**Steam platform review analysis**

Zeng Yuting 24426970

Hu Zexin 24448664

Jiang Hanbing 24456276

Liu Kexin 24420522

**1. Abstract**

This study employs data analysis, natural language processing, and machine learning to systematically analyze over 6.4 million Steam game reviews from Kaggle. After preprocessing, 20,000 samples are selected for in - depth analysis using descriptive analysis, correlation analysis, LDA topic modeling, and sentiment analysis. Findings indicate that players generally give positive evaluations, with gameplay enjoyment and storylines being major positive drivers, while game issues lead to dissatisfaction. Eight topics are identified, showing varied sentiment distributions, and negative reviews are more likely to gain votes. The MultinomialNB model, achieving 84.35% accuracy, is chosen for sentiment classification.This research offers valuable insights but has limitations in topic definition and analysis methods, which will be improved in future work.

1. **Introduction**

In recent years, the gaming industry has experienced rapid growth, with platforms like Steam have become key hubs for player interaction and feedback. Launched by Valve, Steam is one of the largest digital distribution platforms for PC games, featuring over 23,000 games and more than 184 million active users (Galy-onkin, 2018). Users can purchase games on the Steam store and engage with the community by leaving comments after playing. Unlike the common five-star rating system, Steam requires players to express their overall feelings about a game as either "recommended" (good) or "not recommended" (bad). These user reviews not only affect other users' purchasing decisions, but also provide developers with important insights into game performance and player satisfaction. However, the vast number of unstructured reviews, brings many difficulties to research and analysis. Therefore, this study systematically analyzes Steam game reviews through data analysis, natural language processing (NLP) and machine learning methods, aiming to reveal the correlation between user interests, potential topics, sentiment trends and review structure and user behavior.

However, compared to the mature delivery systems and data analysis tools in mobile and online gaming, the Steam platform still has obvious shortcomings in data mining and sentiment analysis. At present, the only available observation tool is SteamDB, which primarily offers basic functions like sales statistics and player activity monitoring. The sentiment analysis of user comments remains at a qualitative research level and lacks systematic analysis methods.

To address this, this study constructs a basic Steam game review analysis framework through data analysis, natural language processing (NLP) and machine learning methods. By systematically analyzing Steam game reviews, revealing the correlation between user interests, potential themes, sentiment trends, and review structure and user behavior, providing game developers with valuable insights to better align with user expectations and enhance platform services. Specifically, through topic classification and sentiment recognition, developers can optimize and iterate the game based on feedback from high-frequency topics, improving user experience. Additionally, tracking emotional fluctuations in comments can prompt game developers to respond in a timely manner and reduce the risk of negative word-of-mouth. For platforms, analyzing comment topics can provide important content support for recommendation systems and optimize algorithm recommendations, while also tracking market preference changes in real-time and provides guidance for investment and distribution.

Based on this, this study discusses the following three core issues: ① What is the focus of users' attention? ② How do users' comments relate to their emotions? ③ How to identify the emotions of user comments?

To answer these questions, this study uses the following framework for analysis:

① Descriptive analysis: explore the characteristics of comment text, user emotions and user preferences; ② LDA topic model: extract clustering topics from text; ③ Correlation analysis: explore the relationship between variables; ④ Sentiment analysis: predict the emotional tendency of comments

# **3. Data Description**

## **3.1 Data Source and Basic Information**

Our data is sourced from the Steam Reviews dataset on the Kaggle platform. The original dataset contains over 6.4 million publicly available English game reviews. Each review record encompasses several key information dimensions: `app\_id`, `app\_name`, `review\_text` (the specific evaluation content of the game by players), `review\_score` (a quantitative evaluation of the game by players), and `review\_votes` (indicating the number of votes the review has received).

## **3.2 Data Preprocessing and Cleaning Process**

At the beginning of the data processing, we first imported a series of Python libraries with diverse functions. Then we unzipped and read the original dataset file in CSV format. After that, we conducted a missing value statistics. To ensure the data quality and the accuracy of subsequent analysis, we used the `dropna()` function to remove the rows containing missing values from the DataFrame.

Given that the original dataset is extremely large , in order to improve the efficiency of subsequent analysis while ensuring the representativeness of the data, we decided to perform a simple random sampling of 20,000 data points. Therefore, we used the `sample` function of the pandas DataFrame and set the random seed `random\_state = 42`. This ensures that we can obtain the same sampling results when running the code multiple times, enhancing the reproducibility of the experiment and the stability of the results.

# **4.Methodology**

## **4.1 Descriptive Analysis Method**

To begin with, we calculated the word count of each review to determine the mean, minimum, and maximum review lengths. Using the Seaborn library, we plotted a histogram to visualize the distribution of review lengths among players. For sentiment analysis, we classified reviews with a score of -1 as negative and those with a score of 1 as positive, then depicted the distribution via a pie chart. We created word clouds to emphasize the frequently occurring words in positive and negative reviews, thus uncovering what players cared about most. By identifying the top ten games with the highest review counts and marking the specific numbers on a bar chart, we presented a clear ranking of popular games. Finally, we calculated the ratios of positive and negative reviews for these top ten games and used stacked bar charts to display their sentiment distributions, enabling easy comparisons of the games' reputations.

## **4.2 LDA topic modeling method**

To convert the preprocessed text into numerical features, we used the TF-IDF algorithm with strict parameters to remove common and rare words, so as to focus on more informative words. We then used the LatentDirichletAllocation class for topic modeling. After testing, we identified 8 topics, setting the topic distribution hyperparameter to 0.1 and the word hyperparameter to 0.01 during training. Finally, we extracted and output the words with the highest weight under each topic, and refined these words based on industry knowledge to categorize them into specific topics. On this basis, we grouped the data according to the dominant topic and calculated the proportions of different sentiment scores for each topic. We intuitively presented the sentiment distribution of different topics through heat maps.

## **4.3 Correlation analysis Method**

## 1. Point-Biserial Correlation

## I used Point-Biserial Correlation analysis because it can be used to study the correlation between a binary variable and a continuous variable. In this dataset, the affective tendency 'review\_score' is a binary variable, and our new comment length variable 'review\_length' is a continuous variable. The traditional Pearson's correlation coefficient is suitable for calculating two continuous variables, so I have applied a pointwise dichotomous correlation coefficient to analyze the relationship between review length and affective tendency.

## 2. Chi-Square Test

## The Chi-Square Test is commonly used to test whether two discrete variables are related. In this case, the chi-square test is used in two parts. The first one is to verify whether there is a significant difference between positive and negative reviews on whether to recommend review. The second is to test whether there is a significant association between two categorical variables: app\_name and review\_score.

## 3. Logistic Regression

## Logistic regression is often used to deal with binary dependent variables, which predicts the probability of the dependent variable taking a certain category. I use logistic regression to better explore the correlation between affective tendencies and whether to recommend a review(review\_votes)

## **4.4 Sentiment analysis Method**

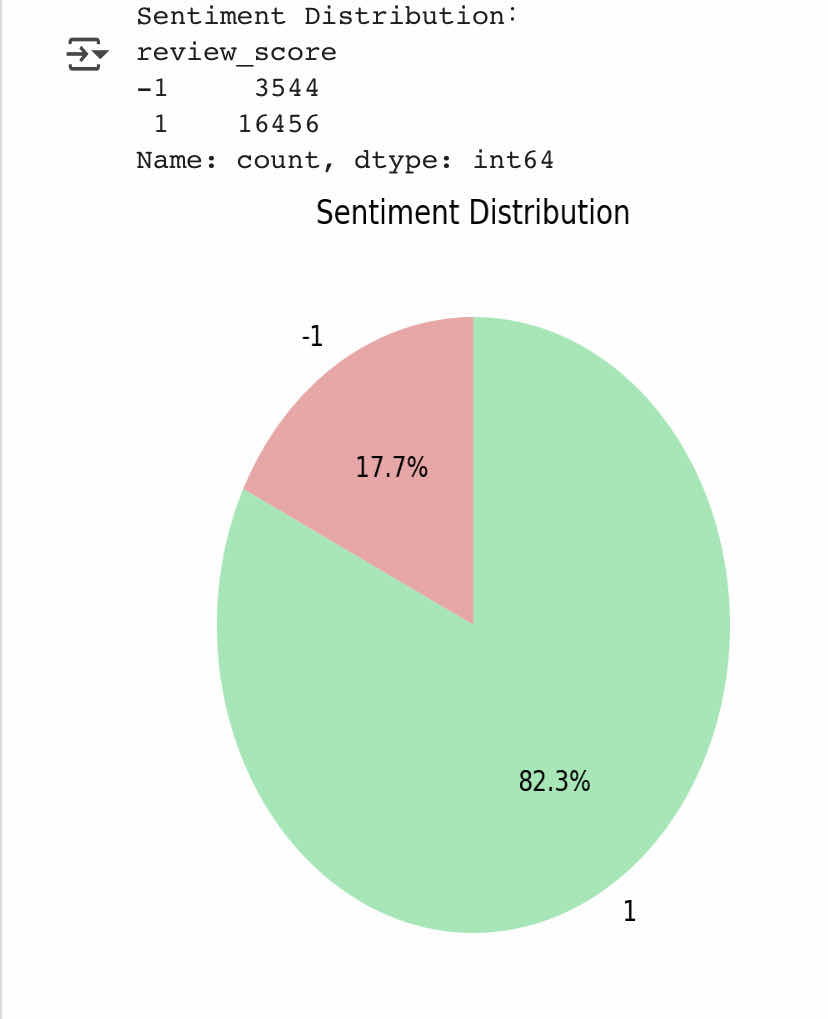
We use CountVectorizer to transform the text into a word frequency matrix, set X as a NumPy array, and define y as the target variable, where 1 represents positive and 0 represents negative.

The dataset is split into 80% training and 20% testing. We import three classic Naive Bayes models: GaussianNB, MultinomialNB, and BernoulliNB. And the evaluation metrics used to test these models are accuracy\_score, precision\_score, and confusion\_matrix. After testing, we select the best-performing model as the sentiment analysis model and define a sentiment prediction function. The final result is as follows: after inputting a string of user comments, the model uses CountVectorizer to vectorize the text, applies the MultinomialNB model to predict the sentiment, and finally returns either "Positive" or "Negative" as the result.

# **5. Results**

## **5.1 Descriptive Analysis**

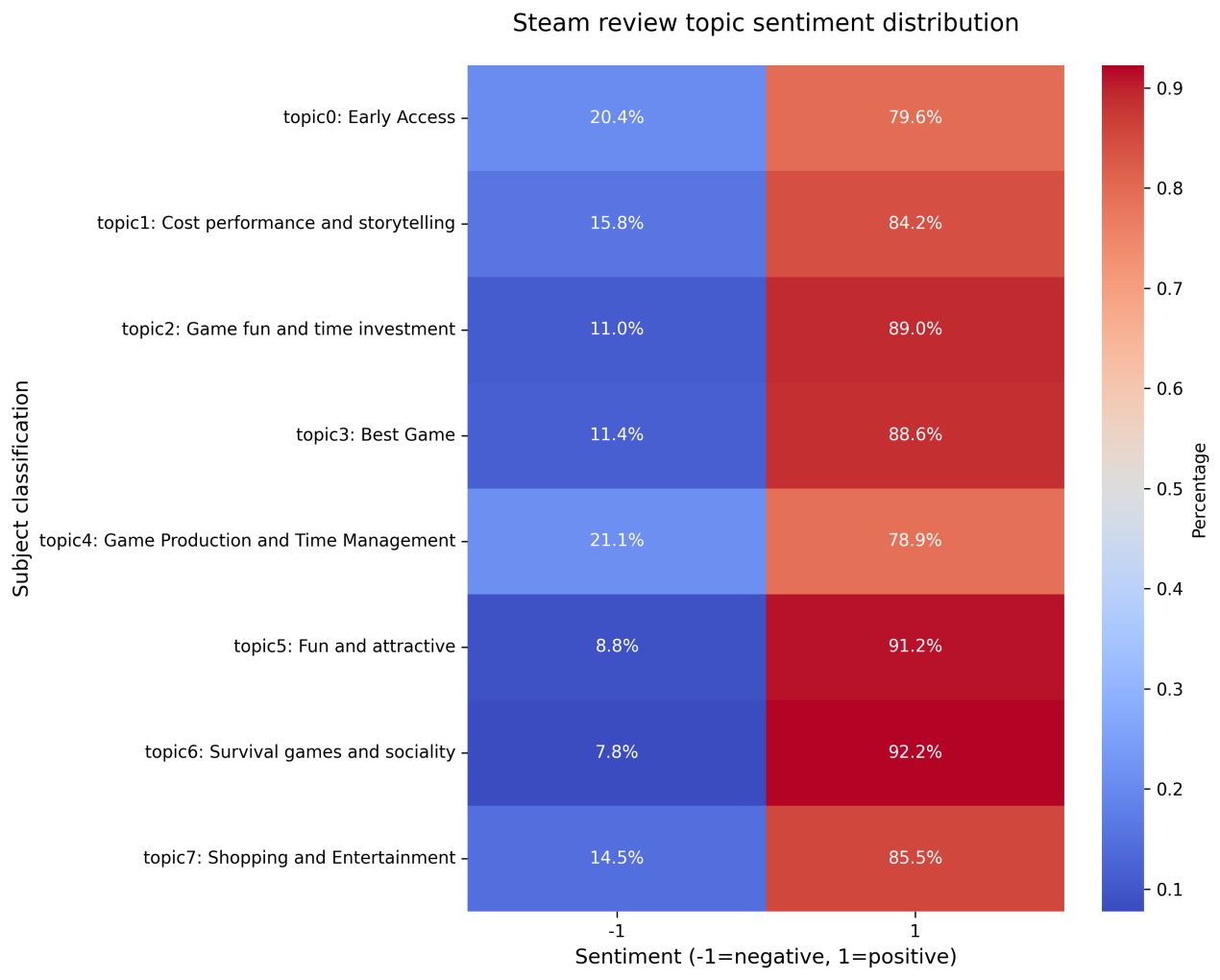
For review length, the calculation shows that the average length is approximately 55.36 words,the histogram reveals a significant variance in the distribution of review lengths, indicating that when expressing evaluations , players vary greatly. The analysis of sentiment distribution shows that the number of positive reviews is 16,456, accounting for a high proportion of 82.3%, while the number of negative accounting for 17.7%. This indicates that in this sampled data, players' overall evaluations of the games are relatively positive. The generation of word clouds further explores players' core concerns. In the word cloud of positive reviews, high - frequency words such as "play", "fun", and "story" highlight that the appeal of the storyline are important factors for winning players' approval.For negative reviews, "problem", "boring", and "early access" frequently appear, suggesting that issues like imperfections during the early access phase are likely to cause player dissatisfaction. In the statistics of popular games, the top ten games with the largest number of reviews have been selected. Among them, PAYDAY 2 ranks first with 308 reviews, and Terraria and DayZ are tied for second place with 275 reviews each. The analysis of the proportions of positive and negative reviews for these top ten games presents differences in the word - of - mouth of each game through stacked bar charts. These results provide intuitive and crucial reference information for game developers and operators, helping them to optimize and promote games in a targeted manner.





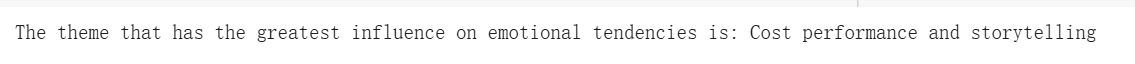
## **5.2 LDA analysis**

After training, we successfully identified eight topics, which were named after the top three most frequent keywords. We listed the top ten words for each topic and, basing on our knowledge of the game industry, we manually refined and labeled these keywords for clarity. In the end, we identified the following eight topics: Early Access, Cost performance and storytelling, Game fun and time investment, Best Game, Game Production and Time Management, Fun and attractive, Survival games and sociality, Shopping and Entertainment. Through the heat map of the sentiment mapping of these topics, we can see that, the positive sentiment in the Survival & Sociality category reaches 92.2%, which shows that users highly recognize the social mechanisms and gameplay design in survival games. Similarly, the Fun & Attractive topic shows a positive sentiment of 91.2%, the importance of enjoyment in players' experiences. In contrast, the negative sentiment of the Production & Time Management topic accounted for 21.1%, the highest among all topics, which suggests that the game may have shortcomings in these areas and needs further optimization to improve user satisfaction. Additionally, the negative sentiment of the Early Access topic also reached 20.4%, which is also worthy of attention. This shows that the experience mechanism of early content can indeed reflect the user's opinions on the game. Therefore, improving communication and feedback with users can lead to necessary enhancements in the game's post-production.

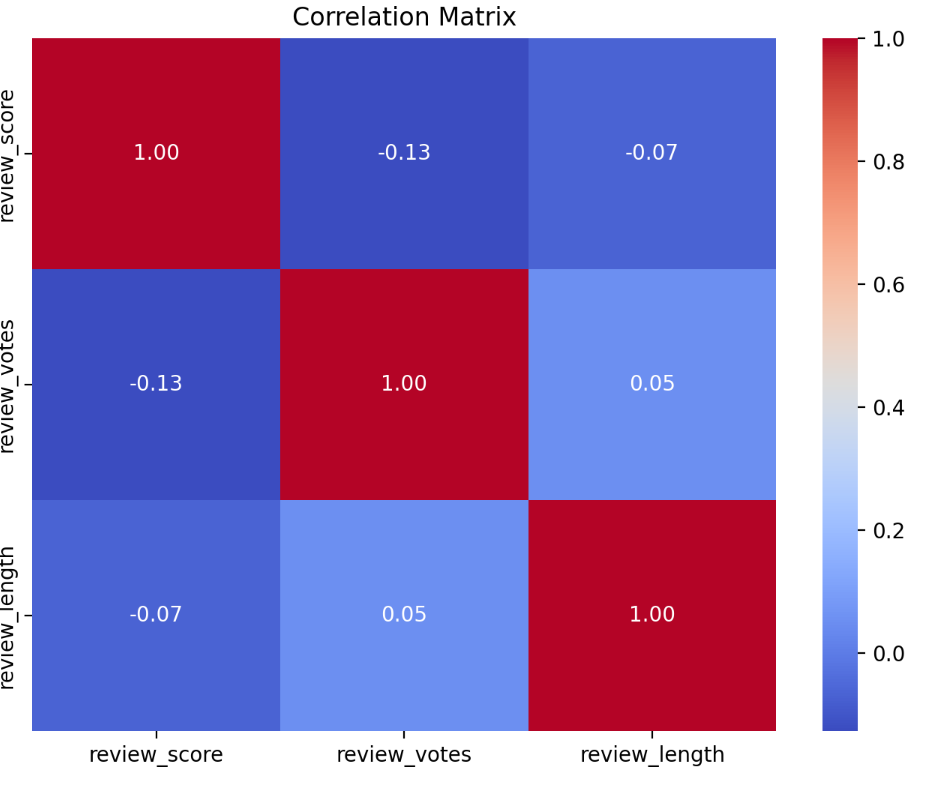


## **5.3 Correlation analysis**

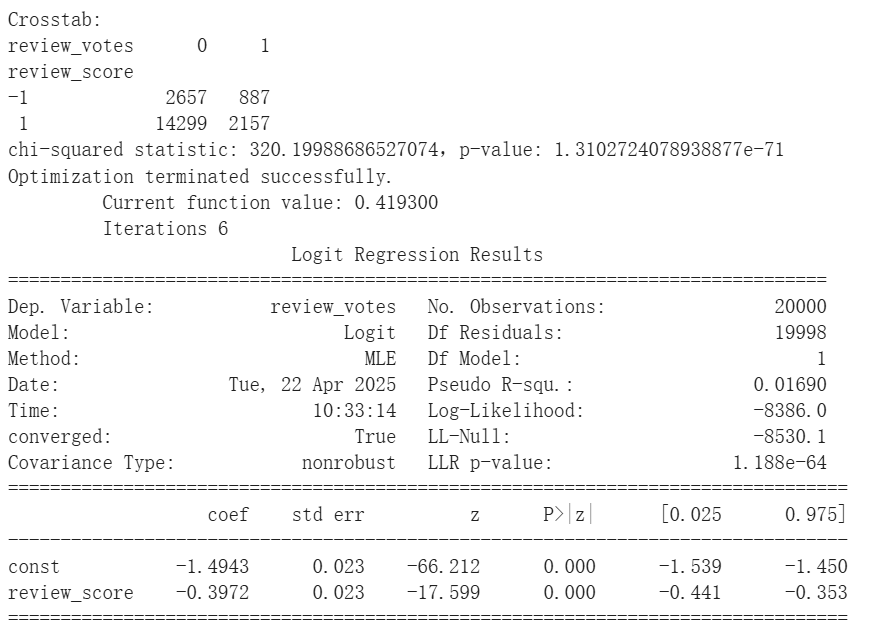
Based on LDA topic modeling, I use these 8 topics to analyse which factors have the greatest influence on emotional tendencies. The result shows that each topic has little impact on emotions and moods, but "Cost performance and storytelling" influence it more.



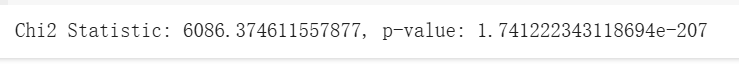
The resulting heat map shows the correlation between the sentiment positives and negatives of a review (review\_score), the number of supports for a review (review\_votes) and the length of a review. It can be seen from the figure that all the values except themselves are around 0, which indicates that the length of the comments independent of positive or negative emotions and whether the comments are recommended or not.



The results of the cross-tabulation produced more positive comments in absolute numbers, but the percentage of votes was lower than the percentage of votes for negative comments, which could mean that negative comments are more likely to get votes. The results of the chi-square test tell us that the p-value is much less than 0.05 and we can reject the original hypothesis, indicating that there is a significant association between review\_score and review\_votes. From the results of the logistic regression, we can find that the logarithmic incidence of getting votes (log-odds) decreases by 0.3972 when review\_score changes from negative comments (-1) to positive comments (1). Those means that the chances of getting votes (odds) are higher for negative comments than for positive comments. Thus, there is a significant association between the positivity of comments and the number of votes.



Finally, I verify the correlation between different apps and review sentiment, and the data show that the distribution of review sentiment (positive/negative) varies very significantly across apps

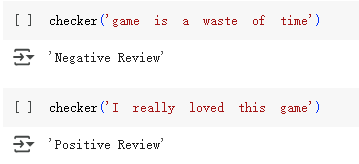


## **5.4 Sentiment analysis**

In this project, we used three Naive Bayes models to classify the sentiment of Steam review texts: GaussianNB, MultinomialNB, and BernoulliNB. After converting the text into word frequency vectors using CountVectorizer with a maximum dimension of 3000, we split the data into a training set and a test set in an 80:20 ratio, and trained and evaluated the three models.

Experimental results show that the MultinomialNB model performs best in terms of accuracy and precision, achieving an accuracy of 84.35% and a precision of 87.26%. In comparison, GaussianNB achieved an accuracy of 74.85%—relatively high, but with many misclassifications. BernoulliNB reached an accuracy of 75.77% and a precision of 84.15%, showing slightly lower overall performance.

Considering various performance metrics and model suitability, we ultimately selected MultinomialNB as the core model for sentiment analysis. This model not only trains quickly but also performs well on discrete text data. We defined a checker function based on this model and completed the training and prediction process by calling the .fit() and .predict() methods of the MultinomialNB model. The checker function can predict the sentiment of any comment and return a "Positive" or "Negative" result. For example, for the comment "game is a waste of time", the system correctly classifies it as negative; for "I really loved this game", it correctly classifies it as positive.



# **Conclusion**

## **6.1 Summary**

Through this analysis, we can systematically answer the questions raised at the beginning.

First, the characteristics of player comments. In the descriptive analysis, we found that players tend to post short and positive comments, mainly focusing on the fun and plot narrative of the game. Classic works such as "PAYDAY 2", "Terraria" and "DayZ", although they have been released for a long time, have a large number of player comments, indicating that high-quality content can support the continued attention of the game. In the correlation analysis part, we found that players are not inclined to post negative comments themselves, but like other negative comments, and many expressions of dissatisfaction are reflected through silent player churn rather than explicit negative feedback.

Based on the understanding of the characteristics of player comments, we further explored the focus of players and the motivations behind them. The Early Access stage is a very important means of publicity in the early stage of game operation, which can accumulate a wave of feedback and word of mouth. At the same time, because the steam platform is a buyout system, players pay great attention to the rationality of pricing and hope that the game is a type that can be invested in for a long time, so "survival" and "construction" games are relatively popular. The social attributes of the game are also quite important, and the multiplayer online mechanism is highly discussed. Others are all relatively conventional player needs, including content narrative, unique innovation, etc.

When exploring the factors that affect players' emotional tendencies, we found that not every topic can clearly distinguish between positive and negative emotions. However, the topic of "Cost Performance and Storytelling" has a greater impact on emotional fluctuations, and game operators need to be more cautious when pricing games. Finally, in order to achieve automatic recognition of comment emotions, we used the MultinomialNB model, combined with text feature extraction, to achieve automatic emotion recognition with an accuracy of 84.35%.

The important inspiration brought by this discovery is that players who leave negative reviews often continue to play, while players who really churn leave more silently. Therefore, when optimizing the player experience, negative comments alone are not enough to fully understand the reasons for churn. It is necessary to combine qualitative research to deeply explore the real motivations of silent churn players.

## **6.2 limitation and future improvement**

We still have limitation in conducting data exploration. In the LDA analysis, although theme labels were generated from high-frequency words and manually summarized, there is still some subjectivity in the definition and interpretation of themes leading to inaccurate theme delineation. Inaccurate delineation of themes also leads to inaccurate subsequent correlation analysis based on the eight topics. In addition, the Naive Bayes analysis based on the assumption of feature independence is inconsistent with the actual semantic correlation of the text, and the feature extraction may also lose important information.

In the future, we will try to add more features lable of the dataset and optimize data cleaning. Use some automated methods to aid in the interpretation of topics, such as incorporating external knowledge bases or domain knowledge to define and describe topics more accurately and reduce the subjectivity of manual interpretation. And improve Naive Bayes for sentiment analysis in a way that can handle feature correlation, like word vectors (e.g. Word2Vec).

# **Workload distribution matrix**

|  |  |  |
| --- | --- | --- |
| Introduction | Research background, research questions, research objectives | HU ZEXIN |
| Data Description | Data sources, processing, statistical description | LIU KEXIN |
| Descriptive Analysis | Method explanation and finding | LIU KEXIN |
| Machine Learning: LDA | Method explanation and finding | HU ZEXIN |
| Correlation Analysis | Method explanation and finding | ZENG YUTING |
| Machine Learning:Sentiment Analysis | Method explanation and finding | JIANG HANBING |
| Summary | Key findings, value of data analysis | JIANG HANBING |
| Limitations and Future Work | Limitations, potential improvements | ZENG YUTING |

1. **Reference**

Galyonkin, S. (2018). SteamSpy - All the data and stats about Steam games.

http://steamspy.com/. (last visited: Oct 16, 2018).