Steam Game Review Sentiment Analysis Utilizing Various Machine Learning and Text Mining Methods

1st Logan Morrison
Department of Data Science
Florida Polytechnic University
Lakeland, U.S.A
lmorrison0938@floridapoly.edu

2nd Henry Marsh
Department of Data Science
Florida Polytechnic University
Lakeland, U.S.A
hmarsh0311@floridapoly.edu

3rd Michael Zaino
Department of Computer Science
Florida Polytechnic University
Lakeland, U.S.A
mzaino0666@floridapoly.edu

***Abstract*—The video game industry has become increasingly reliant on online reviews to gauge player sentiment and inform game development and marketing strategies. Sentiment analysis, a subfield of natural language processing, offers a powerful tool for extracting insights from text data. This study investigates the sentiment of Steam reviews using a comparative analysis of four machine learning methods (VADER, TextBlob, TF-IDF, and BERT) and three clustering algorithms (K-means, DBSCAN, and Hierarchical Clustering)**

***Index Terms*—Machine Learning, Video Games, Sentiment Analysis, Clustering**

I. Introduction

The video game industry has experienced unprecedented growth in recent years, with the global market projected to reach $190 billion by 2025. As the largest digital distribution platform for PC games, Steam has become a hub for gamers to discover, purchase, and discuss their favorite titles. One of the key features of the Steam platform is its review system, which allows users to share their opinions and experiences with others. With millions of reviews written every year, Steam's review dataset offers a unique opportunity to tap into the collective sentiment of the gaming community.

Sentiment analysis, a subfield of natural language processing, has emerged as a powerful tool for extracting insights from text data. By automatically identifying and categorizing the emotional tone of text, sentiment analysis can help game developers, publishers, and marketers better understand player preferences, identify areas for improvement, and make data-driven decisions. Despite its potential, sentiment analysis of Steam reviews remains a relatively understudied area of research.

This paper aims to bridge this gap by applying advanced sentiment analysis techniques to a large dataset of Steam reviews. The goal of our research is to develop a robust sentiment analysis framework capable of accurately identifying the emotional tone of Steam reviews while exploring the relationship between review sentiment, number of reviews, and user engagement. By shedding light on the sentiment of Steam reviews, this research seeks to provide actionable insights for game developers and contribute to the growing body of knowledge on sentiment analysis in the gaming domain.

II. Related Work

Prior research has established sentiment analysis as a valuable tool for understanding user reviews. Simple lexicon-based methods like VADER and TextBlob are often used as a starting point, as they can quickly score text based on word polarity. While VADER has shown utility in gaming contexts and TextBlob has been applied in other domains, these tools have known limitations. They often struggle to correctly interpret nuances common in game reviews, such as sarcasm, community-specific jokes, and mixed-sentiment statements.

To address these shortcomings, supervised machine learning approaches are commonly used. This often involves a two-step process: first, converting text into numerical features, and then training a classifier on those features. A widely adopted feature extraction technique is Term Frequency-Inverse Document Frequency (TF-IDF), which can be paired with a classifier like Random Forest or Logistic Regression to predict review sentiment. More recently, the field of Natural Language Processing (NLP) has shifted towards transformer-based models like BERT (Bidirectional Encoder Representations from Transformers). By pre-training on vast datasets, models like BERT learn to understand language in context, and fine-tuning them for specific tasks enables them to consistently outperform traditional methods in sentiment analysis.

Our study aims to combine these approaches. We will conduct a large-scale comparison of a lexicon model (VADER), a classical machine learning pipeline (TF-IDF with a classifier), and a fine-tuned transformer (BERT) on a massive dataset of Steam reviews. In addition, we will use clustering to discover hidden patterns in the review data and analyze how sentiment connects to player behavior, such as hours played. This will provide a more comprehensive view of what drives player opinions, with all experiments built on standard scientific libraries to ensure our results are reliable.

III. Proposed Approaches

For sentiment analysis, we propose to implement four complementary methods that represent different strands of natural language processing. VADER (Valence Aware Dictionary and sEntiment Reasoner) provides a lexicon-based baseline, relying on rule-driven polarity scoring that is well-suited for the short and informal style of many game reviews. TextBlob offers another lightweight framework, combining basic NLP (Natural Language Processing) techniques to capture both polarity and subjectivity, giving us a straightforward point of comparison.

Next, TF-IDF (Term Frequency–Inverse Document Frequency) with supervised classifiers introduces a statistical perspective. By weighting terms according to frequency and distinctiveness across our dataset, this approach allows models such as Logistic Regression or Random Forest to generate interpretable predictions for sentiment analysis. This serves as a traditional machine learning benchmark, one that highlights the explanatory value of key terms within the dataset.

At the advanced end, BERT (Bidirectional Encoder Representations from Transformers) will be fine-tuned on Steam reviews. Unlike lexicon or statistical methods, BERT interprets words in context, making it more robust to sarcasm, slang, or mixed sentiment that often appear in player feedback. Based on its success in other domains, we expect BERT to deliver higher accuracy and more nuanced results.

Alongside sentiment classification, we will apply clustering techniques to explore hidden structures in reviewer sentiment. K-means will group reviews into distinct clusters based on similarity, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) will detect dense pockets of similar sentiment while filtering outliers, and Hierarchical Clustering will expose nested relationships across review groups. These clustering methods offer perspectives on reviewer communities and engagement behaviors that a single sentiment score cannot reveal.

IV. Planned experiment

To investigate the sentiment of Steam reviews and explore the relationship between review sentiment and game characteristics, we will conduct an experiment involving multiple sentiment analysis methods and clustering algorithms. Our dataset will consist of approximately 15,000,000 reviews from 8,183 games across various genres, including action, adventure, role-playing, strategy, and sports. The dataset was sourced from Kaggle, where it was sourced from the Steam API.

The review text data will be preprocessed to remove stop words, punctuation, and special characters, and will be converted to lowercase. Non-English reviews will be removed from the dataset for ease of use. We will then apply four different sentiment analysis methods to the preprocessed review text data: VADER (Valence Aware Dictionary and sEntiment Reasoner), TextBlob, TF-IDF (Term Frequency-Inverse Document Frequency), and BERT (Bidirectional Encoder Representations from Transformers). VADER is a rule-based sentiment analysis tool that uses a dictionary of words with sentiment scores to calculate the sentiment of text, while TextBlob is a simple API that uses a combination of natural language processing techniques to determine the sentiment of text. TF-IDF is a statistical method that weights word importance based on frequency and rarity across the entire corpus, and BERT is a deep learning-based language model that has achieved state-of-the-art results in sentiment analysis tasks. We will evaluate the performance of each sentiment analysis method using metrics such as accuracy, precision, recall, and F1-score.

To identify patterns and structures in the review sentiment data, we will apply three different clustering algorithms: K-means Clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Hierarchical Clustering. K-means Clustering is a widely used algorithm that partitions the data into clusters based on Euclidean distance, while DBSCAN is a density-based clustering algorithm that groups data points into clusters based on density and proximity. Hierarchical Clustering is a method that builds a hierarchy of clusters by merging or splitting existing clusters. We will evaluate the quality of the clusters using metrics such as silhouette score, calinski-harabasz index, and davies-bouldin index.

The experiment will involve applying each sentiment analysis method to the preprocessed review text data, evaluating the performance of each method, and then applying each clustering algorithm to the sentiment analysis output. We will analyze the relationship between review sentiment and user characteristics such as number of reviews, number of total hours, and number of hours at time of review. We expect to find differences in performance between the sentiment analysis methods, with BERT likely outperforming the other methods. We also expect to find distinct clusters of review sentiment that correspond to those who review many games, versus those who review only a small amount of games. These findings will contribute to the growing body of knowledge on sentiment analysis in the gaming domain and will provide actionable insights for game developers and marketers.

V. Data

The data for this research paper was sourced from the Kaggle Steam Reviews Dataset, which in turn was procured by scraping the Steam API. This dataset is particularly rich and comprehensive, containing approximately fifteen and a half million observations. Each observation includes a variety of information that spans thirteen variables, providing a robust foundation for analyses related to user behavior and sentiment within the Steam gaming platform.

The dataset's variables encompass a diverse range of data types, including boolean, integer, float, and string types. These varied representations are essential for capturing the nuances of user reviews, playtime statistics, and associated metadata. During the preprocessing phase, the dataset was selectively reduced to eight variables of interest, streamlining the scope of analysis to focus on the most pertinent aspects of the data, thereby eliminating unnecessary complexity.

The data preprocessing steps implemented were comprehensive to ensure the integrity and quality of the dataset. Initial stages involved dropping unhelpful columns that did not contribute significantly to the analysis objectives. Following this, a thorough check for any null values was performed in critical fields such as review, playtime_forever, and voted_up. To maintain a high-quality dataset, any rows with missing values in these crucial columns were removed entirely. This measure was taken to ensure that the subsequent analyses were based on complete and reliable information.

To further enhance data quality, a meticulous process was established for identifying and eliminating duplicate reviews. Duplicates were identified by cross-referencing key identifiers including the review, steamid, and appid of the review in question. Additionally, a focus on outlier handling was crucial; reviews were flagged as outliers if they reported zero hours of playtime or contained implausibly high playtime values—such as one million hours—in a single game. Such records were removed to preserve the dataset's overall integrity and utility.

Another key preprocessing step involved filtering the text of reviews to remove any non-UTF8 or non-printable characters, ensuring that the textual analysis would be conducted on clean and standardized text data. Given the substantial size of the dataset, it was split into eleven CSV files, as attempting to consolidate all data into a single file led to prohibitively high computational requirements.

For the entirety of this research paper, all analyses were carried out using Python, leveraging its extensive libraries. The graphical representations of data findings were created utilizing the Seaborn and Matplotlib libraries, which facilitated the visualization of complex relationships within the data.

Ethical considerations were rigorously addressed during the study. Standard ethical concerns regarding consent and anonymization were meticulously evaluated. Since the dataset consists of publicly posted reviews, these were classified as public information. It is important to note that all user reviews were not made under individuals' real names, therefore further mitigating privacy concerns. In cases where users chose to use their actual names as Steam usernames, only the steamid was recorded. This approach effectively anonymized user identities, safeguarding their privacy. Upon completion of the research paper, the dataset accompanying the publication will obfuscate user steamids by replacing them with unique incrementing integer identifiers, thereby enhancing anonymity for individuals.

Upon reviewing the graphical data outputs, Figure 1 and Figure 2 illustrate a compelling trend: the total playtime at the time of review exceeds the total playtime at the point of initial review. This finding suggests that individuals tend to continue engaging with games after posting their reviews, indicating a latent behavior among users to delve deeper into gameplay post-feedback. Such an observation is further substantiated by Figure 3, which reveals that positive reviews are significantly more frequent than negative ones, suggesting an overall favorable sentiment towards the games featured in the dataset.

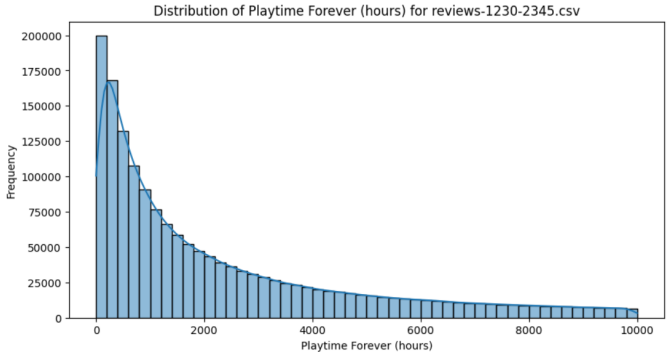

Fig. 1. Graph showing the average total playtime for a given steam review.

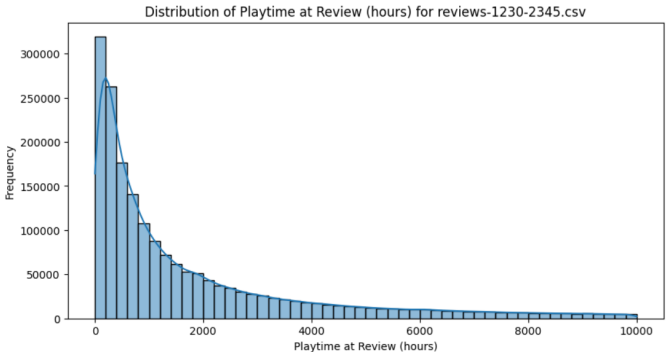

Fig. 2. Graph showing the average playtime at the time that the review was posted.

Moreover, Figure 4 supports this sentiment disparity by demonstrating that users with higher total playtime are more inclined to leave reviews, although there is a notable exception for users with less than five hours of gameplay—who tend to leave fewer reviews. An intriguing aspect of the Steam review system is the review tagging feature, allowing users to react to reviews with specific tags. This research paper pays particular attention to the "funny" tag, as it offers valuable semantic context to user feedback. Notably, as illustrated in Figure 5 for the first dataset segment—and consistently across other segments—all "funny" tags were exclusively assigned to negative reviews.

While the distribution of tags for most segments mirrored the earlier findings in Figures 1 and 2, some notable exceptions appeared in Figures 6 and 7, which depicted distributions that significantly deviated from previous patterns. Despite these variations, the overarching conclusion regarding the total playtime for reviews remaining higher than playtime at the time of review retains strong validation throughout the analysis.

This comprehensive exploration of the data and methodologies employed establishes a solid foundation for the subsequent analyses and discussions presented in this research paper, leading to valuable insights into user behavior within the Steam platform.

Fig. 3. Graph showing the distribution for positive and negative reviews.

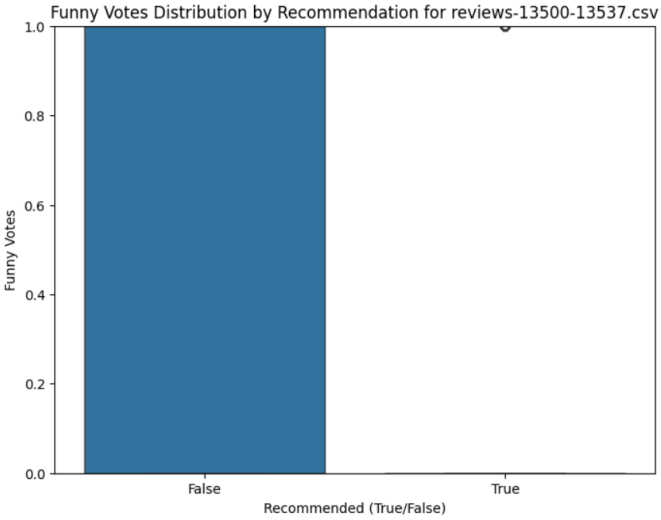

Fig. 5. Graph showing the reviews that steam users considered funny.

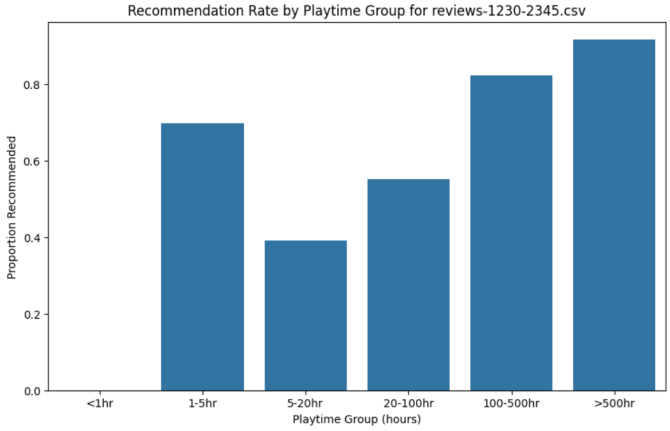

Fig. 4. Graph showing the recommendation rate of a game by the number of total playtime.

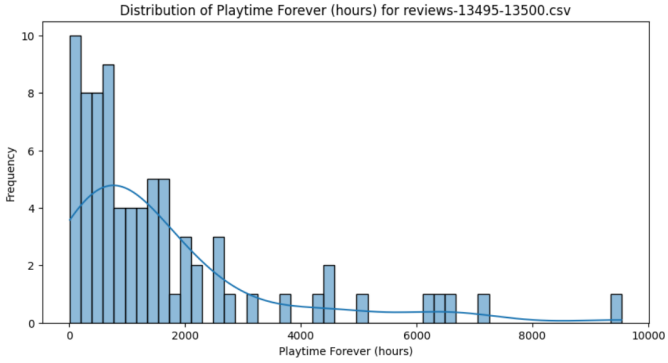

Fig. 6. Graph showing the average total playtime for a given steam review, included for its atypical distribution.

REFERENCES

[1] M. Batumalay, "Harnessing sentiment analysis with VADER for gaming insights: Analyzing user reviews of Call of Duty Mobile through data mining," International Journal Research on Metaverse, vol. 2, no. 2, pp. 121–139, Jun. 2025. doi: 10.47738/ijrm.v2i2.27.

[2] A. Praveen Gujjar and H. Prasanna Kumar, "Sentiment analysis: TextBlob for decision making," International Journal of Scientific Research & Engineering Trends, vol. 7, no. 2, pp. 2395–566, 2021.

[3] P. Guleria, J. Frnda, and P. N. Srinivasu, "NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis," Array, vol. 27, Art. no. 100467, Sep. 2025. doi: 10.1016/j.array.2025.100467.

[4] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data," Heliyon, vol. 10, no. 16, p. e35945, Aug. 2024. doi: 10.1016/j.heliyon.2024.e35945.

[5] Mahendra Dwifebri Purbolaksono, "Sentiment analysis of game review in Steam platform using random forest," International Journal on Information and Communication Technology (IJoICT), vol. 10, no. 2, pp. 161–169, 2024. doi: 10.21108/ijoict.v10i2.1007.

[6] P. Jamjuntr and P. Kaewyong, "Sentiment analysis with a TextBlob

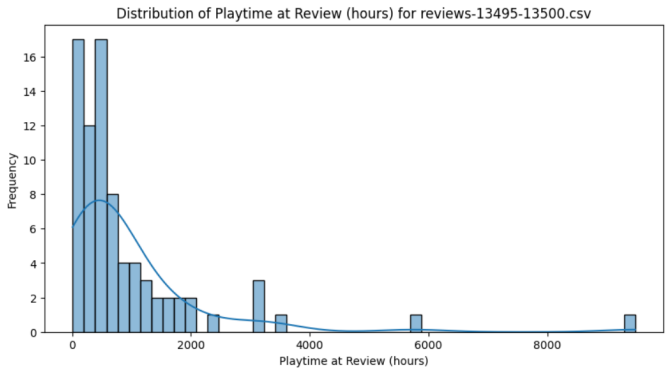

Fig. 7. Graph showing the average playtime at the time that the review was posted, included for its atypical distribution.

package implications for tourism," Journal of Management Information and Decision Sciences, vol. 24, no. S6, pp. 1–9, 2021.
[7] Y. Zhang, Y. Zhou, and J. Yao, "Feature extraction with TF-IDF and game-theoretic shadowed sets," Communications in Computer and Information Science, pp. 722–733, Jun. 2020. doi: 10.1007/978-3-030-50146-4_53.
[8] M. Viggiato, D. Lin, A. Hindle, and C.-P. Bezemer, "What causes wrong sentiment classifications of game reviews," IEEE Transactions on Games, pp. 1–1, 2021. doi: 10.1109/TG.2021.3072545.
[9] ifttt-user, "Sentiment analysis in Python using VADER," Towards AI, Jul. 2023. [Online]. Accessed: Sep. 8, 2025.
[10] P. Virtanen, R. Gommers, and T. E. Oliphant, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," Nature Methods, vol. 17, no. 3, pp. 261–272, Feb. 2020. doi: 10.1038/s41592-019-0686-2.
[11] H. Zhou, "Beyond win rates: A clustering-based approach to character balance analysis in team-based games," arXiv:2502.01250, 2025.
[12] A. R. Aliev, T. A. Aliyev, and R. Eyniyev, "Analyzing price dynamics, activity of players and reviews of popular indie games on Steam post-COVID-19 pandemic using SteamDB," International Journal of Information Technology and Computer Science, vol. 17, no. 3, pp. 26–51, Jun. 2025. doi: 10.5815/ijitcs.2025.03.03.
[13] N. Smairi, H. Abadlia, H. Brahim, and W. L. Chaari, "Fine-tune BERT based on machine learning models for sentiment analysis," Procedia Computer Science, vol. 246, pp. 2390–2399, Nov. 2024. doi: 10.1016/j.procs.2024.09.531.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, 2018.