IADIS International Journal on WWW/Internet
Vol. 22, No. 2, pp. 1-15
ISSN: 1645-7641

PREDICTING REVIEW HELPFULNESS IN VIDEO GAMES: A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS AND NLP INTEGRATION

María Olmedilla[1], Leonardo Espinosa-Leal[2], José Carlos Romero-Moreno[3] and Zhen Li[2]
[1]*SKEMA Business School – Université Côte d'Azur, France*
[2]*Arcada University of Applied Sciences, Graduate Studies and Research, Finland*
[3]*Audencia Business School, Nantes, France*

ABSTRACT

This paper investigates the prediction of video game review helpfulness on the Steam platform using machine learning and natural language processing (NLP) techniques. We applied three models—XGBoost, Extreme Learning Machine (ELM), and Ridge regression—to predict helpfulness scores as both a regression and binary classification problem. XGBoost demonstrated the best performance, while ELM offered a computationally efficient alternative. Text features generated from DistilBERT were incorporated, but their inclusion did not significantly enhance model accuracy. Our findings suggest that non-textual features, such as review length, playtime, and helpful votes, are more influential in determining helpfulness. Early predictions of review helpfulness could benefit users by highlighting valuable feedback and aiding developers in refining their games. Future research will explore fine-tuning NLP models on larger datasets and incorporating additional features, such as sentiment analysis, to improve performance.

KEYWORDS

User Reviews, Extreme Machine Learning, Steam, Binary Classification, NLP, Review Helpfulness Prediction

1. INTRODUCTION

Since the release of the first home console version of Pong in 1977, the video game industry has evolved dramatically, growing into a global entertainment powerhouse. In recent years, this sector has seen exponential growth, with annual global revenues approaching $218 billion by 2021, driven in part by the proliferation of gaming platforms, enhanced game development

technologies, and a surge in digital content consumption during the COVID-19 pandemic (Juniper Research, 2021). Video games are no longer just a source of entertainment; they have become a significant cultural phenomenon and a dynamic industry that continues to push the boundaries of technology and user engagement.

Within this vast digital landscape, user-generated content—specifically, online reviews—has become critical in influencing consumer behavior and shaping market trends. Online platforms like Steam, which serve as digital distribution hubs for video games, provide spaces where users can share their gaming experiences, review games, and offer highly valuable feedback to other users and developers alike. Steam, in particular, dominates the PC gaming market and has a vibrant community of gamers who actively contribute reviews, which are instrumental in shaping purchasing decisions and helping developers make data-driven adjustments to their games (Lin et al., 2019). However, not all reviews are equally helpful, and identifying the most valuable feedback has become a challenging yet essential task. Understanding what makes a review helpful is crucial for both consumers and game developers. Consumers rely on reviews to make informed purchasing decisions, while developers use feedback to refine their games and marketing strategies. Prior studies have explored various factors that influence review helpfulness, including review length, sentiment, and the reviewer's experience level. Despite these efforts, predicting review helpfulness accurately remains a complex problem, particularly when attempting to generalize findings across different types of games and user bases (Wu, 2017).

This study aims to address this gap by leveraging cutting-edge machine learning techniques to predict the helpfulness of Steam user reviews, utilizing a comprehensive dataset comprising nearly one million reviews. Our analysis focuses on two approaches: (1) framing the problem as a regression task, where we predict the helpfulness score on a continuous scale, and (2) modeling it as a binary classification problem, categorizing reviews as either helpful or not based on specific thresholds. We also explore the potential of natural language processing (NLP) models to enhance prediction accuracy by analyzing the textual content of reviews. However, contrary to initial expectations, our findings indicate that incorporating features from pre-trained NLP models does not significantly improve prediction performance, raising important questions about the role of text-based features in predicting review helpfulness. By providing insights into the predictive power of various features—ranging from review length and playtime to review content—this research contributes to the growing field of user-generated content analysis within the video game industry. Our findings offer practical implications for game developers, enabling them to harness review data to optimize game design, improve user engagement, and refine marketing strategies.

The remainder of this paper is structured as follows: Section 2 reviews the background literature and the evolution of the video game industry, mainly focusing on the Steam platform and the factors influencing review helpfulness. Section 3 details the methodology employed in our analysis, including the machine learning models and features used for prediction. Section 4 discusses the results of our experiments, comparing the performance of regression and classification models. Finally, Section 5 presents the conclusions and outlines potential avenues for future research.

2. RESEARCH BACKGROUND

The video game industry has transformed drastically over the past few decades, evolving from its humble beginnings as a niche entertainment medium into a multi-billion-dollar global powerhouse. Significant technological innovations, cultural shifts, and changing consumer behaviors have shaped the industry's growth trajectory. Today, video games are not only a dominant form of entertainment but also a critical component of the digital economy. The convergence of diverse gaming devices—from traditional consoles to PCs and mobile phones—has expanded the accessibility of gaming to an unprecedented degree, bringing it into the daily lives of billions of users worldwide.

Alongside this transformation, the rise of digital distribution platforms has reshaped how games are bought, sold, and consumed. Physical game sales, which once dominated the market, have been rapidly eclipsed by digital sales, a trend accelerated by the global shift towards e-commerce and the increasing convenience of cloud-based services. The growing dominance of digital platforms has opened new avenues for developers to engage with their audiences directly, creating a continuous feedback loop that has significantly altered game development, marketing, and customer engagement strategies. The importance of user-generated content, particularly reviews, cannot be underestimated in this scenario. Online reviews are now pivotal in influencing purchase decisions, driving player engagement, and shaping the reputation of games in competitive markets.

However, while user reviews can be immensely valuable, the sheer volume of content generated across platforms such as Steam presents a significant challenge: how to effectively identify and amplify the most helpful reviews. This research aims to address this challenge by using predictive modeling to analyze review helpfulness, offering insights that can enhance both user experience and developer decision-making processes.

2.1 Video Games and Steam Game Platform

The evolution of video games from simple arcade machines to the immersive, complex experiences available today mirrors broader technological advancements and cultural shifts. Video games have grown into a diverse medium with offerings that span genres, narratives, and gameplay mechanics, reaching a global audience through various platforms. Modern video games are developed for a wide range of devices, from dedicated gaming consoles and high-performance personal computers to the ever-present smartphones, which have brought gaming into the pockets of billions. The increasing ubiquity of these devices, coupled with advancements in internet infrastructure, has significantly broadened the accessibility and appeal of video games.

Digital distribution has revolutionized how consumers access games, shifting the industry from physical copies to more flexible and cost-effective digital formats. Platforms like Steam have played a pivotal role in this transition, allowing developers to reach a global audience without the constraints of physical production or traditional retail channels. The advantages of digital distribution are clear: lower costs, broader reach, and the ability to quickly update and improve games post-launch. According to MarketLine (2015, 2021), the shift to digital aligns with broader e-commerce trends, where convenience and immediacy are prized. By 2020, digital game sales vastly outpaced physical copies in revenue per user, highlighting the economic benefits of the digital ecosystem (Winter, 2021).

This digital shift has also facilitated innovation in game design and monetization. Developers now have the flexibility to experiment with new genres, gameplay mechanics, and monetization models, such as free-to-play games with in-game purchases or downloadable content (DLC). Digital platforms like Steam foster a direct relationship between game developers and players, enabling rapid feedback through reviews, forums, and user-generated content. This creates a more iterative and dynamic development process, where games can evolve over time based on player input. However, this dynamic also presents challenges, particularly in maintaining player engagement in a marketplace where hundreds of new titles are released monthly. In such a competitive environment, user reviews become crucial tools for both discovery and engagement.

Steam, developed by Valve Corporation, has been at the forefront of digital game distribution since its launch in 2003. Initially designed to simplify game updates and combat piracy, Steam has grown into the leading digital marketplace for PC gaming, offering a vast library of games, downloadable content, and community features. It serves not only as a platform for purchasing games but also as a vibrant social hub where users can interact, share content, and participate in discussions. Steam's unique blend of commerce and community engagement has made it an indispensable tool for both players and developers, offering detailed analytics on game performance, sales data, and, crucially, user reviews. These reviews play a significant role in shaping game development strategies and marketing efforts, providing real-time feedback that developers can use to refine their games (Lin et al., 2019).

2.2 Rating Helpfulness of Online Reviews

The rise of user-generated content has redefined how consumers interact with products and services online, and nowhere is this more evident than in the gaming industry. Reviews, ratings, and feedback provided by users are now integral parts of the decision-making process for consumers navigating digital marketplaces. On platforms like Steam, user reviews carry significant weight, influencing purchasing decisions and shaping the broader discourse around video games. For potential buyers, these reviews offer a window into the game's quality, performance, and enjoyment, while for developers, they represent a direct line to their audience's opinions and experiences. However, not all reviews are created equal, and the ability to discern which reviews are genuinely helpful is vital for both consumers and game creators.

The task of predicting review helpfulness is challenging, as it involves parsing not only the textual content of reviews but also contextual factors such as the reviewer's credibility, engagement, and relevance of their feedback. As the volume of reviews on platforms like Steam continues to grow, manually identifying helpful reviews becomes impractical, underscoring the need for automated solutions. Machine learning techniques have emerged as powerful tools in this domain, enabling the prediction of review helpfulness by analyzing large datasets and extracting meaningful patterns. Researchers like Olmedilla et al. (2022) have shown that computational models can provide valuable insights into review effectiveness, and there is a growing body of academic work dedicated to refining these techniques.

Despite advances in predictive modeling, accurately assessing review quality remains a complex problem. One of the main challenges is the subjective nature of helpfulness, which can vary significantly between users based on personal preferences, game knowledge, and expectations. Previous studies have attempted to correlate factors such as review length, sentiment, and the experience level of the reviewer with helpfulness ratings, but the results have often been mixed (Wu, 2017). The diversity of platforms and the differing criteria users apply when judging helpfulness further complicate the task (Krishnamoorthy, 2015). Additionally, simple metrics like keyword frequency or review length may not capture the deeper nuances of what makes a review truly useful to prospective buyers.

In the specific context of video game reviews, the Steam platform offers a wealth of data for analysis. Prior research has identified several key factors that influence review helpfulness, including the length of the review, the amount of time the reviewer has spent playing the game, and the game's genre. However, there is still much to explore regarding the textual content of reviews and how it affects their perceived helpfulness. Moving beyond basic text metrics, advanced semantic analysis techniques, such as sentiment analysis and topic modeling, offer the potential to uncover more meaningful insights into what makes a review stand out (Eberhard et al., 2018; Kasper et al., 2019). By leveraging these sophisticated tools, researchers can develop more accurate models for predicting review helpfulness, ultimately enhancing the user experience and helping developers identify and prioritize valuable feedback.

3. METHODOLOGY

3.1 Data Description and Preprocessing

Our study is based on a publicly available dataset from Kaggle (Kamel, 2022), which provides a comprehensive collection of video game reviews from the Steam platform. The original dataset contains **1,124,903 rows** and **10 columns**, representing reviews of **14,469 different games**. The features included in this dataset are:

- **game_id** (the unique identifier for each game),
- **review_id** (the unique identifier for each review),
- **steamid** (the user's unique identifier),
- **playtime_forever** (the total playtime of the game for the reviewer),
- **playtime_at_review** (the playtime at the moment the review was written),
- **pos_review** (a binary indicator showing if the review was marked as positive or negative),
- **votes_relevant** (the number of users who found the review helpful),
- **review_text** (the actual content of the review),
- **weighted_vote_score** (a numerical measure of the review's perceived helpfulness), and
- **steam_purchase** (indicating whether the game was purchased on Steam or obtained through another means, such as a key from a third-party seller).

To enrich the dataset's utility for machine learning purposes, we created three additional features:

1. **Binarization of 'pos_review'**: We converted the 'pos_review' column into a binary format, marking reviews as either positive (1) or negative (0). This simplification allows for easier integration into classification models.
2. **Binarization of 'steam_purchase':** Similarly, we binarized the 'steam_purchase' feature to indicate whether the game was directly purchased from Steam (1) or obtained elsewhere (0). This feature could potentially capture the differences in review behavior between direct purchasers and those who acquired the game through external sources.
3. **Review length:** We introduced a feature measuring the length of each review in characters. This feature was particularly relevant as previous studies have shown a correlation between review length and perceived helpfulness. To standardize the dataset for natural language processing (NLP) tasks, we filtered out reviews exceeding **512 characters**, a typical maximum input size limitation for most NLP models, including those from the transformer family (e.g., BERT, DistilBERT).

After this preprocessing, the dataset was reduced to **898,326 rows**, and reviews that did not meet the length criteria were filtered out.

3.2 Data Visualization and Inspection

To better understand the structure and distribution of the dataset, we plotted several key metrics (see *Figure 1*). These visualizations included:

- **The frequency of sentence length** (top-left of Figure 1) reveals the distribution of review lengths. This provided insight into the variability of review detail, which is often a proxy for the depth of user engagement.
- **The ordered number of reviews per Game ID** (top-right) highlights that only a small subset of games accounted for the majority of reviews, reflecting a skewed engagement where popular games receive disproportionately more feedback.
- **The distribution of the helpfulness score (weighted_vote_score)** (bottom-left) showed a concentration of reviews with a score of zero, suggesting that many reviews were either not voted on or were considered unhelpful by the community.

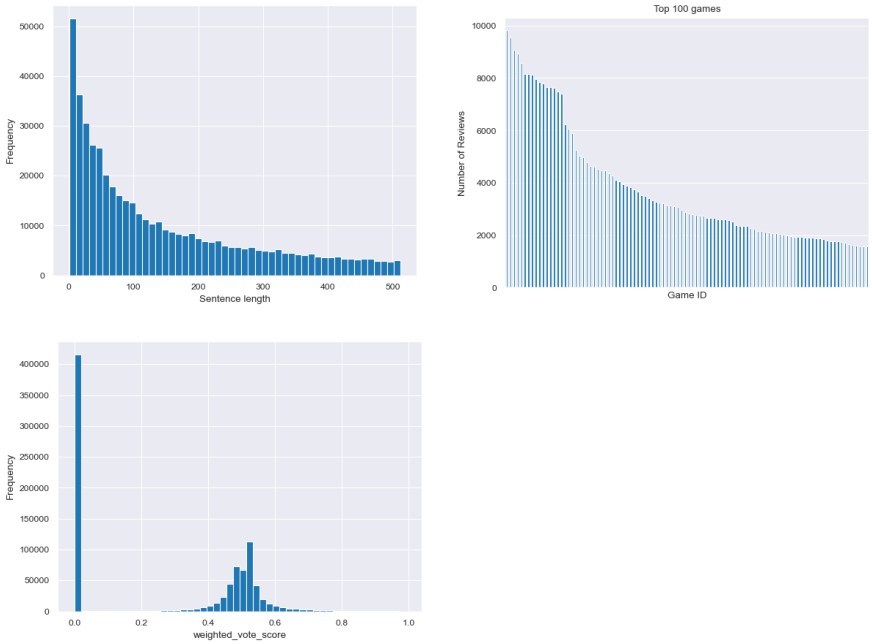

Figure 1. Characteristics of the original dataset used in this work. Top-left: Frequency of the sentence length. Top-right: Ordered number of reviews versus Game ID and Bottom-left: The frequency distribution of the helpfulness score (weighted_vote_score)

3.3 Data Cleaning and Refinement

After thoroughly inspecting the data, we discovered that approximately 50% of the reviews had a helpfulness score of zero. This suggested that half of the reviews either received no votes or were deemed irrelevant by the voting community. Since these reviews provided little to no value for predicting helpfulness, we removed them from the dataset. The final dataset used for training and analysis was reduced to 482,460 rows. This cleaning step helped improve the dataset's quality by focusing on reviews that had received some form of community validation.

Further insights were obtained by analyzing the correlation between different features. We calculated a Pearson correlation matrix to identify the relationships between the key variables (see Figure 2, bottom-right). The correlation analysis showed that:

- **votes_relevant** had the highest correlation with the helpfulness score at 0.36, indicating that the number of relevant votes is the strongest predictor of perceived helpfulness.
- **pos_review_bin** followed with a correlation of 0.21, suggesting that positive reviews tend to be more helpful, likely because they align with user expectations or the community's sentiment.
- **length** also showed a moderate correlation of 0.13, supporting the notion that longer reviews are often seen as more informative or valuable.

Other features, such as **review_id** (0.098), **game_id** (0.072), **playtime_forever** (0.05), and **playtime_at_review_new** (0.047), showed weaker correlations. Interestingly, features like

steam_purchase_bin and **steamid** had very low correlations (0.046 and 0.013, respectively), suggesting that these factors play a relatively minor role in determining review helpfulness.

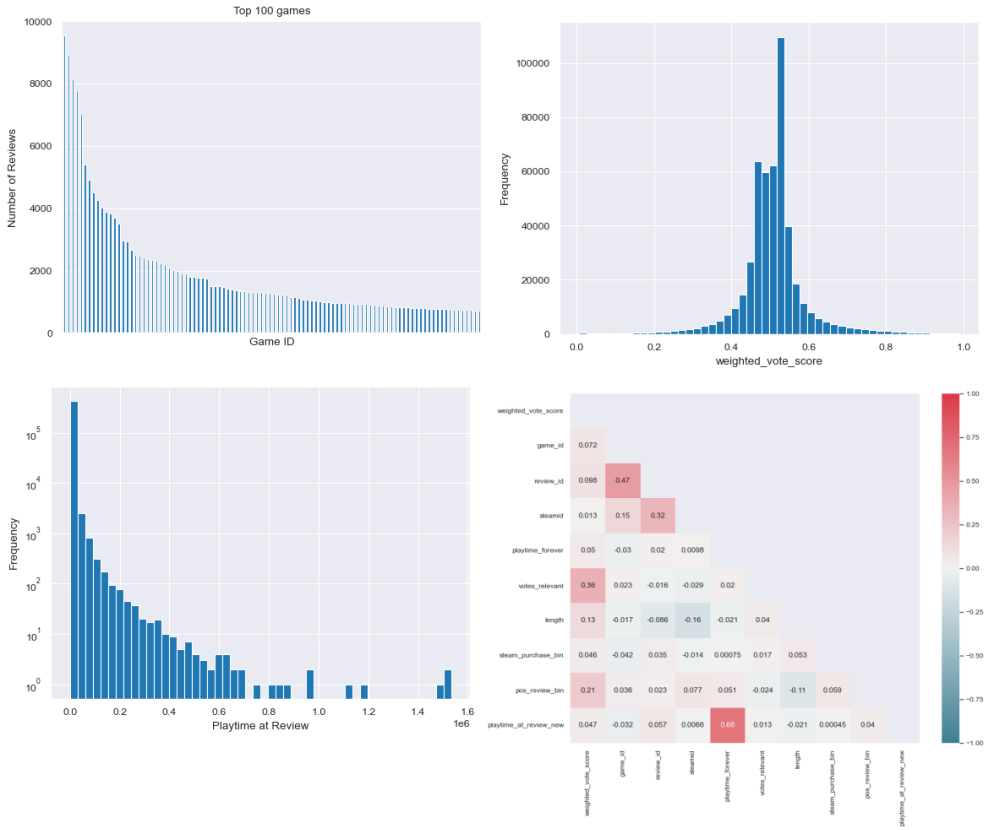

Figure 2. Characteristics of the dataset after filtering used in this work. Top-left: Ordered number of reviews versus Game ID. Top-right: The frequency distribution of the helpfulness score (weighted_vote_score). Bottom-left: Frequency distribution of the play

3.4 Feature Engineering and Model Training:

For the predictive modeling tasks, we used this refined dataset to train both Regression and binary classification models. These models aimed to predict the helpfulness score using a variety of features, with special attention given to user engagement metrics (e.g., playtime and votes) and text-based features (e.g., review length). The inclusion of textual content in combination with user behavior data allowed for a complete approach to predicting helpfulness.

4. RESULTS

In this study, we explored two distinct approaches for predicting the helpfulness of video game reviews on Steam: **Regression** and **binary classification**. We utilized three different machine learning models: **Ridge regression, Extreme Learning Machine (ELM), and Extreme Gradient Boosting (XGBoost).** These models were selected for their ability to handle various complexities of the data, including non-linear relationships and high-dimensional feature spaces.

All models were implemented using well-established Python libraries—**scikit-learn** (Pedregosa et al., 2011) and **scikit-ELM** (Akusok et al., 2021). Each model underwent a rigorous **randomized grid search** of 50 steps for hyperparameter optimization, coupled with a three-fold cross-validation to ensure robust model training. The dataset was split with a **70/30 ratio** between training and testing sets to evaluate the generalization performance of the models. Additionally, the features were **normalized** before training for the Ridge regression and ELM models to ensure that all input variables were on a comparable scale.

4.1 Regression Results

Table 1 presents a summary of the results for the three machine learning models when the problem was treated as a regression task, where the goal was to predict the exact helpfulness score of reviews on a continuous scale.

Table 1. Summary of results for the three machine learning models for predicting the helpfulness score as a regression problem

Models	**Optimized parameters**	**Best-score (train/test) R^2**
Ridge	alpha: 2020.2	0.22/0.21
ELM	alpha: 2.08e-05, density: 0.22, include_original_features: False, n_neurons: 3270, pairwise_metric: cityblock, ufunc: sigm	0.52/0.51
XGBoost	colsample_bytree: 0.51, learning_rate: 0.09, max_depth: 8, n_estimators: 861, scale_pos_weight: 3, subsample: 0.81	0.61/0.60

The regression results indicate that **XGBoost** achieved the highest performance, with an **R^2 score of 0.61 on the training set** and **0.60 on the test set**. This model outperformed the other two in terms of both training and test accuracy, owing to its ability to handle non-linearities and complex feature interactions. XGBoost's advanced tree-based architecture made it particularly suited to capturing the nuanced patterns in the review data. The **ELM model** also performed well, with **R^2 scores of 0.52/0.51** on the train and test sets, respectively. While it did not achieve the same level of accuracy as XGBoost, ELM has the advantage of **computational efficiency**. Due to its fast-training process, ELM presents a more favorable **performance-to-computation ratio**, making it a strong candidate for real-time or large-scale applications.

Finally, **Ridge regression** exhibited the lowest scores, with an R^2 of 0.22/0.21 on the train and test sets. While Ridge regression is effective in handling multicollinearity and linear relationships, its underperformance in this case suggests that the relationship between the features and the target variable (helpfulness score) is likely non-linear, which the Ridge model is not well-equipped to capture.

4.2 Binary Classification Results

In addition to Regression, we approached the problem as a binary classification task, where reviews with helpfulness scores between **0 and 0.5** were labeled as **0** (unhelpful) and those between **0.5 and 1.0** as **1** (helpful). We trained the same three models: **Logistic Regression**, **ELM**, and **XGBoost**, using the same randomized grid search and cross-validation procedures. The results are summarized in Table 2.

Table 2. Summary of results for the three machine learning models for predicting the helpfulness score as a binary classification problem

Models	Optimized parameters	Best score (train/test) Accuracy	F1-score / ROC-AUC
Logistic Regression	C: 98.99	0.79/0.78	0.77/0.71
ELM	alpha: 0.026, density: 0.41, include_original_features: True, n_neurons: 3417, pairwise_metric: cityblock, ufunc: relu	0.88 / 0.87	0.84 / 0.77
XGBoost	colsample_bytree: 0.76, learning_rate: 0.02, max_depth: 13, n_estimators: 990, scale_pos_weight: 2, subsample: 0.60	0.91/0.90	0.85 / 0.79

Once again, **XGBoost** outperformed the other models, achieving the highest accuracy (**0.91 on the training set** and **0.90 on the test set**). The model's **F1-score of 0.85** and **ROC-AUC of 0.79** further highlight its strong classification performance, indicating that it balances precision and recall effectively. XGBoost's success can be attributed to its ability to handle class imbalances, as it allows for parameter tuning (e.g., **scale_pos_weight**) to address the disparity between helpful and unhelpful reviews. The **ELM model** also demonstrated competitive performance, with an **accuracy of 0.88/0.87** and an **F1-score of 0.84/0.77**, closely matching the performance of XGBoost. This result aligns with the findings from the regression task, where ELM offered a strong combination of accuracy and computational efficiency. **Logistic Regression** performed reasonably well, achieving an accuracy of **0.79/0.78** and an **F1-score of 0.77/0.71**. However, it was clearly outclassed by the non-linear models (ELM and XGBoost). While logistic Regression is a robust and interpretable model for binary classification tasks, it struggles with the complexity inherent in the dataset, where linear separability is not guaranteed.

4.3 NLP Modelling

To enhance the predictive performance of our models, we integrated features generated from the text of the reviews using a **state-of-the-art Natural Language Processing (NLP) model** (Wolf et al, 2020). Specifically, we employed **DistilBERT**, a pre-trained transformer model

available in the Huggingface library, which is designed to extract semantic representations from textual data. DistilBERT is a distilled version of the larger BERT model, maintaining approximately 97% of its performance while reducing the model size and computational requirements, making it well-suited for large-scale tasks like ours.

4.3.1 Training and Execution of DistilBERT on Review Text

In addition to integrating the text-based features from DistilBERT into our combined model, we also explored the performance of **DistilBERT** as a standalone model to predict the helpfulness of reviews using only the **review_text**. This experiment aimed to evaluate the predictive power of the textual content alone, without the influence of other features such as playtime, votes, or review length. To prepare for this experiment, we preprocessed the **review_text** feature, tokenizing the text to fit the input requirements of the DistilBERT model. Each review was truncated or padded to a maximum length of **512 tokens**, which aligns with the typical token length limitations of pre-trained transformer models. We utilized the **Huggingface implementation of DistilBERT**, fine-tuning the model for our classification task. Specifically, we modified the model's output layer to predict binary helpfulness (0 or 1), where reviews with a helpfulness score between **0 and 0.5** were labeled as **unhelpful** (0), and those between **0.5 and 1.0** as **helpful** (1).

Given the complexity of fine-tuning transformer models and the risk of overfitting, we implemented several techniques to regularize the training process:

- **Optimizer with Weight Decay**: We used the **AdamW optimizer**, which incorporates weight decay to penalize large weights and improve generalization. This helped ensure that the model does not overfit the relatively small training dataset, especially when fine-tuning a model as powerful as DistilBERT.
- **Learning Rate Scheduler with Warm-up**: A **learning rate scheduler** was employed to gradually ramp up the learning rate during the early stages of training, followed by a decay. This **warm-up period** helps stabilize the training process, particularly when working with pre-trained models. After the warm-up, the learning rate was reduced to fine-tune the model more carefully and avoid overshooting optimal weights.

To further mitigate the risk of overfitting, early stopping was implemented. This technique monitors the validation loss during training, and if the loss does not improve for a pre-defined number of epochs, training is halted. This allowed us to stop training once the model had reached its optimal point on the validation set, preventing unnecessary further training that could lead to overfitting. Due to computational limitations, we were only able to use approx. **1% of the dataset**, corresponding **to 10,000 reviews**. While this subset is a small fraction of the total dataset, it allowed us to fine-tune DistilBERT efficiently while still providing meaningful insights into the model's capacity to understand the textual content of reviews.

Once training was completed, we evaluated the model on the test set. The results of this evaluation are (1) **Accuracy** of 0.62, (2) **Precision**: 0.61, (3) **Recall**: 0.61, and (4) **F1 Score** of 0.58. These results suggest that while the **DistilBERT model** was able to capture some useful information from the review text, its performance was somewhat limited when relying only on textual data to predict helpfulness.

4.3.2 NLP Integration for Regression and Classification

Once the text of each review was converted into this vector representation, we augmented our original dataset by appending these vectors as additional features to the existing 9 features,

which included metadata such as playtime, review length, and votes received. This combined dataset provided a richer set of inputs for both **Regression** and **binary classification** tasks. The goal was to assess whether adding text-based features would improve our models' ability to predict review helpfulness by leveraging the semantic content of the reviews in conjunction with the numerical features. After integrating the features from DistilBERT, we retrained the **Extreme Learning Machine (ELM)** model for the regression task. Table 3 summarizes the performance of the ELM model after incorporating the NLP features.

Table 3. Summary of results for the ELM model for predicting the helpfulness score as a regression problem, including the NLP features

Models	Optimized parameters	Best score (train/test) R^2
ELM	alpha: 0.036, density: 0.65, include_original_features: True, n_neurons: 4025, pairwise_metric: cityblock, ufunc: sigm	0.51/0.50

The results indicate that the inclusion of the **NLP-generated features** did not lead to a significant improvement in the **R^2 scores** of the ELM model, which remained at **0.51/0.50** on the train and test sets, respectively. This suggests that the model is not training well on the training data to predict the expected outcome. A potential reason for this could be the relatively strong influence of non-textual features like the **number of relevant votes** and **playtime**, which already had a substantial correlation with the helpfulness score. The review text might add redundant or less significant information in the context of this task, where the numeric features dominate the predictive landscape.

We also retrained both the **Logistic Regression** and **ELM** models to predict the helpfulness of reviews as a binary classification problem after incorporating the text features generated by DistilBERT. The performance of the models is summarized in Table 4.

Table 4. Summary of results for two machine learning models for predicting the helpfulness score as a binary classification problem, including the NLP features

Models	Optimized parameters	Best score (train/test) Accuracy	F1-score / ROC-AUC
Logistic Regression	C: 52.53	0.80/0.79	0.77/0.72
ELM	alpha: 0.00013, density: 0.135, include_original_features: True, n_neurons: 4040, pairwise_metric: cityblock, ufunc: lin	0.80 / 0.80	0.80 / 0.72

In the binary classification task, the **Logistic Regression model** showed a slight improvement in its performance after adding the NLP features, achieving a **test accuracy of 0.79** (compared to 0.78 without the text features). The **F1-score** also saw a minor increase to **0.77**, suggesting that the semantic information from the review text provided some benefit to the classification task. This improvement, while modest, may indicate that the NLP features help the model distinguish between helpful and unhelpful reviews more effectively by incorporating the sentiment or thematic elements of the text. The **ELM model** also performed similarly to the regression task, with an **accuracy of 0.80** on both the training and test sets. However, there was no significant change in the model's performance metrics (accuracy,

F1-score, or ROC-AUC) compared to the version without NLP features. This consistency across models suggests that the primary factors driving the classification of reviews are still the numerical features, such as the number of relevant votes and the binary sentiment of the review, rather than the text content itself.

5. DISCUSSION AND CONCLUSIONS

5.1 Discussion

The results from both the regression and classification tasks provide valuable insights into the efficacy of machine learning models for predicting review helpfulness. Across both approaches, **XGBoost** consistently outperformed the other models, demonstrating its robustness in handling complex, non-linear data. Its ability to balance bias and variance, as well as to optimize for various performance metrics through hyperparameter tuning, makes it the ideal choice for this task. However, XGBoost's computational requirements are higher, which may limit its practicality in real-time applications or when dealing with extremely large datasets. **ELM**, while slightly less accurate, offers a highly competitive alternative due to its **speed and simplicity**. Its fast training times make it particularly suited for scenarios where computational resources are limited, or where a rapid response is required. The near-parity between ELM and XGBoost in classification tasks further solidifies ELM's potential for real-world applications. **Ridge regression** and **Logistic Regression**, while useful as baselines, were outperformed by the more sophisticated models. Their performance highlights the limitations of linear models in capturing the complex relationships in user-generated content, particularly when text-based features and user interaction metrics are involved.

These findings suggest that combining **tree-based models** like XGBoost, coupled with **fast-learning algorithms** such as ELM, could provide an optimal balance between performance and efficiency in predicting review helpfulness. Future work could explore ensemble methods or hybrid approaches, combining the strengths of these models to enhance predictive accuracy further. Considering the use of NLP features to predict helpfulness, the results demonstrate that while **DistilBERT** is a powerful model for capturing semantic information from text, **relying only on the textual content** of reviews does not achieve a high level of performance in predicting helpfulness. This suggests that other factors—such as user engagement metrics (e.g., playtime or number of helpful votes)—are crucial for making accurate predictions in this context.

Several factors could explain the model's limited performance:

1. **Dataset size**: Using only 1% of the dataset (10,000 reviews) for training constrained the model's ability to generalize. Transformer-based models like DistilBERT typically require large datasets to fine-tune effectively, and the relatively small sample size likely hindered its learning capacity.
2. **Text complexity**: The helpfulness of a review may not always be apparent from the text alone. Other contextual factors, such as the game's popularity or the reviewer's playtime, might provide essential signals that are not captured through text embeddings.
3. **Model constraints**: Although DistilBERT is highly efficient, it is a **distilled version of BERT**, meaning some of the model's capacity to capture complex relationships in the

text has been reduced for speed and scalability. While this makes it practical for tasks with limited computational resources, it may also reduce its overall predictive power in highly nuanced tasks like review helpfulness.

Regarding the addition of NLP features to the regression and classification models, our findings suggest that incorporating features generated from a pre-trained NLP model, such as DistilBERT, did not significantly improve the performance of either regression or classification models in predicting review helpfulness. One possible explanation for this is that the **helpfulness score** on Steam reviews may be more strongly influenced by **quantifiable user interactions**—such as playtime, the number of helpful votes, and whether the review is positive or negative—rather than the linguistic content of the review text itself.

For regression tasks, the marginal contributions of the text features may be overshadowed by the dominance of other highly correlated features, like **votes_relevant** and **pos_review_bin**. In the case of classification, while the logistic regression model benefited slightly from the added textual features, the overall improvement was minimal, suggesting that textual analysis may not be as critical for predicting whether a review is considered helpful.

However, this does not entirely negate the potential utility of NLP models in other contexts. It is possible that more advanced techniques, such as fine-tuning pre-trained models specifically on Steam reviews or incorporating **sentiment analysis** and **topic modeling**, could yield better results. Future research could also explore more sophisticated methods of combining numerical and textual features, such as through **attention mechanisms** in transformer models or hybrid architectures that weigh the importance of each feature type differently.

In conclusion, while the current NLP features did not provide a substantial boost to the model's performance, they open the door for future explorations into how text-based data can complement traditional numerical features in predicting user-generated content metrics like helpfulness.

5.2 Conclusions

This study aimed to predict the helpfulness of Steam video game reviews using **Regression** and **binary classification** approaches. We trained three machine learning models—**XGBoost**, **Extreme Learning Machine (ELM)**, and **Ridge regression**—and found that **XGBoost** outperformed the others in both tasks. Its ability to capture complex patterns made it the best model, while **ELM** offered competitive performance with lower computational costs. **Ridge regression**, a linear model, underperformed, highlighting the non-linear nature of the data.

We also incorporated features generated from the review text using the **DistilBERT** NLP model. However, adding these text features did not significantly improve the predictions. This suggests that **non-textual features**, such as playtime, helpful votes, and review sentiment, are more influential in determining helpfulness. The limited dataset used to train the NLP model may have contributed to its weaker performance. Despite this, early predictions of review helpfulness could provide practical value. For users, accurate predictions would help identify useful reviews, aiding in purchase decisions. For developers, helpfulness predictions could highlight valuable feedback, improving game design and user experience.

There are some limitations to the study. The minimal impact of text-based features suggests a need for **larger datasets** and **fine-tuning** of models like **BERT** for this specific domain. **DistilBERT**, while efficient, may not capture the full complexity needed to predict review helpfulness in this context.

Future research should explore **fine-tuning pre-trained NLP models** on domain-specific data like game reviews. Larger datasets could also enhance the performance of models like DistilBERT. Additionally, combining textual and non-textual features more effectively, possibly through **hybrid models** or **ensemble learning**, could yield better results. Incorporating features like **sentiment analysis**, **topic modeling**, and **temporal dynamics** may also improve predictions. For example, considering how reviews evolve over time could provide deeper insights into helpfulness trends. In summary, while **XGBoost** was the most effective model, there remains room for further exploration of text-based features to predict review helpfulness better.

REFERENCES

Akusok, A., Leal, L. E., Björk, K. M. and Lendasse, A., 2021. Scikit-elm: an extreme learning machine toolbox for dynamic and scalable learning. *Proceedings of ELM2019*, pp. 69-78. Springer International Publishing.

Eberhard, L., Kasper, P., Koncar, P. and Gütl, C., 2018, October. Investigating helpfulness of video game reviews on the steam platform. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 43-50. IEEE.

Juniper Research, 2021. Global video game market value from 2020 to 2025 (in billion U.S. dollars) [Graph]. *Statista*. Available at: https://www-statista-com.skema.idm.oclc.org/statistics/292056/video-game-market-value-worldwide/ (Accessed: 22 February 2023)

Kamel, N. E. L. 2022. Steam reviews. *Kaggle*. Available at: https://www.kaggle.com/datasets/nourelkamel/steam-reviews (Accessed: 5 March 2023)

Kasper, P., Koncar, P., Santos, T. and Gütl, C. 2019. On the role of score, genre and text in helpfulness of video game reviews on metacritic. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 75-82. IEEE.

Krishnamoorthy, S. 2015. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, Vol. 42, No. 7, pp. 3751-3759.

Lin, D., Bezemer, C. P., Zou, Y. and Hassan, A. E. 2019. An empirical study of game reviews on the Steam platform. *Empirical Software Engineering*, Vol. 24, pp. 170-207.

MarketLine, 2021. MarketLine Industry Profile: Games Software in the United States. *Games Software Industry Profile: United States.*

MarketLine, 2015. *MarketLine Case Study: Valve Corporation*.

Olmedilla, M., Martínez-Torres, M. R. and Toral, S., 2022. Prediction and modelling online reviews helpfulness using 1D Convolutional Neural Networks. *Expert Systems with Applications*, Vol. 198, 116787.

Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, Vol. 12, pp. 2825-2830.

Winter, J., 2021. SuperData's 2020 Year in Review: Gaming Revenue Up Overall but Fortnite Takes a Tumble. *MMOBOMB*. Available at: https://www.mmobomb.com/news/superdatas-2020-year-review-gaming-revenue-overall-fortnite-takes-tumble (Accessed: 23 February 2023)

Wolf, T. et al., 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, pp. 38-45.

Wu, J., 2017. Review popularity and review helpfulness: A model for user review effectiveness. *Decision Support Systems*, Vol. 97, pp. 92-103.