ANALYZING PLAYER FEEDBACK IN STEAM REVIEW ACROSS GAME GENRES

SAFIRA NURUL IZZA

UNIVERSITI TEKNOLOGI MALAYSIA

Table of Content

CHAPTER 3 RESEARCH METHODOLOGY

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter starts by explaining the research framework to conduct sentiment analysis on Steam player reviews across genres. Furthermore, the dataset collection method is also explained in this chapter. After the dataset collection process was completed, the data went through a preprocessing stage, where data cleaning and transformation were performed to produce relevant data for analysis. Finally, this chapter provides an explanation of feature engineering to prepare the features needed in machine learning based on the findings from the data processing process that has been carried out.

3.2 Research Framework

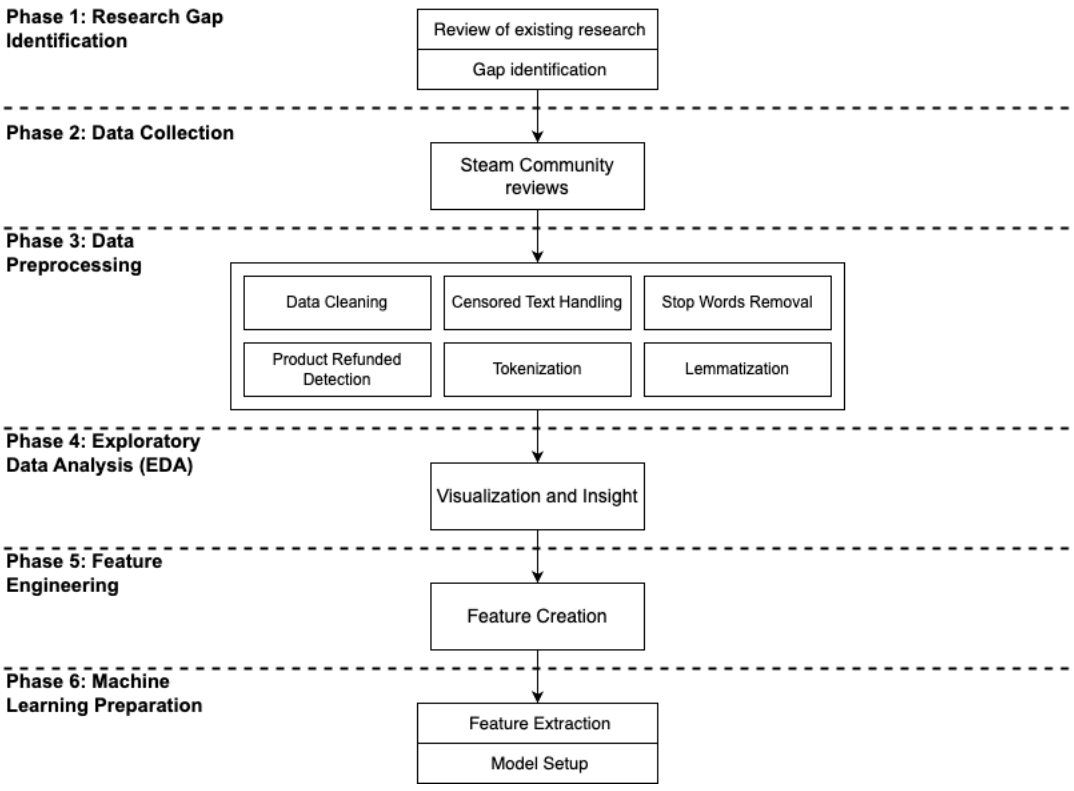

Figure 3.1: Research Framework

There are five phases in the research framework to analyse player review sentiment on Steam games across genres. The first phase is the identification of research gaps, where various studies and papers relevant to this project have been reviewed and studied. Through this phase, insights are obtained to apply the best method or model that can be implemented into this project. In addition, the identification of research gaps also aims to find aspects that have not been widely discussed in previous studies, such as sentiment analysis on reviews containing censored text or the label "product refunded.". Phase 1: Research Gap Identification was conducted and discussed in Chapter 2 where relevant literature was reviewed for this study.

The next phase is data collection, where review data is taken and gathered directly from Steam Community Reviews. This ensures that the dataset used is original data that comes directly from players. The third phase is data preprocessing, where raw data is cleaned and transformed into a consistent and structured format suitable for analysis. The fourth phase is Exploratory Data Analysis (EDA), aimed at understanding the structure, trends, and patterns hidden within the data. The next phase is Feature Engineering, which involves creating or modifying features in the dataset to improve the performance of machine learning models. The final phase is machine learning preparation. This phase consists of two processes: feature extraction and model setup. The goal of feature extraction is to transform the cleaned and pre-processed text into a numerical format. After that, the feature extraction process continues to separate the data and select the machine learning model.

3.3 Data Collection

Continuing the research framework, the next step is Phase 2: Data Collection. Determining the data source is a crucial step in the sentiment analysis process. In this project, it is essential to ensure that the review data used is original and obtained directly from a trusted source, which in this case is the players themselves, without any modifications or manipulation. Once the data source is determined, the next step is to proceed with data collection. During the data collection phase, the relevance of the data

to the project objectives was also considered to ensure that the analysis provided results aligned with the project's goals.

3.3.1 Data Source

The data source for the analysis process was taken from the Steam Community, available on the official Steam website, accessible via the link https://steamcommunity.com/. The Steam Community is a discussion and interaction platform for players, which provides user reviews of games they have played. Each review on the Steam Community includes information such as the username, the content of the review, an indication of whether the player recommends the game (Recommended/Not Recommended), the amount of time spent playing, and the date the review was published.

3.3.2 Data Collection Method

The reviews were collected through web scraping using Microsoft Edge WebDriver. Data for 5 genres (Action, FPS, Indie, RPG, and Strategy) were collected, consisting of approximately 20,000 reviews per genre, resulting in a total dataset of approximately 100,000. The scraping process targeted certain attributes, such as review content, thumb text (recommended or not recommended), play hours, and date posted. To maintain consistency, review filters were applied to only include reviews written in English. Additional filters were also applied, such as "Most Helpful" and "Show All Reviews." The "Most Helpful" filter displays reviews that other players find relevant or relatable. The figure below shows the filters that were set for the data collection process.

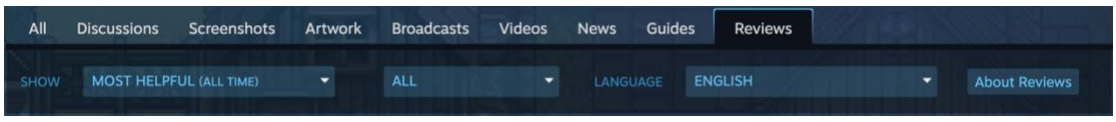

Figure 3.2: Steam Review Filters

Figure 3.3 shows examples of two reviews that received a lot of agreement and engagement from other players. One review, written by a user with the display name **S**

A L S A ⧘, states: "*You play this game every night on your bed when you imagine fake scenarios like having your own house, money, and a wife.*" This review was voted helpful by 1,231 people, showing its impact in the community.

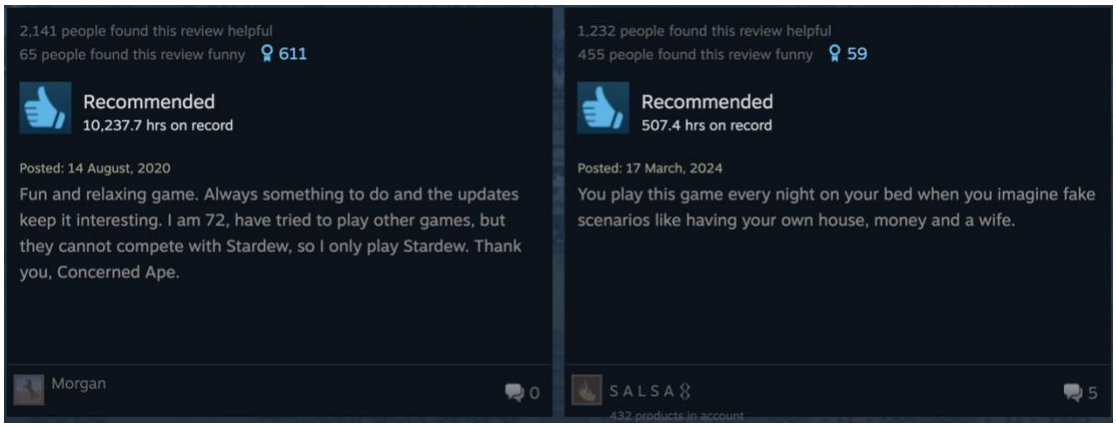

Figure 3.3: Examples of Highly Engaged Reviews

The "Show All Reviews" filter was used to display all types of reviews, regardless of whether the review is Recommended or Not Recommended. By activating this filter, the data collection process becomes more comprehensive because it includes reviews from both sides of the player's perspective, both those who provide positive recommendations and those who provide criticism or complaints.

Initially, the number of reviews required for each game in the dataset was set at 4,000. However, to avoid significant data reduction during the data cleaning process, the number of reviews retrieved was increased to 5,000 per game. Nevertheless, during the scraping process on some games, there was an obstacle where the number of reviews with the "Most Helpful" filter did not meet the target of 5,000 reviews. To overcome this problem, the scraping process was repeated once again on games that did not meet the target using the "Most Recent" filter. In this second scraping process, additional code was applied to ensure that the retrieved reviews were not duplicated with reviews from the "Most Helpful" filter. After the target number of reviews was reached, the two groups of reviews from different filters were merged into one dataset.

3.4 Data Preprocessing

Data preprocessing is a crucial step aimed at preparing raw data for further analysis. Raw data collected through web scraping often contains incomplete data (e.g., missing values or empty data), inconsistencies, and information that is irrelevant to the study's objectives. These issues can reduce the accuracy of the analysis. The data preprocessing process ensures that the dataset is clean and consistent before moving on to the analysis phase.

3.4.1 Data Cleaning

Data cleaning is done to ensure that the dataset is relevant, accurate, and ready for analysis. The image below shows how the review cleaning flows from raw data to data that is ready to be processed in the next stage.

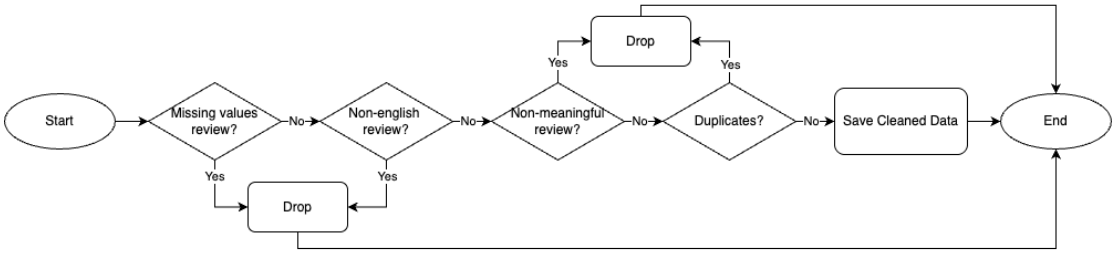

Figure 3.4: Data Cleaning Flow

The process begins by checking for missing values in the Review Content, Thumb Text, Review Length, Play Hours, and Date Posted columns. Missing values in any of these five columns can make the analysis process less optimal. After that, reviews that are not in English are also removed, because this analysis only focuses on English-language sentiment analysis. Although during data collection it was set to only collect English-language reviews, there were still some reviews written in other languages. Therefore, removing reviews that are not in English is an important step to ensure consistency and accuracy in sentiment analysis.

Not all players provide detailed and long feedback; there are some reviews that only contain one word, such as "Great," "Nice," "Bad," or "Foul." This type of review is less

helpful in the analysis process because of the lack of meaningful feedback. Therefore, reviews that are less than six alphanumeric characters long are removed. Because the required dataset must have data variations, a check is carried out on duplicate review content. If the program detects a duplicate review, the review is deleted.

Several reviews were found to contain love emoticons, as seen in Figure 3.5. At first glance, one might think that players are using love emoticons because of the positive experience they got from the game. However, upon further examination by reading the overall context of the review, it can be ascertained that the emoticons are replacements for words that have been censored by Steam. It turns out that Steam replaces inappropriate words or swear words in player reviews with the ♥♥♥ emoticon.

Figure 3.5: Reviews With Censored Words

Not all reviews containing censorship are bad reviews. It could be that it is a way for players to convey their positive feedback with excitement or strong emotions towards their playing experience in the game. To maintain consistency in sentiment analysis, all censored words with emoticons were changed to "[censored]" for further analysis. Some reviews had the label "Product Refunded.", as illustrated in Figure 3.6. Reviews with this label were identified and retained. This is important for analysing player dissatisfaction patterns and understanding the reasons behind game refunds.

Figure 3.6: Review Labelled as “Product Refunded”

3.4.2 Data Transformation

After cleaning the data, certain attributes were standardized to ensure consistency and enable meaningful analysis. To ensure clarity and structured analysis, categories (genres) and game IDs were added to the dataset. This step aimed to group data based on games and genres as unique identifiers, so that data integrity could be maintained properly. Next, the “Play Hours” and “Date Posted” columns were standardized to maintain consistency across datasets. This process was useful because the data became simplified, making the analysis, visualization, and calculation processes easier.

	Review Content	Thumb Text	Review Length	Play Hours	Date Posted
0	the closest we're getting for a bloodborne gam...	Recommended	57	36.7 hrs on record	Posted: 29 October, 2023
1	the children yearn for bloodborne	Recommended	29	26.5 hrs on record	Posted: 3 February
2	Experience the horror of being french	Recommended	32	23.3 hrs on record	Posted: 15 October, 2023
3	all right then. keep your Bloodborne Sony.	Recommended	36	101.6 hrs on record	Posted: 21 December, 2023
4	They really went fine i'll make Bloodborne on ...	Recommended	50	41.5 hrs on record	Posted: 14 October, 2023

Figure 3.7: Raw Review Data Before Standardization

	ID	Category	Review Content	Thumb Text	Review Length	Play Hours	Month-Year
0	1174180	Action	the closest we're getting for a bloodborne gam...	Recommended	57	36.7	10-2023
1	1174180	Action	the children yearn for bloodborne	Recommended	29	26.5	02-2024
2	1174180	Action	Experience the horror of being french	Recommended	32	23.3	10-2023
3	1174180	Action	all right then. keep your Bloodborne Sony.	Recommended	36	101.6	12-2023
4	1174180	Action	They really went fine i'll make Bloodborne on ...	Recommended	50	41.5	10-2023

Figure 3.8: Review Data After Standardization

The next step was to convert all text to lowercase to ensure uniformity and avoid case-sensitivity issues. This process aims to simplify text processing for machine learning models. Commonly used words such as “the,” “a,” “an,” or “in” are removed using the Natural Language Toolkit (NLTK). This is done to ensure the analysis focuses only on important words, thereby improving the accuracy and efficiency of sentiment analysis in a given text.

Next, tokenization was done to break down a text into small units or part, that is similar to what we call tokens. This process aims to change unstructured text into structured data so that it is easier to analyse. After tokenization, the lemmatization step was performed to reduce each word to its basic form (root word or lemma), while maintaining its meaning and context. Both of these processes are very important in sentiment analysis, because computers process data in numerical form, not raw text. After the dataset went through these processes, the transformed dataset was saved for further analysis.

3.4.3 Feature Engineering

Feature Engineering is the process of creating or changing features in datasets to make a machine learning model perform better. To get maximum results, the quality of the features used greatly determines the success of the machine learning model, so that it can find important patterns and relationships in the data so that the model can learn better. (GeeksforGeeks, 2023). In this project, several features are created to extract more meaningful information from the data that has been collected and has gone through various processes. These features can be seen below:

1. Sentiment Tone Score

The sentiment score was taken from the “Content Review” column using one of the sentiment analysis tools, where VADER will be used in this project . This score is

used to determine whether the tone of the review is positive, negative, or neutral, which is the basis for sentiment classification.

2. Review Length

A new feature called Review Length has been created to measure the length of each review based on the number of characters. This feature helps provide insight into the correlation between the number of characters in player feedback and sentiment.

3. Grouping of Playing Hours

The hours spent by players in a game have been categorized into three labels, low, medium, and high. This labelling is useful for providing further understanding regarding the level of player engagement, and getting a comparison between players' satisfaction with the game and their playing time.

4. Time Pattern

The “Date Posted” column is changed to “Month-Year” format for analysing review trends over time. This transformation makes it easier to create visual graphs and compare review patterns over a period of time.

5. Product Refunded

A feature called “Product Refunded” has been created to identify game products that have been returned by players. This feature was created because during the data cleaning process, several games were found to have the label “Product refunded”. This makes the label a valuable feature to get the discovery of player dissatisfaction so that they want to return the product.

6. Censored Text

A unique attribute was found in the dataset, where text containing inappropriate or offensive language in Steam reviews was censored using the ♥♥♥ emoticon. A feature called "Censored Text" was created to capture the intensity and emotional sentiment expressed in these reviews.

The six features mentioned above are important for understanding player behaviour, identifying sentiment trends, and being well prepared for the machine learning process.

3.5 Conclusion

This chapter outlined the methodology employed to conduct sentiment analysis on Steam player reviews across genres. Beginning with Phase 1, the research identified gaps in existing studies through a literature review, as discussed in Chapter 2. Next, the methodology moved to data collection and data preprocessing phases to ensure clean, structured and meaningful datasets. In addition, feature engineering was carried out to enhance the analytical potential of the datasets.