Clustering Analysis of Steam Games by Reviews, Sentiments, and Store Metadata

Detim Zhao
School of Computing
and Augmented Intelligence
Arizona State University

***Abstract*—This study investigates clustering and sentiment analysis on the Steam game dataset to uncover consumer preferences and engagement trends. By combining structured metadata and user reviews, K-Means and Agglomerative Clustering were used to group games based on shared characteristics, with BERT embeddings providing semantic insights from review text. K-Means highlighted challenges with overlapping clusters due to its reliance on Euclidean distance, while Agglomerative Clustering revealed hierarchical relationships but faced issues with cluster dominance. These findings underscore the need for advanced methods like HDBSCAN and UMAP to handle high-dimensional data and improve clustering outcomes. The results could offer actionable insights for personalized recommendations and market strategies in the gaming industry.**

***Index Terms*—Game Review Sentiment Analysis, Steam Game Clustering, Consumer Patterns in Gaming**

1. Introduction

The gaming industry continues to thrive as a dominant force in entertainment, driven by a rapidly expanding digital ecosystem. Steam, a leading digital distribution platform, reflects this growth by offering a vast catalog of games enriched with detailed user-generated reviews and metadata. These reviews offer valuable insights into consumer preferences, satisfaction, and perceptions of gameplay, narrative, and technical quality. However, analyzing Steam's data presents significant challenges. User reviews are diverse and often informal, incorporating slang and typographical errors that complicate sentiment analysis. Additionally, popular games disproportionately dominate review counts, introducing biases that skew the dataset. These complexities highlight the need for robust analytical methods to uncover actionable insights and improve understanding of consumer engagement within Steam's dynamic ecosystem.

Clustering techniques integrated with sentiment analysis offer a novel solution to this problem. Clustering algorithms group games into meaningful categories, while sentiment analysis quantifies the subjective and emotional content of user reviews. Together, these methods enable the exploration of patterns within reviews and metadata, uncovering trends in consumer engagement and preferences. This approach allows consumers to discover games better aligned with their interests and provides developers and marketers with data-driven insights to improve user experiences and design targeted strategies.

Steam's metadata—spanning genres, categories, pricing, and player ratings—offers a rich foundation for clustering. By integrating structured metadata with unstructured insights from sentiment analysis, this study categorizes games in ways that align with user preferences and market trends. This approach addresses a critical need for better organization within the gaming ecosystem, enhancing decision-making for consumers and fostering responsiveness among developers and marketers.

This study seeks to explore the relationship between consumer preferences, as captured by review sentiment and game metadata, and game success. It investigates whether clustering algorithms can uncover meaningful groupings that align with consumer trends and evaluates the effectiveness of sentiment analysis in providing actionable insights for personalized game recommendations and market strategies.

1.1. Related Work

Zuo (2018) [1] conducted a comprehensive sentiment analysis of Steam product reviews, employing Naive Bayes and Decision Tree classifiers. The study detailed the process from data collection to classification, highlighting the effectiveness of these classifiers in predicting sentiments within user-generated content.

Similarly, the "Steam Insider" [2] system utilizes unsupervised machine learning techniques for aspect-based sentiment analysis and text summarization of video game reviews on Steam. By extracting key aspects from user reviews and performing sentiment analysis on each, the system provides valuable insights for developers and marketers, aiding in understanding user feedback at a granular level.

Karabila et al. (2023) [3] proposed an enhancement to collaborative filtering-based recommender systems by integrating sentiment analysis. Their approach leverages the sentiments expressed in user reviews to improve recommendation accuracy and personalization, demonstrating the significance of user sentiment in refining recommendation relevance.

"Sentiment Analysis of Steam Reviews Using Transformer Models" [4] investigates the application of advanced transformer-based models, specifically BERT and

RoBERTa, for sentiment analysis on a substantial dataset of over 6.4 million English-language Steam reviews. The study emphasizes the challenges posed by the informal language, slang, and typographical errors commonly found in user-generated content like game reviews.

Raison et al. (2012) [5] introduced a method to extract fine-grained user opinions from game reviews by employing adjective-context co-clustering. This approach identifies specific game aspects and their associated qualities, providing deeper insights into user sentiments.

Abdul-Rahman et al. (2024) [6] present a novel approach to predicting customer churn by integrating sentiment analysis of user-generated reviews from the Steam platform into churn forecasting models. The study demonstrated that incorporating sentiment analysis of user reviews significantly enhances the accuracy of churn forecasting models, providing valuable insights for businesses aiming to reduce churn and improve customer satisfaction.

These studies collectively underscore the importance of incorporating sentiment analysis and other data into data-driven strategies within the gaming industry. By analyzing user-generated content, platforms can better understand consumer preferences, leading to improved recommendation systems and more targeted marketing efforts.

1.2. Research Objectives

- **RO1**: To analyze consumer preferences and satisfaction as captured in user reviews, as well as store metadata and its connection to a game's success.
- **RO2**: To utilize clustering algorithms for grouping games into meaningful categories, enabling insights into consumer engagement and trends in gaming.
- **RO3**: To explore the integration of sentiment analysis within clustering models, revealing patterns of user feedback that align with game metadata and consumer satisfaction.
- **RO4**: To examine the potential for using clustered data to improve decision-making for personalized game recommendations and market strategy development.

2. Exploratory Data Analysis

The Exploratory Data Analysis (EDA) will be split into two sections for the Steam Store metadata and for the Steam review data.

2.1. Store Data EDA

The first dataset analyzed in this study was sourced from Kaggle. It was compiled using web scraping techniques from the Steam Store, capturing information up to May 19, 2024, and last updated on May 25, 2024. This dataset contains 42,497 records, each representing a unique game available on Steam, making it a comprehensive snapshot of one of the largest digital distribution platforms for video games. Its breadth and diversity ensure broad applicability for understanding consumer preferences, market trends, and game characteristics.

Each entry in the dataset includes a range of interpretable attributes that capture various aspects of each game. Key attributes include *game_title*, *release_date*, *genres*, and *categories*, which provide contextual and categorical information about the game. Developer and publisher details are represented in the *developer* and *publisher* fields, offering insights into the entities responsible for the game.

Pricing information is detailed through attributes such as *original_price_INR*, *percentage_of_original_price*, and *discounted_price_INR*, reflecting both the original and discounted costs of the game in Indian Rupees (INR). The dataset also includes fields like *dlc_available* and *age_restricted*, which highlight specific game characteristics.

Review metrics are captured through *overall_review*, *overall_positive_review_percentage*, and *overall_review_count*, providing quantitative measures of user sentiment and engagement. Furthermore, the *awards_count* column adds additional context regarding the recognition or accolades received by a game. Together, these attributes form a comprehensive dataset for analyzing consumer preferences and gaming trends.

The scatterplot titled "Positive Review Percentage vs. Review Count" (Figure 1) illustrates the relationship between a game's sentiment score (percentage of positive reviews) and its popularity (review count). A clear trend is observed where games with higher review counts tend to cluster at higher positive review percentages, reflecting a consumer bias toward well-reviewed and popular titles. However, the spread of data at lower review counts indicates greater variability in sentiment, suggesting that niche or less popular games receive mixed feedback. This visualization highlights how review count can influence perceived quality, potentially skewing consumer perceptions towards widely played titles.

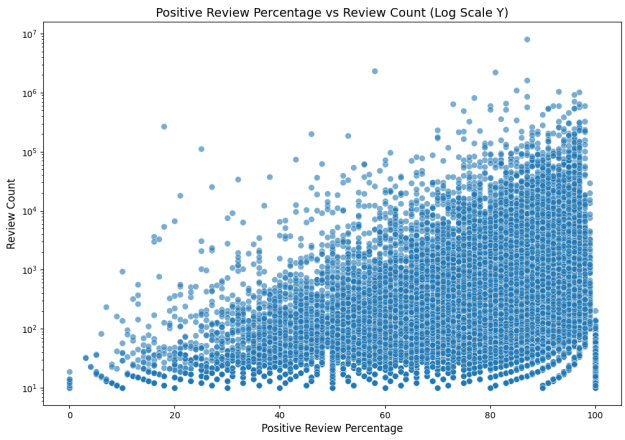

Figure 1. Positive Review Percentage vs. Review Count (Log Scale Y)

The scatter plot of *Awards* vs. *Review Count* (Figure 2) explores the relationship between the number of awards a game has received and the volume of user reviews. As expected, there is a noticeable grouping of games with a low number of awards and high review counts, suggesting that popular games often receive substantial user engagement regardless of critical acclaim. However, games with higher award counts generally have fewer reviews, possibly indicating niche appeal or specialized audiences.

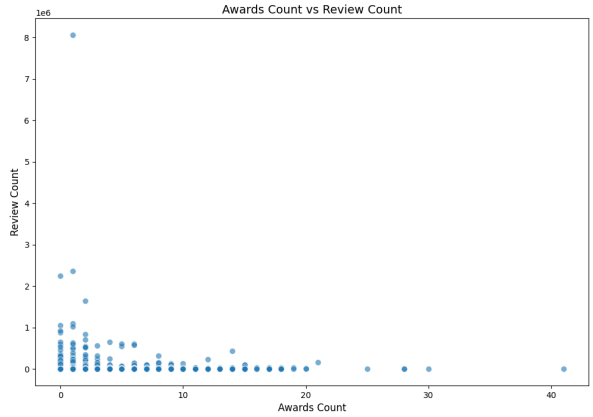

Figure 2. Awards vs. Review Count Scatterplot

2.2. Review Data EDA

To supplement the metadata provided by the Steam Store dataset, we utilized the *steamreviews* Python package to retrieve user review data directly from the Steamworks API. This package provides an efficient interface for querying review information in JSON format for each game listed on Steam. Using *steamreviews*, a random sample of 385 games was selected, based on the median number of reviews across all games in the store dataset. This ensured that our selection captured a representative subset of games with balanced review counts. For each selected game, get english reviews and other data with the request.

The JSON data retrieved from the Steamworks API was subsequently processed and transformed into a structured format to create the Steam Reviews dataset. Key fields include *review_id*, *review_text*, *votes_up*, and *weighted_vote_score*. Sentiment analysis using VADER was performed on the *review_text* to quantify user sentiment on a continuous scale. Note, the data for reviews can only be fetched up to a year from the API, so we had cut off reviews on the date of the store dataset, to eliminate that confounding variable.

This process allowed the integration of structured game attributes with unstructured user feedback, resulting in a comprehensive dataset for machine learning analysis.

A look at the distribution of compound sentiment reveals a heavily right-skewed pattern (Figure 3), with a significant proportion of reviews exhibiting highly positive sentiment scores close to 1. This suggests that the majority of Steam reviews are overwhelmingly positive, reflecting a consumer bias towards expressing favorable opinions or a dataset skewed by popular games with strong user satisfaction. However, the presence of a small, but notable, cluster of negative sentiments (scores near -1) indicates that a subset of games or game features elicit strongly negative reactions, potentially useful for identifying pain points in game design or marketing.

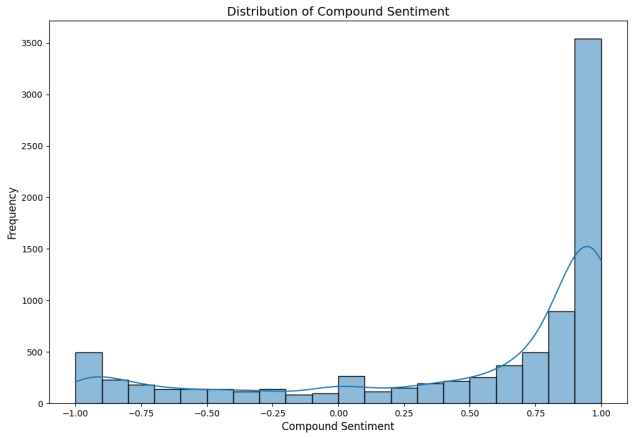

Figure 3. Compound Sentiment Distribution

The scatterplot in Figure 4 illustrates the relationship between the number of *Votes Up* a review receives and its *Weighted Vote Score*. A clear trend is observed where the *Weighted Vote Score* is positively correlated with the number of *Votes Up*, particularly at lower vote counts. However, as the number of votes increases, the *Weighted Vote Score* stabilizes and approaches its upper limit, creating a clustering effect near a score of 1. This suggests that highly popular reviews tend to converge towards a perfect score, regardless of the exact number of votes they accumulate.

Notably, there is a dense concentration of data points at the lower range of *Votes Up*, where a wider spread of *Weighted Vote Scores* is visible. This may indicate variability in how early reviews are received by the community, reflecting differences in user agreement or review quality. Additionally, a small number of outliers with high *Votes Up* but lower *Weighted Vote Scores* could represent controversial reviews or those with polarizing opinions.

3. Methodology

For our clustering analysis, we utilized both the *K-Means* and *Agglomerative Clustering* algorithms, two widely used machine learning techniques for unsupervised learning. While *K-Means* is particularly well-suited for partitioning data into clusters based on similarities, *Agglomerative Clustering* offers a hierarchical approach that provides a more

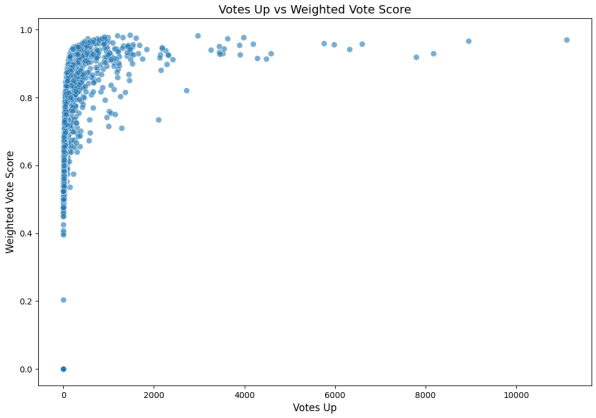

Figure 4. Votes Up vs. Weighted Vote Score Scatterplot

granular view of cluster relationships. By combining these methods, we aimed to leverage their respective strengths to identify patterns in Steam game data.

3.1. K-Means

K-Means partitions the dataset into k clusters by minimizing the sum of squared distances between each data point and the centroid of its assigned cluster. This approach is computationally efficient and scales effectively with large datasets like ours. The number of clusters, k, was determined using the Elbow Method and Silhouette Analysis for *K-Means*.

3.2. Agglomerative Clustering

In contrast, *Agglomerative Clustering* builds clusters hierarchically, starting with each data point as its own cluster and iteratively merging the closest pairs of clusters based on a linkage criterion. This method allows for the construction of a dendrogram, offering a visual representation of the clustering process and enabling us to explore cluster relationships at multiple levels of granularity.

To enhance the performance of *Agglomerative Clustering*, we implemented a custom distance function tailored to the diverse attributes of the dataset. The boolean feature *age_restricted* was treated using a binary distance metric, while numerical features (e.g., pricing, review metrics) were measured using Euclidean distance. This composite distance function ensured that the algorithm appropriately accounted for the distinct nature of different data types. The hierarchical structure generated by this approach provided clusters that were not only meaningful but also interpretable in the context of Steam game attributes.

To prepare the data for clustering, we integrated structured attributes from the Steam Store dataset (e.g., genres, categories, pricing, and review metrics) with BERT embeddings derived from user review text. The structured data was normalized to ensure consistent feature scaling, while the unstructured text embeddings captured a rich semantic representation of user feedback. This integration created a comprehensive feature space encompassing both numerical and contextual information about games. The resulting dataset, comprising 868 features, highlights the potential necessity of dimensionality-aware clustering techniques, which were not employed in this study.

For *Agglomerative Clustering*, we evaluated cluster quality using silhouette plots and metrics such as the Silhouette Score and Calinski-Harabasz Index.

4. Results and Discussion

The results and discussion will be organized into a section for *K-Means* and *Agglomerative Clustering* respectively.

4.1. K-Means Results

By testing various cluster counts (k), we evaluated results using the Silhouette Score and Calinski-Harabasz (CH) Index. The Silhouette Score measures how well-separated and compact clusters are, with higher scores indicating better-defined clusters and minimal overlap between clusters. The Calinski-Harabasz (CH) Index evaluates the ratio of cluster separation to compactness, where higher values reflect better-defined clusters with greater inter-cluster separation and intra-cluster cohesion.

For $k = 3$, the Silhouette Score of 0.0986 and CH Index of 52.65 indicated moderate cluster separation and compactness. Increasing k to 4 and 5 resulted in decreased metrics (Silhouette Scores of 0.0647 and 0.0643, CH Indices of 40.82 and 33.48, respectively), reflecting declining cluster quality as k increased. These results suggest that $k = 3$ offers the most interpretable clustering configuration for the dataset, though the low Silhouette Scores highlight challenges in separating overlapping clusters effectively.

The silhouette plot (Figure 5) evaluates how well-separated clusters are by measuring how similar each data point is to its assigned cluster versus others. A higher silhouette score indicates tighter grouping and better separation. The dashed line shows the average score, summarizing overall clustering performance. Uneven plots suggest overlap or inconsistent cluster definitions. For $k = 3$ (*K-Means*) the plot reveals uneven cluster distributions, with Cluster 0 significantly dominating. This dominance suggests limitations in the *K-Means* algorithm's ability to balance cluster sizes when applied to this dataset. Furthermore, its reliance on Euclidean distance may not fully capture relationships in high-dimensional or mixed-type data.

4.2. Agglomerative Clustering Results

To complement *K-Means*, we applied *Agglomerative Clustering* to leverage hierarchical relationships within the data. For $n_{\text{clusters}} = 3$, we achieved a higher Silhouette Score of 0.5252 and a CH Index of 9.28, indicating better separation and compactness compared to *K-Means*. However,

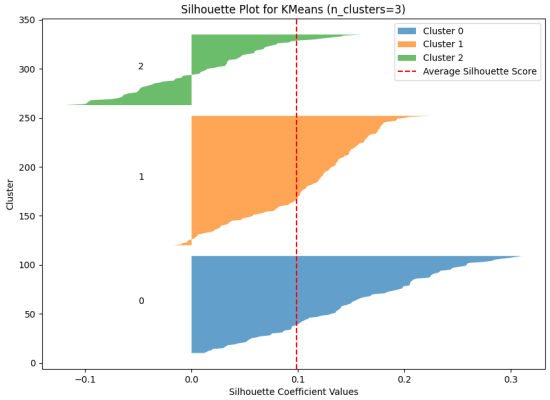

Figure 5. Silhouette Plot of k=3 with K-Means

as shown in the silhouette plot for $n_{\text{clusters}} = 3$ (Figure 6), Cluster 0 dominated the results, similar to *K-Means*. This imbalance suggests that while hierarchical clustering improves certain metrics, it also struggles to distribute points evenly across clusters. For $n_{\text{clusters}} = 4$ and $n_{\text{clusters}} = 5$, metrics declined, with Silhouette Scores of 0.4067 and 0.3455 and CH Indices of 7.81 and 8.02, respectively. These configurations introduced additional clusters but reduced overall cluster cohesion.

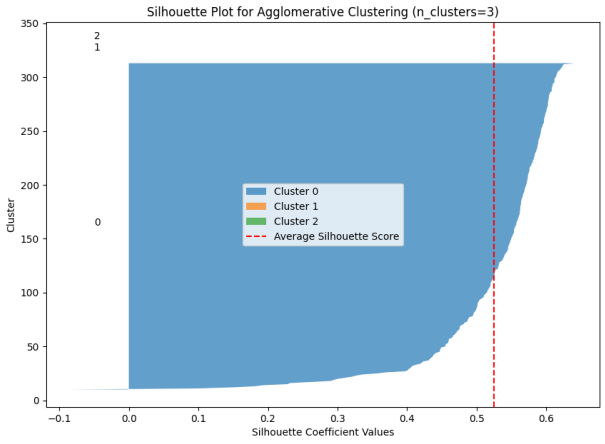

Figure 6. Silhouette Plot of 3 clusters (Agglomerative)

Both methods display the inherent challenges of clustering high-dimensional, heterogeneous datasets like Steam metadata. While metrics such as the Silhouette Score and CH Index provide quantitative evaluations, visualizations (Figures 5 and 6) emphasize the need for additional methods to address cluster dominance and enhance interpretability. These findings guide future exploration of custom distance metrics or alternative algorithms tailored to mixed-type data.

The dendrogram (Figure 7) provides a hierarchical view of the clustering process, illustrating how data points merge into larger clusters at increasing distance thresholds. The structure of the dendrogram is akin to a binary tree, where each node represents a cluster formed by merging two smaller clusters or individual data points. While the root, the merged blue lines, is positioned on the right, it is important to note that the orientation does not affect the hierarchical relationships depicted since it can be graphed in any orientation.

For the $n_{\text{clusters}} = 3$ (Figure 7) configuration (a cyan dotted-line), the dendrogram highlights the dominance of one cluster, reflecting a large portion of similar data points grouped together. This imbalance suggests that certain features heavily influence the clustering process, potentially masking finer substructures within the data.

As the number of clusters increases (e.g., $n_{\text{clusters}} = 4$ and $n_{\text{clusters}} = 5$), smaller clusters emerge, representing finer distinctions or niche patterns within the data. However, the overall dominance of the primary cluster remains evident, indicating the inherent complexity and overlap within the dataset.

The dendrogram complements the silhouette and CH Index evaluations by visualizing hierarchical relationships, providing valuable insights into cluster structures and sizes. It highlights the need to consider feature selection and scaling to achieve a more balanced clustering outcome.

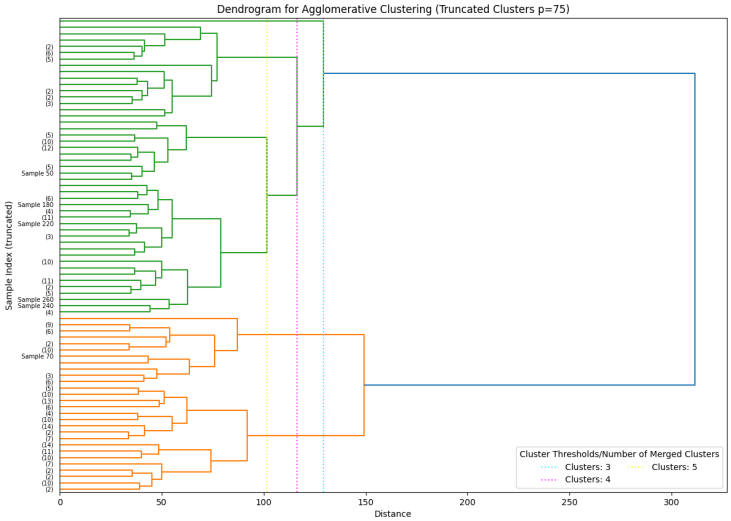

Figure 7. Hierarchical Dendrogram for Agglomerative Clustering, showing cluster relationships and merging thresholds. Dotted markers indicate thresholds for $n_{\text{clusters}} = 3$, $n_{\text{clusters}} = 4$, and $n_{\text{clusters}} = 5$.

Alignment with Research Objectives

RO1: User reviews and store metadata provided insights into consumer preferences and satisfaction. While clustering highlighted broad patterns influenced by genre, pricing,

and sentiment, the dominance of single clusters limited the capture of more diverse consumer behaviors.

RO2: Clustering grouped games into broad categories, but the dominance of large clusters reduced interpretability and masked niche groupings. This suggests a need for refined methodologies to uncover more granular trends.

RO3: The integration of sentiment analysis with clustering revealed some connections between user feedback and game metadata. However, uneven cluster sizes hindered deeper exploration of sentiment patterns.

RO4: Due to the dominance of a primary cluster, the results provided limited actionable insights for personalized recommendations or targeted market strategies. Refinements in clustering techniques are necessary to make findings more applicable.

These results emphasize the need for improved clustering approaches to better address the dataset's complexities and achieve the research objectives.

5. Conclusion

In this study, we applied clustering and sentiment analysis techniques to analyze the Steam game dataset, uncovering trends in consumer engagement and preferences. By utilizing K-Means and Agglomerative Clustering, we identified broad groupings of games based on structured metadata and user feedback. Sentiment analysis using VADER quantified user sentiment, while BERT embeddings provided a semantic representation of review text, enriching the clustering process with contextual insights.

However, limitations such as cluster dominance and the challenges of high-dimensional data underscore the need for more advanced approaches. Future work will explore methods like HDBSCAN for density-based clustering and UMAP for dimensionality reduction, aiming to address these limitations and uncover more granular patterns in user preferences.

Overall, this work highlights the potential of integrating structured and unstructured data for actionable insights, offering valuable guidance for personalized recommendations and market strategies in the gaming industry.

Acknowledgments

Special thanks to Professor Ruben Acuña for his guidance and support.

References

[1] Z. Zuo, "Sentiment analysis of steam review datasets using naive bayes and decision tree classifier," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49561831

[2] N. K. H. Tran, "Development of a machine learning system for aspect-based sentiment analysis and text summarization of video game reviews on steam," 2024. [Online]. Available: http://hdl.handle.net/20.500.12738/15024

[3] I. Karabila, N. Darraz, A. El-Ansari, N. Alami, and M. El Mallahi, "Enhancing collaborative filtering-based recommender system using sentiment analysis," *Future Internet*, vol. 15, no. 7, 2023. [Online]. Available: https://www.mdpi.com/1999-5903/15/7/235

[4] R. Reddy, A. A. Naoman, G. V. S. Charan, and S. N. Fazal, "Sentiment analysis of steam reviews using transformer models," in *Proceedings of the 6th International Conference on Communications and Cyber Physical Engineering*, A. Kumar and S. Mozar, Eds. Singapore: Springer Nature Singapore, 2024, pp. 719–727.

[5] K. Raison, N. Tomuro, S. Lytinen, and J. P. Zagal, "Extraction of user opinions by adjective-context co-clustering for game review texts," in *Advances in Natural Language Processing*, H. Isahara and K. Kanzaki, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 289–299.

[6] S. Abdul-Rahman, M. F. A. M. Ali, A. A. Bakar, and S. Mutalib, "Enhancing churn forecasting with sentiment analysis of steam reviews," *Social Network Analysis and Mining*, vol. 14, no. 1, aug 2024. [Online]. Available: https://link.springer.com/10.1007/s13278-024-01337-3