information

Article

A Data-Driven Approach for Video Game Playability Analysis Based on Players' Reviews

Xiaozhou Li *, **Zheying Zhang** and **Kostas Stefanidis**

Faculty of Information Technology and Communication Sciences, Tampere University, Kalevantie 4, 33100 Tampere, Finland; zheying.zhang@tuni.fi (Z.Z.); konstantinos.stefanidis@tuni.fi (K.S.)
* Correspondence: xiaozhou.li@tuni.fi

Abstract: Playability is a key concept in game studies defining the overall quality of video games. Although its definition and frameworks are widely studied, methods to analyze and evaluate the playability of video games are still limited. Using heuristics for playability evaluation has long been the mainstream with its usefulness in detecting playability issues during game development well acknowledged. However, such a method falls short in evaluating the overall playability of video games as published software products and understanding the genuine needs of players. Thus, this paper proposes an approach to analyze the playability of video games by mining a large number of players' opinions from their reviews. Guided by the game-as-system definition of playability, the approach is a data mining pipeline where sentiment analysis, binary classification, multi-label text classification, and topic modeling are sequentially performed. We also conducted a case study on a particular video game product with its 99,993 player reviews on the Steam platform. The results show that such a review-data-driven method can effectively evaluate the perceived quality of video games and enumerate their merits and defects in terms of playability.

Keywords: playability; player reviews; text classification; sentiment analysis; topic modeling; steam

check for updates

Citation: Li, X.; Zhang, Z.; Stefanidis, K. A Data-Driven Approach for Video Game Playability Analysis Based on Players' Reviews. *Information* **2021**, *12*, 129. https://doi.org/10.3390/info12030129

Academic Editor: Vincenzo Moscato

Received: 27 February 2021
Accepted: 15 March 2021
Published: 17 March 2021

1. Introduction

Playability has been widely acknowledged as the key concept reflecting the overall quality of a video game, in terms of its rules, mechanics, goals, and design within the process of design and analysis [1]. This concept is commonly used in game studies. It reflects the players' degree of satisfaction towards their various ways of interaction with the game system, that is, in a nutshell, *"A good game has good playability"* [2]. It can also be narrowly interpreted as being equal to the quality of "gameplay" or simply the usability of video games, that cannot be balanced by "any non-functional designs" [3]. It is also common to consider both "gameplay" and "usability" as parallel elements of the playability framework [4,5]. Moreover, playability is also seen as the quality in use [6] of video games and represents *"the degree in which specific player achieve specific game goals with effectiveness, efficiency, flexibility, security and, especially, satisfaction in a playable context of use"* [7]. Thus, seeing games as systems and taking into account also the technical, mechanical, or material quality of video games, playability is *"the design quality of a game, formed by its functionality, usability, and gameplay"* [8].

Scholars across domains agree that playability, however measured, can be used to reflect and evaluate the quality of a video game [1,8]. However, regardless of the definition or framework adopted, research on the approaches to analyze the playability of a particular game is limited. The most commonly adopted approach to playability analysis is the use of heuristics [4,5,9,10]. Acknowledgedly, heuristic evaluation has multiple advantages including being cheap, being easy to motivate evaluators, not requiring advanced planning, and importantly it can be done in the early development stage [11]. However, it is also inevitable that such evaluation is biased by the mindset of the evaluators [11,12]. Their experiences and preferences influence the outcome as well [9,13–17]. In addition, it is common that such usability test contains inconsistency due to the evaluator effect [9,18]. Furthermore, the difference in game rule structures, i.e. games of emergence and games of progression [19], has not been considered in playability evaluation using heuristics. It is obviously not possible to detect gameplay issues of the game elements appearing in the later game scenes of progression games (e.g., Witcher 3 [20]) with the limited time spent on testing with game demos [10].

Hence, towards relevantly fair evaluation on the overall playability of any released video game product, the opinions of the players who have played it for a fair amount of time shall be valuable. Players' game reviews are then the target data source for such purposes. For software products, the analysis of end user reviews has been considered important towards evaluation of software quality [21,22]. Meanwhile, text and opinion mining is a well-known way of *"using large text collections to discover new facts"* [23,24]. With such support, many studies have provided various approaches towards effective review analysis to uncover the critical user needs for software products [25–27]. Despite the differences between video games and utilitarian software products and in the review styles, such players reviews can be used towards the improvement of game products [28]. As one of the most popular digital game distribution platforms, Steam (https://store.steampowered.com/, accessed on 16 March 2021) provides an online venue for the players to review games. With such a large amount of players' opinion data at hand and together with the opinion mining techniques, it is then possible to evaluate video game playability from the perspective of players' collective intelligence.

Herein, we propose a data-driven video game playability analysis approach based on the collection of player textual reviews. It answers the following research question: *How can the data-driven approach be used to gain insights into the playability of a game?* The proposed method uses a pre-trained text classifier model to elicit informative reviews from the pre-processed review collection and uses another pre-trained classifier to classify such reviews into pre-defined playability categories. In this paper, we chooose Paavilainen's game-as-system definition of playability as the reference of classification [8]. With such an explicit and simplified framework and the proposed method, we can obtain not only the intuitively quantified evaluation of the overall playability of the target game but also the specified merits and defects of it in every framework-oriented perspective (answering the research question). We also validated the usefulness of the proposed approach by conducting a case study on a real-life video game with 99,993 reviews.

Compared to heuristic evaluation on playability, this approach relies on the collective intelligence of a large number of players instead of a few experts' opinions. Furthermore, this approach evaluates the game by its released versions instead of demos. Thus, it can provide both the overall playability evaluation and the detailed merits and defects on a game-as-system level. Therefore, although acknowledging the usefulness of heuristic evaluation in game development, we emphasize that the contributions of our approach are: (1) to help game developers obtain a quick overall impression of the perceived game playability from players' perspective; and (2) to help game developers understand the collective needs and complaints of players to identify the playability issues for video game maintenance and evolution.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 presents the playability analysis approach, including the series of procedures and details. Section 4 presents the case study on validating the proposed approach. Section 5 provides further discussion. Section 6 concludes the article.

2. Related Work

2.1. Playability Evaluation with Heuristics

Heuristic evaluation, targeting originally usability evaluation, is an informal analysis method where several evaluators are asked to comment on the target design based on pre-defined heuristics/principles [11,29]. It aims at finding the usability problems during the iterative design process so that such problems can be addressed before software products releasing [30]. Despite the rapid development of the video game industry, the methodologies for evaluating game quality and player experience are still limited. Therein, heuristic evaluation is still an effective way of evaluating games compared to other methods for being cheap and fast [31].

Malone proposed the set of heuristics as a checklist of ideas to be considered for designing enjoyable user interfaces which is largely seen as the earliest game heuristics [32]. Therein, three main features are proposed: challenge, fantasy, and curiosity. Federoff's list of game heuristics is based on the observation and interviews with five people from one game development company [33]. For such heuristics, game interface, game mechanics, and game play are the three main aspects. Neither of these early studies provides validation of the respectively proposed heuristics.

Desurvire et al. introduced Heuristic Evaluation for Playability (HEP) towards video, computer, and board game evaluation with four categories: game play, game story, game mechanics, and game usability [4]. HEP is validated via comparison with a user study of a new game at the beginning of its development with four prospective users in 2-h sessions. The authors also emphasized HEP is helpful at the early stage of game design but admitted players' behavior is still unpredictable. In addition, HEP is extended into Game Genre-Specific Principles for Game Playability (PLAY) to adapt usability principles to game design [34]. Forty-eight game design principles from eight categories are proposed.

Korhonen and Koivisto's playability heuristics are designed for mobile games where gameplay, game usability, and mobility are the main categories [5]. It is validated by four experts over a mobile game in the production phase. The authors also admitted that, although heuristics are helpful, the gameplay is much harder to evaluate. Furthermore, they extended the heuristics to mobile multiplayer games with experiments showing the heuristics can be applied to non-mobile games as well [35]. Korhonen et al. also compared their heuristics with HEP finding the respective strength and weakness [9]. The study also detects inconsistency within evaluators in terms of their reported problems due to the potential evaluator effects [18] or different reporting baseline.

Pinelle et al. proposed heuristic evaluation focusing on the usability for video game design based on the analysis of game expert reviews [10]. The heuristic set contains 12 problem categories and 285 individual problems. It is verified via a testing evaluation of the demo of a PC game by five expert evaluators. Thereafter, an extension study is conducted towards heuristics for networked multiplayer games; as a result, five problem categories with 187 problems specially for network multiplayer games are proposed and verified by 10 expert evaluators on two network games [13]. However, Pinelle and colleagues also emphasized *"the heuristics do not address design issues related to how fun and engaging games are for users"*.

Koeffel et al. proposed a three-aspect heuristic set (including game play, game story, and virtual interface) to evaluate the user experience in video games and tabletop games [36]. The authors summarized 29 heuristic items based on extensive literature search and verified the heuristics based on expert evaluation (two experts) on five games of different genres and comparison to game media reviews.

Many other scholars also conduct research on utilizing heuristic evaluation for specific types of games. Röcker and Haar showed that existing heuristics can be transferable to pervasive gaming applications [37]. Tan et al. proposed using heuristic evaluation, the Instructional Game Evaluation Framework, for educational game analysis [38]. Khanana and Law illustrated the use of playability heuristics as design tools for children's games [39]. However, whether these heuristics can be used for video games in general is not verified.

On the other hand, regarding the different ways of using heuristic evaluation towards video game playability, Aker et al. found, based on an extensive literature search, that empirical evaluation, expert evaluation, inspection, and mixed method are the methods used for such purpose [40]. Among the four mentioned, expert evaluation is the most commonly applied with many of the above mentioned studies adopting such a method [5,9,13,33,35,36]. However, the outcomes of such a method rely heavily on the experts' skills and preferences and seldom capture the behaviors and needs of real end users [41]. Empirical evaluation, such as surveys, interviews, and focus groups, is also a relevantly common method of using heuristics [4,32,37,38]. However, with such a method, it is difficult to properly select the correct user sample and reproduce actual usage/play situations within the limited given time [41].

2.2. User Review Studies

Being an important data source, customer feedback is commonly used for companies to understand the market and the needs of their customers so that they could improve products and services accordingly. Regarding software products, it is also critical to facilitate the evolution of software products and services via the analysis of end user reviews [21,22].

Many studies show mining the end user reviews of software products can help reveal the hidden user behaviors, software characteristics, and the relations in between. Vasa et al. conducted a preliminary analysis on 8.7 million reviews of 17,330 mobile apps using statistic methods on user review character counts and ratings [42]. Their results show mobile app reviews tend to be short and both the rating and the category of an app influence the length of a review. With the same data, Hoon et al. showed that the most frequently used words in user reviews are to express sentiment [43]. Harman et al. used customized algorithms to extract app features and correlation analysis on 32,108 non-zero priced apps from Blackberry app store [44]. The results show a strong correlation between customer rating and the app download ranking but no correlation between the app price and either downloads or ratings.

More importantly, many studies also show that the results from mining end user reviews can reflect the positive and negative user experience regarding software products. For example, Vu et al. proposed a keyword-based review analysis method to detect keyword trends and sudden changes that could possibly indicate severe issues [45]. Panichella et al. proposed an approach to extract information from user reviews relevant to the maintenance and evolution of mobile apps using Natural Language Processing (NLP), sentiment analysis, and text analysis techniques [46]. Gu and Kim proposed a method to categorize reviews, extract aspects from sentences, and evaluate the obtained aspects of the mobile apps using NLP techniques [47]. Many other studies also show that opinion mining on end user reviews can help identify user complaints [48], the useful information [26], and the factors for software success [25] and evaluate the experience towards specific software features [49], merits and defects of particular software updates [50,51], and software evolution [27].

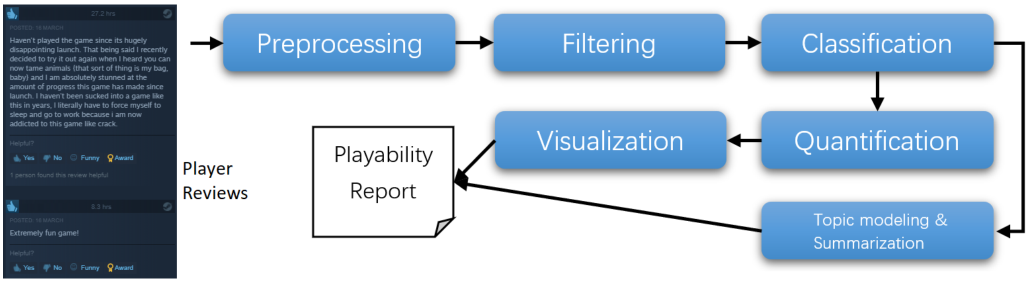
Despite the differences in video games and utilitarian software products, as well as those between the review styles, such end user reviews are considered valuable for game designers and developers towards the improvement of their game products. Lin et al. conducted an empirical study on the reviews of 6224 games on Steam and analyzed the review content and the relation between players' play hours and their reviews [28]. Santos et al. compared the expert and amateur game reviews on Metacritic and found amateur reviews are more polarized and have stronger sentiments than expert reviews [52]. Lu et al. used topic modeling on Steam reviews to investigate the temporal dynamics of player review topics and the influence of updates to such dynamics [53]. Although game reviews form a rich resource for understanding the players' experience and opinion on a particular game, the game playability analysis based on players' reviews is yet under-explored.

3. Method

In this section, we present an overview of our approach towards video game playability evaluation, with further explanation of the key steps.

3.1. *The Framework Overview*

Figure 1 depicts the overview of the video game playability evaluation approach with each key step specified. The framework starts with collecting the player reviews from online platforms (e.g., Steam) via either API or web crawling. Then, we preprocess the obtained raw review data into structured form. The second step is to filter out the "non-playability-informative" reviews via a pre-trained classifier. With the obtained "playability-informative" reviews dataset, the third step is to classify the data into different playability perspectives according to a selected playability framework. For example, when selecting the framework of Paavilainen [8], the reviews are then classified into three perspectives accordingly, i.e., functionality, gameplay, and usability. With each review instance categorized into a specific perspective, the fourth step is to quantify the evaluation result of each perspective. Subsequently, the fifth step is to visualize such a result and present an intuitive summary. Meanwhile, the sixth step is to extract the existing merits and defects from each perspective by modeling and summarizing the topics of the reviews within. The output of both the visualization and topic modeling is then synthesized into a report.



POSTED: 16 MARCH

Haven't played the game since its hugely disappointing launch. That being said I recently decided to try it out again when I heard you can now tame animals {that sort of thing is my bag, baby} and I am absolutely stunned at the amount of progress this game has made since launch. I haven't been sucked into a game like this in years, I literally have to force myself to sleep and go to work because i am now addicted to this game like crack.

1 person found this review helpful

POSTED: 16 MARCH

Extremely fun game!

Figure 1. Video Game Playability Evaluation Framework.

3.2. *Preprocessing*

The preprocessing step encompasses the following key activities. First, we divide each review item from the dataset into sentence-level review instances, due to the fact that each review with multiple sentences can contain multiple topics and various sentiments. In this study, we use the sentence tokenizer feature from the Natural Language ToolKit (NLTK) (http://www.nltk.org/, accessed on 16 March 2021), a popular Python package with text processing libraries, corpora, and lexical resources. Secondly, based on the obtained sentence-level review dataset, we build the bigram and trigram models to identify the phrases within the data. For such a purpose, we use the phrase detection feature of Gensim (https://radimrehurek.com/gensim/models/phrases.html, accessed on 16 March 2021), a popular semantic modeling package. Subsequently, for each review sentence, we perform a series of text processing activities, including transforming text into lowercase, removing non-alpha-numeric symbols, screening stop-words, eliminating extra white spaces, and lemmatization (using the WordNetLemmatizer model of NLTK). Note that the processing is only applied to the text when topic modeling is required. For sentiment analysis, such activities are not only unnecessary but also counter-productive.

3.3. *Filtering*

Herein, the filtering step is to classify the dataset of sentence-level review instances into "playability-informative" and "non-playability-informative". By doing so, we identify the review sentences that contain description regarding the playability of the particular game and screen out those not relevant. Due to the variety in playability definition, the criteria by which review instances are categorized slightly vary. In this study, we adopt the game-as-system playability definition given by Paavilainen [8] as a reference, as this definition provides clear criteria for the identification of playability-related text with a pre-defined playability perspective framework with minimum complexity compared to the other frameworks. Thus, accordingly, we set the two unique class labels as {'Playability-informative (P)', 'Non-Playability-Informative (N)'}. Based on the adopted definition and framework, the criteria for "playability-informative" reviews are listed in Table 1 with explanation and examples attached. On the other hand, a review is accordingly labeled "non-playability-informative" when it contains no information related to such criteria. For example, review sentences such as *"I'm glad I supported this Dev team."* (Development and Publishing), *"Now the game has exceeded my expectations!"* (Feeling Expression), and *"Ive got this game on PS4, XBOX and now Steam."* (Player Self Description) are all seen as "non-playability-informative".

Table 1. "Playability-informative" Criteria based on Paavilainen's Framework [8] and Examples.

Criteria	Explanation	Review Examples
Functionality	the technical, mechanical or material quality of the game that is related to its smooth operation.	*"...the performance in VR mode is absolutely terrible."* *"Crashing and stuttering constantly..."*
Gameplay	the rule dynamics that provide "gameness". e.g., goals, challenge, progress, and rewards.	*"Survival is not challenging unless you play hardcore,..."* *"...doing the same repetitive things over and over again"*
Usability	the user-interface of the game and its ease of use.	*"Controls and menus are bad,..."* *"...the massive improvements to the games graphics..."*

To efficiently identify and filter the "non-playability-informative" review sentences, we herein apply a classifier based on machine learning algorithm. In the study, we compare the Naive Bayes (NB) and the Expectation Maximization for Naive Bayes (EMNB) [54] and adopt the EMNB classifier in the filtering step. EMNB is a well-recognized semi-supervised text classification algorithm, which can build a classifier with high accuracy using only a small amount of manually labeled training data. With EMNB, we thus filter out the review sentences labeled 'N' and build the "playability-informative" review sentence dataset.

3.4. Classification

In this step, we classify the obtained "playability-informative" review sentences into perspectives according to the selected playability framework. As stated above, in this study, we adopt the playability framework that contains three perspectives, i.e., Functionality (F), Gameplay (G), and Usability (U). Targeting the specific objectives of this study when the classes (i.e., playability perspectives) are determined by the existing framework, a supervised learning algorithm is more suitable. On the other hand, it is also frequent that a particular review sentence contains information regarding multiple perspectives. For example, the review sentence *"The gameplay, UI and story are not bad, unfortunately this game has no Beginner friendly and you had to figure out by your own."* describes the players' opinion on both gameplay and usability. Thus, to cope with such a situation, we adopt a multi-label classification algorithm. For such a multi-label classification task, we select from three algorithms: kNN classification method adapted for multi-label classification (MLkNN) [55], Twin multi-Label Support Vector Machines (MLTSVM) [56], and Binary Relevance multi-label classifier based on k-Nearest Neighbors method (BRkNN) [57]. The interfaces of these classification algorithms are provided by the Scikit-multilearn (http://scikit.ml, accessed on 16 March 2021), a BSD-licensed library for multi-label classification built on top of the Scikit-learn ecosystem (https://scikit-learn.org/, accessed on 16 March 2021). The comparison of these algorithms is discussed in Section 4.2.

3.5. Quantification

In this step, with the classified three sets of "playability-informative" review sentences, we evaluate each of the playability perspectives by quantifying the overall opinions extracted from the according set of review sentences.

Herein, we use the average sentiment score of the "playability-informative" review sentences representing the players' collective evaluation towards the playability of the game.

Algorithm 1 depicts the process of quantifying the playability of a particular game. Let R be the set of "playability-informative" review sentences, where each $r_i \in R$ is evaluated with the sentiment score (s_i) assigned via a selected sentiment analysis method, e.g., Valence Aware Dictionary for sEntiment Reasoning (VADER) [58], Sentiment strength [59], etc. Meanwhile, each $r_i \in R$ is labeled by one or more playability perspectives (L_i) using the pre-trained multi-label text classifier (i.e., MLclassifier). Thereafter, for each playability perspective (p), we find the set R_p that contains all the review sentences labeled p and calculate the sentiment value for such perspective as the average of the sum of the sentiment score (see Line 10).

Algorithm 1: Algorithm of Quantifying the Playability on Multiple Perspectives.

```
   Data: A set of "playability-informative" review sentences
   Result: A dictionary of playability scores, each for one perspective
1  $R$ ← set of "playability-informative" review sentences;
2  Let $P$ be the set of all playability perspective labels;
3  for each $r_i \in R$ do
4      $s_i\ (\in S) \leftarrow$ getSentimentScore($r_i$);
5      $L_i\ (\subseteq P, \neq \emptyset) \leftarrow$ MLclassifier.predict($r_i$);
6  end
7  Let $result$ be return dictionary;
8  for each $p \in P$ do
9      $R_p \leftarrow$ any $r_i$ has $p \in L_i$;
10     $v_p \leftarrow \frac{\sum_{r_i \in R_p} s_i}{len(R_p)}$;
11     $result[p] \leftarrow v_p$;
12 end
13 return $result$;
```

3.6. Visualization

In this step, we visualize the output of the quantification of player opinions regarding each playability perspective with a polygon diagram. The number of vertices of the selected polygon is equal to the perspective numbers. For example, when adopting Paavilainen's playability framework of three perspectives, the analysis of playability to a particular game can be depicted as a triangle chart (Figure 2).

As shown in Figure 2, the line segments from the triangle center to each vertex represent the scales measuring each playability perspective. The distance between a green playability triangle vertex and the center represents the playability score in the particular perspective. The central point of each line segment is value *0* indicating the neutrality of the according perspective. The larger the green triangle is, the higher the overall playability score the game has. When the red area is shown in any direction, the playability of that game suffers in that particular perspective. Herein, the scale range indicating the positive and negative of each perspective is determined by the average sentiment score ranging from -1 to 1. For this particular game example, after the analysis of its reviews through Algorithm 1, a result list $[-0.2, 0.5, 0]$ is obtained. Based on such a result, Figure 2 is drawn. It shows the game is good for its gameplay ($G = 0.5$), mediocre for its usability ($U = 0$), and unsatisfactory for its functionality ($F = -0.2$).

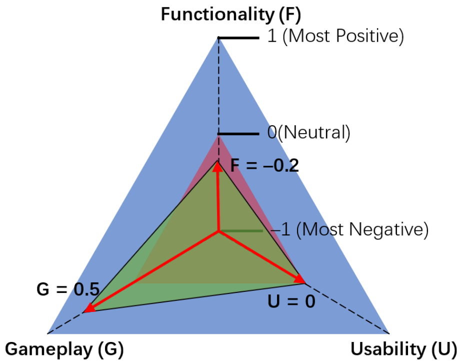


Figure 2. An example of playability triangular chart.

3.7. Topic Modeling and Summarization

As a critical part of the playability evaluation results, finding the players' opinions regarding the pros and cons of a particular game product can convey great value to the developer team. Thus, in this step, besides the quantified outcome of overall playability, we identify the specific issues regarding each playability perspective using a topic modeling algorithm. In this study, we use Latent Dirichlet allocation (LDA) topic modeling, a well-recognized effective topic modeling algorithm that finds the hidden topics from a large set of text data [60].

4. Case Study

In this section, we verify the effectiveness of the proposed playability evaluation method by conducting an experiment on a real-life video game from Steam platform.

4.1. Data Description

The game we select for this case study is No Man's Sky (NMS) [61], a space exploration and survival game developed and published by Hello Games (https://hellogames.org/, accessed on 16 March 2021). The game was first released on 12 August 2016, before which a social media "hype" had been evoked leading to an unprecedentedly high expectation from the players. However, the release of the game was disastrous due to the unfulfilled promises from the developers as well as the performance and gameplay defects. Interestingly, for the last four years, the game has been continuously maintained with its quality gradually increasing, which makes it a unique case where the changes in game quality is observable.

We collected the 99,993 English reviews from 12 August 2016 to 7 June 2020 for NMS. Within the collected review set, the longest contains 116 sentences while the shortest is a single-sentence review. Via tokenization, we obtained 519,195 review sentences. We then manually labeled the sentences with "Playability-informative" (P) and "Non-playability-informative" (N) in a random order, until obtaining 1500 "playability-informative" sentences and 1500 "non-playability-informative" sentences. Therein, 1000 sentences (500 for each label) were saved as training data and 2000 (1000 for each label) as testing data for building the filtering classifier model. Furthermore, adopting Paavilainen's playability framework, we further labeled the 1500 "playability-informative" review sentences in both the training and testing dataset into *Functionality (F)*, *Gameplay (G)*, and *Usability (U)*, where it is possible for one sentence to contain multiple labels. Such dataset was used to train the classifying model. Note that the labeling of the training data is ideally done by three expert evaluators. Two evaluators first label the sentences separately and then each label is confirmed by the agreement of both parties. A third evaluator is invited to provide final verification when agreement cannot be reached.

4.2. Classifier Selection

To evaluate the performance of the proposed method, we conducted experiments testing its key steps, including the filtering and the classification steps.

4.2.1. Filtering Evaluation

To evaluate the performance of filtering, with a series of experiments, we compared the results of the original NB algorithm and the EMNB algorithm with the amount of training and test data from 60 to 3000 with a step of 60. Within the amount of data selected for each experiment iteration, 1/3 was selected as training data with the other 2/3 as test data. The evaluation results show that the performances of NB and EMNB are similar regarding our dataset throughout various data volumes. Throughout the series of experiments, the accuracy (F1-score) difference between NB and EMNB with that same data volume does not exceed 0.04. On the other hand, with a limited number of training data (100 training data and 200 testing data), the accuracy of both algorithms reaches a satisfactory level ($\geqslant$0.7). The level of accuracy does not drop when enlarging the data volume. Furthermore, with the data volume reaches around 1200, both classification algorithms can provide optimal accuracy ($\geqslant$0.8). In this study, considering the large amount of unlabeled review sentence data as well as the according efficiency, we adopted the EMNB algorithm with the full training data volume in order to obtain the best performance (F1-Score = 0.85).

4.2.2. Classification Evaluation

Furthermore, we conducted a series of experiments to compare the performances of three multi-label text classification (MLTC) algorithms, i.e., MLkNN, MLTSVM, and the two versions of BRkNN. With the manually labeled 1500 "playability-informative" training data, we first found the best parameters targeting the best performance for each algorithm. Then, the best accuracy of the three algorithms with the detected parameters were calculated for comparison. The results shown in Table 2 indicate that MLkNN algorithm has the best classification accuracy (0.769) on our training dataset with the detected best parameter. Together with the previous filtering step with EMNB (accuracy of 0.85), the overall accuracy is satisfactory ($0.85 * 0.769 = 0.653$).

Table 2. Comparison of the performance of MLTC algorithms.

Algorithm	Best Parameter	Accuracy
MLkNN	k = 27, s = 0.5	0.769
MLTSVM	c_k = 0.125	0.532
BRkNNaC	k = 19	0.663
BRkNNbC	K = 17	0.712

In addition, to further tune the method, we evaluated both the performance of combining the two individual steps and that of applying only the MLTC algorithm targeting both filtering and classifying tasks. For such purpose, we manually labeled "N" to the 1500 "non-playability-informative" training data and combined them with the 1500 "playability-informative" ones. The performance of the above three algorithms on the enlarged dataset is shown in Table 3.

Table 3. Comparison of the performance of two- and one-step classification.

Two-Step			**One-Step**		
Algorithm	**Best Parameter**	**Accuracy**	**Algorithm**	**Best Parameter**	**Accuracy**
EMNB + MLkNN	k = 27, s = 0.5	0.653	MLkNN	k = 1, s = 0.5	0.121
EMNB + MLTSVM	c_k = 0.125	0.452	MLTSVM	c_k = 0.125	0.349
EMNB + BRkNNaC	k = 19	0.564	BRkNNaC	k = 1	0.121
EMNB + BRkNNbC	K = 17	0.605	BRkNNbC	k = 6	0.276

The results show that a two-step classification, i.e., "playability-informative" review filtering with EMNB and perspective classifying with multi-label text classification, has a much better accuracy rate than one-step classification with only MLTC. In addition, we found that using MLkNN (with the parameter $k = 27$ and $s = 0.5$) for the classifying procedure has the best overall accuracy.

4.3. Results

In this section, we present the results from applying the proposed playability analysis approach on the review dataset of NMS. The results contain two major parts: (1) the overall playability score; and (2) the merits and defects of the game.

4.3.1. Playability Score

With the obtained 519,195 review sentence data, we follow the analysis method procedure by first filtering out the "non-playability-informative" ones. As an outcome, 273,476 review sentences are automatically labeled as "playability-informative" using the pre-trained EMNB classifier with the 3000 training data. Subsequently, we classify them into the three perspectives using the selected MLkNN algorithm receiving 43,110 review sentences on functionality, 20,5474 on gameplay, and 30,176 on usability. To perform sentiment analysis on the review sentences, we select the VADER approach, due to its high classification accuracy on sentiment towards positive, negative, and neutral classes in social media domain [58]. In addition, its overall classification accuracy on product reviews from Amazon, movie reviews, and editorials also outperform other sentiment analysis approaches and run closely with that of an individual human [58]. It is also easy to import and perform using Python as being integrated into the NLTK package. By calculating the sentiment score for each review sentence with VADER and the average score for each review set, we obtain the result as {'Functionality': 0.025, 'Gameplay': 0.111, 'Usability': 0.039} (shown in Figure 3). It indicates that the overall playability of this game is at the level of mediocre in each of the two out of three perspectives, when only performs only slightly better than mediocre in the gameplay perspective. Such results comply with the overall rating of *Mixed* on Steam (https://store.steampowered.com/app/275850/No_Mans_Sky/#app_reviews_hash, accessed on 16 March 2021).

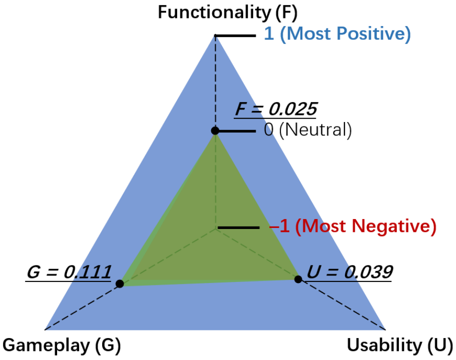


Figure 3. Overall playability score for NMS.

Furthermore, to further verify the results, we follow the major updates of NMS via the information from its patch notes (https://nomanssky.gamepedia.com/Patch_notes, accessed on 16 March 2021). As the release date of the 11th major update is 11th June 2020, the review dataset for the 10th update is incomplete. Thus, focusing on the first nine updates (*Foundation*, *PathFinder*, *Atlas Rises*, *NEXT*, *Abyss*, *Visions*, *Beyond*, *Synthesis*, and *Living Ship*), we divide the "playability-informative" review sentences into 10 subsets based on their release dates. Via the same calculation on each subset, the playability analysis results regarding the original release (marked as Release 1.0) and the nine following major updates, as well as their data volumes, are shown in Table 4.

Table 4. Playability score changes through major updates.

	1.0	**Foundation**	**PathFinder**	**Atlas Rises**	**NEXT**	**Abyss**	**Visions**	**Beyond**	**Synthesis**	**Living Ship**
Date	16.11.27	17.03.08	17.08.11	18.07.24	18.10.29	18.11.21	19.08.14	19.11.28	20.02.18	20.04.07
Count F.	28,251	800	718	1851	4222	135	1858	2680	1052	555
Count G.	120,894	6128	5460	12,582	19,085	554	12,649	10,011	7667	3579
Count U.	18,106	758	751	1698	2876	103	1692	1900	922	449
Score F.	−0.0054	0.0372	0.0973	0.0684	0.0487	0.0091	0.0760	0.0458	0.0966	0.0770
Score G.	0.0765	0.1322	0.1346	0.1534	0.1389	0.1080	0.1686	0.1406	0.2211	0.2113
Score U.	0.0106	0.0708	0.0807	0.0948	0.0608	0.0823	0.0755	0.0481	0.1489	0.1538

Based on such results, we can conclude that the playability of the game increased in terms of all three perspectives through the nine updates, even though it decreased regarding some particular updates (e.g., Beyond). The reason for such a situation is the introduction of new critical features, major interface changes, new vital bugs, etc. Taking the Beyond update as an example, as a Version 2.0.0, it added the Virtual Reality support and a wide range of features to the game (https://nomanssky.gamepedia.com/Update_2.00, accessed on 16 March 2021). It evoked controversy among players regarding its performance and gameplay. Nonetheless, by comparing the playability of Release 1.0 and that of the version after the "Living Ship" update, all three perspectives had been greatly improved.

4.3.2. Playability Merits and Defects

To detect the merits and defects of the game in terms of each playability perspective, we first divide the review sentences into three subsets based on the classification result. For each subset, i.e., the review sentences for each perspective, we further select the positive (sentiment score greater than 0) review sentences and the negative ones (sentiment score smaller than 0) forming six review sentence sets. The volume of each subset is shown in Table 5. To detect the explicit sentiment from the review, we ignore the neutral (sentiment score equals 0) review sentences herein. In addition, to conveniently compare the results to the information extracted from the Metacritic later, we select only the review data concerning the original game release (i.e., between 12 August 2016 and 27 November 2016).

Table 5. Data volume for review subsets for Release 1.0.

	Functionality	**Gameplay**	**Usability**
Positive	10,684	47,780	6537
Negative	10,637	32,627	6396

Subsequently, to find the best topic number for each review subset, we conduct a series of experiments for each set testing with the topic numbers ranging from 2 to 20. We use the topic coherence representing the quality of the topic models. Topic coherence measures the degree of semantic similarity between high scoring words in the topic. A high coherence score for a topic model indicates the detected topics are more interpretable. Thus, by finding the highest topic coherence score, we can decide the most fitting topic number. Herein, we use c_v coherence measure, which is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity [62]. Note that we pick the model that has the highest c_v value before flattening out or a major drop, in order to prevent the model from over-fitting.

With the best topic number (k) values detected for the six review subsets (shown in Table 6), we can continue with building the according topic models and detecting the topics. Table 7 shows the extracted topics for each subset as well as the top 10 keywords that describe each of them.

Table 6. Best fitting topic numbers and c_v values.

	Functionality	**Gameplay**	**Usability**
P	$k = 3, c\_v = 0.552$	$k = 4, c\_v = 0.535$	$k = 3, c\_v = 0.397$
N	$k = 3, c\_v = 0.438$	$k = 3, c\_v = 0.522$	$k = 5, c\_v = 0.374$

Table 7. Detected topics for each review set.

Topic (Positive Functionality)	**Top Words**
+ Load Screen and Crashing	“game”, “play”, “crash”, “time”, “screen”, “start”, “hour”, “go”, “load”, “get”
+ Performance and bugs fixed via update	“issue”, “game”, “performance”, “fix”, “people”, “update”, “would”, “bug”, “problem”, “patch”
+ Running game fine with settings	“run”, “game”, “setting”, “graphic”, “work”, “get”, “fps”, “pc”, “fine”, “high”
Topic (Negative Functionality)	**Top Words**
− Poor Performance, Bugs, Crash, Need Fix	“game”, “issue”, “problem”, “performance”, “fix”, “people”, “crash”, “bad”, “poor”, “bug”
− Lag, Stutter, fps drop, even with low settings	“run”, “setting”, “low”, “game”, “stutter”, “drop”, “pc”, “graphic”, “lag”, “fps”
− Crash at Start screen, try hours	“crash”, “game”, “play”, “time”, “screen”, “get”, “start”, “can”, “try”, “hour”
Topic (Positive Gameplay)	**Top Words**
+ Explore, survival, different planet systems	“planet”, “find”, “new”, “explore”, “system”, “beautiful”, “different”, “look”, “survival”, “thing”
+ Crafting, ship-flying, resource and inventory	“space”, “ship”, “get”,“resource”, “fly”, “well”, “craft”, “upgrade”, “inventory”, “learn”
+ Fun exploration gameplay	“game”, “exploration”, “fun”, “play”, “get”, “hour”, “gameplay”, “good”, “enjoy”, “lot”
+ Need story to make better	“game”, “want”, “make”, “need”, “would”, “give”, “bit”, “people”, “story”, “work”
Topic (Negative Gameplay)	**Top Words**
− Repetitive, boring gameplay	“game”, “get”, “hour”, “feel”, “repetitive”, “start”, “bore”, “boring”, “gameplay”, “people”
− Lack of inventory upgrade	“ship”, “resource”, “make”, “need”, “inventory”, “find”, “upgrade”, “lack”, “craft”, “much”
− Fly, explore, combat	“planet”, “space”, “see”, “look”, “explore”, “combat”, “find”, “fly”, “kill”, “ship”
Topic (Positive Usability)	**Top Words**
+ Control feels with controller, fly ship	“control”, “use”, “ship”, “take”, “feel”, “get”, “controller”, “fly”, “space”, “flight”
+ Beautiful graphics	“game”, “graphic”, “play”, “change”, “setting”, “beautiful”, “look”, “run”, “work”, “good”
+ Music&sound, hold and click button	“hold”, “button”,‘ ‘music”, “menu”, “screen”, “system”, “inventory”, “click”, “sound”, “second”
Topic (Negative Usability)	**Top Words**
− Graphic settings poor, restart	“graphic”, “game”, “setting”, “change”, “run”, “bad”, “start”, “poor”, “get”, “restart”
− Fly control with mouse annoying	“control”, “mouse”, “ship”, “fly”, “game”, “use”, “get”, “annoying”, “make”, “press”
− Terrible texture and sound	“terrible”, “look”, “texture”,‘ ‘game”, “sound”, “pop”, “point”, “require”, “complaint”, “way”
− Horrible flight control, cluncky inventory	“control”, “flight”, “feel”, “people”, “horrible”, “system”, “inventory”, “lack”, “clunky”, “fov”
− Option, click and hold button, bad/awful PC port	“game”, “pc”, “option”, “button”, “hold”, “port”, “menu”, “awful”, “bad”, “click”

From the detected topics, we can easily summarize the merits and defects of the game in terms of each playability perspective. For example, the topics extracted from the “negative-functionality” review set show that users are satisfied with the performance of the game when settings are tackled properly. They are also satisfied with the bugs being fixed and with the game despite the load screen and crashing. On the negative side, players often complain about various issues, including poor performance, bugs, crashes, lagging, stuttering, fps, etc. Regarding gameplay, the exploration and survival through different planet systems, as well as the crafting, spaceship cruising, and resource and inventory management, are well received by the players. They also indicate a better story is needed. On the other hand, the players feel negative about the gameplay being repetitive and boring and frustrated about the lack of inventory upgrade. The flying, exploring, and combat mechanisms also suffer. Regarding usability, the players feel positive regarding the spaceship control using controller and the beautiful graphics. They also like the music and sound effects and the menu interface using a click and hold button to access the inventory. However, players also complain about the following aspects: the graphic setting only changes after restarting, controlling with mouse is annoying, texture and sound being terrible, horrible flight control and clunky inventory, the click-and-hold interaction mechanism, and being an awful PC port. Note that a similar topic shown in both the positive and negative groups (e.g., loading screen and crash) suggests that a relevantly high number of players express different sentiment when talking about ’crash’. For example, *“With a Rift S headset and a gtx1080 graphics card I’m getting great performance out of the game with no crashes.”* and *“This game has a TON of performance problems and has crashed on me far* *too many times to be acceptable."* express sentiment differently when both are about "crash". Such a situation indicates players' opinions diverge regarding this topic.

To verify the correctness of the detected merits and defects of the game via topic modeling, we compare our results to the expert opinions extracted from the critic reviews of Metacritic (https://www.metacritic.com/game/pc/no-mans-sky, accessed on 16 March 2021). Metacritic reviews have been considered valuable in providing insights in evaluating the quality of media products, e.g., movies and games [63,64]. We find the 10 critic reviews (including 'gamewatcher', 'hookedgamers.com', 'ign denmark', 'the games machine', 'mmorpg.com', 'pelit.fi', 'pcgamer.com', 'gamegrin.com', 'games.cz', and 'game-debate.com') on NMS. Their full review contents are accessible online with the "pros and cons" explicitly listed. Due to the fact that all the critic reviews were given soon after the release date, the opinions thus only apply to Release 1.0 of the game. As stated above, such opinions are used to compare with the extracted players' review opinions regarding the same version.

As shown in Table 8, we can easily compare the extracted positive and negative topics from the player reviews and the summarized "playability-informative" "Pros and Cons" from the critic reviews. We can conclude that a great majority of the merits and defects of the game mentioned by the media experts are detected from the player review modeling. For example, regarding functionality, both parties point out the problems of crashing, bugs, frame drops, and performance issues. Note that the critic reviews do not mention the merits regarding functionality, which is reasonable as providing a functional product is clearly a "must-have" instead of an "exciter". Regarding gameplay, the exploration and survival gameplay is praised by both, as well as the different planet systems and spaceship flying. The sense of relaxing that mentioned by the media is not covered by the players' topics. On the negativity of gameplay, the complaints about inventory, repetitive/tedious gameplay (limited options), lack of combat, etc. are mutual. Furthermore, regarding usability, the graphics and sound are praised by both, when the players' reviews additionally give credits to the controlling performance with controllers. On the other hand, both parties reflect negative opinions on the control (with mouse) and menu/option being frustrating, when the players complain more specifically about the "hold and click button" control.

In addition, we also compare these extracted topics to the original game-as-system definition and the according perspective descriptions of Paavilainen's playability framework [8]. Regarding functionality, nearly all the sub-perspectives are covered by the player review topics, except for "error reporting". Apparently, the players are generally not satisfied with functionality from all sub-perspectives, as all such can be related to at least one topic from negative reviews. On the other hand, regarding gameplay, the player reviews reflect positively on the play styles, goals, challenges, and rewards of the game, when convention and consistency are not mentioned enough. Repetitiveness and autonomy (i.e., lack of inventory upgrade -> cannot freely preserve more items) are the gameplay sub-perspectives being complained often. Finally, regarding usability, the negative opinions are about the control with mouse (control), texture (audiovisual), inventory (UI layout), graphic setting, option/menu (Navigation), and click and hold button (feedback). Such results further validate the extracted review topics are "playability-informative".

Table 8. Mapping between extracted player review topics and metacritic reviews pros and cons.

Playability	Players' Review Topic	Metacritic Review Pros and Cons
Functionality	**+ Load Screen and Crashing**	
	+ Performance and bugs fixed via update	
	+ Running game fine with settings	
	− Poor Performance, Bugs, Crash, Need Fix	− Still major technical issues.
		− Deplorable technical condition.
		− The PC version is heavy, buggy, and crashing
	− Lag, Stutter, fps drop, even with low settings	− Random frame rate drops.
		− Poorly optimized.
	− Crash at Start screen, try hours	
Gameplay	**+ Explore, survival, different planet systems**	+ Solid survival gameplay with great freedom.
	+ Crafting, ship−flying, resource and inventory	+ Relaxing exploration
	+ Fun exploration gameplay	+ Massive universe to explore.
		+ It truly is an impossibly huge galaxy.
		+ A sense of majesty and grandeur unlike anything else.
		+ Lots of options to fiddle with.
		+ Near limitless replay value.
		+ Huge scale, infinite content.
		+ Solid survival gameplay with great freedom.
		+ Relaxing exploration.
	+ Need story to make better	− Very little real story.
		− No reason to proceed, lacks a narrative...
	− Repetitive, boring gameplay	− a lack of real discovery
		− Most planets look the same
		− repetitive systems
		− Repetitive
		− Dull, tedious crafting.
		− Planets all hold the same handful of interest points.
		− ... disappoints in almost every way and just has no depth
		− ... gameplay options extremely limited.
		− ... Has too few features to be varied in the long run.
		− soon turns into a routine stereotype...
		− The universe is a lifeless and static backdrop.
	− Lack of inventory upgrade	− Loads of inventory management.
	− Fly, explore, combat	− Not for thrill seekers or combat fans.
		− whilst gathering resources to move on but won't linger.
Usability	**+ Control feels with controller, fly ship**	
	+ Beautiful graphics	+ Beautiful alien worlds.
		+ Breathtaking views.
		+ Stylish in graphics ...
		+ some lovely scenery
		+ An atmospheric walk through beautiful worlds
	+ Music and sound, hold and click button	+ stylish..sound
		+ A successful.. atmospheric audiovisual implementation.
	− Graphic settings poor, restart	
	− Fly control with mouse annoying	
	− Terrible texture and sound	
	− Horrible flight control FOV, cluncky inventory	− Uncomfortable controls
	− Option, click and hold button, bad/awful PC port	− frustrating menus
		+ It may work perfectly as an occasional short distraction
		− Many promises left undelivered

"+" represents positive, "−" represents negative.

5. Discussion

Considering that other factors can also influence the outcome of the playability analysis, we extended the experiments using the *playtime* of the players and the *voted helpfulness* value as the weight to the sentiment score. The playtime value indicates how long each player has been playing the game, i.e., the game experience. It is reasonable to assume that players who spend more time on a particular game with more gaming experience shall provide more trustworthy reviews. On the other hand, the voted helpfulness value indicates how many other players agree with the statement and evaluation in a particular review, i.e., the perceived trustworthiness. According to our review data, among the players who wrote the reviews, the longest playtime of one player is 645,618 min (≈10,760 h) with the shortest being 1 min. The average playtime is 5791.70 min (≈96.53 h). Meanwhile, the highest helpfulness score received by a single review is 12,236 with the lowest being zero. The average helpfulness score is 6.42. By adding the normalized "playtime" and "voted helpfulness" values as weights to the sentiment score of each review sentence, we can obtain the weighted playability score for each perspective as follows: Functionality of -0.1985, Gameplay of -0.1815, and Usability of -0.1983 with the scale of $(-1, 1)$. This result shows that the overall playability of this game is slightly under mediocre. Furthermore, similar experiments with the reviews between updates show that the playability of the game is still increasing through updates, but the values are slightly negative. This phenomenon shows that experienced players and popular reviews can have obvious influence on the playability analysis result when their opinions are credited with more value. However, how to verify the influence of the players' experience and the credibility of their reviews towards the analysis result of playability shall be further investigated in future studies.

On the other hand, as shown in Table 4, the majority (61.2%) of the reviews are given before the first update of the game. Thus, the playability of Release 1.0 likely has a greater influence on the overall score than that of the rest. Therefore, it is reasonable such unevenness is also taken into account. Comparatively, for a similar situation in review-based analysis, the time sequence factor was considered by Chen et al. when evaluating the informativeness of mobile application reviews [26]. However, we are unable to conclude that the newest reviews accurately reflect the current playability of the game without further investigation on the content of such reviews compared to the older ones. A study on the changes of reviewers' opinions regarding the evolution of the target system (similar to the one in [27]) shall be conducted towards tackling such issues.

It is worth emphasizing that the proposed approach can be adapted by considering any proposed playability heuristics when such heuristic-oriented issues are sufficiently mentioned by the players. Due to the nature of heuristics being a checklist of principles [11], it is thus possible to extract players' opinions according to different heuristics via labeling training data accordingly. Although the outcome of applying different heuristics could certainly differ, it is a potentially good practice towards detecting more playability issues. Thus, a comparative study on applying this approach with different playability heuristics shall be conducted in the future work.

Furthermore, an obvious limitation of this approach is the requirement of a large number of player reviews, which is impossible before the release of the game. Heuristic evaluation of playability is an effective way to target such a situation. Comparatively, our approach aims for the continuous maintenance and evolution of games after their releases, where playability evaluation can be conveniently automated through this data mining pipeline with sufficient review data collected. The gap between experts' and end players' opinions is, to a certain extend, inevitable [52]. Hence, our approach can contribute to helping the developers better understand the needs and complaints of the players. Based on that, they can improve the games continuously and effectively.

6. Conclusions

In this paper, we propose a data-driven approach for analyzing the playability of video games based on the players' reviews. Focusing on the collective opinions of a large number of players, this approach provides an effective solution for understanding the overall playability of a particular video game as well as the detailed merits and defects within each pre-defined playability perspective. The results of this study show that the proposed approach can provide fair evaluation and analysis in terms of video game playability with satisfactory accuracy. Compared to the mainstream heuristic evaluation method, our approach contributes specifically to the maintenance and evolution of video games by helping game developers understand the collective needs and complaints of real players. The approach can be improved by taking into account other factors that influence the playability analysis: the playtime, voted helpfulness, player preferences, etc. The different evaluation results by selecting different playability frameworks or using different playability heuristics shall be further investigated comparatively. Furthermore, more video game cases, especially from different genres, shall be used for verification and comparison. We shall also further investigate the credibility of game players as reviewers based on their reviewing behaviors and gaming profiles via computational methods. Such studies shall contribute to the enrichment of the playability and player behavior analysis methodologies.

Author Contributions: Conceptualization, X.L., Z.Z. and K.S.; data curation, X.L.; formal analysis, X.L.; investigation, X.L., Z.Z. and K.S.; methodology, X.L.; software, X.L.; supervision, Z.Z. and K.S.; validation, X.L.; visualization, X.L.; writing—original draft, X.L.; and writing—review and editing, X.L., Z.Z. and K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Faculty of Information Technology and Communication Sciences, Tampere University and the Ulla Tuominen Foundation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.14222531.v1, accessed on 16 March 2021.

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

References

1. Sánchez, J.L.G.; Vela, F.L.G.; Simarro, F.M.; Padilla-Zea, N. Playability: Analysing user experience in video games. *Behav. Inf. Technol.* **2012**, *31*, 1033–1054. [CrossRef]
2. Järvinen, A.; Heliö, S.; Mäyrä, F. *Communication and Community in Digital Entertainment Services*; Prestudy Research Report; University of Tampere: Tampere, Finland, 2002.
3. Fabricatore, C.; Nussbaum, M.; Rosas, R. Playability in action videogames: A qualitative design model. *Hum.-Comput. Interact.* **2002**, *17*, 311–368. [CrossRef]
4. Desurvire, H.; Caplan, M.; Toth, J.A. Using heuristics to evaluate the playability of games. In Proceedings of the CHI'04 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria, 24–29 April 2004; pp. 1509–1512.
5. Korhonen, H.; Koivisto, E.M. Playability heuristics for mobile games. In Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services, Helsinki, Finland, 12–15 September 2006; pp. 9–16.
6. ISO/IEC. *Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Measurement of Quality in Use*; Standard 25022:2016; International Organization for Standardization: Geneva, Switzerland, 2016.
7. Sánchez, J.L.G.; Simarro, F.M.; Zea, N.P.; Vela, F.L.G. *Playability as Extension of Quality in Use in Video Games*; I-USED: 2009. Available online: https://lsi2.ugr.es/juegos/articulos/iused09-jugabilidad.pdf (accessed on 16 March 2021).
8. Paavilainen, J. Defining playability of games: functionality, usability, and gameplay. In Proceedings of the 23rd International Conference on Academic Mindtrek, Tampere, Finland, 29–30 January 2020; pp. 55–64.
9. Korhonen, H.; Paavilainen, J.; Saarenpää, H. Expert review method in game evaluations: Comparison of two playability heuristic sets. In Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era, Tampere, Finland, 30 September–2 October 2009; pp. 74–81.
10. Pinelle, D.; Wong, N.; Stach, T. Heuristic evaluation for games: Usability principles for video game design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, 5–10 April 2008; pp. 1453–1462.
11. Nielsen, J.; Molich, R. Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 1990; pp. 249–256.
12. Pitoura, E.; Koutrika, G.; Stefanidis, K. Fairness in Rankings and Recommenders. In Proceedings of the Advances in Database Technology—EDBT 2020, 23rd International Conference on Extending Database Technology, Copenhagen, Denmark, 30 March–2 April 2020. Available online: OpenProceedings.org (accessed on 16 March 2021).
13. Pinelle, D.; Wong, N.; Stach, T.; Gutwin, C. Usability heuristics for networked multiplayer games. In Proceedings of the ACM 2009 International Conference on Supporting Group Work, Sanibel Island, Fl, USA, 10–13 May 2009; pp. 169–178.
14. Hannula, R.; Nikkilä, A.; Stefanidis, K. GameRecs: Video Games Group Recommendations. In *ADBIS*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 513–524.
15. Gong, J.; Ye, Y.; Stefanidis, K. A Hybrid Recommender System for Steam Games. In *ISIP*; Springer: Cham, Switzerland, 2019; pp. 133–144.
16. Klancar, J.; Paulussen, K.; Stefanidis, K. FIFARecs: A Recommender System for FIFA18. In *ADBIS*; Springer: Cham, Switzerland, 2019; pp. 525–536.
17. Stefanidis, K.; Pitoura, E.; Vassiliadis, P. Managing contextual preferences. *Inf. Syst.* **2011**, *36*, 1158–1180. [CrossRef]
18. Jacobsen, N.E.; Hertzum, M.; John, B.E. The evaluator effect in usability tests. In Proceedings of the CHI 98 Conference Summary on Human Factors in Computing Systems, Los Angeles, CA, USA, 18–23 April 1998; pp. 255–256.

19. Juul, J. *Half-Real: Video Games between Real Rules and Fictional Worlds*; MIT Press: Cambridge, MA, USA, 2011.
20. CDProjekt. The Witcher 3: Wild Hunt [CD-ROM, PC, XBOX, Playstation]. 2015. Available online: https://thewitcher.com/en/witcher3/ (accessed on 16 March 2021).
21. Burnett, M.; Cook, C.; Rothermel, G. End-user software engineering. *Commun. ACM* **2004**, *47*, 53–58. [CrossRef]
22. Ko, A.J.; Abraham, R.; Beckwith, L.; Blackwell, A.; Burnett, M.; Erwig, M.; Scaffidi, C.; Lawrance, J.; Lieberman, H.; Myers, B.; et al. The state of the art in end-user software engineering. *ACM Comput. Surv. (CSUR)* **2011**, *43*, 21. [CrossRef]
23. Hearst, M.A. Untangling text data mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, ML, USA, 1 June 1999; pp. 3–10.
24. Liu, B. Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167. [CrossRef]
25. Fu, B.; Lin, J.; Li, L.; Faloutsos, C.; Hong, J.; Sadeh, N. Why people hate your app: Making sense of user feedback in a mobile app store. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Chicago, IL, USA, 11–14 August 2013; pp. 1276–1284.
26. Chen, N.; Lin, J.; Hoi, S.C.; Xiao, X.; Zhang, B. AR-miner: Mining informative reviews for developers from mobile app marketplace. In Proceedings of the 36th International Conference on Software Engineering, ACM, Hyderabad, India, 1 May 2014; pp. 767–778.
27. Li, X.; Zhang, Z.; Stefanidis, K. Mobile App Evolution Analysis Based on User Reviews. In Proceedings of the 17th International Conference on Intelligent Software Methodologies, Tools, and Techniques, Naples, Italy, 15–17 September 2018; pp. 773–786.
28. Lin, D.; Bezemer, C.P.; Zou, Y.; Hassan, A.E. An empirical study of game reviews on the Steam platform. *Empir. Softw. Eng.* **2019**, *24*, 170–207. [CrossRef]
29. Nielsen, J. Usability inspection methods. In Proceedings of the Conference Companion on Human Factors in Computing Systems, Boston, MA, USA, 23–28 April 1994; pp. 413–414.
30. Nielsen, J. How to conduct a heuristic evaluation. *Nielsen Norman Group* **1995**, *1*, 1–8.
31. Korhonen, H. Comparison of playtesting and expert review methods in mobile game evaluation. In Proceedings of the 3rd International Conference on Fun and Games, Leuven, Belgium, 15–17 September 2010; pp. 18–27.
32. Malone, T.W. Heuristics for designing enjoyable user interfaces: Lessons from computer games. In Proceedings of the 1982 Conference on Human Factors in Computing Systems, Gaithersburg, Ml, USA, 15–17 March 1982; pp. 63–68.
33. Federoff, M.A. Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games. Ph.D. Thesis, Indiana University, Bloomington, IN, USA, 2002.
34. Desurvire, H.; Wiberg, C. Game usability heuristics (PLAY) for evaluating and designing better games: The next iteration. In *International Conference on Online Communities and Social Computing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 557–566.
35. Korhonen, H.; Koivisto, E.M. Playability heuristics for mobile multi-player games. In Proceedings of the 2nd International Conference on Digital Interactive Media in Entertainment and Arts, Perth, Australia, 19–21 September 2007; pp. 28–35.
36. Koeffel, C.; Hochleitner, W.; Leitner, J.; Haller, M.; Geven, A.; Tscheligi, M. Using heuristics to evaluate the overall user experience of video games and advanced interaction games. In *Evaluating User Experience in Games*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 233–256.
37. Röcker, C.; Haar, M. Exploring the usability of videogame heuristics for pervasive game development in smart home environments. In Proceedings of the Third International Workshop on Pervasive Gaming Applications–Pergames, Dublin, Ireland, 7 May 2006; pp. 199–206.
38. Tan, J.L.; Goh, D.H.L.; Ang, R.P.; Huan, V.S. Usability and playability heuristics for evaluation of an instructional game. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*; Association for the Advancement of Computing in Education (AACE): Waynesville, NC, USA, 2010; pp. 363–373.
39. Khanana, K.; Law, E.L.C. Designing children's digital games on nutrition with playability heuristics. In Proceedings of the CHI'13 Extended Abstracts on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 1071–1076.
40. Aker, Ç.; Rızvanoğlu, K.; Bostan, B. Methodological Review of Playability Heuristics. *Proc. Eurasia Graph. Istanb. Turk.* **2017**, *405*. Available online: https://www.researchgate.net/profile/Kerem-Rizvanoglu/publication/321623742_Eurasia_2017_Brave_New_Worlds_Conference_on_Virtual_and_Interactive_Realities/links/5a292400aca2727dd8872361/Eurasia-2017-Brave-New-Worlds-Conference-on-Virtual-and-Interactive-Realities.pdf (accessed on 16 March 2021).
41. Matera, M.; Costabile, M.F.; Garzotto, F.; Paolini, P. SUE inspection: An effective method for systematic usability evaluation of hypermedia. *IEEE Trans. Syst. Man, Cybern. Part A Syst. Hum.* **2002**, *32*, 93–103. [CrossRef]
42. Vasa, R.; Hoon, L.; Mouzakis, K.; Noguchi, A. A preliminary analysis of mobile app user reviews. In Proceedings of the 24th Australian Computer-Human Interaction Conference, Sydney, Australia, 20–24 November 2012; pp. 241–244.
43. Hoon, L.; Vasa, R.; Schneider, J.G.; Mouzakis, K. A preliminary analysis of vocabulary in mobile app user reviews. In Proceedings of the 24th Australian Computer-Human Interaction Conference, Sydney, Australia, 20–24 November 2012; pp. 245–248.
44. Harman, M.; Jia, Y.; Zhang, Y. App store mining and analysis: MSR for app stores. In Proceedings of the 2012 9th IEEE Working Conference on Mining Software Repositories (MSR), Zurich, Switzerland, 2–3 June 2012; pp. 108–111.
45. Vu, P.M.; Nguyen, T.T.; Pham, H.V.; Nguyen, T.T. Mining user opinions in mobile app reviews: A keyword-based approach (t). In Proceedings of the 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), Lincoln, NE, USA, 9–13 November 2015; pp. 749–759.

46. Panichella, S.; Di Sorbo, A.; Guzman, E.; Visaggio, C.A.; Canfora, G.; Gall, H.C. How can i improve my app? Classifying user reviews for software maintenance and evolution. In Proceedings of the 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME), Bremen, Germany, 29 September–1 October 2015; pp. 281–290.
47. Gu, X.; Kim, S. what parts of your apps are loved by users? (T). In Proceedings of the 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), Lincoln, NE, USA, 9–13 November 2015; pp. 760–770.
48. Khalid, H. On identifying user complaints of iOS apps. In Proceedings of the IEEE 2013 35th International Conference on Software Engineering (ICSE), San Francisco, CA, USA, 18–26 May 2013; pp. 1474–1476.
49. Guzman, E.; Maalej, W. How do users like this feature? A fine grained sentiment analysis of app reviews. In Proceedings of the 2014 IEEE 22nd International Requirements Engineering Conference (RE), Karlskrona, Sweden, 25–29 August 2014; pp. 153–162.
50. Li, X.; Zhang, Z.; Stefanidis, K. Sentiment-Aware analysis of mobile apps user reviews regarding particular updates. In Proceedings of the Thirteenth International Conference on Software Engineering Advances, Nice, France, 14–18 October 2018; p. 109.
51. Li, X.; Zhang, B.; Zhang, Z.; Stefanidis, K. A Sentiment-Statistical Approach for Identifying Problematic Mobile App Updates Based on User Reviews. *Information* **2020**, *11*, 152. [CrossRef]
52. Santos, T.; Lemmerich, F.; Strohmaier, M.; Helic, D. What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic. *Proc. ACM Hum.-Comput. Interact.* **2019**, *3*, 1–22. [CrossRef]
53. Lu, C.; Li, X.; Nummenmaa, T.; Zhang, Z.; Peltonen, J. Patches and Player Community Perceptions: Analysis of No Man's Sky Steam Reviews. DiGRA. In Proceedings of the 2020 DiGRA International Conference: Play Everywhere, Lüneburg, Germany, 14–17 May 2020.
54. Nigam, K.; McCallum, A.K.; Thrun, S.; Mitchell, T. Text classification from labeled and unlabeled documents using EM. *Mach. Learn.* **2000**, *39*, 103–134. [CrossRef]
55. Zhang, M.L.; Zhou, Z.H. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* **2007**, *40*, 2038–2048. [CrossRef]
56. Chen, W.J.; Shao, Y.H.; Li, C.N.; Deng, N.Y. MLTSVM: A novel twin support vector machine to multi-label learning. *Pattern Recognit.* **2016**, *52*, 61–74. [CrossRef]
57. Spyromitros, E.; Tsoumakas, G.; Vlahavas, I. An empirical study of lazy multilabel classification algorithms. In *Hellenic Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 401–406.
58. Gilbert, C.; Hutto, E. Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). 2014; Volume 81, p. 82. Available online: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf (accessed on 16 April 2020).
59. Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2544–2558. [CrossRef]
60. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
61. HelloGames. No Man's Sky [CD-ROM, PC, XBOX, Playstation]. 2016. Available online: https://www.nomanssky.com/ (accessed on 16 March 2021).
62. Syed, S.; Spruit, M. Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation. In Proceedings of the 2017 IEEE International conference on data science and advanced analytics (DSAA), Tokyo, Japan, 19–21 October 2017; pp. 165–174.
63. Greenwood-Ericksen, A.; Poorman, S.R.; Papp, R. On the validity of Metacritic in assessing game value. *Eludamos. J. Comput. Game Cult.* **2013**, *7*, 101–127.
64. Bossert, M.A. *Predicting Metacritic Film Reviews Using Linked Open Data and Semantic Technologies*; KNOW@ LOD: Portorož, Slovenia, 2015.