Sentiment Analysis on Game Reviews

Analysis Theoretical Report

Jessica Christine Erasmus
July 2023

Table of Contents

1 Introduction 3

2 Background 3

2.1 The dataset 3

2.2 Text Processing 3

2.3 Descriptive Statistics 4

2.4 Exploratory Data Analysis 4

2.5 Predictive analysis 5

2.6 Sentiment Analysis 5

2.7 Usability 5

2.7.1 Dataset suitability 5

2.7.2 Legitimacy 5

2.7.3 Purpose of analysis 5

3 Pre-processing Data 6

3.1 Loading Data 6

3.2 Data Cleaning 7

4 Exploratory data analysis 8

4.1 Popularity 8

4.2 Average Rating per year 10

4.3 Game Engagement and Performance 10

4.4 Correlation 11

4.5 Target Variable 13

5 Natural language processing 14

5.1 Stop words 14

5.2 Language detection 14

5.3 Term frequency-inverse document frequency method 15

5.4 Advanced tokenisation 15

5.4.1 Word tokenisation 15

5.4.2 Multi-word expression tokenisation 15

5.4.3 Custom tokenisation 16

5.5 Stemming 16

5.6 Lemmatisation 16

5.6.1 Part-of-Speech (POS) Tagging 16

5.6.2 SpaCy 17
5.7 Latent Dirichlet Allocation 17
6 Data Modelling 17
6.1 Splitting the data 17
6.2 Pipelines 18
6.2.1 Classifiers 18
6.2.2 NLP pipeline 20
7 Model evaluation and improvement 20
7.1 Cross validation 21
8 Predictions 21
9 Conclusion 21
10 References 22

1 Introduction

The aim of the assignment is to create a model using a specific dataset made from text data. For this task, sentiment analysis took place to determine whether each overall game review was positive or negative. Reviews like these can then help gaming companies improve their current games or create even better ones in the future. A task as specific as this would also save busy companies' time instead of having to manually scroll comment for comment on what players like about their, the dislikes on the game and how frequently the game is be played or was played. This provides a competitive advantage towards these companies as well due to being able to manage large quantities of data in an effective, fast, and accurate way by combining text processing with machine learning.

This assignment contains analysing the gaming reviews from 1980 to 2023, which includes descriptive analysis, predictive analysis, and sentiment analysis. The assignment also includes text analysis techniques such as tokenisation and stemming. To find a suitable model, many models will be tested and chain multiple steps together with the use of a pipeline. Furthermore, what might make this even more difficult to achieve is that it requires the tasks which were performed in the earlier assignments for this module.

2 Background

2.1 The dataset

The dataset makes use of gaming reviews from 1980 to the current year (2023) which later is used to analyse treads and predict the overall success of the mentioned games. The original dataset will be used to gain overall exploratory statistics and a subset of the data will be used to predict if these reviews are positive or negative. To determine or target feature [Rating Status], two X variables. [Reviews] and [Rating] will be used.

The reviews column holds the reviews of many players and is also one of the most important features to pay attention to. This feature has many unique values and will be manipulated often throughout the assignment to process data and conclude the analysis.

The rating column will be used to estimated which reviews were good or bad based on its rating score. Hence this column helps to train sentiment analysis, which will be explained in detail later, and helps to predict the rating status column.

The rating status column is determined using one-hot binary encoding to determine positive and negative ratings. Therefore, data is split and grouped into positive reviews and negative reviews, which includes text sentiment.

2.2 Text Processing

Text processing is the automated method of analysing and categorizing unstructured text input to extract useful information (Roldós, 2019). This unstructured data contains insights and views on numerous topics, goods, and services, but businesses must first organize, analyse, and measure text data to access this important information (Roldós, 2019). It includes extracting, transforming, and interpreting textual information for various purposes.

Text processing activities can range from simple tasks such as finding, replacing, and formatting text to complex tasks such as NLP (natural language processing) and text mining. Text processing may be used by product teams to extract insights from customer input to build their product roadmap, while customer service teams may use it to automate operations. Text processing apps manage vast quantities of text to do categorization or translation, which requires a significant amount of effort on the server side. (Nabi, 2018) It is a challenging process to convert text into something that an algorithm can understand. There are several meanings associated with various words that our current algorithms cannot comprehend. Therefore, the data must be cleaned and processed before it can be fed into an algorithm, such as stemming, lemmatization, and tokenization.

2.3 Descriptive Statistics

The process of summarizing and characterizing a dataset to acquire insights into its main traits and qualities is referred to as descriptive analysis or descriptive statistics (Muller & Guido, 2018). It entails employing a variety of statistical metrics and techniques to organize, display, and summarize data in a relevant manner. The goal of descriptive analysis is to provide a brief and useful overview of the dataset without drawing any conclusions or making any generalizations beyond the data itself (Muller & Guido, 2018). It is often the first phase in examining a dataset and serves as the basis for further statistical analysis or data exploration. Therefore, descriptive statistics will be used to help gain initial insight to the data through exploratory data analysis.

2.4 Exploratory Data Analysis

EDA is the process of evaluating and summarizing a dataset to discover patterns and find potential correlations between variables (Madhugiri, 2023). According to the characteristics of the data, the specific aims of the analysis may vary. Data visualization tools such as histograms, scatter plots, bar charts, box plots, and heatmaps are often used.

EDA has a vast set of goals. EDA includes summarizing the key characteristics of a dataset using descriptive statistics, visualizations, and data exploration techniques such as calculating metrics such as central tendency, dispersion, and correlation and developing plots, charts, and graphs that represent the data (Madhugiri, 2023). EDA aids in identifying patterns, trends, or anomalies within a dataset by studying variable distributions, recognizing outliers, and exploring links or dependencies between variables (IBM, 2023). EDA includes assessing the dataset's quality and credibility, as well as looking for missing values, discrepancies in the data, and probable mistakes. Taking care of data quality concerns is critical for maintaining the dependability of further analysis. EDA aids in selecting which variables or characteristics are the most important and instructive for further analysis or modelling. Researchers can focus on the most significant aspects by analysing the correlations and relevance of different variables (Madhugiri, 2023). Therefore, EDA plays a great role in terms of understanding the data, generating insights, and guiding subsequent analyses or decision-making processes.

2.5 Predictive analysis

Predictive analytics is the process of making predictions or forecasts about future events or outcomes using historical data and statistical tools (Hastie, et al., 2017). It involves constructing mathematical models that can predict future behaviour or trends based on relationships and patterns discovered in data (Stevens, 2022). Predictive analysis plays a crucial role in preparing and selection model to make accurate and sufficient predictions. Once the model is fully trained, the model is then deployed and makes predictions on the target feature. In this case, predictive analysis enables the involved organizations to make data-driven decisions and foresee customer behaviour. Predictive analysis uses past data to reveal hidden patterns, correlations, and trends that may be utilized to foresee and plan (Stevens, 2022).

2.6 Sentiment Analysis

Sentiment analysis is a type of text mining that seeks to understand people's views, feelings, and attitudes toward something (Pascual, 2022) Sentiment analysis may be used on a variety of text data, including customer reviews. It gives businesses and organizations important insights into consumer attitudes, identifies new trends, and makes decisions based on the data. It is however crucial for one to understand that sentiment analysis may not always be accurate and can be impacted by issues like sarcasm, irony, cultural distinctions, and sentiment based on context. However, with the use of other analysis approaches, sentiment analysis can provide valuable insights expressed through sentiment.

2.7 Usability

2.7.1 Dataset suitability

Gaming reviews capture user feedback, preferences, and experiences. Text processing techniques enable useful information from reviews to be extracted, such as the most loved or disliked features, popular game genres, or typical player concerns. The text data for the game dataset can provide valuable insights into the strengths, weaknesses, and overall sentiment surrounding specific games or game features. It is feasible to establish whether the reviews are positive, negative, or neutral by using sentiment analysis algorithms. This data may help game developers, marketers, and gaming industry decision-makers evaluate user satisfaction, identify areas for development, and modify their plans appropriately.

2.7.2 Legitimacy

This dataset is based on multiple existing games however the data legitimacy cannot be confirmed. The model must be trained on real-world data to effectively extract sentiment from it. Different people's speech patterns and mannerisms must be considered, and data derived from one individual is unlikely to be included in these variations. The credibility of the data is taken from the Backlogged websites.

2.7.3 Purpose of analysis

The purpose of the analysis is to find interesting trends within game review data and predict positive and negative reviews within the dataset. Textual analysis of game reviews gives insights into market trends, player preferences, and the competitive environment. Understanding player feelings and preferences stated in reviews may help gaming firms gain

a competitive advantage, uncover new trends, and evaluate their games' performance in comparison to competitors.

Text processing also assists in the monitoring and management of game and game developer reputations (Nabi, 2018). Companies may detect and handle player complaints, communicate with their player community, and establish a positive relationship with their player base by monitoring game reviews. This increases both player pleasure and loyalty. Therefore, text processing in gaming reviews is used to extract important information, attitudes, and preferences from textual data to influence decision-making, improve game development, increase player contentment, and remain competitive in the gaming business.

3 Pre-processing Data

3.1 Loading Data

Initially the data needs to be loaded and it needs to be checked that all the data from the source is properly loaded.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1512 entries, 0 to 1511
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         1512 non-null   int64
 1   Title              1512 non-null   object
 2   Release Date       1512 non-null   object
 3   Team               1511 non-null   object
 4   Rating             1499 non-null   float64
 5   Times Listed       1512 non-null   object
 6   Number of Reviews  1512 non-null   object
 7   Genres             1512 non-null   object
 8   Summary            1511 non-null   object
 9   Reviews            1512 non-null   object
 10  Plays              1512 non-null   object
 11  Playing            1512 non-null   object
 12  Backlogs           1512 non-null   object
 13  Wishlist           1512 non-null   object
dtypes: float64(1), int64(1), object(12)
memory usage: 165.5+ KB
```

	Unnamed: 0	Title	Release Date	Team	Rating	Times Listed	Number of Reviews	Genres	Summary	Reviews	Plays	Playing	Backlogs	Wishlist
595	595	Star Wars: The Force Unleashed	Sep 16, 2008	['THQ Wireless', 'Aspyr Media']	3.2	417	417	['Adventure', 'Brawler', 'RPG']	Star Wars: The Force Unleashed is a Star Wars ...	['okay!', 'Really makes you\n ...	7.4K	50	823	350
963	963	Castlevania: Aria of Sorrow	May 06, 2003	['Konami Computer Entertainment Tokyo', 'Konami']	4.2	1K	1K	['Adventure', 'Platform', 'RPG']	The year is 2035 and Soma Cruz is about to wit...	['An immensely satisfying game with a rich, da...	5.1K	114	1.7K	871

To prepare for data cleaning, inspecting the data to find null and duplicate values are important as well.

```
There are 0 duplicate values in this dataset.

Team        1
Rating     13
Summary     1
dtype: int64
```

As shown the data output show that there are no duplicates however there are a few columns with null values.

3.2 Data Cleaning

The initial step is to collect data and ensure that it is clean and ready to be put into the model. Any null or inaccurate numbers must be repaired or discarded since they cause inconsistencies in the data, potentially affecting accuracy. Dropping duplicate and unnecessary columns is important as well.

It is also important to ensure all data is has the same format, especially in dates for this case.

	Unnamed: 0	Title	Release Date	Team	Rating	Times Listed	Number of Reviews	Genres	Summary	Reviews	Plays	Playing	Backlogs	Wishlist
644	644	Deltarune	releases on TBD	['tobyfox']	4.3	313	313	['Adventure', 'Indie', 'Music', 'Puzzle', 'RPG']	UNDERTALE's parallel story, DELTARUNE. Meet ne...	['Spamton is so hot, I want to kiss him in the...	1.3K	83	468	617
649	649	Death Stranding 2	releases on TBD	['Kojima Productions']	NaN	105	105	['Adventure', 'Shooter']	NaN	[]	3	0	209	644
1252	1252	Elden Ring: Shadow of the Erdtree	releases on TBD	['FromSoftware', 'Bandai Namco Entertainment']	4.8	18	18	['Adventure', 'RPG']	An expansion to Elden Ring setting players on ...	['I really loved that they integrated Family G...	1	0	39	146

Three release dates are found which are undetermined. To solve this problem, the columns are dropped, and the datatype of the column is changed to datetime.

	Title	Team	Rating	Number of Reviews	Genres	Summary	Reviews	Plays	Playing	Backlogs	Wishlist	Release Year
0	Elden Ring	['Bandai Namco Entertainment', 'FromSoftware']	4.5	3.9K	['Adventure', 'RPG']	Elden Ring is a fantasy, action and open world...	["The first playthrough of elden ring is one o...	17K	3.8K	4.6K	4.8K	2022
1	Hades	['Supergiant Games']	4.3	2.9K	['Adventure', 'Brawler', 'Indie', 'RPG']	A rogue-lite hack and slash dungeon crawler in...	['convinced this is a roguelike for people who...	21K	3.2K	6.3K	3.6K	2019
2	The Legend of Zelda: Breath of the Wild	['Nintendo', 'Nintendo EPD Production Group No...	4.4	4.3K	['Adventure', 'RPG']	The Legend of Zelda: Breath of the Wild is the...	['This game is the game (that is not CS:GO) th...	30K	2.5K	5K	2.6K	2017
3	Undertale	['tobyfox', '8-4']	4.2	3.5K	['Adventure', 'Indie', 'RPG', 'Turn Based Stra...	A small child falls into the Underground, wher...	['soundtrack is tied for #1 with nier automata...	28K	679	4.9K	1.8K	2015
4	Hollow Knight	['Team Cherry']	4.4	3K	['Adventure', 'Indie', 'Platform']	A 2D metroidvania with an emphasis on close co...	["this games worldbuilding is incredible, with...	21K	2.4K	8.3K	2.3K	2017

Next, the columns become cleaned in terms of fixing column names, formats, and naming constructs.

	Title	Team	Rating	Number of Reviews	Genres	Summary	Reviews	Number of Plays	Active Players	Backlogs	Wishlist	Release Year
0	Elden Ring	['Bandai Namco Entertainment', 'FromSoftware']	4.5	3.9K	['Adventure', 'RPG']	Elden Ring is a fantasy, action and open world...	["The first playthrough of elden ring is one o...	17000	3800	4600	4800	2022
1	Hades	['Supergiant Games']	4.3	2.9K	['Adventure', 'Brawler', 'Indie', 'RPG']	A rogue-lite hack and slash dungeon crawler in...	['convinced this is a roguelike for people who...	21000	3200	6300	3600	2019
2	The Legend of Zelda: Breath of the Wild	['Nintendo', 'Nintendo EPD Production Group No...	4.4	4.3K	['Adventure', 'RPG']	The Legend of Zelda: Breath of the Wild is the...	['This game is the game (that is not CS:GO) th...	30000	2500	5000	2600	2017
3	Undertale	['tobyfox', '8-4']	4.2	3.5K	['Adventure', 'Indie', 'RPG', 'Turn Based Stra...	A small child falls into the Underground, wher...	['soundtrack is tied for #1 with nier automata...	28000	679	4900	1800	2015
4	Hollow Knight	['Team Cherry']	4.4	3K	['Adventure', 'Indie', 'Platform']	A 2D metroidvania with an emphasis on close co...	["this games worldbuilding is incredible, with...	21000	2400	8300	2300	2017

After cleaning the data and dropping all null values contain the following columns and information on the data.

```
Title                object
Team                 object
Rating              float64
Number of Reviews    object
Genres               object
Summary              object
Reviews              object
Number of Plays       int64
Active Players        int64
Backlogs              int64
Wishlist              int64
Release Year          int64
dtype: object
```

The values change within specific columns to look neater which later table results will display.

4 Exploratory data analysis

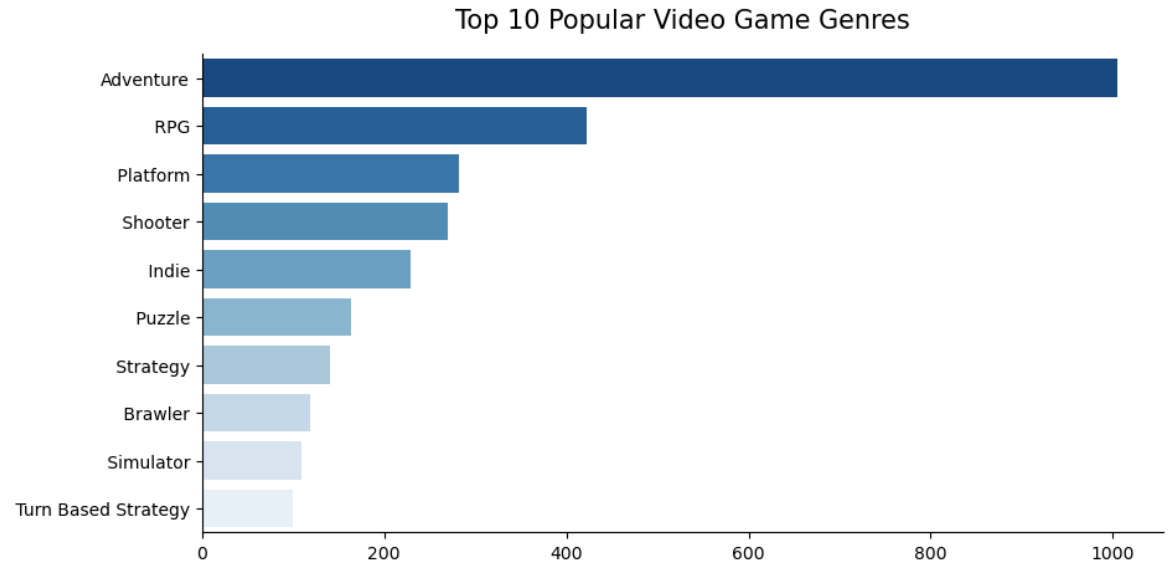
Exploratory data analysis helps us to explore and examine the dataset to gain a better understanding of its main characteristics. It entails a variety of techniques to summarize, visualize, and interpret data to discover relevant insights and identify potential problems biases, patterns, and correlations (Madhugiri, 2023). To explore the data, we will use data visualisation, summary statistics and correlation analysis.

4.1 Popularity

The popularity of a feature or trend on a dataset could correspond to its influence or understand the impact of the data distribution. Assuming this data is valuable to marketing analysis as well, tracking popularity can reveal patterns and help in making informed decisions and providing comprehensive analysis. First, we look at the popularity of each game.

```
Adventure    1005
 RPG          422
 Platform     282
 Shooter      270
 Indie        229
dtype: int64
```

The adventure genre seems to be popular among gamers. Let's look at the top 10 popular genres.

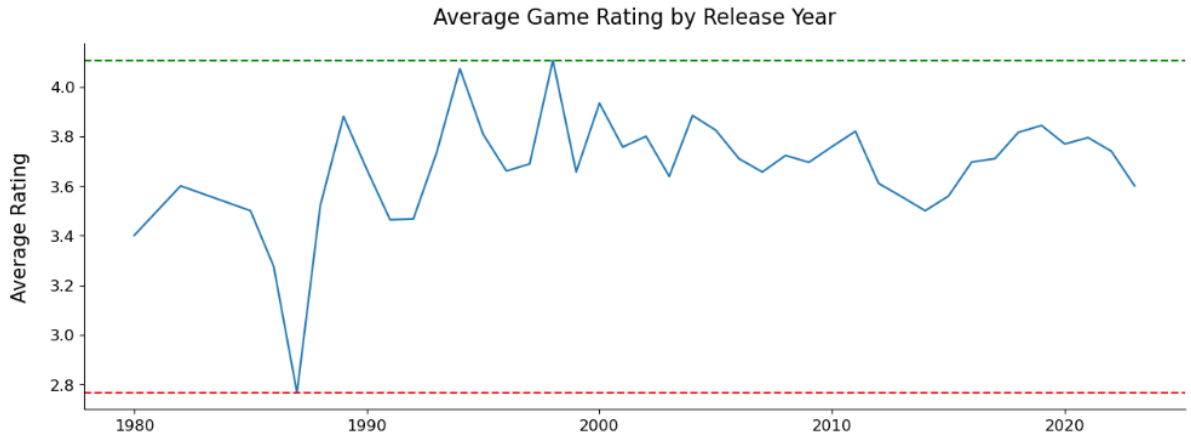


Now that we know what the most popular genres are, let's look at each genre most popular game.

	Game Genre	Most Popular Game
0	Adventure	Metal Gear Solid 2: Sons of Liberty
1	Arcade	Chrome Dino
2	Brawler	Devil May Cry
3	Card & Board Game	Mario Party Superstars
4	Fighting	MultiVersus
5	Indie	Friday the 13th: The Game
6	MOBA	Pokémon Unite
7	Music	Friday Night Funkin'
8	Platform	Super Mario Bros. 3
9	Puzzle	Tetris
10	RPG	Xenoblade Chronicles: Definitive Edition
11	Racing	Mario Kart Wii
12	Shooter	Doom
13	Simulator	Animal Crossing
14	Sport	Mario Strikers: Battle League
15	Visual Novel	Saya no Uta

4.2 Average Rating per year

After looking at the popularity, we consider the average rating produced per year. This helps to get an idea of if game production has improved since earlier or not. Here are the results.



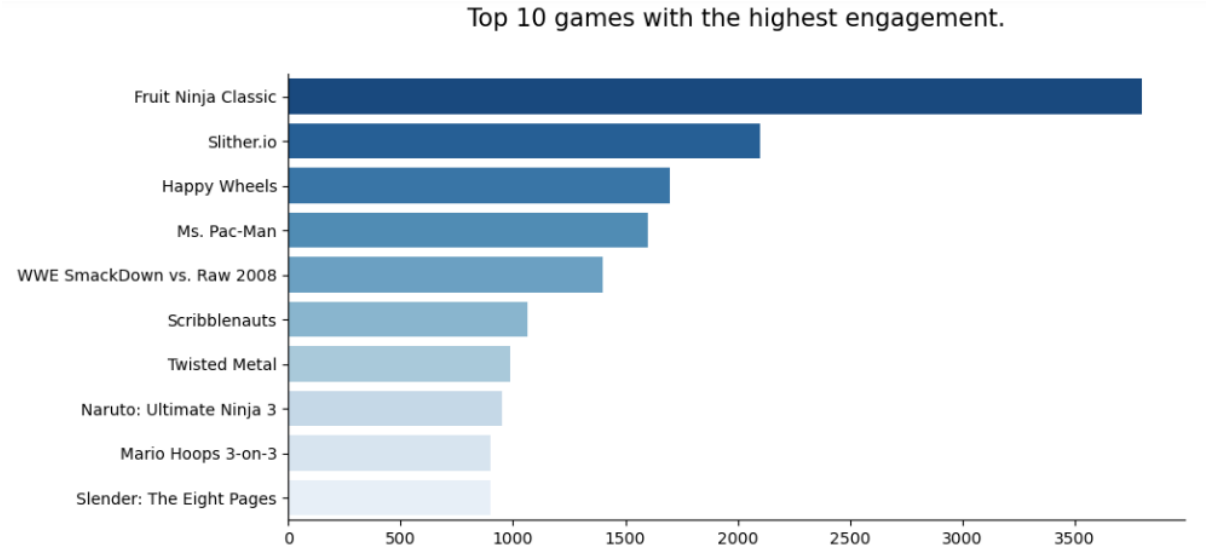
The maximum rating in the dataset seems to be above 4 (in the 1990s), and the lowest rating seems to be under 2,8 (in the 1980's). After 2000, majority game remained at average rating, under 4, and took a slight dip after 2010 before rising back to average again. However, it can also be taken in consideration that the amount of people rating games per year might not have been popular before 1990 and probably lost interest in providing ratings after 2000. Game graphics might have increase over time as well, but game quality could be a debatable based on gamer preference. Overall, it is clear to say that the 'peak' of gaming was in the 90's.

4.3 Game Engagement and Performance

Exploratory data analysis may give significant insights into player behaviour and game success by analysing game engagement. Therefore, the engagement of players can be analysed based on features such as player behaviour, frequency of play, number of plays, period of decreased plays as well as the effect of monetisation strategies.

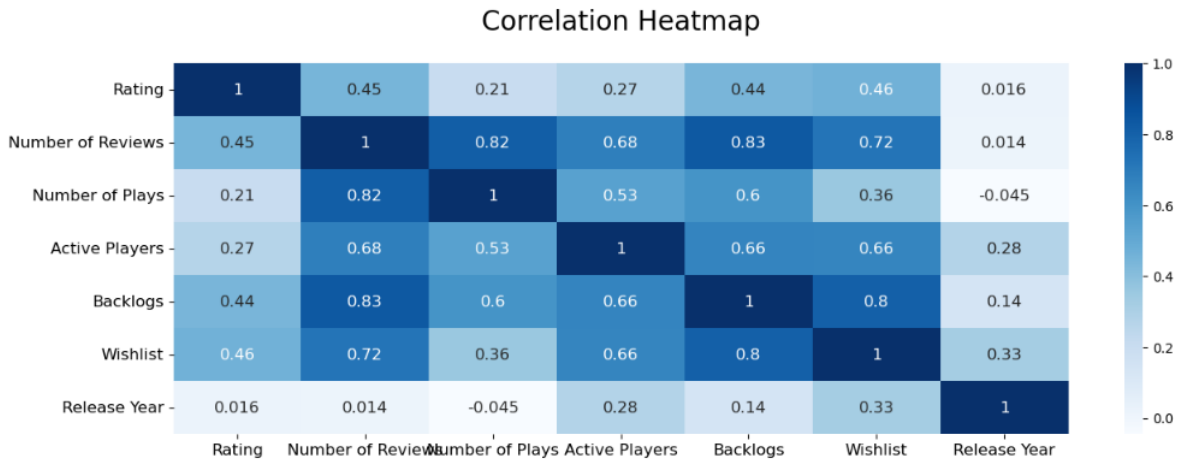
	Title	Team	Rating	Number of Reviews	Genres	Summary	Reviews	Number of Plays	Active Players	Backlogs	Wishlist	Release Year	Average Plays per Player
1331	Fruit Ninja Classic	Halfbrick Studios	2.7	54	Arcade	Slice fruit, don't slice bombs – that is all y...	This was one of the upper-level mobile games f...	3800	1	32	10	2010	3800.0
1476	Slither.io	Lowtech Studios, Steve Howse	2.2	46	Arcade, Simulator	Slither.io is a massively multiplayer browser ...	Snake, computer lab edition., cancer. fun duri...	2100	1	21	4	2016	2100.0
1189	Happy Wheels	Fancy Force	2.8	124	Adventure, Platform, Racing	Try to get to the finish line in this twisted ...	i always wanted to play as santa but santa suc...	5100	3	61	21	2010	1700.0
1454	Ms. Pac-Man	Atari, Inc., General Computer Corporation (GCC)	3.6	198	Arcade	In 1982, a sequel to the incredibly popular Pa...	Do you really need a review of Ms. Pac- Man?, ...	1600	1	45	38	1982	1600.0
1499	WWE SmackDown vs. Raw 2008	THQ, YUKES Co., Ltd.	3.1	62	Fighting, Sport	The 2008 edition in the Smackdown vs. Raw seri...	The earliest memory I have of this game was wi...	1400	1	34	16	2007	1400.0

This sample shows the averages plays by player which is calculated based on the number of plays and active players. It also needs to be taken in consideration the less players, that is greater than zero, will provide a greater average play per player, which also provides a greater chance to be included in the top 10, such as Fruit Ninja Classic.

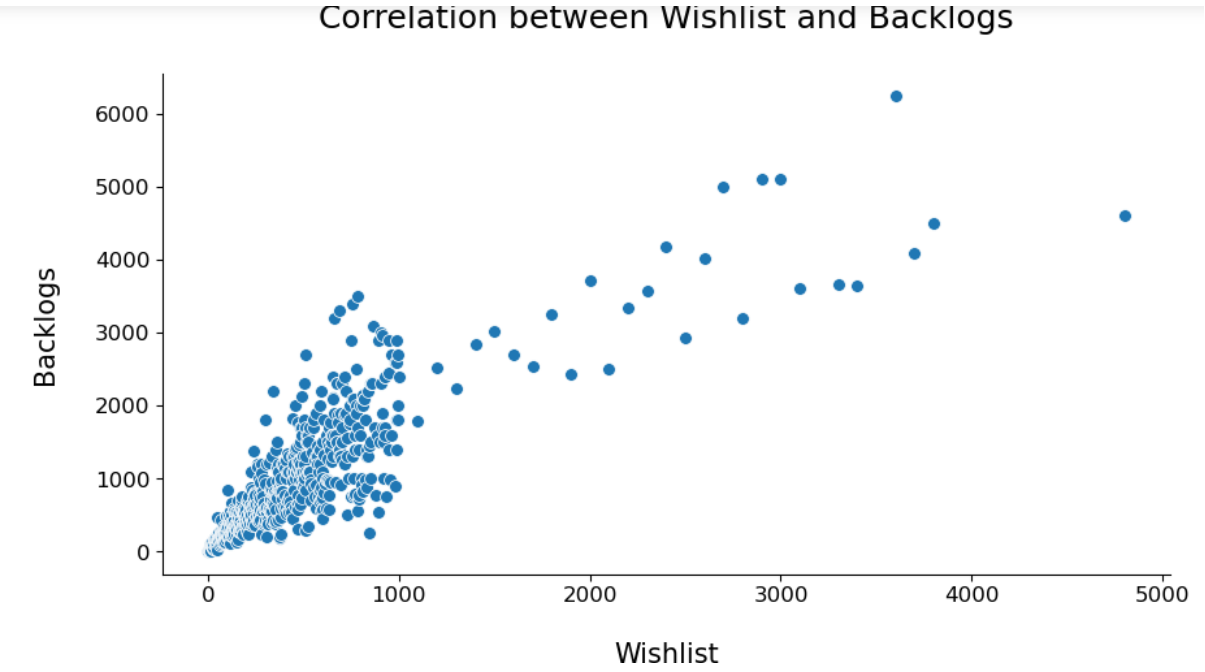


4.4 Correlation

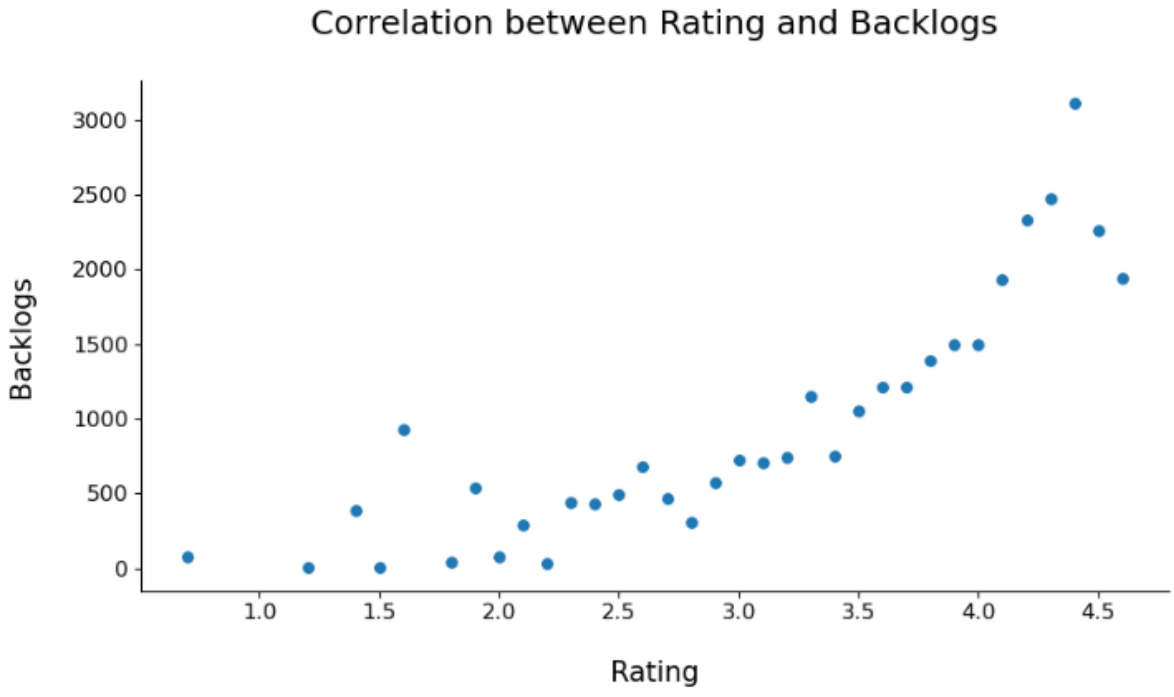
A metric used in statistics that measures the association between two variables is known as correlation (Hastie, et al., 2017). It assesses the extent to which the variables interact linearly to one another. Correlation is frequently used in exploratory data analysis to determine the degree and direction of relationships between variables in a dataset. Here are the results for this dataset.



It appears the number of plays and backlogs have strong correlation towards the number of reviews. Let's consider the backlog variable correlation to Wishlist and ratings.



The correlation between Wishlist and backlogs shows clustering of points towards lower values implies a concentration or grouping of data points in that position, however the general trend of the points reveals an increasing pattern. Hence, the correlation remains strong.



Data points are scattered and there is no clear pattern, hence it suggests a weaker correlation between the two variables with no consistent response between the variables. It suggests the relationship between the variables has an extensive degree of fluctuation or randomness.

4.5 Target Variable

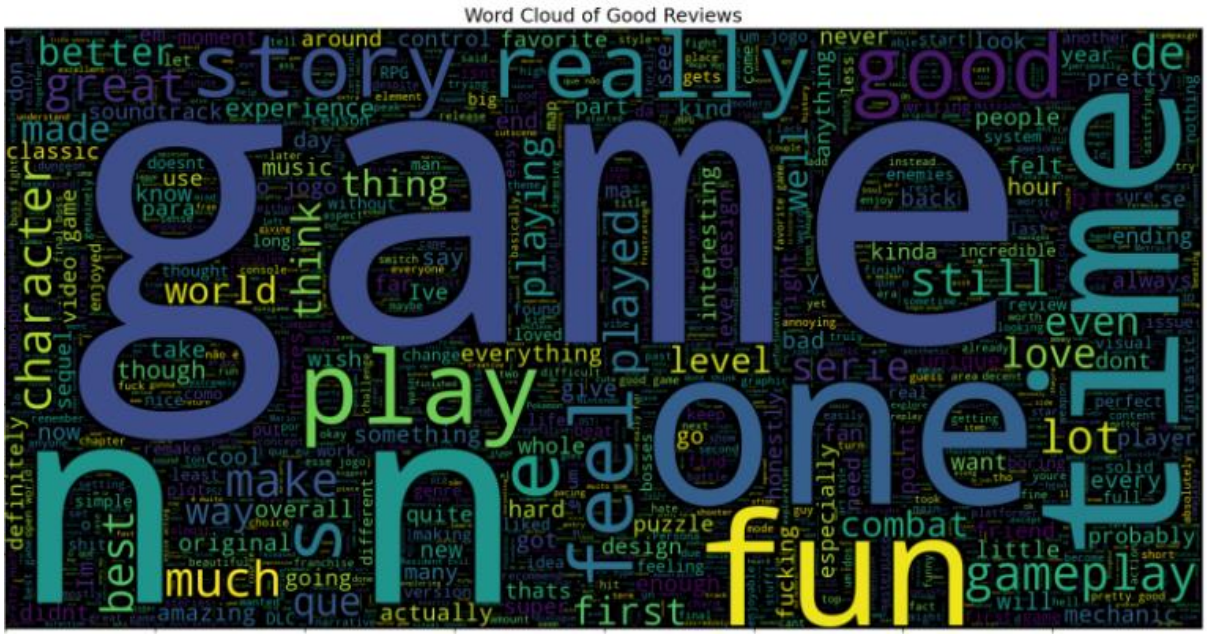
The variable of interest which is being predicted or estimated in a model is referred to as the target variable. It acts as the focus of the analysis or modelling process. For this analysis, we will create our target value to determine which reviews in the dataset are good and which reviews are bad. The rating system measure from 0 to 5 as shown within the statistics below with the average rating of 3,7.

```
count    1496.000000
mean        3.718917
std         0.531543
min         0.700000
25%         3.400000
50%         3.800000
75%         4.100000
max         4.600000
Name: Rating, dtype: float64
```

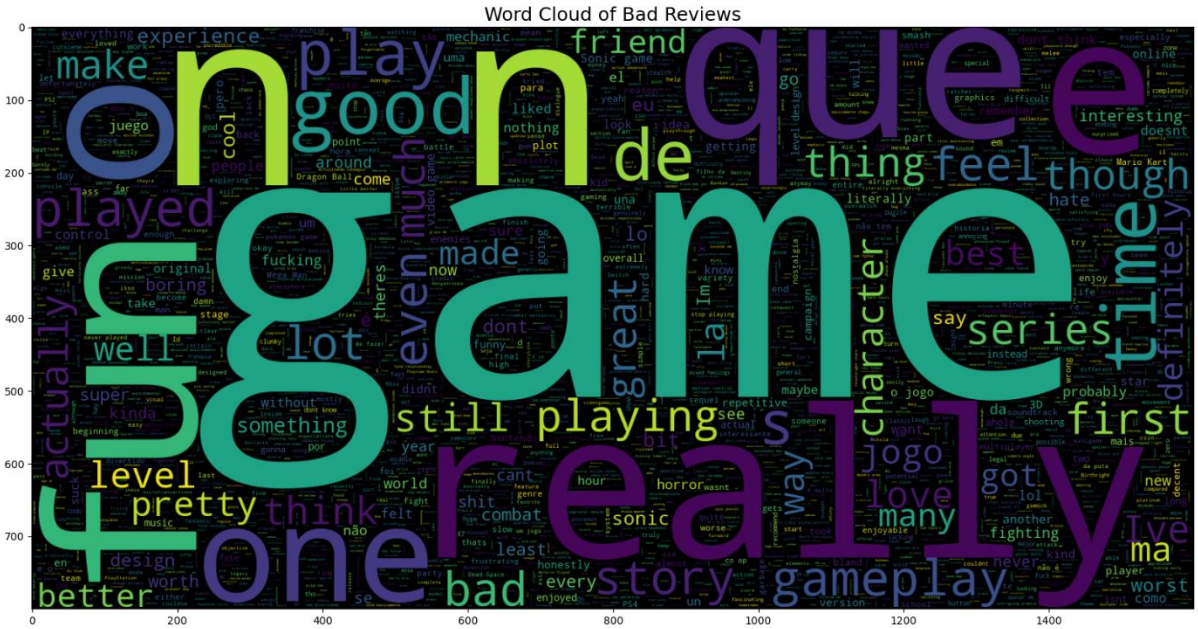
Sentiment analysis is then used to determine the rating result based on the rating, 0(bad, ratings less than 3) or 1 (good, ratings greater than three).

	Title	Team	Rating	Number of Reviews	Genres	Summary	Reviews	Number of Plays	Active Players	Backlogs	Wishlist	Release Year	Rating Status
377	The Witcher 3: Wild Hunt	WB Games, CD Projekt RED	4.3	2000	Adventure, RPG	RPG and sequel to The Witcher 2 (2011), The Wi...	One of my favorite games of all time. Incredib...	17000	1200	6100	1600	2015	1
1196	Call of Duty: Advanced Warfare	Activision, Sledgehammer Games	2.5	415	Shooter	Call of Duty: Advanced Warfare envisions the p...	Me julguem mas eu gosto desse jogo, a campanha...	5200	13	308	151	2014	0

The following word clouds are developed to represent good and bad views as well.



In the good reviews, it seems positive reviews towards combat, characters, and story.



The words provided by comments seem positive however ratings are negative. Negative reviews might include average ratings below 3. Speech factors such as sarcasm, connotations and euphemism can also be a cause for this as well as several mixed reviews contained in each value. This issue might become a problem in later analysis.

5 Natural language processing

The concept of Natural Language Processing (NLP) is a branch of artificial intelligence (AI) and linguistics which examines how computers interact with human language (Oracle, 2023). It involves the research and use of computer algorithms that allow computers to understand, interpret, and synthesize relevant spoken language (Oracle, 2023). In this section, we focus on the text analysis sector to analyse the review data alone through exploring different methods used to process text data to filter reviews.

5.1 Stop words.

Stop words are frequently used words that tend to be viewed insignificant in interpreting the meaning of a text. These words are typically filtered out or removed from text analysis and search queries because they are thought to provide little value to the overall comprehension of the content (Muller & Guido, 2018). The objective of deleting stop words is to minimize the dimensionality of text data, enhance processing efficiency, and focus on more important and relevant terms (Muller & Guido, 2018). By deleting these terms, NLP algorithms can focus more attention to the more linguistically significant content words, such as nouns, verbs, and adjectives.

5.2 Language detection

Our analysis' main language is English, hence needing to filter out as much non-English language data as possible to make results easier to find. By examining character patterns and statistical data, the langdetect library can automatically determine the language of a given text. Langdetect examines the text by analysing the frequency distribution of character sequences and making predictions using statistical models (Lee, 2020). To choose

the most probable language, it considers several linguistic patterns such as letter combinations, word frequencies, and punctuation usage. In the below sample English is represented by ‘en’.

	Title	Team	Rating	Number of Reviews	Genres	Summary	Reviews	Number of Plays	Active Players	Backlogs	Wishlist	Release Year	Rating Status	language
322	Bloons TD 6	Ninja Kiwi	3.8	241.0	Strategy, Tactical	The Bloons are back and better than ever! Craf...	Irei logar isso aqui e nunca mais tocar nesse ...	4100.0	331.0	343.0	73.0	2018.0	1.0	pt
792	Yakuza 0	Ryū Ga Gotoku Studios, Sega	4.4	2700.0	Adventure, Brawler, RPG, Simulator	The glitz, glamour, and unbridled decadence of...	O amor platónico entre um muro de cimento e um...	15000.0	1800.0	6400.0	2000.0	2015.0	1.0	pt
533	Outlast	Red Barrels	3.3	547.0	Adventure, Indie	Hell is an experiment you can't survive in Out...	(Played 2023)\n \n This probably best non-comb...	7500.0	90.0	1400.0	401.0	2013.0	1.0	en

5.3 Term frequency-inverse document frequency method

The TF-IDF is a statistical measure used in data extraction and text mining to assess the significance of a phrase inside a document or corpus of documents. It combines a keyword's local relevance (Term Frequency) with its worldwide value (Inverse Document Frequency) to calculate a weight that represents the term's value in each document within a corpus (Hastie, et al., 2017). TF-IDF assists in emphasizing the most important and distinctive terms in a document by allocating larger weights to such phrases (Muller & Guido, 2018).

5.4 Advanced tokenisation

Advanced tokenization consists of more complex and advanced methods of dividing text into smaller components called tokens for additional analysis or processing (Muller & Guido, 2018). Advanced tokenization approaches are essential for improving NLP task accuracy and performance. These approaches provide improved analysis and processing of text data by taking into consideration linguistic rules, morphology, domain knowledge, or task-specific requirements.

5.4.1 Word tokenisation

The purpose of word tokenization is to separate text into unique words. For languages with complex word structures or synthetic languages, this technique could prove problematic. As a result, word tokenization may not always be a solution that fits all circumstances (Muller & Guido, 2018). Depending on the task or language, multiple tokenization procedures may be necessary to produce the desired results.

5.4.2 Multi-word expression tokenisation

These refer to idioms, phrasal verbs, compound nouns and fixed expressions which can be found in various languages with cultural, linguistic, or historical significance. Because of their distinctive qualities, multi-word expressions create significant problems in natural language processing (Prakash, 2021). Identifying multi-word phrases with great accuracy is essential for tasks such as machine translation, sentiment analysis, information retrieval, and other language processing applications (Prakash, 2021).

5.4.3 Custom tokenisation

When traditional tokenization approaches do not effectively meet the specific demands of an event, especially when dealing with specific fields or languages, custom tokenization might be beneficial. Because of their unique properties, certain languages might require the use of specific tokenization approaches. When creating a custom tokenization method, it is essential to evaluate its performance to ensure that it matches with the objectives of adhering to NLP tasks (Briggs, 2021). Custom tokenization provides more control and flexibility when handling text data, resulting in improved analysis, modelling, and comprehension of the base information.

5.5 Stemming

Stemming is a natural language processing method which aims to reduce words to their root form, referred to as the stem (Muller & Guido, 2018). The purpose of stemming is to simplify word variants such that diverse versions of the same word may be considered as similar tokens, allowing for improved text data processing. However, stemming algorithms may not consistently produce linguistically acceptable stems and can sometimes generate stems that are not genuine words, hence the errors within some text value outputs (Muller & Guido, 2018). Stemming algorithms fail to consider context or meaning into consideration, which may contribute to over- or under-stemming. Over-stemming happens when several words are reduced to the same stem, resulting in information loss (Srinidhi, 2020). When words with the same stem fail to appear as related, this is known as under-stemming (Srinidhi, 2020). Therefore, for irregular phrases or words from languages with complex morphological systems, stemming may not perform effectively.

5.6 Lemmatisation

Lemmatization is a substitute to stemming which aims to identify the root or dictionary form of a word, known as the lemma, while considering its part of speech and context (Muller & Guido, 2018). Lemmatization, compared to stemming, generates legitimate words and is more linguistically accurate. In the code we will use lemmatisation to fix our stemmed data. Lemmatization takes into consideration a word's context and part of speech, enabling more accurate and insightful analysis. It assists in capturing word meanings, retaining grammatical integrity, and the maintenance of consistency in following NLP tasks.

5.6.1 Part-of-Speech (POS) Tagging

Part-of-speech tagging, a technique that assigns grammatical tags to each word in a phrase, is a crucial phase in many NLP tasks since it offers insight into the role and function of each word in a phrase, allowing for more accurate text analysis and comprehension (Oracle, 2023). Since the fundamental structure of a word could differ based on its part of speech, the POS tags assist in locating the appropriate lemma. Due to word ambiguity, whereby a word might have numerous alternative parts of speech depending on the context, POS tagging can be difficult (Oracle, 2023). Disambiguation techniques can however be used in this case scenario. Hence, the predicted tags are compared to manually labelled gold-standard tags to determine the accuracy of POS tagging.

5.6.2 SpaCy

SpaCy is a popular open-source library for natural language processing (NLP) in Python which offers efficient and operational tools for a wide range of NLP activities, making it a popular choice among researchers and developers (Oracle, 2023). Tokenization, part-of-speech tagging, named entity recognition, dependency parsing, text classification, lemmatization, sentence segmentation, and other NLP features are available in SpaCy (Oracle, 2023). SpaCy is built with performance and efficiency in consideration, which contributes to its efficient execution speed and integration. SpaCy's efficient implementation makes it appropriate for large-scale text data processing. SpaCy enables users to customize its capabilities and behaviour. It additionally supports rule-based matching, allowing the development of unique patterns for data extraction. Therefore, SpaCy provides a strong and adaptable toolbox to aid many parts of natural language comprehension and analysis, regardless of text processing, information extraction, or designing NLP applications (Oracle, 2023).

5.7 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a statistical model that identifies hidden concepts inside a collection of documents and statistically assigns those topics to each document (Muller & Guido, 2018). The aim of LDA is to identify these latent concepts and the word distributions corresponding to them. This would imply that LDA assumes the entire document collection has a fixed number of topics, that each document appears as a mixture of topics, with a probability distribution indicating the relevance of each topic in the document, and that each topic is defined by a probability distribution over words, corresponding to the likelihood of a word appearing in that topic (Muller & Guido, 2018).

Unfortunately, LDA ignores word order and document structure, seeing documents as unorganized bags of words (Muller & Guido, 2018). Furthermore, the number of themes must be set in advance, which might be difficult in some circumstances. Therefore, LDA aids in organizing, summarizing, and comprehending textual material in a more interpretable manner by uncovering the fundamental structure and topics in texts.

6 Data Modelling

Data modelling focuses on understanding the relationships between various data items, identifying their properties, and building a logical framework to assist analysis and insights. It is the process of structuring and organizing data to facilitate effective analysis. Data modelling also addresses data integrity and quality issues such as ensuring data consistency, accuracy, and completeness to support reliable analysis (Muller & Guido, 2018). We will analyse the accuracies of each data model with the help of implementing pipelines and using classification reports.

6.1 Splitting the data

Splitting the data is a basic step in machine learning and data analysis that ensures an unbiased evaluation of the model's capabilities. Using the specified X and Y variables, the data is separated into training and test sets. The X feature will contain the reviews data

content whereas the Y target variable will be the status variable which evaluates the type of variable(0 or 1, bad or good).

Splitting the data is essential for accurately evaluating model performance, modifying hyperparameters, and measuring the trained model's generalization capability. Reliable estimates of the model's performance make informed choices regarding model selection and optimization by using proper data splitting procedures (Muller & Guido, 2018).

6.2 Pipelines

A pipeline is a sequential workflow that connects multiple data processing and modelling steps to allow for the fast and repetitive execution of complicated machine learning algorithms (Muller & Guido, 2018). A pipeline facilitates and automates data preparation, feature engineering, model training, and assessment processes. Pipelines automate the end-to-end workflow, decreasing manual intervention and risks. Pipelines are simple to modify and expand to accommodate new data, additional pre-processing steps, or other modelling approaches. We use two pipelines; one pipeline contains classifiers while the other pipeline contains NLP method.

6.2.1 Classifiers

6.2.1.1 Logistic Regression pipeline

Logistic Regression is a common linear model used for binary classification tasks, employing a logistic function (Muller & Guido, 2018). A logistic regression pipeline often consists of a series of logistic regression-specific pre-processing and modelling steps. The pipeline uses the pre-processed data to train a logistic regression model and may include steps for hyperparameter tuning or model evaluation.

6.2.1.2 K-Nearest Neighbours pipeline

K-Nearest Neighbours is a non-parametric classification algorithm that labels data points based on their neighbours and operates by estimating the distance in feature space between the data point and its k nearest neighbours (Muller & Guido, 2018). It can handle both classification and regression problems. A K-Nearest Neighbours pipeline consists of a set of KNN-specific pre-processing and modelling processes. It uses the pre-processed data to train a KNN model and can include phases for hyperparameter shifting or model evaluation.

6.2.1.3 Naïve Bayes pipeline

Naive Bayes is a class of probability classifiers based on Bayes' theorem with a strong assumption of feature independence. Considering the "naive" premise, Naive Bayes classifiers are common for their ease of use, efficiency, and efficacy in a wide range of text categorization and spam filtering applications (Muller & Guido, 2018). A Naive Bayes pipeline involves the pre-processing and modelling steps specific to Naive Bayes algorithms to train a Naive Bayes model and can incorporate steps for hyperparameter tuning or model evaluation.

6.2.1.4 Support Vector Classifier pipeline

Support Vector Classifier is a supervised learning technique that generates a hyperplane or series of hyperplanes in the feature set to improve the division between different groups (Muller & Guido, 2018). It is used for both classification and regression tasks. The pipeline uses the pre-processed data to train an SVC classifier and may include steps for hyperparameter modification or model evaluation.

6.2.1.5 Random Forest pipeline

Random Forest is an ensemble learning approach that makes predictions by combining multiple decision trees. Each tree is trained on a random segment of the data, and predictions are created by combining individual tree predictions (Muller & Guido, 2018). It can handle complex relationships and handle high-dimensional data. A Random Forest pipeline uses pre-processed data to train a Random Forest model and may include stages for hyperparameter modification or model assessment.

6.2.1.6 XGBoost pipeline

Extreme Gradient Boosting is an enhanced version of gradient boosting, which progressively develops an ensemble of weak prediction models (in most cases decision trees), with each new model trained to fix the mistakes of the previous models (Muller & Guido, 2018). Like all the other classifier pipelines, data preparation procedures such as feature scaling or normalization, feature selection, or missing value handling may be included in the pipeline and uses the pre-processed data to train an XGBoost model. It can include phases for hyperparameter modification or model assessment.

6.2.1.7 Results

```
LogisticRegression Accuracy: 0.9230769230769231
LogisticRegression Classification Report:
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00        20
         1.0       0.92      1.00      0.96       240

    accuracy                           0.92       260
   macro avg       0.46      0.50      0.48       260
weighted avg       0.85      0.92      0.89       260

-------------------------------------------
KNeighborsClassifier Accuracy: 0.9192307692307692
KNeighborsClassifier Classification Report:
              precision    recall  f1-score   support

         0.0       0.33      0.05      0.09        20
         1.0       0.93      0.99      0.96       240

    accuracy                           0.92       260
   macro avg       0.63      0.52      0.52       260
weighted avg       0.88      0.92      0.89       260

-------------------------------------------
```

```
MultinomialNB Accuracy: 0.9230769230769231
MultinomialNB Classification Report:
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00        20
         1.0       0.92      1.00      0.96       240

    accuracy                           0.92       260
   macro avg       0.46      0.50      0.48       260
weighted avg       0.85      0.92      0.89       260

-------------------------------------------
SVC Accuracy: 0.9346153846153846
SVC Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.15      0.26        20
         1.0       0.93      1.00      0.97       240

    accuracy                           0.93       260
   macro avg       0.97      0.57      0.61       260
weighted avg       0.94      0.93      0.91       260

-------------------------------------------
```

```
RandomForestClassifier Accuracy: 0.9423076923076923
RandomForestClassifier Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.25      0.40        20
         1.0       0.94      1.00      0.97       240

    accuracy                           0.94       260
   macro avg       0.97      0.62      0.68       260
weighted avg       0.95      0.94      0.93       260

-----------------------------------------
XGBClassifier Accuracy: 0.9384615384615385
XGBClassifier Classification Report:
              precision    recall  f1-score   support

         0.0       0.75      0.30      0.43        20
         1.0       0.94      0.99      0.97       240

    accuracy                           0.94       260
   macro avg       0.85      0.65      0.70       260
weighted avg       0.93      0.94      0.93       260

-----------------------------------------
```

Considering the results above, the random forest classifier seems to have the highest accuracy. The analysis will continue with the use of random forest classifier alone.

6.2.2 NLP pipeline

As earlier discussed, natural language processes have great importance to text data. Let's look at the results of NLP methods' accuracies while using the random forest classifier.

```
No Method - RandomForestClassifier Accuracy: 0.9423076923076923
No Method - RandomForestClassifier Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.25      0.40        20
         1.0       0.94      1.00      0.97       240

    accuracy                           0.94       260
   macro avg       0.97      0.62      0.68       260
weighted avg       0.95      0.94      0.93       260

-----------------------------------------
Tokenization - RandomForestClassifier Accuracy: 0.9423076923076923
Tokenization - RandomForestClassifier Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.25      0.40        20
         1.0       0.94      1.00      0.97       240

    accuracy                           0.94       260
   macro avg       0.97      0.62      0.68       260
weighted avg       0.95      0.94      0.93       260

-----------------------------------------
```

```
Stemming - RandomForestClassifier Accuracy: 0.9423076923076923
Stemming - RandomForestClassifier Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.25      0.40        20
         1.0       0.94      1.00      0.97       240

    accuracy                           0.94       260
   macro avg       0.97      0.62      0.68       260
weighted avg       0.95      0.94      0.93       260

-----------------------------------------
Lemmatization - RandomForestClassifier Accuracy: 0.9423076923076923
Lemmatization - RandomForestClassifier Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.25      0.40        20
         1.0       0.94      1.00      0.97       240

    accuracy                           0.94       260
   macro avg       0.97      0.62      0.68       260
weighted avg       0.95      0.94      0.93       260

-----------------------------------------
```

The accuracies and results appear the same in each separate method. It is important to consider that over multiple runs results may differ however.

7 Model evaluation and improvement

Model evaluation is the process for evaluating a predictive model's performance and effectiveness, assisting in analysing the model's behaviour, optimizing its performance, and gaining insights into the data and its root cause, such as its strengths, weaknesses, and suitability for the given task (Muller & Guido, 2018). It assists in making informed choices about model selection, hyperparameter modification, understanding model behaviour, and implementation in practical applications.

7.1 Cross validation

Cross-validation is a resampling approach used to evaluate a model's performance and adaptability (Muller & Guido, 2018). Cross-validation improves the model's performance by minimizing the risk of bias and variation resulting from a single train-test split. Cross-validation is an essential step in model evaluation as it enables informed decision making regarding the model's performance and predictability.

8 Predictions

This section predicts if a review was positive or negative (0 or 1). This section will need further testing as prediction accuracy cannot be guaranteed and is experiencing a slight prediction error.

9 Conclusion

The analysis was quite insightful for learning more about the gaming sector despite running into a few inaccuracies. This could have been due to the inaccuracies found while conducting sentiment analysis or loss of information from over-stemming. These possibilities are not confirmed but are likely. Overall, this analysis has made use of multiple concepts and techniques found within NLP, data pre-processing, model evaluation, pipelines, classifiers, and cross-validation. Understanding these concepts is important for effectively working with textual data, developing machine learning models, and evaluating their performance.

It is also important to take in consideration the processing of multiple reviews within the dating which could contain mixed reviews of the overall rating. The accuracy of the data can also not be confirmed true outside of its original source. Assuming the website receives multiple reviews a day therefore ratings and data may change quite often in real time.

10 References

argha_c14, 2023. *Detect an Unknown Language using Python.* [Online]
Available at: https://www.geeksforgeeks.org/detect-an-unknown-language-using-python/
[Accessed 21 Jun 2023].

Briggs, J., 2021. *How-to Build a Transformer Tokenizer.* [Online]
Available at: https://towardsdatascience.com/transformers-from-scratch-creating-a-tokenizer-7d7418adb403
[Accessed 23 Jun 2023].

Hastie, T., Tibshirani, R. & Friedman, J., 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. NY: Springer.

IBM, 2023. *What is exploratory data analysis?.* [Online]
Available at: https://www.ibm.com/topics/exploratory-data-analysis#:~:text=Exploratory%20data%20analysis%20(EDA)%20is,often%20employing%20data%20visualization%20methods.
[Accessed 20 Jun 2023].

Jitender_1998, 2023. *Python NLTK | nltk.tokenize.mwe().* [Online]
Available at: https://www.geeksforgeeks.org/python-nltk-nltk-tokenize-mwe/
[Accessed 22 Jun 2023].

Lee, J., 2020. *Benchmarking Language Detection for NLP.* [Online]
Available at: https://towardsdatascience.com/benchmarking-language-detection-for-nlp-8250ea8b67c
[Accessed 21 Jun 2023].

Madhugiri, D., 2023. *Exploratory Data Analysis (EDA): Types, Tools, Process.* [Online]
Available at: https://www.knowledgehut.com/blog/data-science/eda-data-science
[Accessed 20 Jun 2023].

Muller, A. C. & Guido, S., 2018. *Introduction to Machine Learning with Python.* Forth ed. Sebastopol, CA: O'Reilly Media, Inc.

Nabi, J., 2018. *Machine Learning — Text Processing.* [Online]
Available at: https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958
[Accessed 4 Jun 2023].

Oracle, 2023. *What is Natural Language Processing?.* [Online]
Available at: https://www.oracle.com/za/artificial-intelligence/what-is-natural-language-processing/
[Accessed 22 Jun 2023].

Pascual, F., 2022. *Getting Started with Sentiment Analysis using Python.* [Online]
Available at: https://huggingface.co/blog/sentiment-analysis-python
[Accessed 21 Jun 2023].

Prakash, A., 2021. *Top 5 Word Tokenizers That Every NLP Data Scientist Should Know.* [Online]
Available at: https://towardsdatascience.com/top-5-word-tokenizers-that-every-nlp-data-scientist-should-know-45cc31f8e8b9
[Accessed 22 Jun 2023].

Roldós, I., 2019. *Text Processing: What Is It?.* [Online]
Available at: https://monkeylearn.com/blog/text-processing/
[Accessed 5 Jun 2023].

Srinidhi, S., 2020. *Stemming of words in Natural Language Processing, what is it?.* [Online]
Available at: https://towardsdatascience.com/stemming-of-words-in-natural-language-processing-what-is-it-41a33e8996e2
[Accessed 23 Jun 2023].

Stevens, E., 2022. *The 4 Types of Data Analysis [Ultimate Guide].* [Online]
Available at: https://careerfoundry.com/en/blog/data-analytics/different-types-of-data-analysis/
[Accessed 18 Mar 2023].