information

Article

Analysis of Unsatisfying User Experiences and Unmet Psychological Needs for Virtual Reality Exergames Using Deep Learning Approach

Xiaoyan Zhang [1], Qiang Yan [1,2,*], Simin Zhou [1], Linye Ma [1] and Siran Wang [1]

[1] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2013213376@bupt.edu.cn (X.Z.); zhousimin@bupt.edu.cn (S.Z.); linye_ma@bupt.edu.cn (L.M.); wangsiran@bupt.edu.cn (S.W.)
[2] School of Modern Post (School of Automation), Beijing University of Posts and Telecommunications, Beijing 100876, China
* Correspondence: yan@bupt.edu.cn

Abstract: The number of consumers playing virtual reality games is booming. To speed up product iteration, the user experience team needs to collect and analyze unsatisfying experiences in time. In this paper, we aim to detect the unsatisfying experiences hidden in online reviews of virtual reality exergames using a deep learning method and find out the unmet psychological needs of users based on self-determination theory. Convolutional neural networks for sentence classification (textCNN) are used in this study to classify online reviews with unsatisfying experiences. For comparison, we set eXtreme gradient boosting (XGBoost) with lexical features as the baseline of machine learning. Term frequency-inverse document frequency (TF-IDF) is used to extract keywords from every set of classified reviews. The micro-F1 score of textCNN classifier is 90.00, which is better than 82.69 of XGBoost. The top 10 keywords of every set of reviews reflect relevant topics of unmet psychological needs. This paper explores the potential problems causing unsatisfying experiences and unmet psychological needs in virtual reality exergames through text mining and makes a supplement for experimental studies about virtual reality exergames.

Keywords: virtual reality; exergame; user experience; online reviews

check for updates

Citation: Zhang, X.; Yan, Q.; Zhou, S.; Ma, L.; Wang, S. Analysis of Unsatisfying User Experiences and Unmet Psychological Needs for Virtual Reality Exergames Using Deep Learning Approach. *Information* **2021**, *12*, 486. https://doi.org/10.3390/info12110486

Academic Editor: Willy Susilo

Received: 30 September 2021
Accepted: 19 November 2021
Published: 22 November 2021

1. Introduction

With the development of virtual reality technology, the number of consumers who play virtual reality games is booming. Referring to the statistics of Valve published lately in 2021, there were 1.7 million new users from SteamVR last year, and the sales of virtual reality games increased by 32% year-over-year [1]. The rapid growth of the virtual reality market has prompted the user experience team of a game to collect more feedback on user experiences and accelerate product iteration.

According to ISO 9241-210:2019, user experience refers to "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" [2]. Psychological need is an important dimension to measure user experience, so unsatisfying experiences may contain unmet psychological needs of users [3–5]. A widely used theory of psychological needs in the game area is the self-determination theory (SDT) by Deci and Ryan [6]. In SDT, needs "are essential for ongoing psychological growth, integrity, and well-being". Deci and Ryan pointed out that users will complete tasks with sufficient intrinsic motivation if their autonomy, competence, and relatedness needs are met [6].

Since SDT is a classical psychological theory, some studies [7,8] in recent years applied experimental methods to explore the relationship between users' psychological needs and the intention to use gamification exercise applications. However, the research of virtual reality exergames based on SDT is still in its infancy. Exergame is a game with exercise goals or tasks. It is also suitable to explore whether users' psychological needs are met by experimental methods. In exergames, users need to move their parts of the body and achieve goals with movements [9]. However, limited by the experimental environment, existing research in experimental scenes may be different from the condition in reality.

Online reviews reflect users' experience of using products in their daily life. Accordingly, the text analysis of online reviews of exergames can help us understand user experiences in the non-experimental state. To the best of our knowledge, only two studies [9,10] analyzed online consumer reviews of virtual reality exergames from e-commerce platforms. Less than 500 consumer reviews in the two qualitative studies were selected for theme analysis.

The main purpose of this paper is to determine in the actual environment what kind of potential problems can lead to unsatisfying experiences and unmet psychological needs of virtual reality exergames. Consumer online reviews in this paper are collected from Steam, a large-scale e-commerce platform selling digital video games. We use textCNN as a deep learning approach to classify reviews with unmet needs. For comparison, we set XGBoost with lexical features as the baseline of machine learning. Therefore, we can enumerate the main research questions that this work addresses:

- RQ1: How are the unmet psychological needs in virtual reality exergames reflected in online reviews? Which design elements of virtual reality exergames lead to that when we consider the extracted keywords?
- RQ2: Are there any differences between the results of consumer online review analysis and experimental results? If so, why do these differences occur?

2. Background

2.1. Evaluation of User Experience from Psychological Needs

User experience is a "dynamic, highly context-dependent, and subjective" account of human–computer interaction [11]. Hassenzahl and Tractinsky pointed out that the research of user experience focuses on satisfying experiences which can improve the cognition level of users and make effective decisions [12]. However, some researchers believe that unsatisfying experiences are also worth studying because unsatisfying experiences can provide suggestions for product iterative development [3]. This paper focuses on unsatisfying experiences hidden in online reviews. Before text mining, we need to consider the following questions: How to measure satisfying/unsatisfying experiences? What dimensions can we refer to? To measure user experiences, a series of research have proposed theories about psychological needs.

SDT proposed by Ryan and Deci has been used extensively in research related to motivation for playing games and exercise. The big three elements of SDT are autonomy, competence, and relatedness. Autonomy means users' sense of controlling over their own choices using the product. Competence refers to users' sense of knowledge and skills required to achieve a goal. In addition, relatedness means users' sense of community and psychological connection with others. When the three needs are met, users will feel good at the psychological level, and they will have sufficient motivation to use the product to achieve the goal [6,13,14]. In other words, when the needs are not satisfied, the users' willingness to play the game will be weakened, which means that unsatisfying experiences are related to unsatisfied needs.

2.2. Virtual Reality Exergames

Different from people who play common video games, virtual reality games players need to wear head mounted displays (HMDs) during games. HMD insulates users from the real world and displays a three-dimensional graph to enhance immersion [15]. HMD also enhances immersion through a high degree of freedom (DoF). DoF describes how an object moves in space. The early HMDs were 3DoF, which can detect the rotation motion of head on X-Y-Z axes to realize head tracking. Oculus Rift, the HMD used in the experiment by Tan et al., is a 3DoF device [15]. Before the popularity of HMD, previous studies have shown that head tracking can bring players pleasure and immersion in game, but it reduces the accuracy of user operation and is not suitable for fast-paced games [16,17]. Based on head tracking, the 6DoF device can detect users' movements on X–Y–Z axes, and realize more interaction modes, such as squatting, which provides technical support for the development of virtual reality exergames.

In exergames, users need to move their parts of the body and achieve goals with movements, do exercise, and have entertainment at the same time [9]. Exergames without HMDs are motion-sensing games. When playing motion-sensing games, users keep a distance from the screen, and the movement is recognized by a camera (e.g., the Kinect of Microsoft) or handled devices (e.g., the Ring-Con of Nintendo). Jennett et al. pointed out that separation from the real world is an important factor of immersion [16]. Therefore, virtual reality exergames are more immersive than motion-sensing games, which is more conducive to users to extend the exercise time. The results of Lee and Kim showed that the body fat percentage of the participants decrease and the density of skeletal muscle increase after continuous virtual reality training, which significantly build a physique [18]. Virtual reality exergames can also improve users' exercise skills [19]. From a positive view, the improvement of exercise levels and skills means that virtual reality exergames may meet the competence need.

2.3. Sentiment Analysis for Online Consumer Reviews

Sentiment analysis is a typical task of text classification to evaluate the sentiment polarity or extract opinions from consumer online reviews [20]. According to the number of classes, the classification tasks related to sentiment analysis can be defined as emotion recognition and polarity detection. Emotion recognition focuses on multi-class labeling, while polarity detection is usually a binary classification task [21].

As for the dimension of sentiment analysis, researchers have proposed emotion models. Shaver et al. [22] made earlier efforts in developing emotion models. They developed an abstract-to-concrete emotion hierarchy and discovered six emotions: joy, love, surprise, sadness, anger, and fear. Another widely used emotion model is the positive and negative affect schedule (PANAS) proposed by Watson et al. [23]. This model consists of 10 psychometric scales for both positive and negative emotions, including scared, hostile, inspired, and proud. Plutchik constructed a wheel-like diagram of emotions to visualize the eight basic emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation [24]. Cambria et al. proposed "The Hourglass of Emotions" model [25]. In this model, emotions are categorized by four dimensions: sensitivity, aptitude, pleasantness, and attention. Each dimension of emotion can be subdivided into six sentic levels.

Machine learning algorithms play an important role in sentiment analysis. Traditional algorithms such as support vector machine (SVM), naive Bayes, and decision tree are common methods for text classification. Ensemble learning algorithms have impressive performance recently, typified by XGBoost [26]. Deep learning algorithms, including recurrent neural networks (RNN), convolutional neural networks (CNN), graph neural networks, and transformers, have also achieved good results in the field of sentiment analysis [20].

In recent years, some new efforts have emerged. Li et al. redefined the formulation of conversational sentiment analysis based on emotional recurrent units and provided a compact structure to better encode the context information [27]. Peng et al. incorporated the phonetic feature of Chinese (i.e., Pinyin) into the reinforcement learning method, which provides a novel perspective for Chinese sentiment analysis [28]. Aiming at the imbalance of data sets, Lin et al. proposed a classification method combining reinforcement learning and deep learning [29]. Ofek et al. enriched SenticNet, a public knowledge base, with domain-level concepts, aspects, and sentiment word pairs. The enriched SenticNet had better performance in sentence-level sentiment classification than the original one [30].

3. Materials and Methods

3.1. Data Sampling and Cleaning

The data processing workflow of this paper is shown in Figure 1. We collected post-purchase consumer reviews from January 2016 to June 2021 from Steam. All the 124 related games were labeled with "virtual reality only" and "sports" on Steam. We screened consumers' language preferences through the platform's review system to make sure that the reviews were written by Chinese consumers.

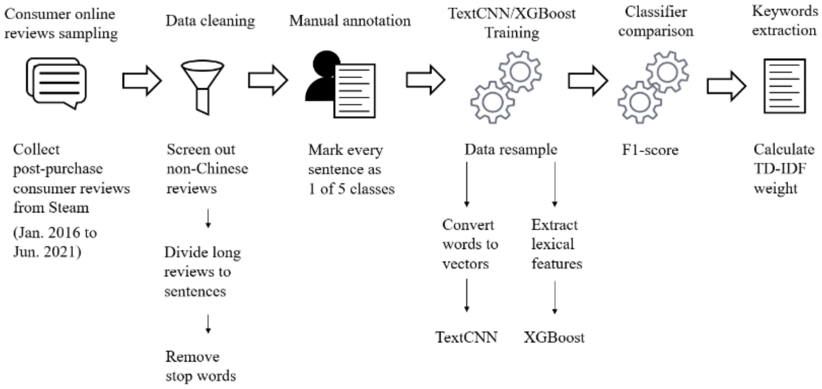


Figure 1. The data processing workflow.

A good pre-processing step can reach better precision in text mining tasks [31]. We first screened the language of all the reviews. Although the language preference of consumers was simplified Chinese, some reviews were written in traditional Chinese, English, or Japanese. We applied Python's langid library to filter reviews most in Simplified or Traditional Chinese. Then we used Python's opencc library to convert traditional Chinese to simplified Chinese. Therefore, the processed comments are mainly in simplified Chinese but may contain some English words, such as Oculus Quest (a kind of virtual reality HMD). Online reviews before and after being pre-processed can be accessed via uploaded Supplementary Materials.

After that, we cut long reviews into sentences. This step is based on two following considerations. Firstly, textCNN is a sentence-oriented convolutional neural network, which is suitable for processing short texts. In addition, in an online review, a user may have multiple unmet psychological needs. Take this comment as an example: "The AI difficulty setting is unreasonable, making me not be willing to continue playing. The online matching experience is poor, and there are few people." The first sentence of the review indicates that the user's competence need is not met, while the other sentence reflects that the relatedness need is not met. If the long review is not segmented, the performance of the classifier may be reduced.

The third step was to remove the stop words. Stop words refer to the words that often appear in the text but are not helpful for information retrieval [32]. To remove stop words in Chinese, we used a stop words list, which is commonly applied in the field of Chinese natural language processing and developed by the Harbin Institute of Technology [33]. After that, we screened out repetitive and blank reviews. Finally, 4507 sentences were left.

3.2. Manual Annotation

In this paper, a deep learning method was used to classify unlabeled data. Since statistical text classifiers work with acceptable accuracy only when given a sufficiently large text input [21], we randomly selected 2000 sentences from the corpus for manual annotation before the automatic classification. The classification task in this study was relevant to emotion recognition [21]. A set of emotion labels described the satisfying or unsatisfying experience hidden in consumer online reviews. As shown in Table 1, each sentence will be labeled as one of the five categories. Two of the authors participated in the manual tagging work. The annotation guideline is shown in Appendix A.

Table 1. Categories and labelling rules.

Category	Label	Examples
Not relevant to unsatisfying experience	0	This is my first review.
Unsatisfying experience not relevant to psychological needs	1	Have bugs and cannot contact to developers.
Unsatisfying experience with unmet autonomy needs	2	Few favorite songs, and not open to custom music editing.
Unsatisfying experience with unmet competence needs	3	After repeatedly squatting and standing up for a period of time, players will soon feel the pain in their legs, resulting in the failure of the game.
Unsatisfying experience with unmet relatedness needs	4	Hope you can release multi-player mode in the future.

The unsatisfying experiences were determined by the results of Partala and Kallinen's study [3], including the lack of guidance of operation, unattractive interaction interface, incorrect use of product functions, or arousing negative emotions. The judgment of autonomy, competence, or relatedness need was according to [6–8]. The annotation results showed that 618 (30.9%) sentences in the annotation set contained unsatisfying experiences. In the 618 sentences, 300 of them are with unmet autonomy needs, 133 with unmet competence needs, and 28 with unmet relatedness needs. Because the annotation dataset was unbalanced, we resampled it to obtain the same proportion of five types of annotation data. The resampled dataset included 6910 sentences, and then we reordered the data randomly. The resampled and reordered dataset can be accessed via uploaded Supplementary Materials.

3.3. Classifiers

3.3.1. TextCNN

TextCNN combines n-gram language model with convolutional neural network, which is a representative model for using convolutional neural network in short text natural language processing tasks [34]. The basic principle of textCNN is shown in Figure 2. First, the input text is transformed into word vectors, and then the context features in multiple dimensions are extracted through different convolution layer kernel sizes. After that, the extracted word vectors are strengthened through the maximum pooling layer, and the output of the max-pooling layer is Softmax classified. With the above processing, the ability of text feature extraction can be strengthened, so as to improve the effect of text classification.

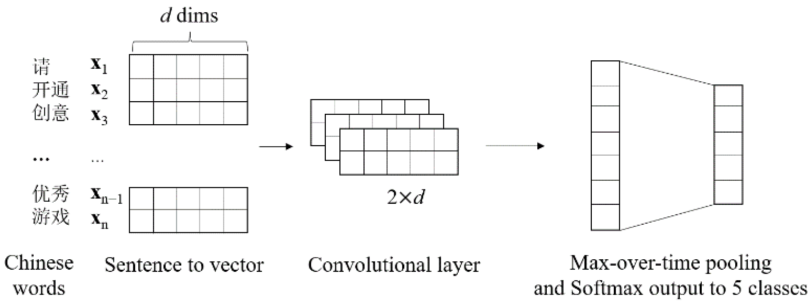


Figure 2. The structure of textCNN.

Specifically, the input of the convolution layer is a sentence $x_{1:n}$ with length n can be expressed as

$$\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \ldots \oplus \mathbf{x}_n, \quad (1)$$

where $\mathbf{x}_{\mathrm{i}}$ is the word vector of the ith word in text, and $\oplus$ is conjunction operator. In this paper, we adopted 2-g model and set the feature dimension as 128, so the size of convolution kernel is 2 × 128. Then, the text local features are extracted according to the following formula:

$$c_i = f(\mathbf{w}\cdot\mathbf{x}_{i:\mathrm{i+h-1}} + \mathrm{b}) \tag{2}$$

where w is the parameter of kernel size, h is the height of kernel, and b is bias. The activation function is rectified linear unit (ReLU).

c, the output of the convolution layer inputs through the pool layer to obtain the enhanced feature $\hat{c}$ (as shown in Formula (3)):

$$\hat{c} = max\{\mathrm{c}\} = max\{[\mathrm{c}_1, \mathrm{c}_2, \ldots, \mathrm{c}_{\mathrm{n-h+1}}]\} \tag{3}$$

The parameters of textCNN in this paper is shown in Table 2. Furthermore, the epoch in training is 50.

Table 2. TextCNN parameters from torch.nn library.

Parameters	Value	Description
in_channels	1	Number of information input channels of convolution layer.
out_channels	9	Number of information output channels of convolution layer.
kernel_size	2 × 128	Size of convolution layer kernel.
stride	1	Steps per move.
batch_size	512	Number of samples used for each training.
learning_rate	0.001	The learning speed of the model.
dropout	0.5	The probability of abandoning the activation of neurons (to avoid over fitting).

3.3.2. XGBoost

XGBoost is an improved algorithm based on gradient boosting decision tree. There are a series of decision trees in the model, and the sum of the prediction results of each tree is the final prediction result of the model.

The objective function of XGBoost is:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{4}$$

Here w is the score in the corresponding leaves, l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. The second term Ω penalizes the complexity of the model (i.e., the regression tree functions).

The second-order approximation is carried out for Formula (4) to optimize the objective and the optimal prediction score of each leaf node can be obtained, then calculate the corresponding optimal value $\widetilde{\mathcal{L}}^{(t)}(q)$ [26]. When splitting decision tree nodes, XGBoost uses the exact greedy algorithm. In order to improve the calculation speed, XGBoost also supports global variant and local variant approximate algorithm.

Because XGBoost can effectively deal with sparse data (such as one-hot code), we use the lexical feature of sentences as the input of XGBoost. Lexical feature, which means the distribution of frequency of top N words in the corpus, is a kind of common text classification feature [35]. In this study, we set N = 1000. The top 1000 words with high frequency can be accessed via uploaded Supplementary Materials.

The parameters of XGBoost in this paper is shown in Table 3. In addition, the number of iterations for each round of cross-validation is 50.

Table 3. XGBoost parameters from xgboost library.

Parameters	Value	Description
objective	multi:softmax	Multiclass classification using the softmax objective.
gamma	0.1	Minimum loss reduction required to make a further partition on a leaf node of the tree.
max_depth	6	Maximum depth of a tree.
learning_rate (eta)	0.1	Step size shrinkage used in update to prevents overfitting.
subsample	0.5	Subsample ratio of the training instance. Setting it to 0.5 means that XGBoost randomly collected half of the data instances to grow trees and this will prevent overfitting.
min_child_weight	1	Minimum sum of instance weight needed in a child.

3.4. Classifiers Evaluation

F1-score is used to evaluate the performance of a classifier. Comparing the results of the classifier on the test set with manually labeled data, we can obtain the number of samples of true positive (TP), false positive (FP), and false negative (FN). Thus, we can measure the classification effect by following indexes:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{5}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{6}$$

$$\text{F1-score} = 2 \times \text{Precision} \times \text{Recall}/(\text{Precision} + \text{Recall}) \tag{7}$$

The above indexes are the evaluation of the binary classification task, while micro or macro-F1 score can be considered for the multi-classification task. From the uneven distribution of the annotation dataset, the classification task in this paper is applicable to micro-F1 measurement, and indexes are:

$$\text{Precision}_{mi} = \sum \text{TP}_i / \sum (\text{TP}_i + \text{FP}_i) \tag{8}$$

$$\text{Recall}_{mi} = \sum \text{TP}_i / \sum (\text{TP}_i + \text{FN}_i) \tag{9}$$

$$\text{micro F1-score} = 2 \times \text{Precision}_m \times \text{Recall}_{mi} / (\text{Precision}_{mi} + \text{Recall}_{mi}) \tag{10}$$

We applied the 5-fold cross-validation to evaluate the training performance of the classifier. In Section 3.2, we mentioned that we resampled the data and obtained 6910 randomly sorted sentences. Four parts of data (i.e., 4952 sentences) were used for training and one (i.e., 1238 sentences) for testing in every evaluation. $\overline{\text{micro-F1}}$, which is the average of micro-F1 in all five rounds of training, was regarded as the evaluation index. The higher value of $\overline{\text{micro-F1}}$ means the better performance of the classifier.

3.5. Keywords Extraction by TF-IDF

TF-IDF can effectively reduce the dimension of text and extract keywords from it. The principle of TF-IDF is that the higher the frequency of a word in a specific document, the higher the discrimination of the word. On the contrary, it is difficult for the words appearing in most documents to show the characteristics of the subject and make a low contribution to the keyword extraction task. A word with high weight w_{ij} can be regard as a keyword, and the formula is:

$$\text{w}_{ij} = \text{tf}_{ij} \times \log(\text{N}/\text{n}_j) \tag{11}$$

Here tf_{ij} means the frequency of characteristic item t_j in document d_i. N refers to the sum of all documents, and n_j is the number of documents including item t_j.

Jieba, a Python library that is commonly used in Chinese natural language processing tasks, is used to complete word tokenization and TF-IDF weight calculation. As we stated in Section 3.2, the number of sentences in each category was imbalanced. If we calculated TF-IDF weight for all the documents (i.e., sentences), we could not determine the importance of a keyword for a certain class. Therefore, we extracted the top 10 keywords of each category of sentences. The results of keyword extraction reflect the topic information of the text. Generally, only nouns are regarded as keywords [36].

4. Results

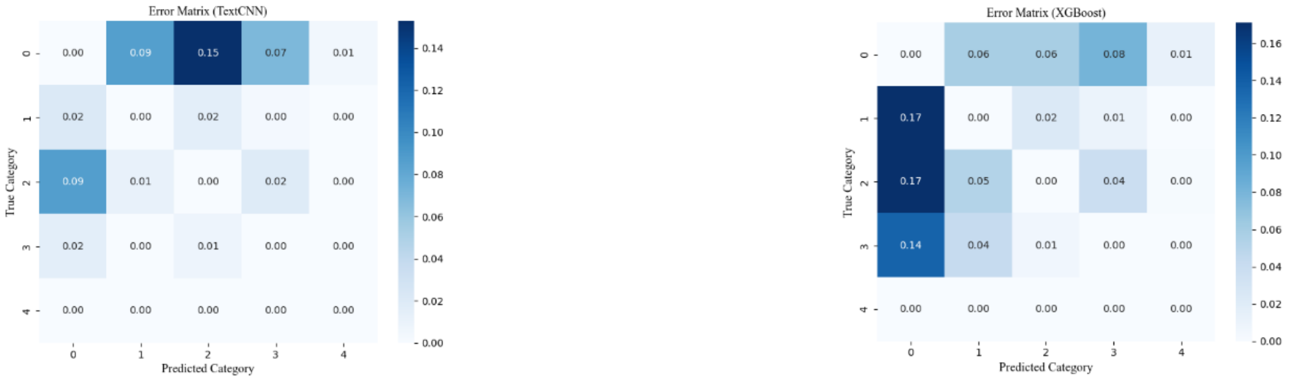
4.1. Classifier Performance and Selection

The $\overline{\text{micro-F1}}$ scores of the two classifiers are shown in Table 4.

Table 4. Classifier performance of textCNN and XGBoost.

Micro-F1 Score (%) in Each Cross-Validation	1st	2nd	3rd	4th	5th	$\overline{\text{micro-F1}}$
textCNN	91.90	87.99	89.94	86.03	94.13	90.00
XGBoost	82.85	81.11	84.37	82.49	82.63	82.69

The error matrix of textCNN and XGBoost are shown in Figure 3. The error matrix is calculated based on the comprehensive results of five cross-validation processes. It can be seen that textCNN has better classification performance. In addition, XGBoost is more likely to misclassify sentences containing unmet needs to category 0. Therefore, we applied textCNN to classify the unlabeled online consumer reviews.



(**a**) (**b**)

Figure 3. Error matrix (**a**) of textCNN; (**b**) of XGBoost.

After merging with the annotated text, we obtained a total of 1479 (32.8%) sentences with unmet autonomy, competence or relatedness needs. The distribution of different categories in the annotation set and the overall dataset is shown in Figure 4. The proportion of Class 2 in the whole data set is greater than that in the annotation set because textCNN could predict sentences of Class 0 as Class 2. The annotation and prediction results of all the 4507 sentences can be accessed via uploaded Supplementary Materials.

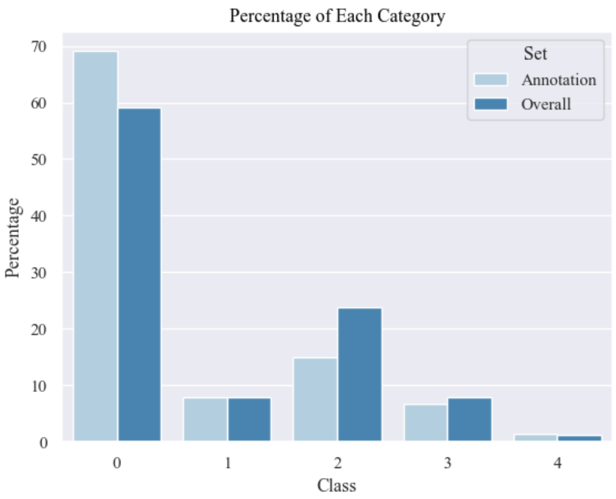


Figure 4. The distribution of five categories in different datasets.

4.2. Keywords Extraction Results

For sentences with unsatisfying experiences, the keywords extracted by TF-IDF are shown in Table 5. We also marked the words which are the only keywords of a certain category. The keywords translated from Chinese to English are determined by a volunteer with a Master's Degree in Teaching Chinese to Speakers of Other Languages.

Table 5. Keywords of sentences related to the unsatisfying experience.

Class 1: Unsatisfying Experience without Unmet Psychological Needs		**Class 2: Unmet Autonomy Needs**		**Class 3: Unmet Competence Needs**		**Class 4: Unmet Relatedness Needs**	
Word	TF-IDF	Word	TF-IDF	Word	TF-IDF	Word	TF-IDF
game	1.6613	game	0.9641	game	0.8371	**on-line cooperation**	1.2314
experience	0.1874	**song**	0.3099	feeling	0.1546	game	0.8912
controller	0.1445	**workshop**	0.1806	experience	0.1527	mode	0.6977
player	0.1388	player	0.1643	player	0.1440	experience	0.2457
feeling	0.1264	controller	0.1451	**hour**	0.1377	**skin**	0.2148
frame	0.1092	music	0.1381	**arm**	0.1291	official	0.1744
problem	0.0944	experience	0.1278	frame	0.0979	**foreigner**	0.1720
hot-air balloon	0.0912	mode	0.1272	**sensation**	0.0848	**human and computer**	0.1705
gameplay	0.0872	official	0.1264	music	0.0832	**social activity**	0.1547
interface	0.0857	**music game**	0.0841	mode	0.0826	**function**	0.1367

5. Discussion

When collecting user experiences on exergames or gamification exercise platforms, it is rare to use consumer online reviews while experimental or qualitative methods are very common. We compared the performance of textCNN and XGBoost, the two state-of-the-art models for natural language processing tasks. In addition, we selected textCNN as the classifier for unsatisfying experiences and unmet psychological needs discovery.

5.1. Research Findings

As shown in Table 5, although Category 1 represents unsatisfying experiences, it is unrelated to psychological needs. It can be seen that words such as "problem", "gameplay" and "interface" reflect the interaction problems caused by bugs or improper game design, resulting in unsatisfying experiences. In addition, we checked sentences containing "hot-air balloon" and found that when players were running the game, they often became stuck near the hot-air balloon and could not move.

Categories 2–4 are related to psychological needs. The feature of unmet autonomy needs is that its keywords are relevant to exergames concerning music. "Workshop" probably refers to the workshop part of Steam. Users can upload their own game content in the workshop. Taking music exergames as an example, the contents uploaded by users are mainly songs and corresponding action sequence files. It can be seen that if the customized function of a music exergame, such as a workshop, is not open to users, the user's autonomy needs may not be met. We also noticed that many sentences of Category 2 point to *Beat Saber*, a popular music exergame. This is a representative case. *Beat Saber* did not offer the workshop but released the licensed songs in the form of paid downloadable content. Although the workshop could attract many players, the founder of *Beat Saber* was worried about the copyright risks brought by it [37]. The current study shows that customized avatars and running routes meet the autonomy needs of users [8]. In this paper, the extracted keywords such as "song" and "workshop" are also related to the customized function, which means providing the customized function is an effective way to meet the autonomy needs of virtual reality exergame users.

Keywords of Category 3 are relevant to the exercise task of virtual reality exergames. "Arm" and "sensation" indicate that the user needs to move his/her body parts to complete the exercise task, while "hour" reflects the duration of the task. When the user's fitness does not match the exercise task (e.g., they feel tired), their competence needs cannot be met. In order to understand the specific expression of this mismatch, we looked at the sentences marked as Category 3 and found that it was difficult for users to aware the passage of time in the game. They may feel that the intense workout is beyond their tolerance, but the immersion experience brought by virtual reality makes them reluctant to quit the game. Finally, only when they take off the HMD, they know that several hours have passed in the real world, feeling very tired. The distraction from thinking about physical activity through immersion is consistent with the research of Faric et al. [9]. Experimental studies [7,8] did not reflect that, because the user's exercise duration (less than 1 h) in the experiment is much less than that in daily playing, the stamina consumption is also less.

For relatedness needs, the extracted keywords (e.g., on-line cooperation, foreigner, social activity) reveal that virtual reality exergame users hope to compete and communicate with other players, but it is difficult to find fixed partners. Our results are consistent with the research of Faric et al. [9]. Users also think they are alone in the virtual world because some games have poor online matching functions. The experimental study [8] assigned "runner-spectator" partners, so there is no lack of mismatch in the experiment.

5.2. Limitations and Directions for Future Research

There are several limitations of this research, which can be overcome in future work. Firstly, the sample is limited to a small dataset, because we selected consumer reviews from only one country. Although our sample size is larger than that of relevant studies [9,10], it is not enough to fully reflect the unsatisfying experiences of virtual reality exergames. Secondly, the heterogeneous nature of human language also introduced noise into text mining. In the future, a larger range of virtual reality games and applications might be considered. Furthermore, we need to collect online reviews in different languages and design the text mining process for multiple language styles of consumer reviews. Moreover, a series of newly excellent methods for classification tasks, such as deep reinforcement learning [28] and domain knowledge incorporated text mining [30], need to be considered for future research.

6. Conclusions

Qualitative or experimental research are often limited by the number of samples or experimental design, so it is difficult to obtain comprehensive results. This study provides a swift way to extract acceptable classification results by labeling consumer reviews less than half of that in the corpus. To the best of our knowledge, it is the first time that a text mining method is proposed to analyze unsatisfied user experiences and unmet psychological needs from virtual reality exergame users' online reviews.

This study has theoretical contributions. SDT has been widely used in the research of gamification exercise platforms [7,8]. However, there is no research on using SDT to explore virtual reality exergames with stronger immersion. This paper finds that compared with gamification exercise platforms, virtual reality exergame users have a longer playing duration, and it is easy for the intense workout to exceed the range that users can tolerate, which brings unmet competence needs. This is caused by the stronger immersion of virtual reality exergames. In virtual reality exergames, it is difficult for users to be aware of the passage of time. In addition, in the real environment, virtual reality exergame users are more difficult to find online matching opponents or partners, bringing dissatisfying relatedness needs. The experimental study of Tsai et al. [8] assigned partners to participants in each group, so that participants can fully engage in the social interaction process in exercise. This paper provides a novel perspective for the comparative study of virtual reality exergames and gamification exercise applications. The text mining method we applied can also be used to extract user experiences in other fields.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/info12110486/s1.

Author Contributions: Conceptualization, X.Z. and Q.Y.; methodology, X.Z.; software, X.Z.; validation, X.Z.; investigation, X.Z. and S.Z.; resources, X.Z.; data curation, X.Z.; writing—original draft preparation, X.Z. and S.W.; writing—review and editing, X.Z., S.Z., L.M. and S.W.; visualization, X.Z.; supervision, Q.Y.; project administration, Q.Y.; funding acquisition, Q.Y. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Social Science Foundation of China, Grant Number 17AGL026, and BUPT Excellent Ph.D. Students Foundation, Grant Number CX2019128.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors thank Z.L. from Shanghai Pinghe Bilingual School for his checking and deciding the keywords translation in Table 5.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

A.1. The Annotation Guideline

Definitions

- Unsatisfying experiences: users feel upset, hostile, ashamed, distressed, irritable, scared, or guilty when or after playing the virtual reality exergame. These emotions can be perceived in online reviews.
- Autonomy: users' sense of control over their own choices using the product.
- Competence: users' sense of knowledge and skills required to achieve a goal, or feeling capable, effective in users' actions.
- Relatedness: users' sense of community and psychological connection with others. Annotation examples are shown in Table A1.

Table A1. Annotation examples.

Category	Label	Examples
Not relevant to unsatisfying experiences	0	Good game/this is my first review/worthy to buy.
Unsatisfying experiences not relevant to psychological needs	1	The game is running on SteamVR, but not on my HMD/have bugs and cannot contact to developers/why my license expired?
Unsatisfying experiences with unmet autonomy needs	2	Few favorite songs, and not open to custom music editing/we need mods!
Unsatisfying experiences with unmet competence needs	3	After repeatedly squatting and standing up for a period of time, players will soon feel the pain in their legs, resulting in the failure of the game.
Unsatisfying experiences with unmet relatedness needs	4	Hope you can release multi-player mode in the future/no one in online game.

References

1. SteamVR Logged 104m Sessions and 1.7m New VR Users in 2020. Available online: https://www.vrfocus.com/2021/01/steamvr-logged-104m-sessions-and-1-7m-new-vr-users-in-2020 (accessed on 15 January 2021).
2. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part 11: Guidance on Usability. Available online: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en (accessed on 13 September 2021).
3. Partala, T.; Kallinen, A. Understanding the most satisfying and unsatisfying user experiences: Emotions, psychological needs, and context. *Interact. Comput.* **2012**, *24*, 25–34. [CrossRef]
4. Partala, T.; Saari, T. Understanding the most influential user experiences in successful and unsuccessful technology adoptions. *Comput. Hum. Behav.* **2015**, *53*, 381–395. [CrossRef]
5. de Saenz-Urturi, Z.; Zapirain, B.G.; Zorrilla, A.M. Elderly user experience to improve a Kinect-based game playability. *Behav. Inf. Technol.* **2015**, *34*, 1040–1051. [CrossRef]
6. Deci, E.; Ryan, R.M. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychol. Inq.* **2000**, *11*, 227–268. [CrossRef]
7. Ijaz, K.; Ahmadpour, N.; Wang, Y.; Calvo, R.A. Player Experience of Needs Satisfaction (PENS) in an immersive virtual reality exercise platform describes motivation and enjoyment. *Int. J. Hum. Comput. Interact.* **2020**, *36*, 1195–1204. [CrossRef]
8. Tsai, T.-H.; Chang, Y.-S.; Chang, H.-T.; Lin, Y.-W. Running on a social exercise platform: Applying self-determination theory to increase motivation to participate in a sporting event. *Comput. Hum. Behav.* **2020**, *114*, 106523. [CrossRef]
9. Faric, N.; Potts, H.W.W.; Hon, A.; Smith, L.; Newby, K.; Steptoe, A.; Fisher, A. What Players of Virtual Reality Exercise Games Want: Thematic Analysis of Web-Based Reviews. *J. Med. Internet Res.* **2019**, *21*, e13833. [CrossRef]
10. McMichael, L.; Faric, N.; Newby, K.; Potts, H.W.W.; Hon, A.; Smith, L.; Steptoe, A.; Fisher, A. Parents of adolescents perspectives of physical activity, gaming and virtual reality: Qualitative study. *JMIR Serious Games* **2020**, *8*, e14920. [CrossRef] [PubMed]
11. Hassenzahl, M.; Diefenbach, S.; Göritz, A. Needs, affect, and interactive products—Facets of user experience. *Interact. Comput.* **2010**, *22*, 353–362. [CrossRef]
12. Hassenzahl, M.; Tractinsky, N. User experience—A research agenda. *Behav. Inf. Technol.* **2006**, *25*, 91–97. [CrossRef]
13. Ryan, R.M.; Deci, E.L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* **2000**, *55*, 68–78. [CrossRef]
14. Reis, H.T.; Sheldon, K.M.; Gable, S.L.; Roscoe, J.; Ryan, R.M. Daily Well-Being: The Role of Autonomy, Competence, and Relatedness. *Pers. Soc. Psychol. Bull.* **2000**, *26*, 419–435. [CrossRef]
15. Tan, C.T.; Leong, T.W.; Shen, S.; Dubravs, C.; Si, C. Exploring Gameplay Experiences on the Oculus Rift. In Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play, London, UK, 5–7 October 2015; pp. 253–263.
16. Jennett, C.; Cox, A.L.; Cairns, P.; Dhoparee, S.; Epps, A.; Tijs, T.; Walton, A. Measuring and defining the experience of immersion in games. *Int. J. Hum. Comput. Stud.* **2008**, *66*, 641–661. [CrossRef]
17. Ilves, M.; Gizatdinova, Y.; Surakka, V.; Vankka, E. Head movement and facial expressions as game input. *Entertain. Comput.* **2014**, *5*, 147–156. [CrossRef]
18. Lee, H.T.; Kim, Y.S. The effect of sports VR training for improving human body composition. *EURASIP J. Image Video Process.* **2018**, *2018*, 148. [CrossRef]
19. Michalski, S.C.; Szpak, A.; Saredakis, D.; Ross, T.; Billinghurst, M.; Loetscher, T. Getting your game on: Using virtual reality to improve real table tennis skills. *PLoS ONE* **2019**, *14*, e0222351. [CrossRef] [PubMed]
20. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning–based text classification: A comprehensive review. *ACM Comput. Surv.* **2021**, *54*, 62. [CrossRef]
21. Cambria, E. Affective Computing and Sentiment Analysis. In *IEEE Intelligent Systems*; IEEE: Piscataway, NJ, USA, 2016; Volume 31, pp. 102–107. [CrossRef]

22. Shaver, P.; Schwartz, J.; Kirson, D.; O'Connor, C. Emotion knowledge: Further exploration of a prototype approach. *J. Pers. Soc. Psychol.* **1987**, *52*, 1061–1086. [CrossRef]
23. Watson, D.; Tellegen, A.; Clark, L. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Pers. Soc. Psychol.* **1988**, *54*, 1063–1070. [CrossRef]
24. Plutchik, R. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* **2001**, *89*, 344–350. [CrossRef]
25. Cambria, E.; Livingstone, A.; Hussain, A. The Hourglass of Emotions. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 144–157. [CrossRef]
26. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data, San Francisco, CA, USA, 13–17 August 2016.
27. Li, W.; Shao, W.; Ji, S.; Cambria, E. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing* **2022**, *467*, 73–82. [CrossRef]
28. Peng, H.; Ma, Y.; Poria, S.; Li, Y.; Cambria, E. Phonetic-enriched text representation for Chinese sentiment analysis with reinforcement learning. *Inf. Fusion* **2021**, *70*, 88–99. [CrossRef]
29. Lin, E.; Chen, Q.; Qi, X. Deep reinforcement learning for imbalanced classification. *Appl. Intell.* **2020**, *50*, 2488–2502. [CrossRef]
30. Ofek, N.; Poria, S.; Rokach, L.; Cambria, E.; Hussain, A.; Shabtai, A. Unsupervised Commonsense Knowledge Enrichment for Domain-Specific Sentiment Analysis. *Cogn. Comput.* **2016**, *8*, 467–477. [CrossRef]
31. Eler, D.M.; Grosa, D.; Pola, I.; Garcia, R.; Correia, R.; Teixeira, J. Analysis of Document Pre-Processing Effects in Text and Opinion Mining. *Information* **2018**, *9*, 100. [CrossRef]
32. Lo, T.W.; He, B.; Ounis, I. Automatically building a stopword list for an information retrieval system. *J. Digit. Inf. Manag.* **2005**, *3*, 3–8.
33. Che, W.; Li, Z.; Liu, T. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstrations, Beijing, China, 23–27 August 2010; pp. 13–16.
34. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1746–1751.
35. Wang, Y.; Dang, Y.; Xu, Z. Mining automobile quality problems based on the characteristics of forum data. *Chin. J. Manage. Sci.* **2021**, *29*, 201–212.
36. Wang, J.; Zhao, Z.; Liu, Y.; Guo, Y. Research on the Role of Influencing Factors on Hotel Customer Satisfaction Based on BP Neural Network and Text Mining. *Information* **2021**, *12*, 99. [CrossRef]
37. Beat Saber CEO Talks Hacks, Mods and Getting Artists Paid. Available online: https://uploadvr.com/gdc-beat-saber-ceo-mods/ (accessed on 25 March 2019).