_{i=0}^{M-1} \sum_{k=0}^{K-1} y_{i,k} log(p_{i,k}) \tag{15}$$

where M denotes the number of labeled texts, K is the number of persuasive tactics, $y_{i,k} \in \{0,1\}$ refers to the human label in the sentence s_i.
(ii) The mean absolute error loss for effectiveness prediction on labeled data $\mathcal{L}_{label}$ and (iii) unlabeled data $\mathcal{L}_{un}$:

$$\mathcal{L}_{label} = \frac{1}{M} \sum_{y_1 \in D_L} |y_1 - \hat{y}_1| \tag{16}$$

$$\mathcal{L}_{un} = \frac{1}{N} \sum_{y_2 \in D_U} |y_2 - \hat{y}_2| \tag{17}$$

where M is the number of labeled texts and N is the number of unlabeled texts. y_1 is the ground truth of effectiveness score and $\hat{y}_1$ is the prediction result in the labeled data, while y_2 is the ground truth of effectiveness score and $\hat{y}_2$ is the prediction result in unlabeled data.
(iv) The covariance regulization loss $\mathcal{L}_C$ to disentangle different persuasive tactics as described in Eq. (8):

$$\mathcal{L}_C = \frac{1}{2}(\|C\|_F^2 - \|diag(C)\|_F^2) \tag{18}$$

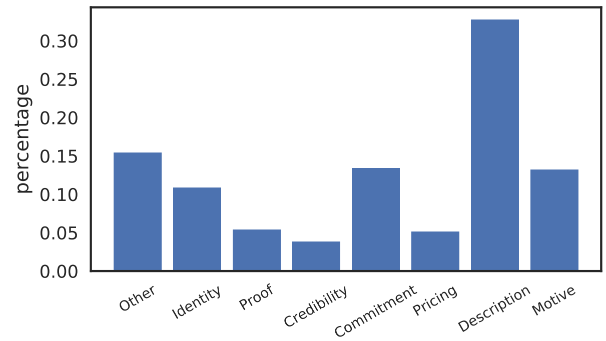


Fig. 5. The proportion of persuasive tactics in labeled data.

(v) In addition, we also add l_2 regularization loss $\mathcal{L}_r$ to prevent overfitting.

$$\mathcal{L}_r = \sum_{\mathbf{\Theta} \in \mathcal{P}} \|\mathbf{\Theta}\|_2 \tag{19}$$

where $\mathcal{P}$ denotes the set that contains all the parameters in the model.
Therefore, the combined loss function is as follows:

$$\mathcal{L} = \beta\mathcal{L}_p + \eta\mathcal{L}_{label} + (1-\eta)\mathcal{L}_{un} + \lambda\mathcal{L}_C + \mu\mathcal{L}_r \tag{20}$$

where β, η, λ, and μ are hyper parameters. We use hyper parameter η to control the penalization for labeled data and unlabeled data, a hyper parameter β for trading off the two tasks, and a hyper parameter λ to control the covariance regularization. During the training phase, we use Adam optimizer to optimize the model parameters by minimizing the loss function in a mini-batch mode.

4 EXPERIMENT

4.1 Dataset

The dataset is collected from one of the largest social e-commerce platforms in China, which provides the opportunity of studying the problem. On the platform, users are motivated to recommend products to others, which offers rich research data for the problem. Textual advertisement is a typical method for promotion, which exploiting posted text to introduce and recommend products. Here we use the click-through rate (CTR) to represent the promotion effectiveness, which is a commonly used index to measure the effectiveness of advertisement. We calculate the CTR of each text as the clicks divided by impressions in three months. The dataset covers posted advertisements in a period of six months, and we finally obtain 8823 advertising texts in total. In terms of product attributes, we have the information on category, price, popularity, brand level, gender and age.

We sampled 150 messages for each fixed message length ranging from one sentence to eight sentences and obtain 1200 texts for persuasive tactics labeling. Then we ask two annotators to label each sentence in the text and provide detailed descriptions and examples of the persuasive tactics. Specifically, we label each sentence as the most matching one of the operationalized tactics, and label it to *other* if none of the tactics are used. We obtain an intra-class correlation coefficient (ICC) of 0.758, which shows good consistency among annotators [36]. For the samples that two annotations disagreed, we asked another annotators to judge them and labeled the sentence as two of them agreed. Figure 5 illustrates the tactic distribution in labelled data, and we can see that tactics are not uniformly distributed. It is observed that *Product Description* is the most dominant tactic, which is in line with our intuitions. Besides, the percentage of *Social Identity*, *Commitment and Consistency*, and *Motive* are all around 15%, which is moderately leveraged. The remaining tactics are less used relatively but still have adequate samples considering the total annotation amount. In order to train the model, we randomly split the data into the training set, validation set, and test set by 8:1:1.

4.2 Baselines

We compare the performance of our model with the state-of-the-art methods from three research lines to show its effectiveness, including those that only focus on tactic classification, those that complete the two tasks with multi-task learning, and those that incorporate features of product attributes to assist effectiveness measurement. The first group is proposed to predict sentence-level persuasive tactics and consists of:

- **SVM** [37] *Support Vector Machine* (SVM) is a competitive classic classifier, which is widely-adopted in NLP tasks [15]–[17]. Here we use RBF [38] kernel and TF-IDF [39] features.
- **GRU** [33] *Gated Recurrent Unit RNN* (GRU) is an effective tool to process input sequential data. Here we use the last hidden states of the sentence as features to classify persuasive tactics.
- **bi-GRU** [33] It is similar to GRU, and obtains the last hidden states with bi-directional GRU.

The methods in the second group utilize a hierarchical network, which is widely used in multi-task learning, to classify persuasive tactics and predict the promotion effectiveness in the meantime, and they are as follows:

- **HAN** [40] *Hierarchical Attention Networks* (HAN) is famous for first using hierarchical networks and attention mechanism to obtain the text embedding. We utilize this method as a baseline, where the word encoder generates embeddings for tactic classification, and the sentence encoder outputs the text embedding.
- **Semi-Att Net** [41] *Semi-supervised attention net* (Semi-Att Net) is state-of-the-art method to model persuasion strategies, which utilizes semi-supervised learning to improve the performance.

In addition, for a fair comparison, we also use FM and Semi-Att Net that incorporates features of product attributes:

- **FM** [42] *Factorization Machine*(FM) is a competitive baseline for feature interactions through a pairwise inner product. Here we use features of text and product attributes as input to predict the promotion effectiveness.
- **Semi-Att Net-C** *Combined* (-C) denotes that it takes in the concatenated embedding of persuasive tactics and product attributes for predicting promotional effectiveness.

Finally, we also summarize our complete model and its variant as follows:

- **CAPE-base** It is a basic version of our model, where we do not disentangle different persuasive tatics.
- **CAPE** It is the complete version of our model as shown in Figure 2.

4.3 Experimental Settings

We split the text into sentences and then tokenize each sentence at the char level [43]. Here, the data structure contains two sequence levels: the word sequence into a sentence and the sentence sequence into a text. Specifically, we use the Tencent AILab Chinese Embedding model [35] to initialize the word embeddings, which is widely adopted in academia and industry [44]–[46]. In the adopted model, over 8 million Chinese words are pre-trained and 200-dimension word vectors are generated.

Specifically, to maximize the utility of the dataset and improve the model's performance, the labeled data and unlabeled data are both utilized with semi-supervised learning. Following the existing work [41], for the multi-class classification, we use Accuracy, F1 score, Precision, and Recall to evaluate the performance, and the metrics are computed as the arithmetic mean of per-class results. For the promotion effectiveness prediction task, we evaluate the prediction results by Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which are widely used in regression tasks [47]–[49].

The dimension of the pre-trained word embeddings is 200. Specifically, we fix the hidden size of GRU as 64 and the embedding size of each product attribute as 32. We train the model parameters with Adam optimizer regularized by early stop and set the mini-batch size as 32. In addition, we set the balancer β to be 0.1 via searching in a set of $\{0.01,0.1,1,10,100\}$, and search the regularization weight λ from $\{1e\text{-}3,1e\text{-}4,1e\text{-}5\}$. Besides, we also search for the balancer η and set it to be 0.5. We initialize the learning rate as 1e-2 and reduce it dynamically by 0.1. The complete implementation of our model are available at https://github.com/tsinghua-fib-lab/CAPE.

4.4 Overall Performance

We present the overall performance results in Table 1 concerning Accuracy, F1 score (Macro), Precision (Macro), Recall (Macro), MAE, and RMSE. We compare our proposed CAPE model with baselines in three groups. From these results, we summarize key observations and insights as follows:

- **Model Effectiveness** From results in Table 1, we can find that our proposed model CAPE significantly outperforms over baseline methods in both tasks and achieves state-of-the-art performance in terms of all metrics. Compared to the nearest baseline competitor, Semi-Att Net-C, our complete model CAPE improves the accuracy by 14.0%, and the F1 score by 9.3%, and reduces the MAE by 11.3%, and the RMSE by 10.0%, which proves its effectiveness.
- **Multi-task learning works better than single-task learning, and semi-supervised learning is effective.** Baselines in the group 1 cannot perform well, which can be explained as the information on promotion effectiveness has

TABLE 1
The overall performance of baseline methods and our model.

Model	Persuasive Tactic Classification				Effectiveness Prediction	
	Accuracy	Macro F1	Macro Precision	Macro Recall	MAE	RMSE
SVM + TF-IDF	0.335	0.331	0.345	0.295	-	-
GRU	0.475	0.317	0.311	0.343	-	-
bi-GRU	0.381	0.491	0.380	0.395	-	-
HAN	0.397	0.474	0.382	0.429	0.0826	0.1041
Semi-Att Net	0.477	0.516	0.469	0.501	0.0800	0.0997
FM	-	-	-	-	0.0774	0.0956
Semi-Att Net-C	0.456	0.515	0.452	0.469	0.0773	0.0947
CAPE-base	0.516	0.536	0.521	0.516	0.0707	0.0881
CAPE	**0.544**	**0.564**	**0.592**	**0.530**	**0.0685**	**0.0852**

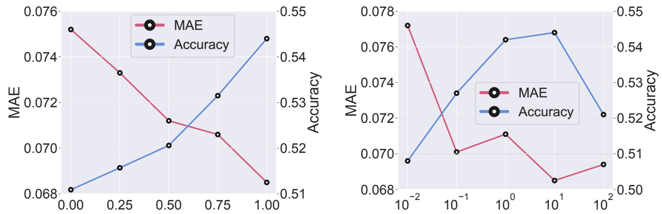
the potential to benefit the recognization of persuasive tactics. Despite this, bi-GRU improves the classification performance from 0.335 to 0.381 of accuracy, demonstrating the advantage of neural networks in feature extraction. Compared with group 1, baselines of the group 2 give better experiment results of the classification of persuasive tactics, which validate that the supervision from the global label (promotion effectiveness) assists the local task (persuasive tactic classification). In addition, the semi-supervised learning method improves the performance in both tasks, which increases the F1 score from 0.474 to 0.513 and reduces the MAE from 0.0826 to 0.0789, further indicating a promising solution for tasks only with small amounts of the expensive human-labeled dataset. Though the attention mechanism's improvement is miner, it contributes to understanding the association between words and persuasive tactics.

- **Our context-aware attention module effectively captures the relationships between persuasive tactics and product attributes.** Despite the moderate performance of persuasive tactic classification, the results of the effectiveness prediction task in the group 2 are not satisfying. We guess the reason for such results is that the outcome of persuasive tactics is correlated with the context, which can be described as product attributes. We present experimental results of baselines in the group 3, which consider features of product attributes. On the one hand, we observe that incorporating the information of product attributes contributes to promotion effectiveness prediction, which verifies the necessity of considering the context of persuasion. On the other hand, the results indicate that just combining the features (Semi-Att Net-C) or exploring the second-order feature interaction (FM) are not significantly useful, so the challenge of modeling the complicated relationship remains unresolved by existing works. Particularly, the CAPE-base model not only produces better results of effectiveness prediction, which achieves an 11.3% MAE reduction compared with the best baseline method, but also improves the classification task by 14.0% accuracy increase. This confirms the utility of our context-aware attention module, which can effectively capture the relationships and interactions.
- **Disentanglement of different persuasive tactics is essential.** CAPE-base, a basic version of our proposed model, generates representations without disentangling different tactics. Although CAPE-base outperforms baselines, it is worth exploring the utility of the disentanglement of persuasive tactics. We observe that, with disentangled representation learning, our whole model (CAPE) achieves better performance than the CAPE-base model, which improves the accuracy from 0.516 to 0.544, and reduces the RMSE from 0.0881 to 0.0852. On the one hand, the improvement validates that it is significant to leverage separate embeddings for different persuasive tactics, which facilitates better prediction outcomes. Besides, the disentanglement of different persuasive tactics also makes our model highly interpretable, which is crucial to persuasion modeling.

4.5 Hyper-parameter Sensitivity Analysis

To explore how different hyper-parameters affect the model's performance, we further study the sensitivities of several key hyper-parameters by varying them in different scales. Specifically, we investigate the proportion of used unlabeled data in the training process and the parameter β, which balances the two tasks. The result of different proportions of used unlabeled data is shown in Figure 6(a). Note that we use all the labeled data and vary the proportion of used unlabeled data. We observe that as more unlabeled texts are included in the training process, the performances on both tasks are improved as well, which shows the utility of the semi-supervised learning method. Figure 6(b) presents the influence of β. Among all the parameter settings, we observe that as β increases from 0.01 to 10, performances of both tasks are improved, and $\beta = 10$ delivers the best performance. Besides, we find that the prediction error decreases as the classifier is more accurate, which validates the important role of persuasive tactics. An interesting observation is that the accuracy drops as β increases to 100, denoting assigning more weight to the classification task. It suggests a mutually reinforcing relationship between the two tasks.

In conclusion, experiment results validate the effectiveness of our proposed model. Moreover, the utility of our specially designed method, including learning disentangled representations and context-aware attention module, is verified. Further analysis of hyper-parameters shows the advantage of semi-supervised learning and demonstrates the essential roles of persuasive tactics in promotion effectiveness prediction.

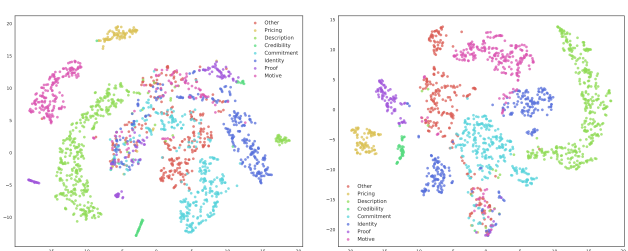


(a) proportion of used unlabeled data (b) β

Fig. 6. Results of hyper-parameter sensitivity analysis on the proportion of used unlabeled data and the balancer β.

4.6 Disentangled Representations Analysis

The performance of CAPE over CAPE-base illustrates the effectiveness of disentangled representation learning. To better understand how this task contributes to the modelling of persuasive tactics, we carry out further studies to investigate the effect of disentangled learning. Specifically, we visualize the learned persuasive tactic embeddings in our model using t-SNE [50]. Figure 7 shows the results of the learned embeddings with and without the disentanglement task. We can observe that with this task, embeddings are better clustered according to persuasive tactics. However, when we remove it, the distribution of tactic embeddings becomes messy and there are obvious overlapping phenomena. The visualization results demonstrate that the disentanglement design can separate embeddings better for different tactics. In summary, the embedding visualization in Figure 7 and the performance improvements of our prediction model both confirm the effectiveness of the disentanglement on learning high-quality representations.



(a) w/o disentanglement (b) w/ disentanglement

Fig. 7. Visualizations of the learned tactic embeddings with and without the disentanglement task.

5 Causal Analysis with Large-scale Observational Data

Though our CAPE model achieves remarkable performance on persuasive tactics identification and promotion effectiveness prediction, it remains unknown how it can be applied in practice. This subsection builds on the revealed insights in performance analysis that the context-aware mechanism contributes to the prediction, and desires to provide deeper insights into applications, which can also disentangle the explanatory power of the proposed model. As a result, the research is based on discovering knowledge from large-scale observational data. Moreover, we are interested in the causality that analysis reveals, which serves great robustness and guarantees high interpretability. To achieve such a goal, we conduct a causal analysis to investigate the mechanism of context-aware. Specifically, for persuasive tactics, the context is product attributes, and for product attributes, the context is the employment of persuasive tactics. Therefore, we conduct two paralleled analyses: one is to explore the roles of persuasive tactics with different product attributes; the other focuses on analyzing how employing specific tactics affects the sensitivity to product attributes. In summary, the causal analysis aims to answer the following research questions effectively:

- **RQ1**: How the context influences the effect of persuasive tactics?
- **RQ2**: How using persuasive tactics affects consumers' sensitivity to product attributes?
- **RQ3**: Can our CAPE model extract useful information from unlabeled behavioral data?

To answer the above research questions, we investigate the attention weights that our model has learned. By grouping the results according to the value of the studied variable A, we investigate the differences of attention weights of target variable B between groups, where we could explore A's influence on B. Note that product attributes and the use of persuasive tactics can both serve as the grouping variable. However, the problem is that the data we collected is observational data. Thus, the effect of a specific attribute is likely confounded by other attributes. Similarly, the effect of using one persuasive tactic might be influenced by other tactics' employment. As a result, a simple grouping method might give rise to biased results. One solution is to conduct controlled experiments. However, it is not infeasible to conduct controlled experiments in our scenario because we cannot control one's decision to use persuasive tactics, and we also cannot control the product attributes. To tackle this issue, we apply Coarsened Exact Matching (CEM) [19], a recently developed causal effects estimation method that controls for the confounding influence. Its basic idea is to control covariates that affect both the treatment and the outcome and rebalance the covariates' data distribution between the treated and control groups. Theoretically, after matching, the only difference between the treated and the control group is the specific grouping variable, such as the product attribute or whether the advertising text uses a given persuasive tactic. Therefore, the difference we observe between the two groups must result from the factor we focus on.

In particular, we conduct causal analysis on both unlabeled data and labeled data to validate our model's generalization ability, which is of valuable significance since human annotation is expensive and time-consuming. For the labeled dataset, we have accurate sentence-level labels of persuasive tactics. For the unlabeled dataset, we utilize the proposed model to predict the result, where the values denote the probabilities of employing persuasive tactics in the text. To represent the use of persuasive tactics in the text with predicted values, we take the top 10% values as the treated group since the persuasive tactic is most likely employed in this advertisement text, and the remaining are treated as the control group.

TABLE 2
Results of percentage change of persuasive tactics' attention weights with product attributes as context. We treat the first line in Groups as the control group and the second line as the treated group. The values in the table denote the percentage change of the treated group compared with the control group. * indicates that the difference between groups is significant for p<0.05, and ** indicates that the difference is significant for p<0.01. The significant results are bolded.

Attributes	Groups	Matched/all	Pricing	Description	Credibility	Commitment	Identity	Proof	Motive
Price	low high	395/464 186/254	**-8.0%***	+7.1%	**-10.6%***	+6.5%	**+10.8%***	-1.0%	**+12.5%***
Category	food cleaning	169/205 192/227	**-12.6%***	**-66.4%****	**+16.7%***	**-35.8%****	**-32.7%****	+6.8%	**+51.7%****
Brand level	low high	229/386 382/657	-0.95%	**-24.2%****	+6.9%	-10.1%	**+13.1%***	**+14.5%***	**+13.4%***
Popularity	low high	370/488 204/299	+2.3%	**-18.6%****	-6.5%	-8.8%	**-14.8%****	**+15.3%****	**+13.6%***
Gender	male female	111/134 279/282	**+23.7%***	+1.5%	**-8.73%***	-4.6%	**-8.7%***	**+14.6%***	**-18.7%***
Age	old young	119/123 305/322	**+35.2%***	-9.5%	-7.1%	**-10.8%***	**+46.5%****	**+13.5%***	+2.7%

5.1 Context-aware Persuasive Tactics (RQ1)

In the previous section, we introduce modeling the persuasion in a context-aware manner, where we consider the product attributes' influence on tactic importance. Here we aim to better understand the roles that product attributes play. To achieve the target, we divide the data into several groups according to attribute values, and the first line in Groups in Tabel 2 serves as the control group, and the second line is the treated group. The percentage values denote the percentage difference of the treated group compared with the control group. The following are the key insights that we observe.

Shall we leverage *social proof* in all cases?

Social proof [25] tactic is employed to encourage behaviors of following others by showing the endorsements from other people, where consumers' conformity is exploited. Differently, *social identity* [24] tactic focuses on emphasizing unique characteristics to inspire the consumers' sense of identity, and sequentially increases the click propensity. Obviously, the two tactics promote products from different aspects, and we compare their roles, considering the popularity of products (see Table 2). An interesting observation is that *social proof* tactic becomes more critical with popular products (15.3% increase). In contrast, *social identity* captures more attention in products with low sales (high sales products decrease the attention weight by 14.8%). It is probably because high sales products are appealing to more consumers. Due to the product's characteristics, such as the latest fashion, it is easier to trigger and leverage people's conformity [51]. However, things are different in niche products whose sales are relatively low, as they are likely provided for specific crowds. The attention weight rise of *social identity* under this circumstance can be explained that the tactic describes the characteristics of groups that would most benefit from the product, and it exactly meets niche products' specialty, which accounts for its benefits. Therefore, *social proof* tactic is suggested for popular products, which emphasizes others' behaviors to convince consumers, and *social identity* tactic shows the potential to be utilized when promoting niche products whose attributes could not inspire consumers' conformity.

Is the *pricing* strategy always useful?

It is traditionally acknowledged that *pricing* strategy is effective in attracting customers as they need to pay less money. However, we observe that low prices are not so important in some cases. For example, in the context of high-price products, the *pricing* tactic's attention weight is reduced by 8.0%, indicating a less important role in prediction. On one hand, *pricing* tactic maybe less employed for products at high prices. Besides, the observation could be supported by prior studies, which found a significant negative relationship between price and perceived value of products [52], [53]. It is probably because the perceived value serves as a vantage point for high-price products, and the price could affect the perceived value to a certain degree as it sometimes reflects the quality. As a result, the low price strategy may fade into insignificance in these cases as it reduces the perceived value. Although a high price does not guarantee high-quality perception, a low price could form low-quality impressions. Therefore, *pricing* tactic contributes more to low-price products, where the price could be treated as a selling point. But for high-price products, the message that high price delivers has demonstrated the product's positioning, so emphasizing lower prices than others becomes eclipsed. Besides, we also find that the attention weight of *pricing* tactic is increased in products for women (23.7% increase) and products for the young (35.2% increase), implying that these groups are more likely to be attracted by low prices, and it provides insights to leverage *pricing* tactic to promote such products.

Do you have motives to click?

Customers' motivation is prominent in clicking behaviors, and giving reasons for purchasing contributes to arousing consumers' interests. We observe that *motive* tactic becomes significantly more important in higher price and higher brand level products. Following our intuitions, people tend to perform a more in-depth analysis under these circumstances; therefore, persuading through reasoning makes sense. Interestingly, we find that *motive* tactic is also more important in products with high sales(13.6% increase), suggesting that even for popular products, in addition to leveraging the conformity, reasoning also benefits and exerts effects.

In summary, we verify the fact that effect of persuasive tactics varies by context, and explore how the context in-

TABLE 3
Results of percentage change of product attributes' attention weights with persuasive tactics as context. Note that *no* and *yes* in Groups denotes whether the persuasive tactic is used. The settings of treated and control groups and the significance indication is the same as Table 2.

Persuasive tactics	Groups	Matched/all	Price	Category	Brand	Gender	Age	Popularity
Pricing	no yes	595/1111 117/125	**+16.8%****	+0.94%	-0.90%	0.0%	**-15.4%***	-0.96%
Description	no yes	581/677 504/559	**-23.5%***	**+18.6%***	-4.4%	**-16.5%***	**-23.8%***	**+19.4%***
Credibility	no yes	449/1165 66/71	+3.4%	-4.2%	+1.8%	+1.4%	+2.3%	-3.1%
Commitment	no yes	721/913 297/323	**-10.3%***	+6.2%	+1.8%	-7.7%	+6.8%	+5.0%
Identity	no yes	717/929 290/307	**-16.3%***	**+9.6%***	0.91%	**-10.6%***	**-14.5%***	**+15.3%***
Proof	no yes	622/1107 120/129	+2.8%	-1.2%	+0.90%	+1.6%	+2.0%	-2.0%
Motive	no yes	6682/1072 162/164	**-17.2%****	**+11.5%***	-2.7%	**-12.7%***	**-17.6*%**	**+12.1%***

fluences persuasive tactics. For example, *social proof* and *social identity* tactics should be utilized in the consideration that if the product benefits most consumers or just for specific crowds. We also find that low prices are not so effective under some circumstances as the possibly formed low-quality impressions. In addition, *motive* tactic is also useful for popular products which are always promoted by leveraging consumers' conformity.

5.2 Context-aware Product Attributes (RQ2)

It is also meaningful to investigate the causal effects of employing persuasive tactics, which can reveal the persuasion's influence on attribute sensitivity. For product attributes, the context is the use of persuasive tactics. Specifically, we obtain each text's employment of persuasive tactics from the model. Similarly, we calculate the treatment effects of using persuasive tactics based on CEM. Table 3 shows the results, and the values denote the relative change of product attributes' attention weight with the persuasive tactic used compared with the control group, where the tactic is not used. The following are the key observations.

Persuasive tactics can change the price sensitivity

Price sensitivity can be defined as the weight attached to price when consumers evaluate the product's attractiveness [54]. We observe that employing the *pricing* tactic leads to a higher price sensitivity (the attention weight on price attribute has a 16.8% improvement). We may explain this from two aspects. One is that the price sensitivity of existing consumers might be increasing as consumers receive more price information; thus, they compare the attribute more actively and increase its importance in decision making [55]. Another is that low prices attract more price-sensitive customers. In addition, we also find that using some nonprice tactics reduces the price sensitivity, including *commitment and consistency*, *description*, *socia identity*, *motive* tactics, implying that consumers exposed to other information become less price sensitive. A possible explanation is that consumers simply no longer pay attention to the price attribute without price-related information. In summary, persuasive tactics could be employed to change consumers' price sensitivity. If the price attribute is expected to be emphasized as an advantage, practitioners may use the *pricing* tactic. On the other hand, when expecting a low price sensitivity, nonprice strategies lead to the goal.

Persuasive tactics' effects are correlated with categories.

We observe that the usage of some persuasive tactics leads to higher attention weight on the category attribute. It is consistent with results in Q1 that the attention weights of persuasive tactics are different across categories. Tabel 3 show that when using *descirption*, *commitment and consistency*, *social identity* or *motive*, the attention weight on category attribute has been increasing, indicating that the impacts of persuasive tactics are not consistent across different categories, that is to say, whether the use of the tactic benefits the promotion effectiveness depends on the product category.

The above analysis reveals that employing persuasive tactics can change consumers' sensitivity to product attributes. Moreover, it suggests that practitioners can utilize appropriate persuasive tactics to increase or reduce consumers' attention to the target product attribute.

5.3 Generalization Analysis (RQ3)

The model's application in datasets without annotation is of great significance since human labeling is costly and time-consuming. Therefore, we compare the causal analysis results between ground truth and model predictions to assess the model's reliability and generalization ability. For the labeled dataset, we know the ground truth that whether a persuasive tactic is used in an advertising text; for dataset without annotation, the employment of persuasive tactics is predicted by the model. Specifically, if the changes are significant and have the same sign, or are both insignificant, we take the pair as a consistent case. We have obtained 42 results in total (note that there are seven persuasive tactics and six product attributes), among which only five cases are inconsistent. Meanwhile, we examine that the key insights we provide in the causal analysis are all consistent between ground truth and model predictions. The experiment indicates that the causal analysis results are largely consistent, which validates that the analysis is reliable and the obtained insights from it are valid. Furthermore, it demonstrates the effectiveness and generalization ability of our proposed model, which could be generalized to datasets without label annotation.

To summarize, the causal analysis explains why the context-aware mechanism works and how the context affects the persuasion process, which illustrates the superior interpretability of the proposed method.

6 Related Work

This study touches on three research strands, including classic persuasion theories, NLP technologies, and computational analysis of textual advertisement as follows:

Persuasion Theories Persuasion has long been studied in a wide range of fields, which motivates a desirable attitude or behavior change. There are two classic persuasion models, which are widely used and recognized. One is Chaiken's systemic-heuristic model, which explains how recipients process persuasive messages [56]. The other is the Elaboration Likelihood Model (ELM) [57], which is proposed by Pretty and Cacioppo and uses central-peripheral processing to explain how people process social information. They both suggest that people perceive persuasive messages in the dual process. Systemic processing and central processing refer to an in-depth analysis of the quality of the argument. In contrast, persuasion occurs with heuristic processing and peripheral processing, in which the message recipients exert less effort in assessing the argument. Based on classic persuasion models, researchers find that processing information in both routes influences people's reactions to persuasion [58], [59]. Studies on persuasion also find that the resistance of attitudes, which is formed through two processing, could be different [60]. In this work, we borrow persuasion theories and operationalize a set of persuasive tactics based on the classic persuasion models.

Deep Linguistic Methods Recently, the rapid development of natural language processing technologies contributes to modeling textual features. For example, Long Short Term Memory (LSTM) networks [61] replaced the Vanilla Recurrent Neural Network to solve the problem of gradient vanishing and explosion. Besides, other deep linguistic models, such as bidirectional lstm [62], which processes the context in two directions, and Gated Recurrent Unit RNN (GRU) [33], which extracts the necessary components in LSTM learning with fewer parameters, also show their advantages in processing sequential data. In addition, attention mechanism [34], which is one of the core and practical innovations in NLP, improves computing efficiency compared with RNN-based models. We base our persuasion encoder module on GRU, and leverage the attention mechanism to associate words with persuasive tactics.

Computational Analysis of Textual Advertisement It is a typical problem to predict the effectiveness of the online textual advertisement. Odds ratios are used for selecting important features [10], and researchers also induce persuasion lexicons from product descriptions that are predictive of sales [63]. In addition, the predictive relationship between textual advertisement and behavioral responses is studied from different perspectives. Sentiment information is extracted to assist sales prediction [64], and semantic product descriptions are investigated for the design of recommender systems. Apart from these, Pryzant et al. [65] have studied how the deconfounded writing style impacts an ad's performance. There are also a wide range of studies on automated text generation, including health-related advertisement derived from web pages [66], product descriptions [1], [2] and review response [3] on e-commerce websites. Importantly, little prior work in this problem seeks to investigate the textual influence on consumers' behaviors from a perspective of persuasive tactics with reference and principled method. Different from prior studies, we conduct a computational analysis of persuasive tactics in the textual advertisement in a context-aware manner, which provides a new analytical framework for computational studies in this field.

7 Conclusion

In this paper, we operationalize a set of persuasive tactics that are widely used in textual advertisements based on persuasion theories. We present a novel framework to study persuasion process in online advertisement by quantifying used persuasive tactics in advertisement, predict promotional effectiveness and examining causal effects of the context. To the best of our knowledge, our work is the first to model context-aware persuasion.

Experiments on a large-scale real-world dataset show that our proposed model CAPE greatly outperforms existing baselines, which not only classifies sentence-level persuasive tactics more accurately, but also better predicts the advertisement's promotion effectiveness. In addition to showing the prediction power of our model, we further demonstrate that it provides insightful guidelines for utilizing persuasive tactics in real-world advertisements by performing causal analysis with large-scale observational data. Our study suggests a number of further directions. In particular, the specially designed disentangled learning generates high-quality and highly interpretable representations, which shed light on persuasion studies using the disentangled representations. In addition, the context-aware persuasion we study also motivates the further development of personalized advertisements that incorporate customers' features as the context in the meantime. Moreover, a particularly promising direction for future work is to maximize advertisement effectiveness by applying persuasion theories in automatic text generation.

Acknowledgement

This work was supported in part by The National Key Research and Development Program of China under grant SQ2018YFB180012, the National Nature Science Foundation of China under 61971267, 61972223, 61861136003, and 61621091, Beijing Natural Science Foundation under L182038, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] T. Zhang, J. Zhang, C. Huo, and W. Ren, "Automatic generation of pattern-controlled product description in e-commerce," in *The World Wide Web Conference*, 2019, pp. 2355–2365.

[2] S. Novgorodov, I. Guy, G. Elad, and K. Radinsky, "Generating product descriptions from user reviews," in *The World Wide Web Conference*, 2019, pp. 1354–1364.

[3] L. Zhao, K. Song, C. Sun, Q. Zhang, X. Huang, and X. Liu, "Review response generation in e-commerce platforms with external product information," in *The World Wide Web Conference*, 2019, pp. 2425–2435.

[4] M. Kokkodis, "Learning from positive and unlabeled amazon reviews: Towards identifying trustworthy reviewers," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 545–546.

[5] E.-J. Lee and S. Y. Shin, "When do consumers buy online product reviews? effects of review quality, product type, and reviewer's photo," *Computers in Human Behavior*, vol. 31, pp. 356–366, 2014.

[6] M. Limayem, M. Khalifa, and A. Frini, "What makes consumers buy from internet? a longitudinal study of online shopping," *IEEE Transactions on systems, man, and Cybernetics-Part A: Systems and Humans*, vol. 30, no. 4, pp. 421–432, 2000.

[7] L. A. Jiang, Z. Yang, and M. Jun, "Measuring consumer perceptions of online shopping convenience," *Journal of Service Management*, vol. 24, no. 2, pp. 191–214, 2013.

[8] N. Archak, A. Ghose, and P. G. Ipeirotis, "Deriving the pricing power of product features by mining consumer reviews," *Management science*, vol. 57, no. 8, pp. 1485–1509, 2011.

[9] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE transactions on knowledge and data engineering*, vol. 23, no. 10, pp. 1498–1512, 2010.

[10] A. Ghose, A. Sundararajan *et al.*, "Evaluating pricing strategy using e-commerce data: Evidence and estimation challenges," *Statistical Science*, vol. 21, no. 2, pp. 131–142, 2006.

[11] D. Jurafsky, V. Chahuneau, B. R. Routledge, and N. A. Smith, "Narrative framing of consumer sentiment in online restaurant reviews," *First Monday*, 2014.

[12] N. Farra, S. Somasundaran, and J. Burstein, "Scoring persuasive essays using opinions and their targets," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, pp. 64–74.

[13] D. Ghosh, A. Khanam, Y. Han, and S. Muresan, "Coarse-grained argumentation features for scoring persuasive essays," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 549–554.

[14] D. Yang and R. E. Kraut, "Persuading teammates to give: Systematic versus heuristic cues for soliciting loans," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–21, 2017.

[15] I. Habernal and I. Gurevych, "Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1589–1599.

[16] E. Simpson and I. Gurevych, "Finding convincing arguments using scalable bayesian preference learning," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 357–371, 2018.

[17] M. Gleize, E. Shnarch, L. Choshen, L. Dankin, G. Moshkowich, R. Aharonov, and N. Slonim, "Are you convinced? choosing the more convincing evidence with a siamese network," *arXiv preprint arXiv:1907.08971*, 2019.

[18] R. Pryzant, Y. Chung, and D. Jurafsky, "Predicting sales from the language of product descriptions." in *eCOM@ SIGIR*, 2017.

[19] S. M. Iacus, G. King, and G. Porro, "Causal inference without balance checking: Coarsened exact matching," *Political analysis*, vol. 20, no. 1, pp. 1–24, 2012.

[20] F. Xu, G. Zhang, Y. Yuan, H. Huang, D. Yang, D. Jin, and Y. Li, "Understanding the invitation acceptance in agent-initiated social e-commerce," 2020.

[21] H. Cao, Z. Chen, F. Xu, T. Wang, Y. Xu, L. Zhang, and Y. Li, "When your friends become sellers: An empirical study of social commerce site beidian," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 83–94.

[22] J. O'Shaugnessy and N. O'Shaughnessy, *Persuasion in advertising*. Routledge, 2003.

[23] D. J. Terry, M. A. Hogg, and K. M. White, "The theory of planned behaviour: self-identity, social identity and group norms," *British journal of social psychology*, vol. 38, no. 3, pp. 225–244, 1999.

[24] N. Ellemers, R. Spears, and B. Doosje, "Self and social identity," *Annual review of psychology*, vol. 53, no. 1, pp. 161–186, 2002.

[25] R. B. Cialdini and R. B. Cialdini, "Influence: The psychology of persuasion," 1993.

[26] C. I. Hovland and W. Weiss, "The influence of source credibility on communication effectiveness," *Public opinion quarterly*, vol. 15, no. 4, pp. 635–650, 1951.

[27] T. C. Brock, "Cognitive restructuring and attitude change." *The Journal of Abnormal and Social Psychology*, vol. 64, no. 4, p. 264, 1962.

[28] A. Fagerstrøm and G. Ghinea, "On the motivating impact of price and online recommendations at the point of online purchase," *International Journal of Information Management*, vol. 31, no. 2, pp. 103–110, 2011.

[29] C. Koyuncu and G. Bhattacharya, "The impacts of quickness, price, payment risk, and delivery issues on on-line shopping," *The Journal of Socio-Economics*, vol. 33, no. 2, pp. 241–251, 2004.

[30] W. K. Ng, G. Yan, and E.-P. Lim, "Heterogeneous product description in electronic commerce," *ACM SIGeCom Exchanges*, vol. 1, no. 1, pp. 7–13, 2000.

[31] O. C. Schultheiss, "An information processing account of implicit motive arousal," *Advances in motivation and achievement*, vol. 12, pp. 1–41, 2001.

[32] G. Packard, A. D. Gershoff, and D. B. Wooten, "When boastful word of mouth helps versus hurts social perceptions and persuasion," *Journal of Consumer Research*, vol. 43, no. 1, pp. 26–43, 2016.

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[35] Y. Song, S. Shi, J. Li, and H. Zhang, "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 175–180.

[36] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[37] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research*, vol. 2, no. Nov, pp. 45–66, 2001.

[38] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin, "Training and testing low-degree polynomial data mappings via linear svm," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1471–1490, 2010.

[39] K. S. Jones, "Index term weighting," *Information storage and retrieval*, vol. 9, no. 11, pp. 619–633, 1973.

[40] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[41] D. Yang, J. Chen, Z. Yang, D. Jurafsky, and E. Hovy, "Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3620–3630.

[42] S. Rendle, "Factorization machines," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 995–1000.

[43] J. Lee, K. Cho, and T. Hofmann, "Fully character-level neural machine translation without explicit segmentation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 365–378, 2017.

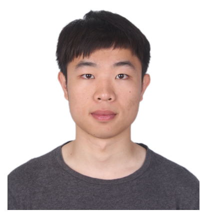
[44] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao, "Weak supervision for fake news detection via reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 516–523.

[45] Y. Tian, Y. Song, F. Xia, T. Zhang, and Y. Wang, "Improving chinese word segmentation with wordhood memory networks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8274–8285.

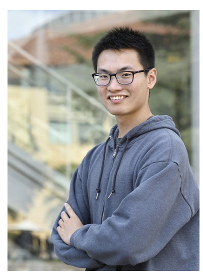
[46] H. Nie, X. Han, B. He, L. Sun, B. Chen, W. Zhang, S. Wu, and H. Kong, "Deep sequence-to-sequence entity matching for heterogeneous entity resolution," in *Proceedings of the 28th ACM*

International Conference on Information and Knowledge Management, 2019, pp. 629–638.
[47] H. Choi and H. Varian, "Predicting the present with google trends," *Economic record*, vol. 88, pp. 2–9, 2012.
[48] L. Shan, L. Lin, C. Sun, and X. Wang, "Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization," *Electronic Commerce Research and Applications*, vol. 16, pp. 30–42, 2016.
[49] X. Xiong, C. Xie, R. Zhao, Y. Li, S. Ju, and M. Jin, "A clickthrough rate prediction algorithm based on users' behaviors," *IEEE Access*, vol. 7, pp. 174 782–174 792, 2019.
[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.
[51] J. Tang, S. Wu, and J. Sun, "Confluence: Conformity influence in large social networks," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 347–355.
[52] P. E. Boksberger and L. Melsen, "Perceived value: a critical examination of definitions, concepts and measures for the service industry," *Journal of services marketing*, 2011.
[53] R. Kashyap and D. C. Bojanic, "A structural analysis of value, quality, and price perceptions of business and leisure travelers," *Journal of travel research*, vol. 39, no. 1, pp. 45–51, 2000.
[54] T. Erdem, J. Swait, and J. Louviere, "The impact of brand credibility on consumer price sensitivity," *International journal of Research in Marketing*, vol. 19, no. 1, pp. 1–19, 2002.
[55] A. Kaul and D. R. Wittink, "Empirical generalizations about the impact of advertising on price sensitivity and price," *Marketing Science*, vol. 14, no. 3_supplement, pp. G151–G160, 1995.
[56] S. Chaiken, "Heuristic versus systematic information processing and the use of source versus message cues in persuasion." *Journal of personality and social psychology*, vol. 39, no. 5, p. 752, 1980.
[57] R. E. Petty and J. T. Cacioppo, "The elaboration likelihood model of persuasion," in *Communication and persuasion*. Springer, 1986, pp. 1–24.
[58] R. E. Petty, D. W. Schumann, S. A. Richman, and A. J. Strathman, "Positive mood and persuasion: Different roles for affect under high-and low-elaboration conditions." *Journal of personality and social psychology*, vol. 64, no. 1, p. 5, 1993.
[59] J. D. Morris, C. Woo, and A. Singh, "Elaboration likelihood model: A missing intrinsic emotional implication," *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 14, no. 1, pp. 79–98, 2005.
[60] R. E. Petty and D. T. Wegener, "Matching versus mismatching attitude functions: Implications for scrutiny of persuasive messages," *Personality and Social Psychology Bulletin*, vol. 24, no. 3, pp. 227–240, 1998.
[61] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[62] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.
[63] R. Pryzant, K. Shen, D. Jurafsky, and S. Wagner, "Deconfounded lexicon induction for interpretable social science," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1615–1625.
[64] H. Yuan, W. Xu, Q. Li, and R. Lau, "Topic sentiment mining for sales performance prediction in e-commerce," *Annals of Operations Research*, vol. 270, no. 1-2, pp. 553–576, 2018.
[65] R. Pryzant, S. Basu, and K. Sone, "Interpretable neural architectures for attributing an ad's performance to its writing style," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 125–135.
[66] B. Youngmann, E. Yom-Tov, R. Gilad-Bachrach, and D. Karmon, "The automated copywriter: Algorithmic rephrasing of health-related advertisements to improve their performance," in *Proceedings of The Web Conference 2020*, 2020, pp. 1366–1377.

Yuan Yuan received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree in the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her research interests include data mining, user behavior modeling, urban computing.



Fengli Xu received the B.Eng. degree from Huazhong University of Science and Technology in 2015, and the Ph.D. degree in Electronic Engineering from Tsinghua University in 2020. He held a postdoc position in Hong Kong University of Science and Technology in 2020. Currently, he is a Mansueto Institute Fellow and Postdoctoral Researcher at the University of Chicago Knowledge Lab. His research interests include data science, computational social science, and urban science.

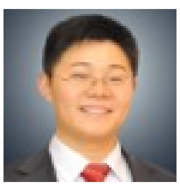


Hancheng Cao received the B.S. degree in electronic information science and technology from Tsinghua University, Beijing, China, in 2018. He is currently a Stanford Interdisciplinary Graduate Fellow and a PhD candidate in Computer Science Department, Stanford University, Stanford, California. His research has received best paper honorable awards at ACM CHI 2021, ACM CSCW 2020, and been covered by various media (e.g. Wired, Forbes). His research interests include data mining, ubiquitous computing, and computational social science.

Guozhen Zhang received the B.S. degree in electronic information science and technology from Tsinghua University, Beijing, China, in 2018, and he is currently working toward the PhD degree in the Department of Electronic Engineering, Tsinghua University, Beijing, China. His research interests include social computing, social network and big data mining.

Pan Hui (SM'14-F'18) received his Ph.D. degree from the Computer Laboratory at University of Cambridge, and both his Bachelor and MPhil degrees from the University of Hong Kong.

He is the Nokia Chair Professor in Data Science and Professor of Computer Science at the University of Helsinki. He is also the director of the HKUST-DT Systems and Media Lab at the Hong Kong University of Science and Technology. He was a senior research scientist and then a Distinguished Scientist for Telekom Innovation Laboratories (T-labs) Germany and an adjunct Professor of social computing and networking at Aalto University. His industrial profile also includes his research at Intel Research Cambridge and Thomson Research Paris. He has published more than 300 research papers and with over 17,500 citations. He has 30 granted and filed European and US patents in the areas of augmented reality, data science, and mobile computing. He has been serving on the organising and technical program committee of numerous top international conferences including ACM SIGCOMM, MobiSys, IEEE Infocom, ICNP, SECON, IJCAI, AAAI, ICWSM and WWW. He is an associate editor for the leading journals IEEE Transactions on Mobile Computing and IEEE Transactions on Cloud Computing. He is an IEEE Fellow, an ACM Distinguished Scientist, and a member of the Academia Europaea.



Yong Li (M'09-SM'16) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Faculty Member of the Department of Electronic Engineering, Tsinghua University.

Dr. Li has served as General Chair, TPC Chair, SPC/TPC Member for several international workshops and conferences, and he is on the editorial board of two IEEE journals. His papers have total citations more than 6900. Among them, ten are ESI Highly Cited Papers in Computer Science, and four receive conference Best Paper (run-up) Awards. He received IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers, Young Talent Program of China Association for Science and Technology, and the National Youth Talent Support Program.

Depeng Jin (M'2009) received his B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999 respectively both in electronics engineering. Now he is an associate professor at Tsinghua University and vice chair of Department of Electronic Engineering. Dr. Jin was awarded National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design and future internet architecture.