# Identification of Key Features for VR Applications with VREVIEW: A Topic Model Approach

Yang Qian*
School of Management,
Hefei University of Technology

YingQiu Xiong†
School of Management,
Hefei University of Technology

Yuyang Wang‡
Arts et Metiers Institute of Technology
LISPEN, HESAM Universite, UBFC

Yuanchun Jiang§
School of Management,
Hefei University of Technology

Yezheng Liu**
School of Management,
Hefei University of Technology

Yidong Chai††
School of Management,
Hefei University of Technology

## Abstract

These are a series of online platforms that allow users to rate and comment on VR virtual reality applications. In this paper, we develop a topic model, namely the general and sparse topic model, that automatically identifies a set of features of VR applications from user reviews. In our context, we overcome two severe challenges (i.e., internal noise and limited features mentioned in each review) to successfully learn the features of VR applications. Specifically, we introduce a general topic and a "spike and slab" prior. In addition, we design a collapsed Gibbs sampling algorithm for model inference. We apply this topic model to a dataset from Oculus (namely VREVIEW), and show that our model can identify some distinct, economically meaningful features for VR applications, e.g., "entertainment and fun," "challenge," "immersive," and "sickness." Our research provides implications for VR consumer behavior analysis, optimizing user experience in virtual environments, and VR application recommendation.

**Keywords**: Virtual Reality, Feature Identification, Topic Model, User Reviews, VR Game.

**Index Terms**: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Computing methodologies—Artificial intelligence—Knowledge representation and reasoning

## 1 Introduction

Recently, the metaverse has gained great attention from both academia and industry with the development of new technologies, e.g., 3D graphics and stereo display technology. Virtual Reality (VR), an important designed technology for metaverse, offers users immersive experiences in the virtual world through the head-mounted display (HMD). In the consumer market, VR applications have become very popular, due to their virtuality, entertainment, and enjoyment. For example, according to the report released by Harvard Business Review[1] in 2021, Roblox[2], an online VR game platform, had 37 million daily active users and 20 million multiplayer games. This company reaches more than $48 billion in market value at its peak, which even exceeds the market value of gaming giant EA (Electronic Arts). Similarly, MMO Zenith[3], a VR game, obtain $10 million in funding in 2021. An interesting phenomenon is that some VR applications after they are released achieve good sales in the market, whereas others attracted much less attention. In this paper, we attempt to analyze the reasons for this phenomenon via mining key features for VR applications.

Many previous studies have focused on the feature extraction of products and can be classified into three categories from the perspective of data: product descriptions, user-item interactions (e.g., purchase records), user-generated contents (e.g., online reviews and hashtags). For example, Toubia et al. [1] integrate positive psychology with the guided Latent Dirichlet Allocation (LDA) model to identify the features of entertainment products (i.e., movies) based on their textual descriptions, automatically. Pan et al. [2] propose a deep learning model for testing product aesthetics based on their images. However, these studies based on product descriptions are unable to obtain consumer perception of the products. In terms of studies based on user-item interactions, Zhang et al. [3] use a matrix factorization model to learn the representations of products and users, and then explore the effects of different consumer consumption strategies on the longitudinal performance of recommender systems. Barkan and Koenigstein [4] propose an item2vec method to produce representations for products, and then use these features to infer product relationships for recommendations. However, these features are implicit and hard to explain, thus leading to that managers can not improve or design their products by using these features. Nowadays, a series of researchers aim to identify interpretable product features commented in reviews. The most commonly used method to obtain the product features is to apply the conventional topic models such as Latent Dirichlet Allocation (LDA) [5, 6]. Unfortunately, these classical topic models yield poor results due to their liberal assumptions that a review is generated from a mixture of multiple topics.

This paper aims to identify key features for VR applications by using online consumer reviews, which has not been explored in previous studies. Identification of key features for VR applications

* e-mail: soberqian@hfut.edu.cn
† e-mail: yqxiong@mail.hfut.edu.cn
‡ e-mail: yuyang.wang@ensam.eu (Corresponding author)
§ e-mail: ycjiang@hfut.edu.cn
** e-mail: liuyezheng@hfut.edu.cn
†† e-mail: chaiyd@hfut.edu.cn
1 https://hbr.org/2021/03/where-brands-are-reaching-gen-z
2 https://www.roblox.com/
3 https://zenithmmo.com/

978-1-6654-8402-2/22/$31.00 ©2022 IEEE
DOI 10.1109/VRW55335.2022.00046

will help optimize user experience in virtual environments [7]. Figure 1 shows the sampled examples of VR applications with reviews in our dataset. In our context, many consumers share their opinions on different aspects of VR applications after using the products purchased from the online store. This provides us with a unique opportunity to explore the VR application features perceived by consumers. Because of factors of consumer reviews, the identification of key features for VR applications faces its own challenge. The first challenge lies in internal noise. Different from the product descriptions, reviews of VR applications include massive general or noise information (e.g., background words and non-feature words) that may lead to poor quality in mining application features. A second challenge is the limited features discussed in each review. For example, a gamer of a VR game often mentions only a few features in their comments such as immersion and cybersickness. This challenge is caused by the sparsity of reviews, which affects the performance of the experiments.

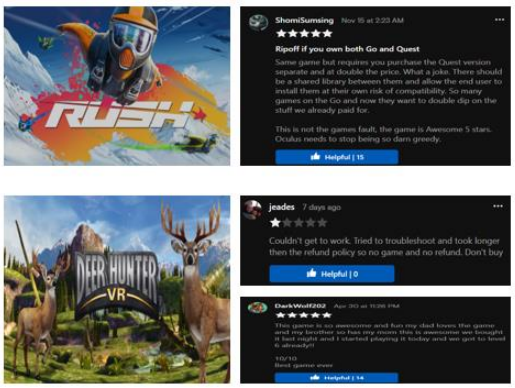



Figure 1: Sampled examples of VR applications with reviews in our dataset. Each VR application includes a set of online comments from multiple consumers.

Motivated by the above challenges, we present a novel topic model, namely the general and sparse topic model (GSTM), to mine key features for VR applications and derive useful insights into VR application design and improvement. In our GSTM, we introduce a general topic to filter out the general or noise information, and select words relevant to VR application features. To achieve this goal, we use a Bernoulli variable to decide whether a word in each review is a feature-related word or a general word. In addition, we assume that each review describes a narrow range of topics related to VR applications. To achieve this goal, in GSTM, we introduce a "spike and slab" prior [7-11] over the review-topic distribution. The "spike and slab" prior is a common method for Bayesian variable selection， which relies on binary variables to control whether the variables in different groups are active or not. In our paper, we use this prior to generate focused topics for each review.

The main contributions of this paper can be summarized as threefold:

(1) To our best knowledge, our research is the first attempt so far to identify a set of key features for VR applications via using large-scale consumer reviews.
(2) We develop a novel general and sparse topic model (GSTM) for feature extraction. The key of GSTM is the introduction of a "spike and slab" prior over the review-topic distribution. Through this process, we allow each review to focus on a limit of topics (features). In addition, we introduce a general topic to filter out the noise information and then obtain the idiosyncrasies of specific reviews.
(3) We present a novel dataset about VR applications. We hope that our dataset will foster research in different fields including VR product design, VR consumer behavior analysis, VR product recommendation, and other machine/deep learning methods for VR applications.

The remainder of this paper proceeds as follows. Section 2 gives the related work. Then, we explain the proposed GSTM model, the generative process, and inference in Section 3. Section 4 describes our dataset in detail and then gives the results of our experiments. Finally, we conclude our paper in Section 5.

## 2 METHOD

In this part, we propose a novel general and sparse topic model (GSTM) for feature extraction of VR applications. We first give problem formulation. Then, we describe the generative process of our model. Finally, we show how to estimate the latent parameters via Gibbs sampling.

### 2.1 Problem Formulation

Suppose there is a collection of user reviews of VR applications, indexed by $d = 1,2,\cdots,D$ . $D$ denotes the number of all user reviews. For the document $d$, we use $\mathbf{w}_d = \{w_{di}\}_{i=1}^{N_d}$ to denote a set of words in this document, where $w_{di}$ is the $i$th word in this document and $N_d$ is the number words in this document. Let $v = 1,2,\cdots,V$ denote the words in the vocabulary. We assume that these are a set of specific topics in VR reviews, indexed by $k = 1,2,\cdots,K$. Similar to the LDA model, we define the $\boldsymbol{\phi}_k$ as a $1 \times$ V vector that denotes the topic-word distribution for specific topic $k$. In addition, to weaken the influence of noise information on feature extraction of VR applications, we define a $\boldsymbol{\varphi}$ as a $1 \times$ V vector that denotes the topic-word distribution for the general topic. To determine topic selection (specific topic or general topic) for the word $w_{di}$ in the document $d$, we also define a binary variable $y_{di}$, where $y_{di} = 1$ means that the word $w_{di}$ is related to the specific topic, whereas $y_{di} = 0$ means that the word $w_{di}$ is related to the general topic.

Different from the LDA model, our model assumes that each review contains a limited of specific topics. Thus, we use a "spike and slab" prior [8, 12] to constrain the topic space for each review. Specifically, we define a specific topic selector $\boldsymbol{c}_d = \{c_{dk}\}_{k=1}^{K}$ as a $1 \times$ K vector that denotes whether a specific topic is related to the review $d$. For each topic $k$, $c_{dk} = 1$ if this topic is related to review $d$; $c_{dk} = 0$, otherwise. We define a topic distribution $\boldsymbol{\theta}_d$ as a $1 \times$ K vector that denotes the review-topic distribution for document $d$. Unlike LDA, $\boldsymbol{\theta}_d$ is constrained by the topic selector $\boldsymbol{c}_d$.

### 2.2 Generative Process

The generative process of the GSTM is given in Figure 2. In our model, we assume that each review of a VR application can be viewed as a mixture of a limited of specific topics and a general topic. To be more specific, we first sample the specific topic $\boldsymbol{\phi}_k$ from a Dirichlet prior with parameter $\beta_0$. And, we sample the general topic $\boldsymbol{\varphi}$ from a Dirichlet prior with parameter $\beta_1$. Then, we sample the Bernoulli parameter μ from a Beta distribution with parameter $\eta_0$ and $\eta_1$. The value of μ characterizes the weight of a word belonging to a specific topic.

For a review $d$, we sample the Bernoulli parameter $\gamma_d$ from a Beta distribution with parameter $\varepsilon_0$ and $\varepsilon_1$ . The value of $\gamma_d$ characterizes the weight of a specific topic selected by the review $d$. Then, we generate a set of binary variables $\boldsymbol{c}_d$ for the specific topic selector, and the element $c_{dk}$ is sampled from Bernoulli distribution with parameter $\gamma_d$. Based on the specific topic selector, we sample the review-topic distribution $\boldsymbol{\theta}_d$ from a Dirichlet distribution with parameter $\alpha_0\boldsymbol{c}_d + \alpha_1\mathbf{1}$. In the "spike and slab," $\alpha_0$ is smoothing prior and $\alpha_1$ the weak smoothing prior. Because the $\alpha_1$ is close to zero and $\alpha_0$ is a routine parameter, thus we see that when $c_{dk} = 1$, the topic $k$ is selected to describe the review $d$, whereas the topic $k$ with $c_{dk} = 0$ do not appear in review $d$.

To generate the word in review $d$, we sample the binary variable $y_{di}$ from a Bernoulli distribution with parameter μ. Now if $y_{di} = 1$, we use the topic distribution $\boldsymbol{\theta}_d$ to pick a specific topic $z_{di}$. Afterward, the word $w_{di}$ is sampled from the topic-word distribution $\boldsymbol{\phi}_{z_{di}}$ using a Multinomial distribution. When $y_{di} = 0$, we use the general topic-word distribution $\boldsymbol{\varphi}$ to generate the word $w_{di}$ by using a Multinomial distribution.

```
Algorithm 1: Generative process of GSTM
for each specific topic k ∈ {1,2,··· ,K} do
  | Draw the specific topic ϕ_k ~ Dirichlet(β_0)
end
Draw the general topic φ ~ Dirichlet(β_1)
Draw weight on specific topic μ ~ Beta(η_0, η_1)
for each review d ∈ {1,2,··· ,D} do
  | Draw weight on the specific topic selector γ_d ~ Beta(ε_0, ε_1)
  | for each specific topic k ∈ {1,2,··· ,K} do
  |   | Draw specific topic selector c_dk ~ Bernoulli(γ_d)
  | end
  | Draw review-topic distribution θ_d ~ Dirichlet(α_0 c_d + α_1 1)
  | for each word i ∈ {1,2,··· ,N_d} do
  |   | Draw the binary variable y_di ~ Bernoulli(μ)
  |   | if y_di = 1 then
  |   |   | Draw the specific topic z_di ~ Multinomial(θ_d)
  |   |   | Draw the word from the specific topic
  |   |   | w_di ~ Multinomial(ϕ_{z_di})
  |   | else
  |   |   | Draw the word from the general topic w_di ~ Multinomial(φ)
  |   | end
  | end
end
```

Figure 2: The generative process of GSTM

## 2.3 Model Inference

Because the exact inference for GSTM is intractable, we thus select a collapsed Gibbs sampling algorithm for approximate estimation. Because of space constraints, we omit the detailed derivation and only give the final formulas used in the collapsed Gibbs sampling algorithm.

(1) Sampling the topic selector $\boldsymbol{c}_d$:

The joint probability distribution of $\boldsymbol{c}_d$ and $\gamma_d$ can be derived as:

$$P(\boldsymbol{c}_d, \gamma_d|\text{rest}) \propto \prod_k P(c_{dk}|\gamma_d)P(\gamma_d|\varepsilon_0, \varepsilon_1) \\ \frac{\mathrm{I}[B_d \in A_d]\Gamma(|A_d|\alpha_0 + K\alpha_1)}{\Gamma\left(n_d^{(*)} + |A_d|\alpha_0 + K\alpha_1\right)} \tag{1}$$

where $A_d = \{k: c_{dk} = 1, k = 1,2,\cdots,K\}$ denotes a set of specific topics discussed by the review $d$. $B_d = \{k: n_d^k > 0, k = 1,2,\cdots,K\}$ denotes the specific topics that have assignments in review $d$. $n_d^k$ denotes the number of words assigned to the specific topic $k$ in the review $d$. $n_d^{(*)}$ denotes the size of review $d$ and $n_d^k$ denotes the number of words assigned to the specific topic $k$ in review $d$. To ensure fast convergence, we integrate out $\gamma_d$ to sample $c_{dk}$:

$$P(c_{dk} = 0|\text{rest}) \propto \left(n_{d,0}^{DC} + \varepsilon_0\right)\frac{I[B_d \in A_d]\Gamma(|A_d|\alpha_0 + K\alpha_1)}{\Gamma\left(n_d^{(*)} + |A_d|\alpha_0 + K\alpha_1\right)} \tag{2}$$

$$P(c_{dk} = 1|\text{rest}) \propto \left(n_{d,1}^{DC} + \varepsilon_1\right)\frac{I[B_d \in A_d]\Gamma(|A_d|\alpha_0 + K\alpha_1)}{\Gamma\left(n_d^{(*)} + |A_d|\alpha_0 + K\alpha_1\right)} \tag{3}$$

where $n_{d,c}^{DC}$ is the times of the specific topic selector $c$ that is assigned to review $d$.

[4] https://www.oculus.com/

(2) Sampling the topic assignment $z_{di}$:

$$P\left(z_{di} = k|\boldsymbol{z}_{-(di)}, w_{di} = v, \mathbf{w}_{-(di)}, \boldsymbol{y}\right) \propto \\ \frac{n_{k,y_{di}=1,-(di)}^v + \beta_0}{n_{k,y_{di}=1,-(di)}^{(*)} + V\beta_0}\left(n_{d,-(di)}^k + c_{dk}\alpha_0 + \alpha_1\right) \tag{4}$$

where $_{-(di)}$ denotes the count excluding $w_{di}$. $n_{k,y_{di}=1}^v$ denotes the number of times word $v$ assigned to the specific topic $k$. $n_{k,y_{di}=1}^{(*)} = \sum_v n_{k,y_{di}=1}^v$ is the total number of words assigned to the specific topic $k$.

(3) Sampling the binary variable $y_{di}$:

$$P\left(y_{di} = 0|w_{di} = v, \mathbf{w}_{-(di)}, \boldsymbol{y}_{-(di)}, \boldsymbol{z}\right) \propto \\ \frac{n_{0,-(di)}^{GV} + \eta_0}{\sum_{y=0}^1 n_{y,-(di)}^{GV} + \eta_0 + \eta_1}\frac{n_{y_{di}=0,-(di)}^v + \beta_1}{n_{y_{di}=0,-(di)}^{(*)} + V\beta_1} \tag{5}$$

$$P\left(y_{di} = 1|w_{di} = v, \boldsymbol{w}_{-(di)}, \boldsymbol{y}_{-(di)}, z_{di} = k, \boldsymbol{z}_{-(di)}\right) \propto \\ \frac{n_{1,-(di)}^{GV} + \eta_1}{\sum_{y=0}^1 n_{y,-(di)}^{GV} + \eta_0 + \eta_1}\frac{n_{k,y_{di}=1,-(di)}^v + \beta_0}{n_{k,y_{di}=1,-(di)}^{(*)} + V\beta_0} \tag{6}$$

where $n_0^{GV}$ denotes the total number of words assigned to the general topic; $n_1^{GV}$ denotes the total number of words assigned to the specific topic; $n_{y_{di}=0}^v$ denotes the number of word $v$ assigned to the general topic; $n_{k,y_{di}=1}^v$ presents the number of word $v$ assigned to the specific topic.

(4) Estimating the latent parameters $\boldsymbol{\theta}_d$, $\boldsymbol{\phi}_k$ and $\boldsymbol{\varphi}$:

$$\theta_{dk} = \frac{n_d^k + c_{dk}\alpha_0 + \alpha_1}{n_d^{(*)} + |A_d|\alpha_0 + K\alpha_1} \tag{7}$$

$$\phi_{kv} = \frac{n_{k,y=1}^v + \beta_0}{n_{k,y=1}^{(*)} + V\beta_0} \tag{8}$$

$$\varphi_v = \frac{n_{y=0}^v + \beta_1}{n_{y=0}^{(*)} + V\beta_1} \tag{9}$$

# 3 RESULTS

In this section, we give extensive experimental results on a real-world dataset about VR games. We first describe the dataset in detail. Then, we investigate the topics and label these topics as VR application features for qualitative analysis. Finally, we present the feature distributions for VR applications.

## 3.1 Data Descriptions

We use Oculus[4] as our empirical application, as it is one of the most representative VR application platforms. We focus on the VR games on this platform because it covers a wide range of categories and user groups. In this platform, only customers who have purchased and experienced VR games can comment, which ensures the authenticity and credibility of the data. To collect the data from Oculus, we design personalized algorithms written by Python.

For the preprocessing for the reviews, we first transform all content into lowercase, then remove the non-English characters and words (e.g., URLs, Chinese). To avoid the influence of irrelevant information on the results, we then eliminate the stopwords and frequent words. Finally, we select the reviews that contain at least 5 words.

Table 1 shows descriptive statistics of our final dataset. In total, our dataset contains 2,845 unique VR games, covering 122,323 consumers and 198,301 reviews. On average, each VR game consists of 69.70 reviews. In addition, this corpus contains a total of 123,056 reviews with “helpful”. On average, each review contains 19.43 words, with a standard deviation of 25.34. For the convenience of researchers, we name the dataset, VREVIEW. The dataset can be downloaded from our Github repository[5].

Table 1: Data description and statistics

| Statistic | Number |
|---|---|
| Number of VR applications | 2,845 |
| Number of users | 122,323 |
| Number of reviews | 198,301 |
| Number of reviews per VR applications | 69.70 |
| Number of reviews with "helpful" | 123,056 |
| Number of words per review | |
| Mean | 19.43 |
| Standard deviation | 25.34 |
| Maximum | 5,371 |

### 3.2 Topic Results

We implement our method using Java 8. We run our experiments on a Linux server with a 2.10 GHz Intel(R) Xeon(R) E5-2620 CPU, 128 GB of memory. In terms of the setting of hyperparameters, we set $\alpha_0 = 50/K$, $\alpha_1 = 10^{-8}$, $\beta_0 = \beta_1 = 0.01$, $\varepsilon_0 = \varepsilon_1 = 1$ and $\eta_0 = \eta_1 = 1$. To decide the optimal topic number $K$, we select the perplexity score [13] as the metric and then use the cross-validation to evaluate the robustness with different choices of $K$. Based on the results of the perplexity score, we set the topic number to 50.

Due to the space limitations, this paper just presents 15 topics. Figure 3 presents the labeled topic learned by our GSTM model. From the topic results, we note that our model can identify distinct, economically meaningful topics. Through our experiments, we find a series of VR features that are often mentioned in consumer reviews. For example, topic 2 is related to "entertainment and fun," with the related semantic words "fun," "enjoy," and "entertain." Topic 6 is related to "immersive," with the related semantic words "immersive," "immersion," and "experience." The frequent appearance of the words "motion," "sickness," "fly," "sick" in topic 44 suggests that this topic relates to the "sickness" of VR application. More importantly, our model can find some additional features of VR applications that are important for VR application design and consumer behavior analysis, e.g., "challenge," "colorful," "chat and social," "music and rhythm," and "usability."

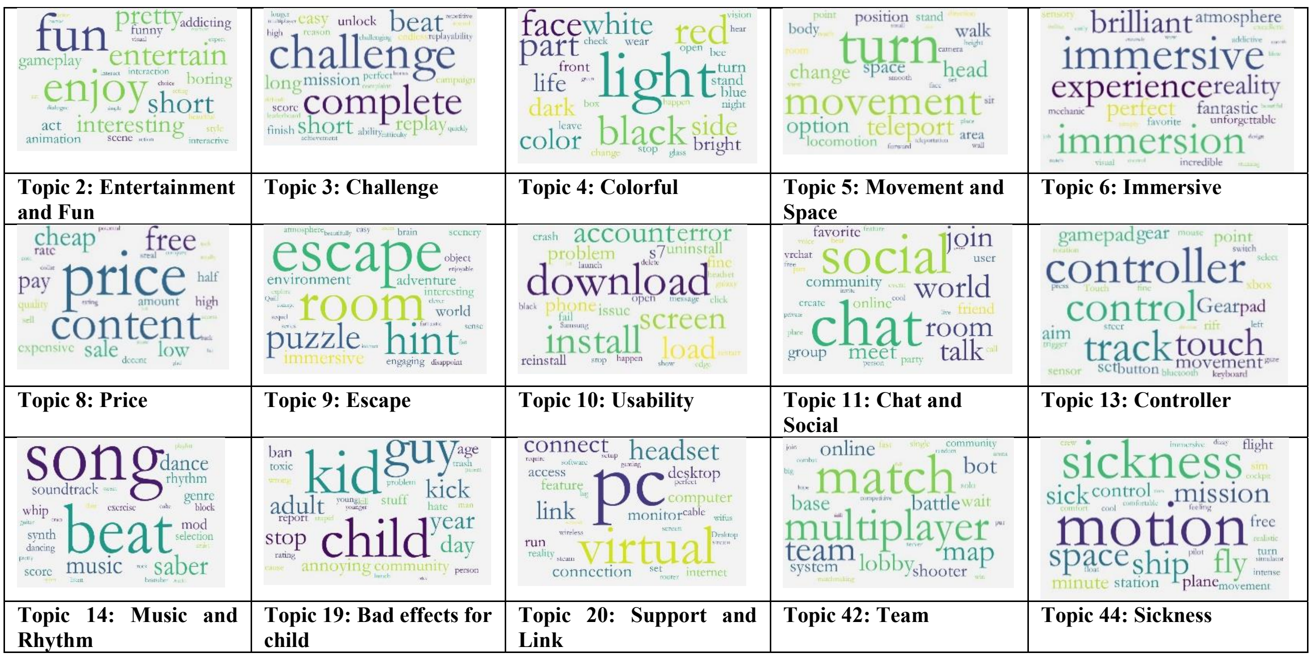
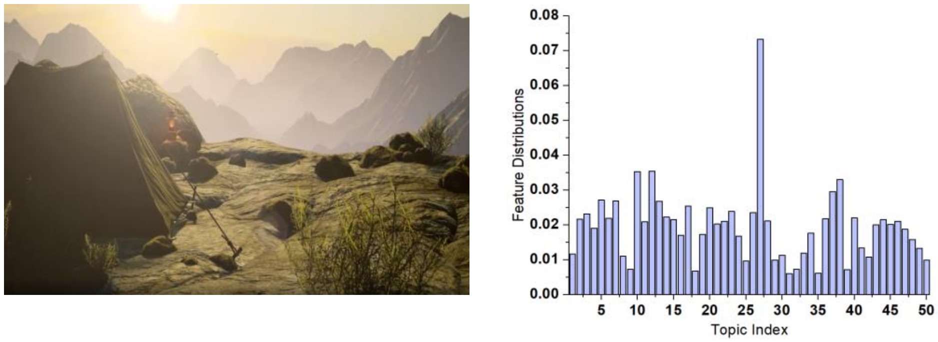
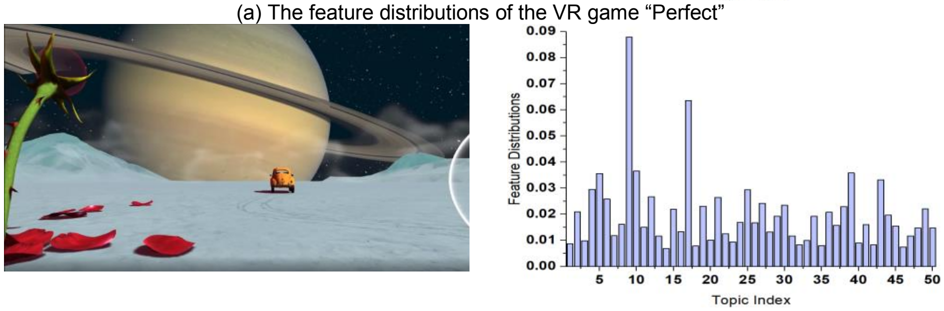



Figure 3: Topic results and its label

[5] https://github.com/YingqiuXiong/VR/tree/main/dataset




(a) The feature distributions of the VR game “Perfect”




(b) The feature distributions of the VR game “Annie Amber”

Figure 5: Feature distributions of two VR games

### 3.3 Feature Distributions for VR Applications

Using our model, we can obtain the review-topic distribution. We use a t-SNE method [14] to visualize these reviews by a 2D map, as shown in Figure 4. Qualitatively, we note that the reviews with similar topic distribution are clustered into the same group, which proves that the features hidden in reviews learned by our model are highly interpretable.

Because each VR application contains a set of reviews, we also use empirical estimation, by integrating the topic distribution of all reviews of a VR application, to analyze the feature distributions for VR applications. Figure 5 gives two cases of VR applications. The VR game “Perfect” in Figure 5 (a), provides beautiful locations for gamers, involving sun-drenched beaches, peaceful mountains, and northern lights. We note that gamers often talk about topic 27 that describes the application feature of “leisure and relaxing.” The VR game “Annie Amber” in Figure 5 (b) is about the journey across the solar system. We find that gamers often talk about topic 9 and topic 17. From Figure 3, we note that topic 9 describes the application feature of “escape.” In addition, topic 17 describes the application feature of “imagine.” Thus, we conclude that “escape” and “imagine” can be regarded as core features of the VR game “Annie Amber.” Our case studies represent that our model can extract VR applications features well.

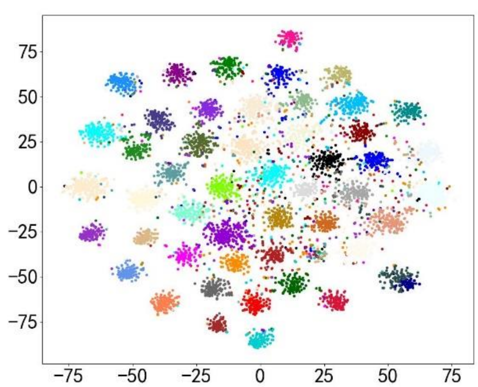



Figure 4: The t-SNE visualizations for reviews

## 4 Conclusion

In this paper, we develop a topic model to automatically identify the key features of VR applications from user reviews. Considering the internal noise and limited features in each review, we introduce a general topic and a “spike and slab” prior to the proposed model. To learn the latent parameters of the model, we also design a collapsed Gibbs sampling algorithm. For model applications, we analyze a VR game dataset including 2,845 unique VR games, 122,323 consumers, and 198,301 reviews. With the experimental results, we find that the proposed model can identify a set of meaningful features related to VR games. We also analyze the feature distributions for VR games by integrating the topic distribution of all reviews of each VR game. In addition, we provide a dataset, namely VREVIEW, for researchers.

For future work, we close by emphasizing additional areas. First, future research might explore the evaluation method to validate our approach for inferring VR application features. Second, our model constructs a set of features, thus future research may use these features as input to analyze consumers’ preferences for VR applications, recommend VR applications for consumers, identify substitutes for a given VR applications.

## ACKNOWLEDGEMENTS

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

## References


[1] Toubia, O., et al., *Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption.* Journal of Marketing Research, 2019. **56**(1): 18-36.

[2] Pan, Y., et al. *Deep design: Product aesthetics for heterogeneous markets*. in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017. p. 1961-1970.

[3] Zhang, J., et al., *Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework.* Information Systems Research, 2020. **31**(1): 76-101.

[4] Barkan, O. and N. Koenigstein. *Item2vec: neural item embedding for collaborative filtering*. in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2016. IEEE. p. 1-6.

[5] Ma, B., et al., *An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews.* Journal of Electronic Commerce Research, 2013. **14**(4): 304.

[6] Guo, Y., S.J. Barnes, and Q. Jia, *Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation.* Tourism management, 2017. **59**: 467-483.

[7] Qian, Y., et al., *Segmenting market structure from multi-channel clickstream data: A novel generative model.* Electronic Commerce Research, 2020. **20**(3): 509-533.

[8] Lin, T., et al. *The dual-sparse topic model: mining focused topics and focused terms in short text*. in *Proceedings of the 23rd international conference on World wide web*. 2014. p. 539-550.

[9] Rakesh, V., et al. *A sparse topic model for extracting aspect-specific summaries from online reviews*. in *Proceedings of the 2018 World Wide Web Conference*. 2018. p. 1573-1582.

[10] Williamson, S., et al. *The IBP compound Dirichlet process and its application to focused topic modeling*. in *ICML*. 2010.

[11] Liu, Y., et al., *Using favorite data to analyze asymmetric competition: Machine learning models.* European Journal of Operational Research, 2020. **287**(2): 600-615.

[12] Wang, C. and D. Blei, *Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process.* Advances in neural information processing systems, 2009. **22**: 1982-1989.

[13] Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* the Journal of machine Learning research, 2003. **3**: 993-1022.

[14] Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(11).