Analyzing Sentiment to Predict Indie Game Success on Steam using Sentiment Analysis and Machine Learning

Derek W. Graves

Northwest Missouri State University, Maryville MO 64468, USA
S573443@nwmissouri.edu and derek.graves4@outlook.com

Abstract. This project analyzes the sentiment of user reviews and metadata of indie games on Steam to predict their success. Using data collected through the Steam API, we apply sentiment analysis techniques alongside machine learning models, such as Random Forest and Logistic Regression, to identify key factors contributing to a game's popularity. We explore user sentiment, gameplay features, and other relevant data to predict success metrics, offering valuable insights that can help developers optimize their games for better audience reception.

Keywords: sentiment analysis · machine learning · data analytics · steam · indie games

1 Introduction

This project analyzes sentiment from user reviews and metadata of indie games on Steam to predict success. Using data collected through the Steam API, we apply sentiment analysis and machine learning models, such as Random Forest and Logistic Regression, to identify key factors contributing to popularity. The findings offer insights to help developers optimize their games for better audience reception.

1.1 Goals of this Research

The primary goal of this research is to analyze the relationship between user sentiment in game reviews and the commercial success of indie games on Steam. Specifically, we aim to:

- Identify key sentiment features in user reviews that correlate with game success.
- Apply machine learning models to predict game success based on these features.
- Provide actionable insights for developers to optimize game reception.

2 Data Collection

Data is sourced from the Steam API, which provides access to user reviews and game metadata. Key data points include user sentiment, gameplay features, and game popularity indicators. We used the Steam API Documentation as a reference to understand the API parameters and format.

3 Data Preprocessing and Cleaning

Data preprocessing includes handling missing values and text processing for sentiment analysis. Techniques such as tokenization and lemmatization are applied to prepare the review text for our machine learning models. For review-based analysis, methods like those found in Guzsvinecz and Szűcs [1] prove foundational in guiding our approach.

4 Model Training and Evaluation

This section covers training models using the processed dataset and evaluating predictive performance through accuracy, precision, recall, and F1-score. We plan to use models like Random Forest and Logistic Regression, with hyper-parameter tuning conducted through grid search for optimization [3].

The choice of Random Forest is motivated by its robustness in handling datasets with high variance across explanatory and noise variables, as highlighted in the findings of Kirasich et al. [2]. Conversely, Logistic Regression offers greater interpretability and stability for smaller, lower-dimensional datasets, allowing for deeper insights into the sentiment features that most significantly impact a game's success.

The data is split into training and testing sets to effectively evaluate each model's accuracy and overall performance.

5 Results and Discussion

This section will discuss the results of model evaluations, insights, and potential implications for indie game developers. Once model training is complete, specific findings and data visualizations will be added.

6 Limitations and Future Work

Limitations include potential biases in user reviews, data constraints for more niche games, and challenges in modeling complexity behind user behavior. Future work may involve incorporating data from other platforms such as the Epic Games store, and exploring more advanced deep learning methods.

7 Conclusion

This project highlights the value of sentiment analysis and machine learning in predicting indie game success on Steam. By identifying features and sentiments associated with popular games, we aim to provide actionable insights for indie developers looking to improve game reception and engagement.

Additional Resources

For more details, please refer to the project resources below:

- Overleaf Report
- GitHub Repository
- Steam API Documentation

References

1. Guzsvinecz, T., Szűcs, J.: Length and sentiment analysis of reviews about top-level video game genres on the steam platform. Computers in Human Behavior (2022), bluehttps://www.sciencedirect.com/science/article/pii/S0747563223003060
2. Kirasich, K., Smith, T., Sadler, B.: Random forest vs logistic regression: Binary classification for heterogeneous datasets. SMU Data Science Review **1**(3), Article 9 (2018), bluehttps://scholar.smu.edu/datasciencereview/vol1/iss3/9, creative Commons License
3. Lounela, K.: On identifying relevant features for a successful indie video game release on steam. Master's Programme in Department of Information and Service Management (2024), bluehttps://aaltodoc.aalto.fi/items/d578980e-71fa-4618-b500-dff30bbac490